FLAN: Fine-tuned Language Models Are Zero-Shot Learners

Author: Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

Outline

- Introduction to FLAN
- Instruction Tuning
- Tasks and Templates
- Training Datasets
- Performance Evaluation
- Conclusion

Introduction to FLAN

- GPT-3 performed few-shot learning remarkably well
- Zero-shot performance is much worse than few-shot performance on tasks such as
    - Reading comprehension, Question answering and Natural language inference

Without few-shot exemplars, performance drops on **prompts differing from pretraining data**

Finetuned **La**nguage **N**et

Instruction Tuning
FLAN

Decoder only model

137B parameter **LaMDA-PT**
using 60 NLP dadtasets

- Finetuning a Language model on a collection of NLP tasks **describe using instructions**
- FLAN **performs tasks** that it hasn’t been **seen before** via instructions

Introduction to FLAN

- Clustered NLP datasets by task type
- Hold out each cluster for evaluation while instruction tuning FLAN on all other clusters

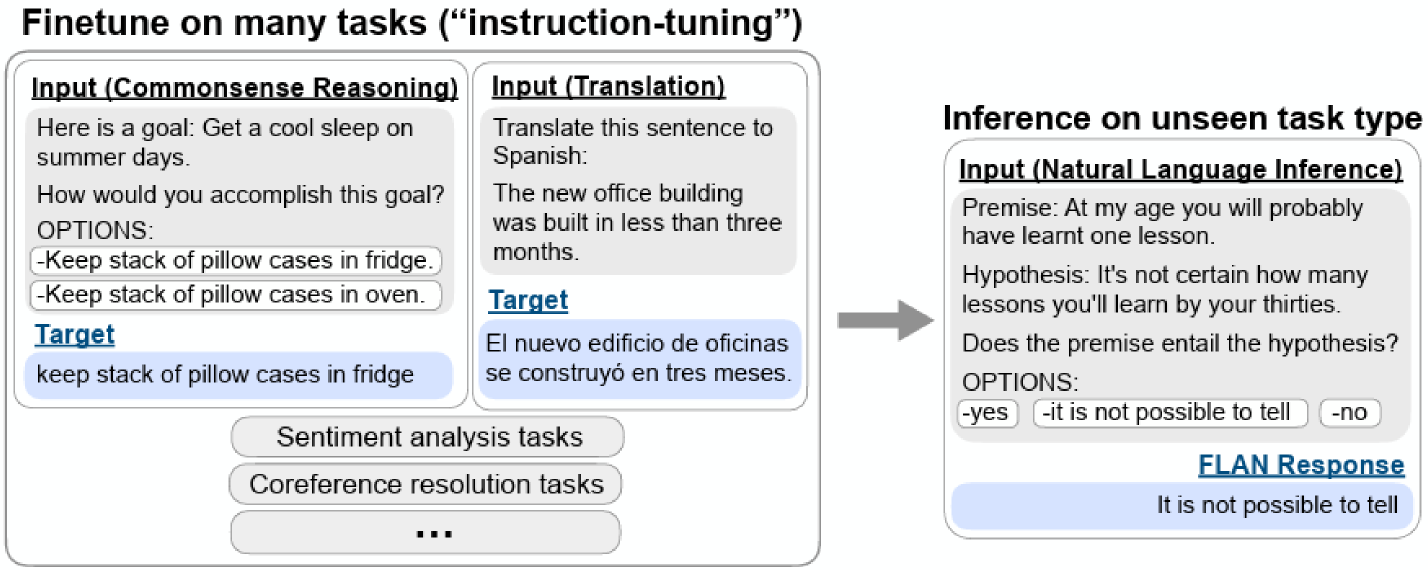
FLAN > 0-shot LaMDA-PT

FLAN > 0-shot 175B GPT-3 on 2-/25 datasets

FLAN > Few-shot GPT-3

FLAN > 0-shot GLaM on 13/19 datasets

FLAN > 1-shot GLaM on 11/19 datasets



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:

