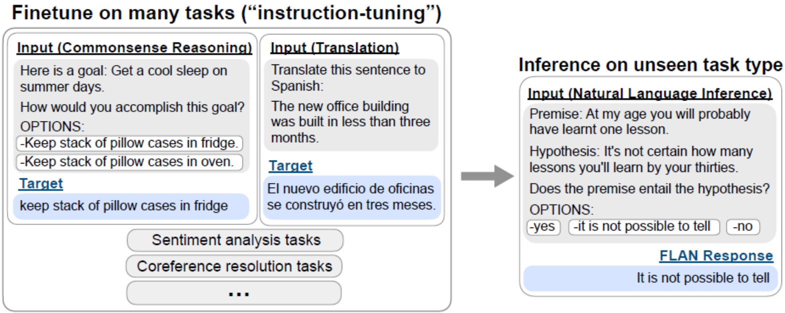
Finetuned Language Models Are Zero-Shot Learners[1]

September 17, 2021

Abstract

This paper aims to improve the zero-shot learning abilities of language models and proposes **instruction tuning**–finetuning language models on a collection of tasks described via instructions.



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

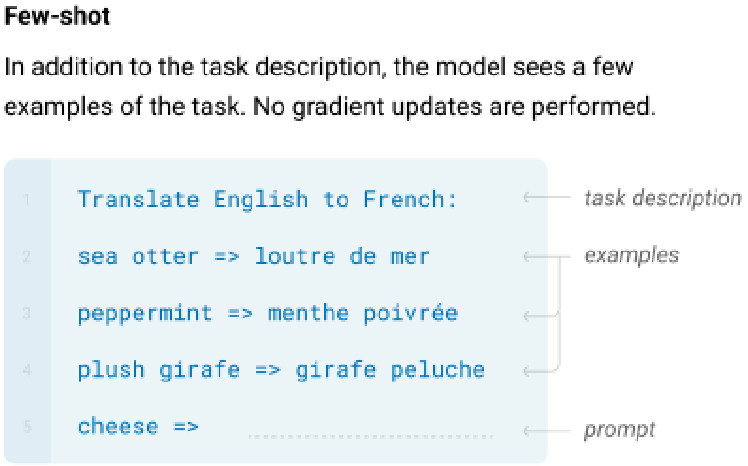
Figure 1: Framework

Overview

- Introduction
  - GPT-3's few-shot performance
  - FLAN
- Methodology
  - Tasks
  - Templates
  - Input
- Experiment
  - Model and Tuning
  - Results
- Comparison
- Conclusion

GPT-3

GPT-3[2] has been shown to perform few-shot learning remarkably well but less successful at zero-shot learning.



Few-shot

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

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese => ...............................
```

FLAN

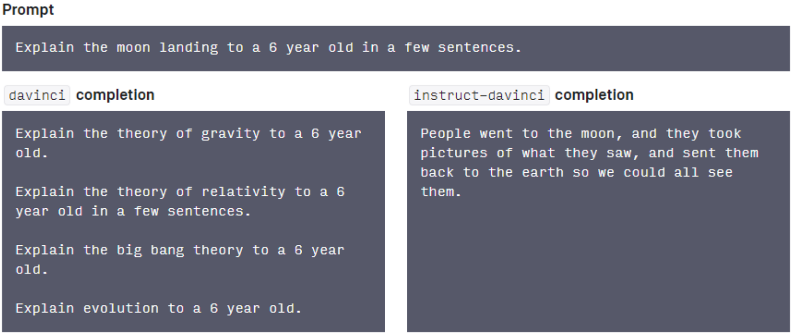
The paper leverages the intuition that NLP tasks can be described via natural language instructions, such as:

- “Is the sentiment of this movie review positive or negative?”
- “Translate ‘how are you’ into Chinese.”

The process is as follows:

- pretraining a language model of 137B parameters
- grouping NLP tasks into clusters based on their task types and hold out one for evaluation;
- instruction tuning the LM on all other clusters; the resulting model is called Finetuned LAnguage Net(FLAN).

FLAN



Prompt

```
Explain the moon landing to a 6 year old in a few sentences.
```

`davinci` **completion**

```
Explain the theory of gravity to a 6 year
old.

Explain the theory of relativity to a 6
year old in a few sentences.

Explain the big bang theory to a 6 year
old.

