Finetuned Language Models are Zero-Shot Learners

By - Wen Zhang
Hemanth Reddy

Motivation

Language models (LMs) at scale, such as GPT-3 (Brown et al., 2020), have been shown to perform few-shot learning remarkably well. They are less successful at zero-shot learning, however.

For example, GPT-3’s zero-shot performance is much worse than few-shot performance on tasks such as reading comprehension, question answering, and natural language inference.

Finetuned Language Net(FLAN):

Authors leverage the intuition that NLP tasks can be described via natural language instructions, such as “Is the sentiment of this movie review positive or negative?” or “Translate ‘how are you’ into french.”

We take a pretrained language model of 137B parameters and perform instruction tuning—finetuning the model on a mixture of more than 60 NLP datasets expressed via natural language instructions.

Few shot learning

Few shot learning uses a small group of examples from new data to learn new tasks. The procedure of small sample learning deals with a class of machine learning problems, which consists of a finite number of examples and supervisory information for target T.

Few shot learning is commonly used by OpenAI as GPT3 is a few-shot learner.

Zero-shot learning

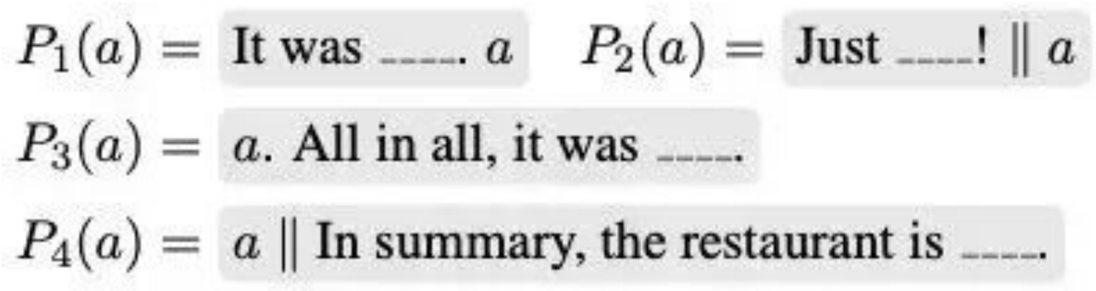
Zero-sample learning is the challenge of learning modeling without using data labels. Zero-sample learning requires little human intervention, and the model relies on previously trained concepts and additional existing data. This approach reduces the time and effort required to tag data. Instead of providing training examples, zero-sample learning provides high-level descriptions of new categories so that the machine can relate them to existing categories that the machine already knows.

Zero-shot learning

Zero sample learning essentially consists of two stages: training and reasoning. In training, intermediate layers of semantic attributes are captured, and then in the reasoning phase, this knowledge is used to predict categories within a new set of categories. At this level, the second layer models the relationship between attributes and classes and uses the initial property signature of the class to anchor the class.

Prompt

Prompt is designed to prompt completion of the language model, such as giving the first half of a sentence to generate the second half of a sentence, or filling in the blanks, just like doing the language Model task, and its template looks like this:



$$P_1(a) = \text{It was } \underline{\hspace{2em}}.\ a \qquad P_2(a) = \text{Just } \underline{\hspace{2em}}! \parallel a$$

$$P_3(a) = a.\ \text{All in all, it was } \underline{\hspace{2em}}.$$

$$P_4(a) = a \parallel \text{In summary, the restaurant is } \underline{\hspace{2em}}.$$

Prompt vs Instruction

Let's get rid of all the concepts in our head and think of ourselves as a model. I give you two tasks:

