A better way to get language models to do what you ask

Colin Raffel
A better way to get language models to do what you ask

... and many other awesome collaborators
The cabs were charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were those, internal, conspicuous, cab.

The cabs charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were those, internal, conspicuous, cab.

horse-drawn cabs and were quite popular, the Prince of Wales (the King Edward VII) travelled in. The cabs quickly known as "hummingbirds" for noise made by their motors and their distinctive black and livery. Passengers the interior fittings were when compared to cabs but there were some complaints the lighting made them too to those outside.
"translate English to German: That is good."

"cola sentence: The course is jumping well."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."

"translate German to English: Das ist gut."

"not acceptable"

"six people hospitalized after a storm in attala county."

"3.8"
The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the internal lighting made them too conspicuous to those outside the cab. The fleet peaked at around 75 cabs, all of which needed to return to the single depot at Lambeth to switch batteries.

Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?

yes
Language Models are Few-Shot Learners

TriviaQA zero-shot performance

Parameters (Billions)

Exact match score
1. In what year was the first-ever Wimbledon Championship held? **Answer: 1877.**
2. Hg is the chemical symbol of which element? **Answer: Mercury.**
3. Which email service is owned by Microsoft? **Answer: Hotmail.**
4. Which country produces the most coffee in the world? **Answer: Brazil.**
5. In which city was Jim Morrison buried? **Answer: Paris.**
6. Which song by Luis Fonsi and Daddy Yankee has the most views (of all time) on YouTube? **Answer: “Despacito.”**
7. What was the first state? **Answer: Delaware.**
8. What is the capital city of Spain? **Answer: Madrid.**
9. What is the painting “La Gioconda” more usually known as? **Answer: The Mona Lisa.**
UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi\textsuperscript{1}  Sewon Min\textsuperscript{2}  Tushar Khot\textsuperscript{1}  Ashish Sabharwal\textsuperscript{1}  
Oyvind Tafjord\textsuperscript{1}  Peter Clark\textsuperscript{1}  Hannaneh Hajishirzi\textsuperscript{1,2}  

\textsuperscript{1} Allen Institute for AI, Seattle, U.S.A.  
\textsuperscript{2} University of Washington, Seattle, U.S.A.

**Extractive [SQuAD]**

**Question:** At what speed did the turbine operate?  
**Context:** (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 \textsuperscript{rpm} bladeless turbine. ...  
**Gold answer:** 16,000 \textsuperscript{rpm}

**Multiple-Choice [ARC-challenge]**

**Question:** What does photosynthesis produce that helps plants grow?  
**Candidate Answers:** (A) water (B) oxygen (C) protein (D) sugar  
**Gold answer:** sugar

**Abstractive [NarrativeQA]**

**Question:** What does a drink from narcissus's spring cause the drinker to do?  
**Context:** Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...  
**Gold answer:** fall in love with themselves

**Yes/No [BoolQ]**

**Question:** Was America the first country to have a president?  
**Context:** (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...  
**Gold answer:** no
## Measuring Massive Multitask Language Understanding

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dan Hendrycks</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>Collin Burns</td>
<td>Columbia University</td>
</tr>
<tr>
<td>Steven Basart</td>
<td>UChicago</td>
</tr>
<tr>
<td>Andy Zou</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>Mantas Mazeika</td>
<td>UIUC</td>
</tr>
<tr>
<td>Dawn Song</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>Jacob Steinhardt</td>
<td>UC Berkeley</td>
</tr>
</tbody>
</table>

### Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23
Answer: B

Compute $i + i^2 + i^3 + \cdots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) $i$ (D) $-i$
Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
(A) 28 (B) 21 (C) 40 (D) 30
Answer: C