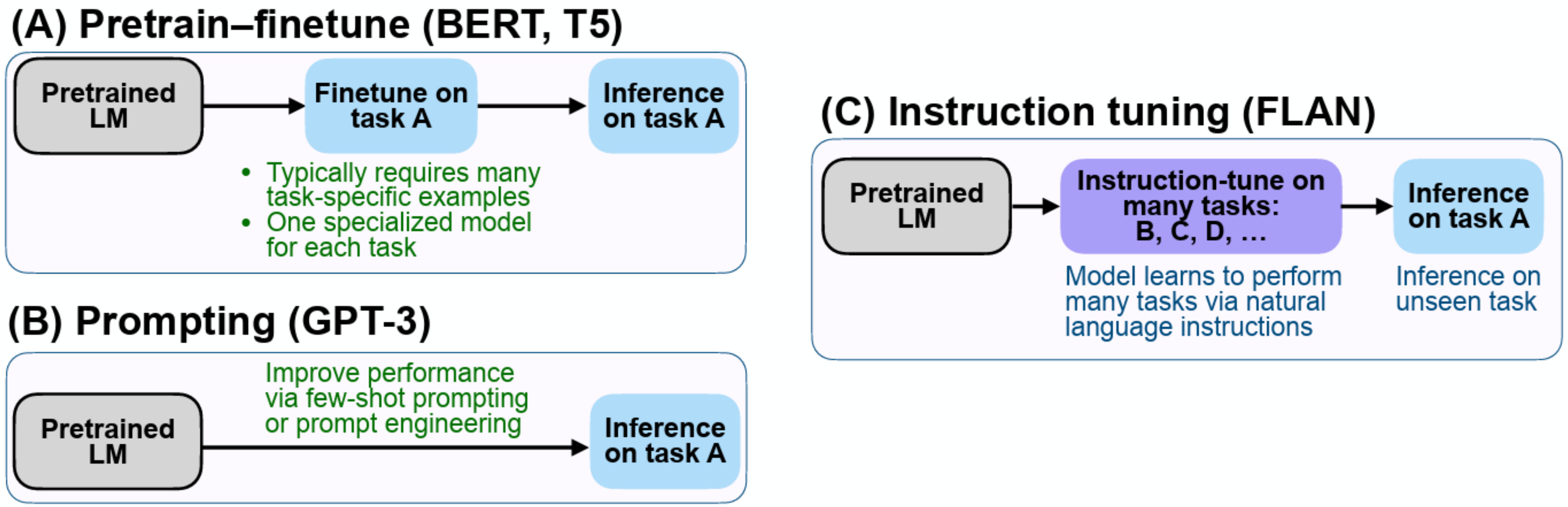
-yes -it is not possible to tell -no

FLAN Response

It is not possible to tell

Instruction Tuning

- Combines strengths of **pretrain–finetune** and **prompting** paradigms
- Utilizes **supervision via finetuning** to enhance model performance
- Improves **language model responses** during inference-time interactions



(A) Pretrain–finetune (BERT, T5)

Pretrained LM → Finetune on task A → Inference on task A

- Typically requires many task-specific examples
- One specialized model for each task

(B) Prompting (GPT-3)

Pretrained LM → Inference on task A

Improve performance via few-shot prompting or prompt engineering

(C) Instruction tuning (FLAN)

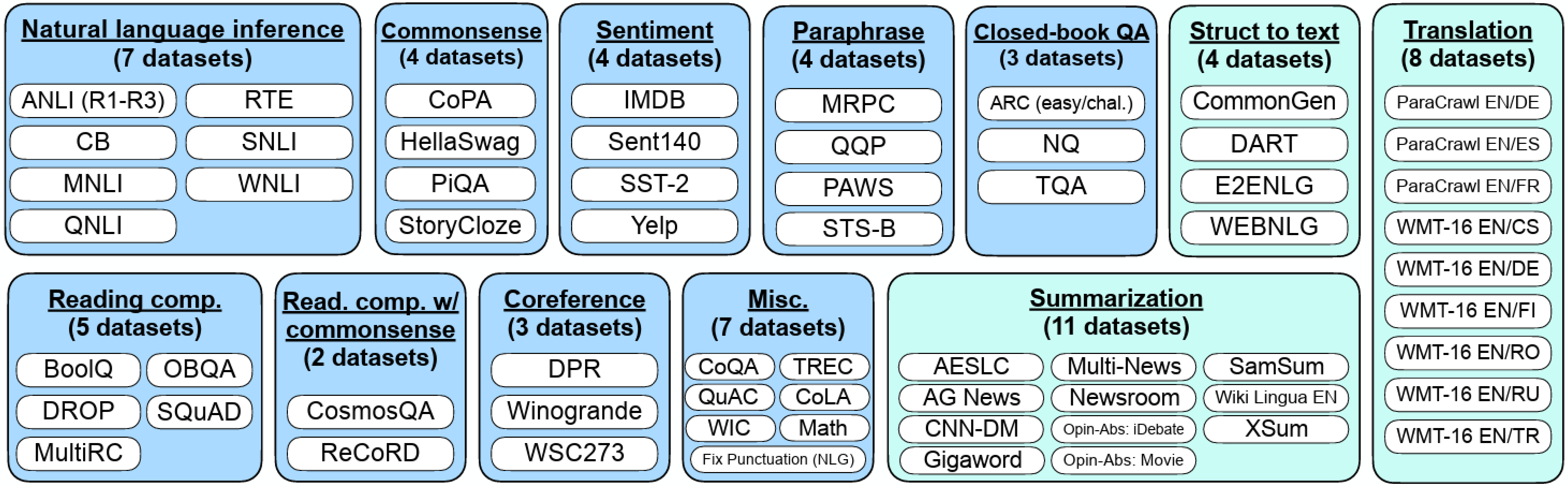
Pretrained LM → Instruction-tune on many tasks: B, C, D, … → Inference on task A