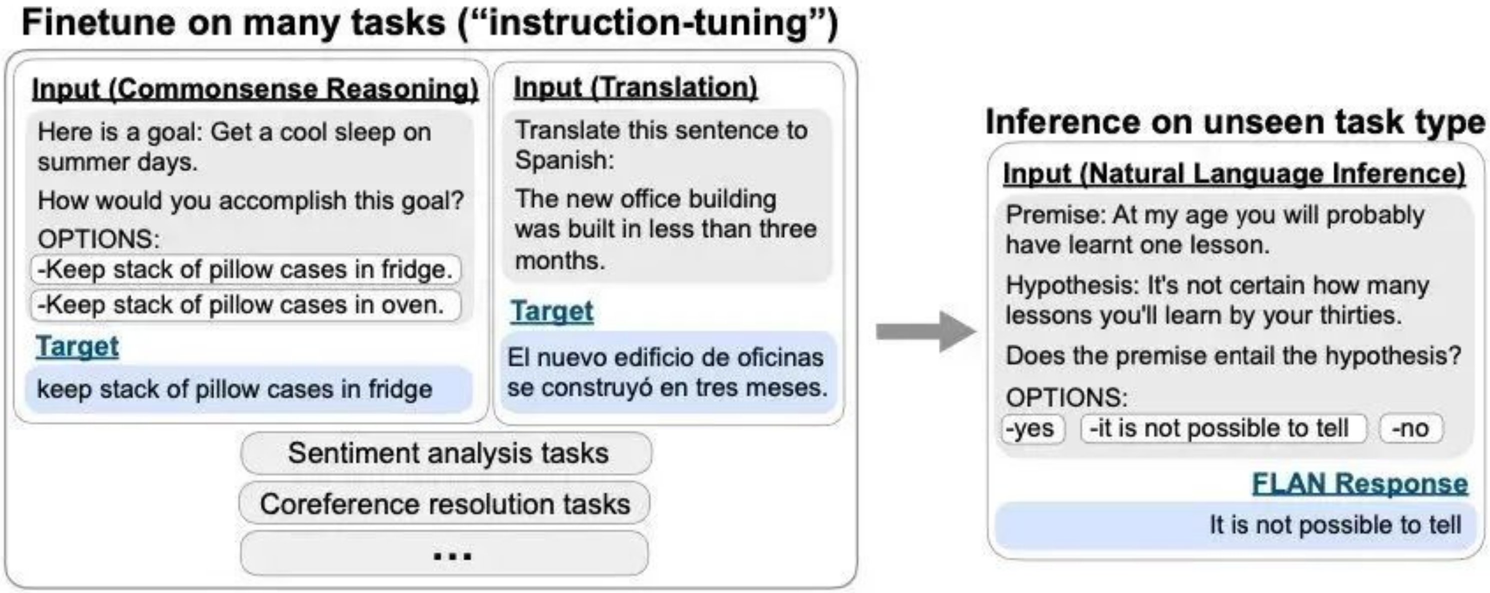
1. I took my girlfriend to a restaurant and she enjoyed it very much, because the restaurant was so __!

2. Identify the emotion of this sentence: You took your girlfriend to a restaurant and she enjoyed the meal. Choice: A= good, B= fair, C= poor

Prompt is the first, Instruction is the second.

Instruction Tuning

Instruction Tuning is to stimulate the understanding ability of the language model, and through giving more obvious instructions/instructions, the model can understand and make correct action. For example, NLI/ sorting task:



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.

Sentiment analysis tasks

Coreference resolution tasks

...