<table>
<thead>
<tr>
<th>Model</th>
<th>Humanities</th>
<th>Social Science</th>
<th>STEM</th>
<th>Other</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>27.9</td>
<td>28.8</td>
<td>27.0</td>
<td>27.7</td>
<td>27.9</td>
</tr>
<tr>
<td>ALBERT</td>
<td>27.2</td>
<td>25.7</td>
<td>27.7</td>
<td>27.9</td>
<td>27.1</td>
</tr>
<tr>
<td>GPT-2</td>
<td>32.8</td>
<td>33.3</td>
<td>30.2</td>
<td>33.1</td>
<td>32.4</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>45.6</td>
<td>56.6</td>
<td>40.2</td>
<td>54.6</td>
<td>48.9</td>
</tr>
<tr>
<td>GPT-3 Small (few-shot)</td>
<td>24.4</td>
<td>30.9</td>
<td>26.0</td>
<td>24.1</td>
<td>25.9</td>
</tr>
<tr>
<td>GPT-3 Medium (few-shot)</td>
<td>26.1</td>
<td>21.6</td>
<td>25.6</td>
<td>25.5</td>
<td>24.9</td>
</tr>
<tr>
<td>GPT-3 Large (few-shot)</td>
<td>27.1</td>
<td>25.6</td>
<td>24.3</td>
<td>26.5</td>
<td>26.0</td>
</tr>
<tr>
<td>GPT-3 X-Large (few-shot)</td>
<td>40.8</td>
<td>50.4</td>
<td>36.7</td>
<td>48.8</td>
<td>43.9</td>
</tr>
</tbody>
</table>
Meta-tuning Language Models to Answer Prompts Better

Ruiqi Zhong  Kristy Lee*  Zheng Zhang*  Dan Klein
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

Standard Classification Format

A total waste of time.  classify 0
Great movie, must see!  classify 1

Convert

Question Answering Format

[Question] Is the review positive?
[Context] A total waste of time.  answer “No”
[Context] Great movie, must see!  answer “Yes”

Hate Speech Detection
Question Categorization
Topic Classification
Sentiment Classification
Stance Classification

Meta-tune
Evaluate on Unseen Tasks
CROSSFIT 🦅: A Few-shot Learning Challenge for Cross-task Generalization in NLP

Qinyuan Ye  Bill Yuchen Lin  Xiang Ren
University of Southern California
{qinyuany, yuchen.lin, xiangren}@usc.edu
Can we obtain good zero-shot task generalization by training a large language model on a massively multitask mixture of diversely prompted datasets?
Summarization
The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification
“How is air traffic controlled?” “How do you become an air traffic controller?” Pick one: these questions are duplicates or not duplicates.

Question answering
I know that the answer to “What team did the Panthers defeat?” is in “The Panthers finished the regular season [...]”. Can you tell me what it is?

Multi-task training
Zero-shot generalization

Natural language inference
Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?
QQP (Paraphrase)

<table>
<thead>
<tr>
<th>Question1</th>
<th>How is air traffic controlled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question2</td>
<td>How do you become an air traffic controller?</td>
</tr>
<tr>
<td>Label</td>
<td>0</td>
</tr>
</tbody>
</table>

{Question1} {Question2} Pick one: These questions are duplicates or not duplicates.

{Choices[label]}

{Choices[label]}

XSum (Summary)

Document: The picture appeared on the wall of a Poundland store on Whymark Avenue...

Summary: Graffiti artist Banksy is believed to be behind...

{Document} How would you rephrase that in a few words?

{Summary}

{Summary}

First, please read the article: {Document} Now, can you write me an extremely short abstract for it? {Summary}
I am hesitating between two options. Help me choose the more likely cause:

- The sun was rising.
- The grass was cut.

Y

The sun was rising.
Number of prompted datasets: 170
Number of prompts: 1939
More prompts are better than one
Adding datasets (usually) helps
FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*  Maarten Bosma*  Vincent Y. Zhao*  Kelvin Guu*  Adams Wei Yu
Brian Lester  Nan Du  Andrew M. Dai  Quoc V. Le
Google Research

Finetune on many tasks ("instruction-tuning")

Input (Commonsense Reasoning)
Here is a goal: Get a cool sleep on summer days.
How would you accomplish this goal?
OPTIONS:
-Keep stack of pillow cases in fridge.
-Keep stack of pillow cases in oven.
Target
keep stack of pillow cases in fridge

Input (Translation)
Translate this sentence to Spanish:
The new office building was built in less than three months.
Target
El nuevo edificio de oficinas se construyó en tres meses.