Explain evolution to a 6 year old.
```

`instruct-davinci` **completion**

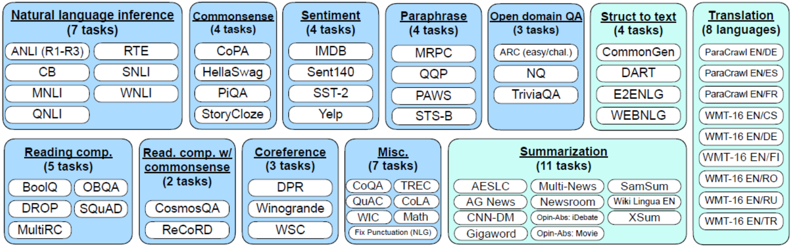
```
People went to the moon, and they took
pictures of what they saw, and sent them
back to the earth so we could all see
them.
```

Figure 3: Instruct models[1]

[1] https://beta.openai.com/docs/engines/instruct-series-beta

Tasks

They aggregate 62 text datasets that are publicly available on Tensorflow Datasets. Each dataset is categorized into one of twelve task clusters.



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

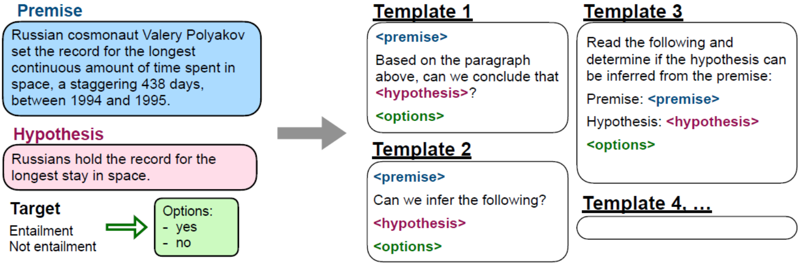
Misc. (7 tasks): CoQA, TREC, QuAC, CoLA, WIC, Math, Fix Punctuation (NLG)

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

Figure 4: Tasks and clusters

Templates

For each task, they manually compose ten unique templates that describe the task using natural language instructions.

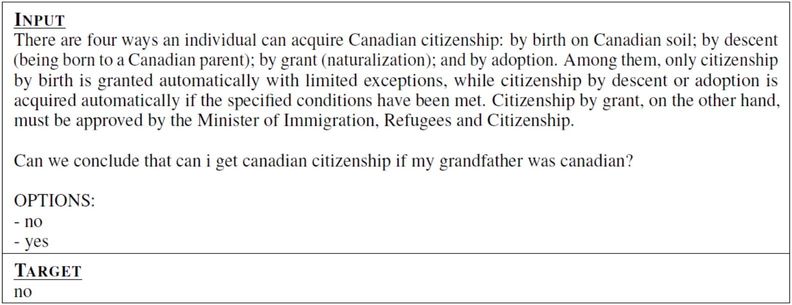


Premise
Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.
Hypothesis
Russians hold the record for the longest stay in space.
Target
Entailment
Not entailment
Options:
- yes
- no
Template 1
<premise>
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, ...

Figure 5: Templates

Input

For classification tasks, they include an options suffix.



INPUT

There are four ways an individual can acquire Canadian citizenship: by birth on Canadian soil; by descent (being born to a Canadian parent); by grant (naturalization); and by adoption. Among them, only citizenship by birth is granted automatically with limited exceptions, while citizenship by descent or adoption is acquired automatically if the specified conditions have been met. Citizenship by grant, on the other hand, must be approved by the Minister of Immigration, Refugees and Citizenship.

Can we conclude that can i get canadian citizenship if my grandfather was canadian?

OPTIONS:
- no
- yes

TARGET

no

Figure 6: Input

Model and Tuning

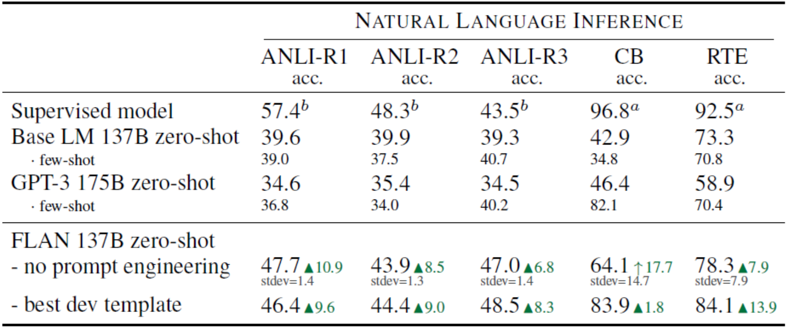
Model:

- Arch: a dense left-to-right, decoder-only transformer language model of 137B parameters
- Data: 2.81T BPE tokens with a vocabulary of 32K tokens (SentencePiece); Approximately 10% is non-English.