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

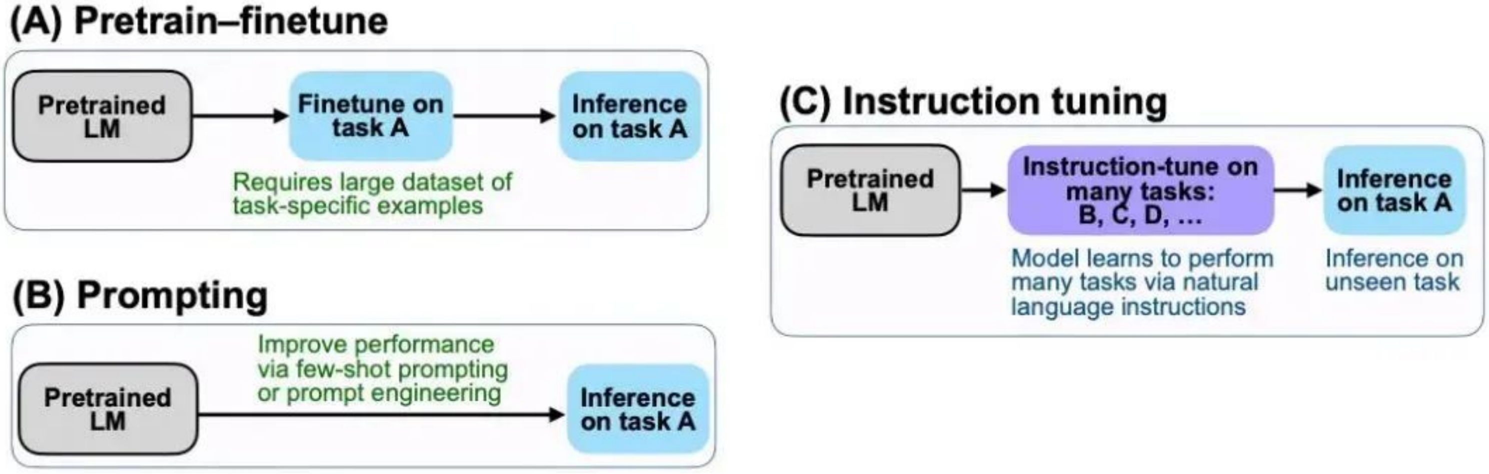
Prompt vs Instruction

Prompt also has fine tuning. After Prompt tuning, the model learns this Prompt mode.

What is the difference between Prompt and Instruction Tuning after fine tuning?

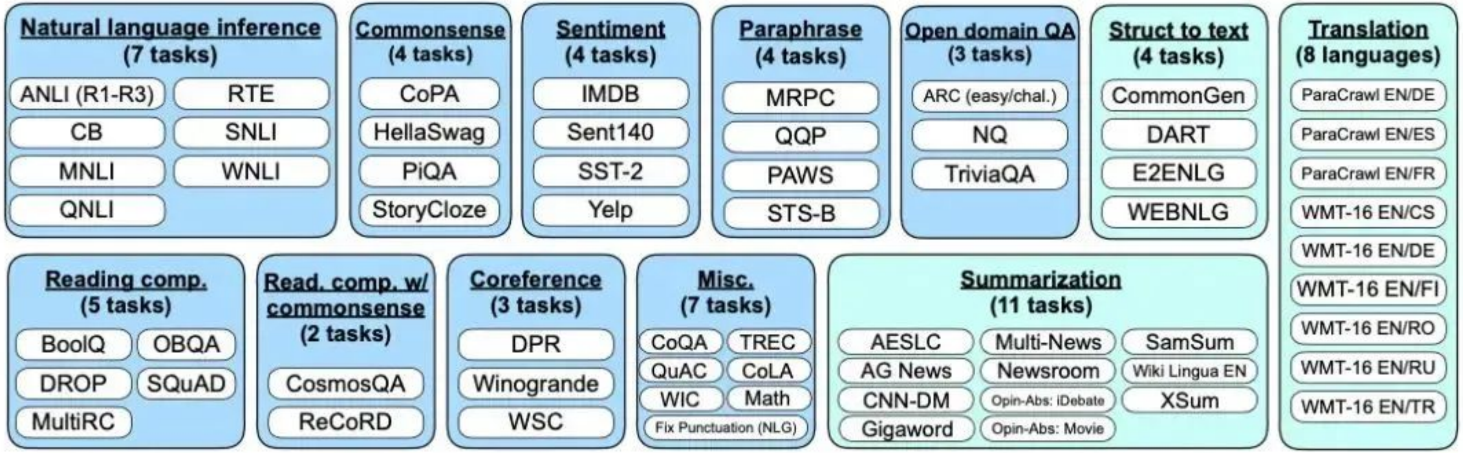
Prompt vs Instruction

This is the advantage of Instruction Tuning, prompt tuning is for a task, for example, prompt tuning of emotion analysis task, the refined model can only be used for emotion analysis task. After Instruction Tuning multi-task fine-tuning, you can use the zero-shot for other tasks!