Inference on unseen task type

Input (Natural Language Inference)
Premise: At my age you will probably have learnt one lesson.
Hypothesis: It’s not certain how many lessons you’ll learn by your thirties.
Does the premise entail the hypothesis?
OPTIONS:
yes  it is not possible to tell  no

FLAN Response
It is not possible to tell

Performance on unseen task types

- GPT-3 175B zero-shot
- GPT-3 175B few-shot
- FLAN 137B zero-shot

Natural language inference:
- GPT-3 175B: 42.9
- GPT-3 175B few-shot: 53.2
- FLAN 137B: 56.2

Reading Comprehension:
- GPT-3 175B: 63.7
- GPT-3 175B few-shot: 72.6
- FLAN 137B: 77.4

Closed-Book QA:
- GPT-3 175B: 49.8
- GPT-3 175B few-shot: 55.7
- FLAN 137B: 56.6
Performance on held-out tasks

Average zero-shot accuracy on 13 held-out tasks (%)

- Instruction tuning
- Untuned model

Model Size (# parameters)

- 0.4B
- 2B
- 8B
- 68B
- 137B
Performance on **held-out** tasks

- **Instruction tuning**
- **Untuned model**

Average zero-shot accuracy on 13 held-out tasks (%)

T0 = 11B parameters
<table>
<thead>
<tr>
<th></th>
<th>GPT-3</th>
<th>FLAN</th>
<th>T0 (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>175B</td>
<td>137B</td>
<td>11B</td>
</tr>
<tr>
<td>Multitask supervision</td>
<td>Implicit</td>
<td>Explicit</td>
<td>Explicit</td>
</tr>
<tr>
<td>Zero-shot performance</td>
<td>Decent</td>
<td>Better</td>
<td>Better</td>
</tr>
<tr>
<td>Architecture</td>
<td>Decoder</td>
<td>Decoder</td>
<td>Encoder/decoder</td>
</tr>
<tr>
<td>Pre-training</td>
<td>LM</td>
<td>LM</td>
<td>MLM→LM</td>
</tr>
<tr>
<td>Multiple prompts?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prompt diversity</td>
<td>N/A</td>
<td>Some</td>
<td>Lots</td>
</tr>
<tr>
<td>Public</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Toolkit for collecting and applying templates of prompting instances

Readme

Apache-2.0 License

https://github.com/bigscience-workshop/promptsourc
A one-year long research workshop on large multilingual models and datasets

https://bigscience.huggingface.co/
## MultiTask Prompted Training Enables Zero-Shot Task Generalization

<table>
<thead>
<tr>
<th>Victor Sanh*</th>
<th>Albert Webson*</th>
<th>Colin Raffel*</th>
<th>Stephen H. Bach*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hugging Face</td>
<td>Brown University</td>
<td>Hugging Face</td>
<td>Brown University</td>
</tr>
</tbody>
</table>

| Lintang Sutawika      | Zaid Alyafeai          | Antoine Chaffin        | Arnaud Stiegler        |
| BigScience            | KFUPM                  | IRISA, IMATAG          | Hyperscience           |

| Manan Dey              | M Saiful Bari          | Canwen Xu              | Urmish Thakker         |
| SAP                   | NTU                    | UCSD/HF                | SambaNova Systems      |

| Eliza Szczeslha        | Taewooon Kim           | Gunjan Chhablani       | Shanya Sharma          |
| BigScience             | VU Amsterdam           | BigScience             | Walmart Labs           |

| Debajyoti Datta        | Jonathan Chang         | Mike Tian-Jian Jiang   | Han Wang               |
| University of Virginia | ASUS                   | NYU                    | Matteo Manica          |

| Sheng Shen             | Zhong-Xin Yong         | Harshit Pandey         | Michael McKenna        |
| UC Berkeley            | Brown University       | BigScience             | Parity                 |

| Thomas Wang            | Trishala Neeraj        | Jos Rozen              | Abheesht Sharma        |
| Inria, France          | BigScience             | BigScience             | BITS Pilani, India     |

| Thibault Fevry         | Jason Alan Fries       | Ryan Techan            | Teven Le Scao          |
| BigScience             | Stanford University    | Charles River Analytics| Hugging Face           |

| Stella Biderman        | Leo Gao                | Tali Bers              | Thomas Wolf            |
| EleutherAI             | EleutherAI             | Brown University       | Hugging Face           |

|                          |                        |                        | Alexander M. Rush      |
|                          |                        |                        | Hugging Face           |