Tuning:

- $>$10m: limit the number to 30000
- low-resource: examples-proportional mixing scheme

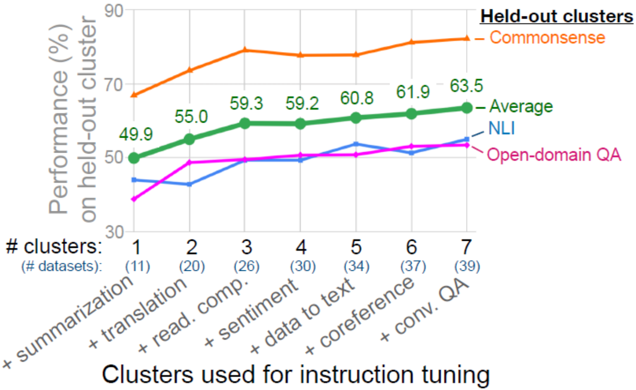
Results



	Natural Language Inference				
	ANLI-R1 acc.	ANLI-R2 acc.	ANLI-R3 acc.	CB acc.	RTE acc.
Supervised model	57.4^b	48.3^b	43.5^b	96.8^a	92.5^a
Base LM 137B zero-shot	39.6	39.9	39.3	42.9	73.3
· few-shot	39.0	37.5	40.7	34.8	70.8
GPT-3 175B zero-shot	34.6	35.4	34.5	46.4	58.9
· few-shot	36.8	34.0	40.2	82.1	70.4
FLAN 137B zero-shot					
- no prompt engineering	47.7 ▲10.9 stdev=1.4	43.9 ▲8.5 stdev=1.3	47.0 ▲6.8 stdev=1.4	64.1 ↑17.7 stdev=14.7	78.3 ▲7.9 stdev=7.9
- best dev template	46.4 ▲9.6	44.4 ▲9.0	48.5 ▲8.3	83.9 ▲1.8	84.1 ▲13.9

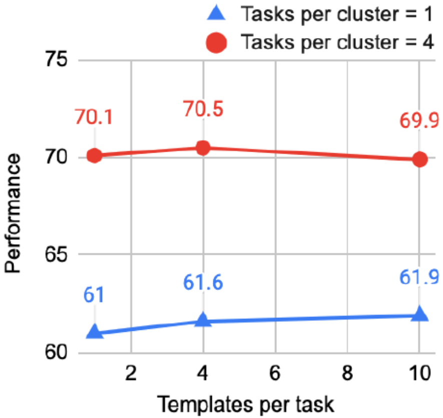
Figure 7: Results on natural language inference.

Cluster



Adding additional task clusters to instruction tuning improves zero-shot performance on held-out task clusters.

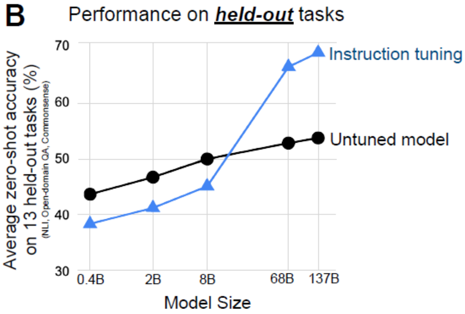
Tasks and Templates



Using more tasks per cluster improved performance.

Using more templates per task, however, had a comparatively negligible effect on performance.

Model Size



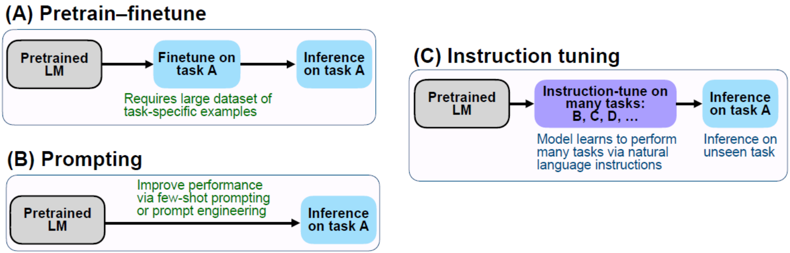
Instruction tuning actually hurts performance on held-out tasks for small-scale models.

Overall

- FLAN substantially improves the performance of its unmodified counterpart and surpasses zero-shot 175B GPT-3 on 19 of 25 tasks.
- FLAN even outperforms few-shot GPT-3 by a large margin on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, and StoryCloze.
- Number of tasks and model scale are key components to the success of instruction tuning.

Comparison

- vs T5 and Prompt learning


(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

- vs GPT-3 in-context learning

Conclusion

This paper has explored instruction tuning.

- They presented FLAN, a 137B parameter language model that performs NLP tasks described using instructions.
- The performance of FLAN compares favorably against both zero-shot and few-shot GPT-3

Bibliography

[1] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” *arXiv preprint arXiv:2109.01652*, 2021.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.