(A) Pretrain–finetune
Pretrained LM
Finetune on task A
Inference on task A
Requires large dataset of task-specific examples
(B) Prompting
Pretrained LM
Improve performance via few-shot prompting or prompt engineering
Inference on task A
(C) Instruction tuning
Pretrained LM
Instruction-tune on many tasks: B, C, D, ...
Inference on task A
Model learns to perform many tasks via natural language instructions
Inference on unseen task

Experiment

After understanding the concept of Instruction Tuning, it is much clearer to look at the experimental method. The authors divided 62 NLP tasks into 12 classes. They trained on 11 of them and tested the zero-shot effect on 1, so as to ensure that the model had never seen such a task before, and to see if the model could really understand the "instruction" :



Natural language inference (7 tasks)
- ANLI (R1-R3)
- RTE
- CB
- SNLI
- MNLI
- WNLI
- QNLI

Commonsense (4 tasks)
- CoPA
- HellaSwag
- PiQA
- StoryCloze

Sentiment (4 tasks)
- IMDB
- Sent140
- SST-2
- Yelp

Paraphrase (4 tasks)
- MRPC
- QQP
- PAWS
- STS-B

Open domain QA (3 tasks)
- ARC (easy/chal.)
- NQ
- TriviaQA

Struct to text (4 tasks)
- CommonGen
- DART
- E2ENLG
- WEBNLG

Translation (8 languages)
- ParaCrawl EN/DE
- ParaCrawl EN/ES
- ParaCrawl EN/FR
- WMT-16 EN/CS
- WMT-16 EN/DE
- WMT-16 EN/FI
- WMT-16 EN/RO
- WMT-16 EN/RU
- WMT-16 EN/TR

Reading comp. (5 tasks)
- BoolQ
- OBQA
- DROP
- SQuAD
- MultiRC

Read. comp. w/ commonsense (2 tasks)
- CosmosQA
- ReCoRD

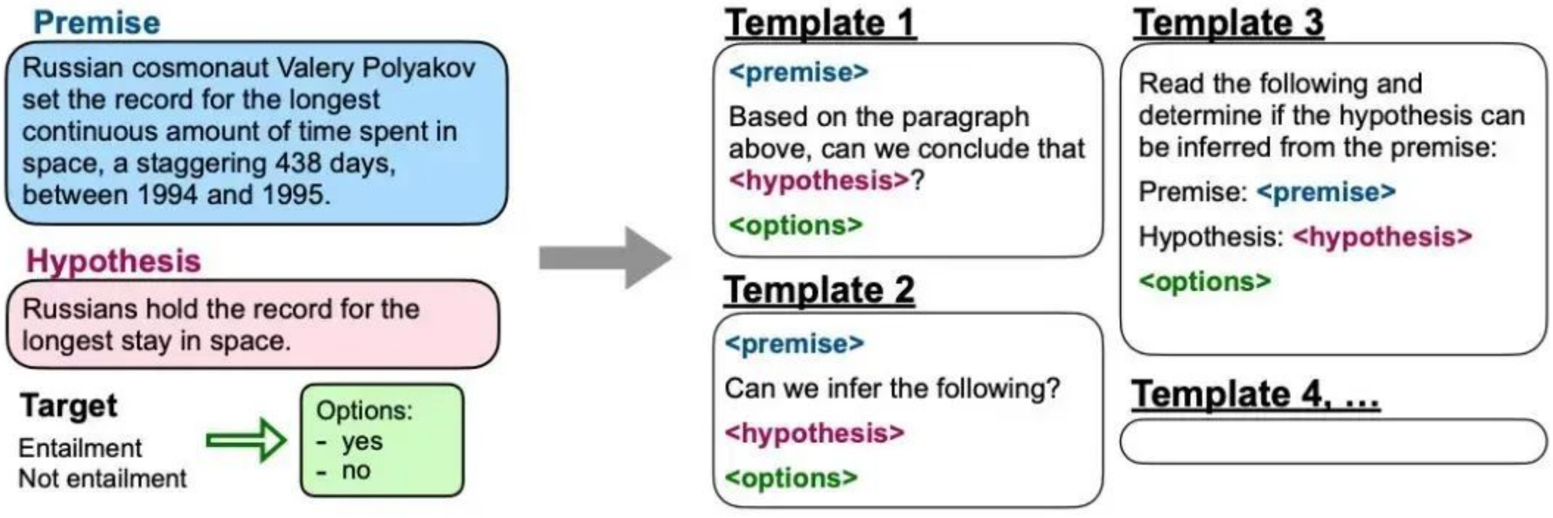
Coreference (3 tasks)
- DPR
- Winogrande
- WSC

Misc. (7 tasks)
- CoQA
- TREC
- QuAC
- CoLA
- WIC
- Math
- Fix Punctuation (NLG)

Summarization (11 tasks)
- AESLC
- Multi-News
- SamSum
- AG News
- Newsroom
- Wiki Lingua EN
- CNN-DM
- Opin-Abs: iDebate
- XSum
- Gigaword
- Opin-Abs: Movie

Experiment

Like Prompt, the authors design 10 instruction templates for each task, and test to see average and best performance



Premise

Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.

Hypothesis

Russians hold the record for the longest stay in space.

Target

Entailment
Not entailment

Options:
- yes
- no

Template 1

<premise>

Based on the paragraph above, can we conclude that <hypothesis>?

<options>

Template 2

<premise>

Can we infer the following?

<hypothesis>

<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

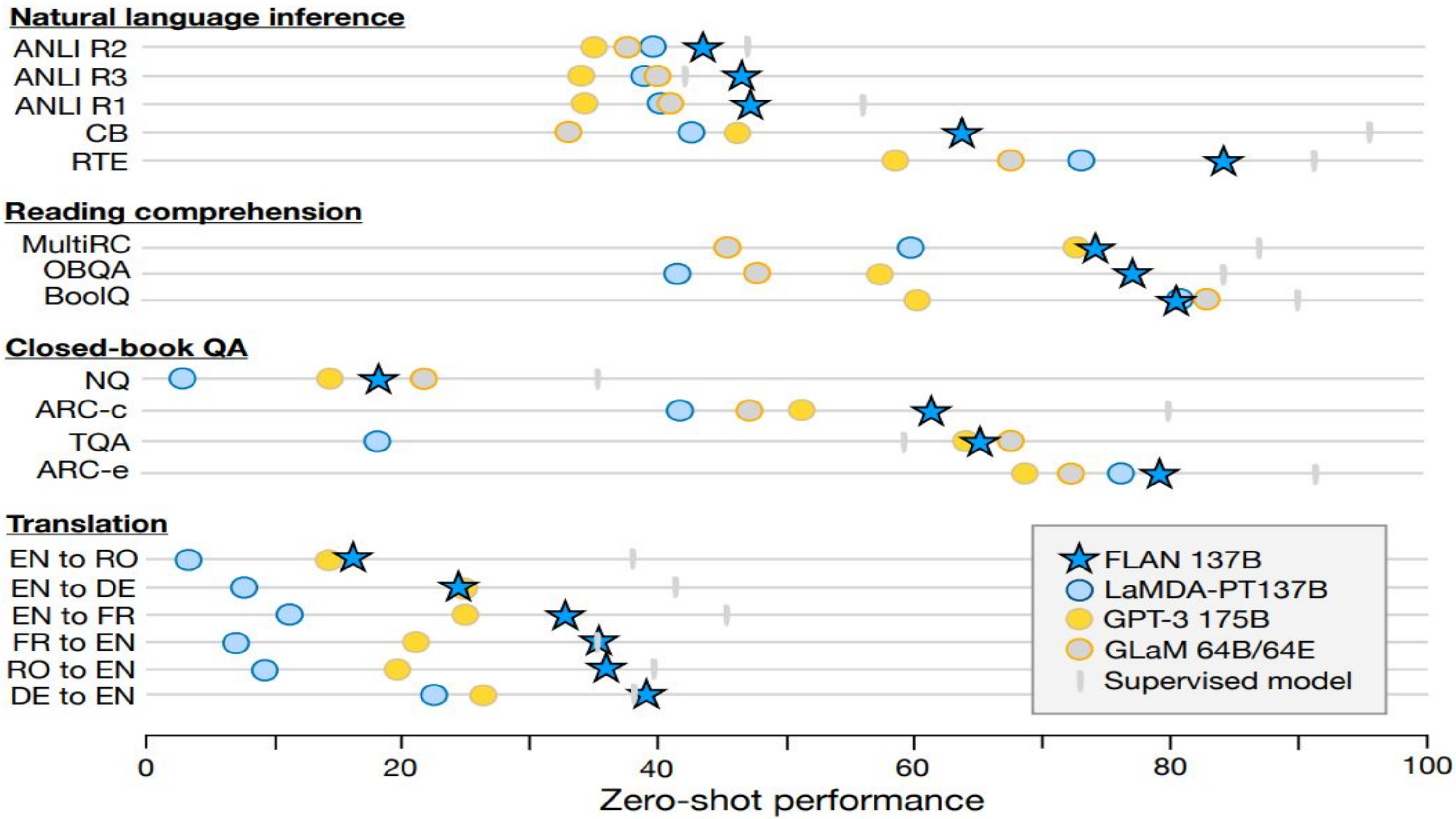
Premise: <premise>

Hypothesis: <hypothesis>

<options>

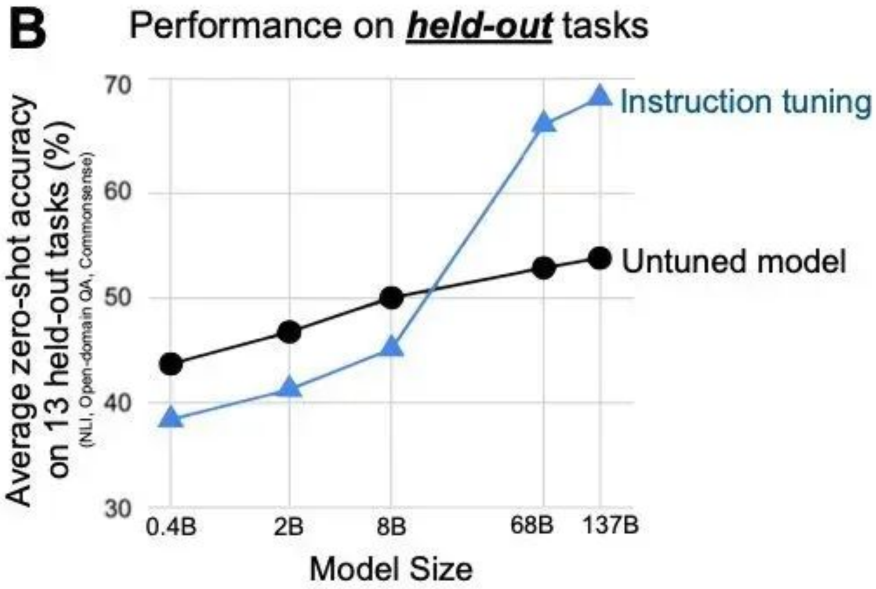
Template 4, ...

Results


Natural language inference
ANLI R2
ANLI R3
ANLI R1
CB
RTE
Reading comprehension
MultiRC
OBQA
BoolQ
Closed-book QA
NQ
ARC-c
TQA
ARC-e
Translation
EN to RO
EN to DE
EN to FR
FR to EN
RO to EN
DE to EN
FLAN 137B
LaMDA-PT137B
GPT-3 175B
GLaM 64B/64E
Supervised model
0
20
40
60
80
100
Zero-shot performance

Result

Unfortunately, this method only works on large models, and even degrades performance on small models. The author thinks that because the capacity of small models is limited, it is not easy to learn the knowledge of only one task:


B
Performance on ***held-out*** tasks
Average zero-shot accuracy on 13 held-out tasks (%)
(NLI, Open-domain QA, Commonsense)
0.4B
2B
8B
68B
137B
Model Size
Instruction tuning
Untuned model