Model learns to perform many tasks via natural language instructions

Inference on unseen task

Tasks and Templates

- 62 Datasets, Each dataset is categorized into one of 12 task clusters (**Blue** = NLU, **Teal**= NLG)
- Datasets in a given cluster belong to the same task type



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

Commonsense (4 datasets)
CoPA, HellaSwag, PiQA, StoryCloze

Sentiment (4 datasets)
IMDB, Sent140, SST-2, Yelp

Paraphrase (4 datasets)
MRPC, QQP, PAWS, STS-B

Closed-book QA (3 datasets)
ARC (easy/chal.), NQ, TQA

Struct to text (4 datasets)
CommonGen, DART, E2ENLG, WEBNLG

Translation (8 datasets)
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 datasets)
BoolQ, OBQA, DROP, SQuAD, MultiRC

Read. comp. w/ commonsense (2 datasets)
CosmosQA, ReCoRD

Coreference (3 datasets)
DPR, Winogrande, WSC273

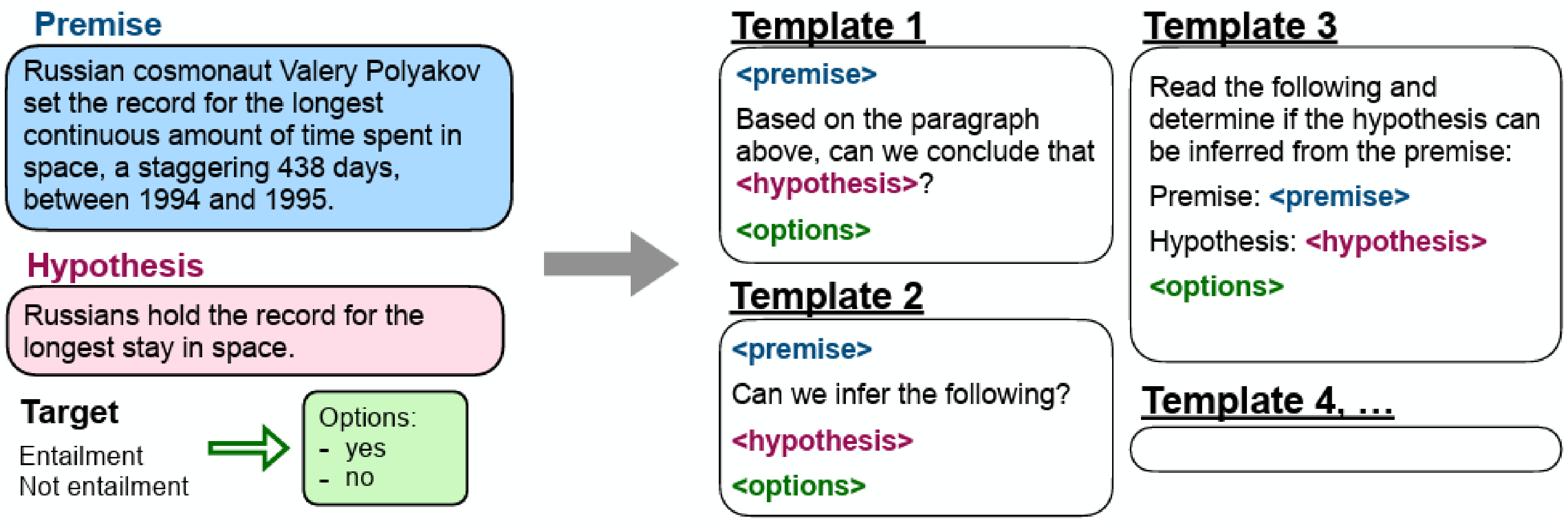
Misc. (7 datasets)
CoQA, TREC, QuAC, CoLA, WIC, Math, Fix Punctuation (NLG)

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

Tasks and Templates

- Each dataset has **10** unique templates using NLI to describe the task
  - Most templates describe the original task to increase diversity
  - ≤ **3** templates modify the task (“**turned the task around**”)

    (e.g., for sentiment classification author includes templates asking to generate a movie review)
- Each dataset's examples are formatted using a **randomly selected instruction template**



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, …

Evaluation Splitting and Classification Options

Evaluation Splits:

- Consider datasets **D unseen at evaluation time** if **no datasets from any task clusters that D belongs to** were seen **during instruction tuning**
- **Example:** Evaluate on NLI tasks while finetuning on translation and sentiment analysis

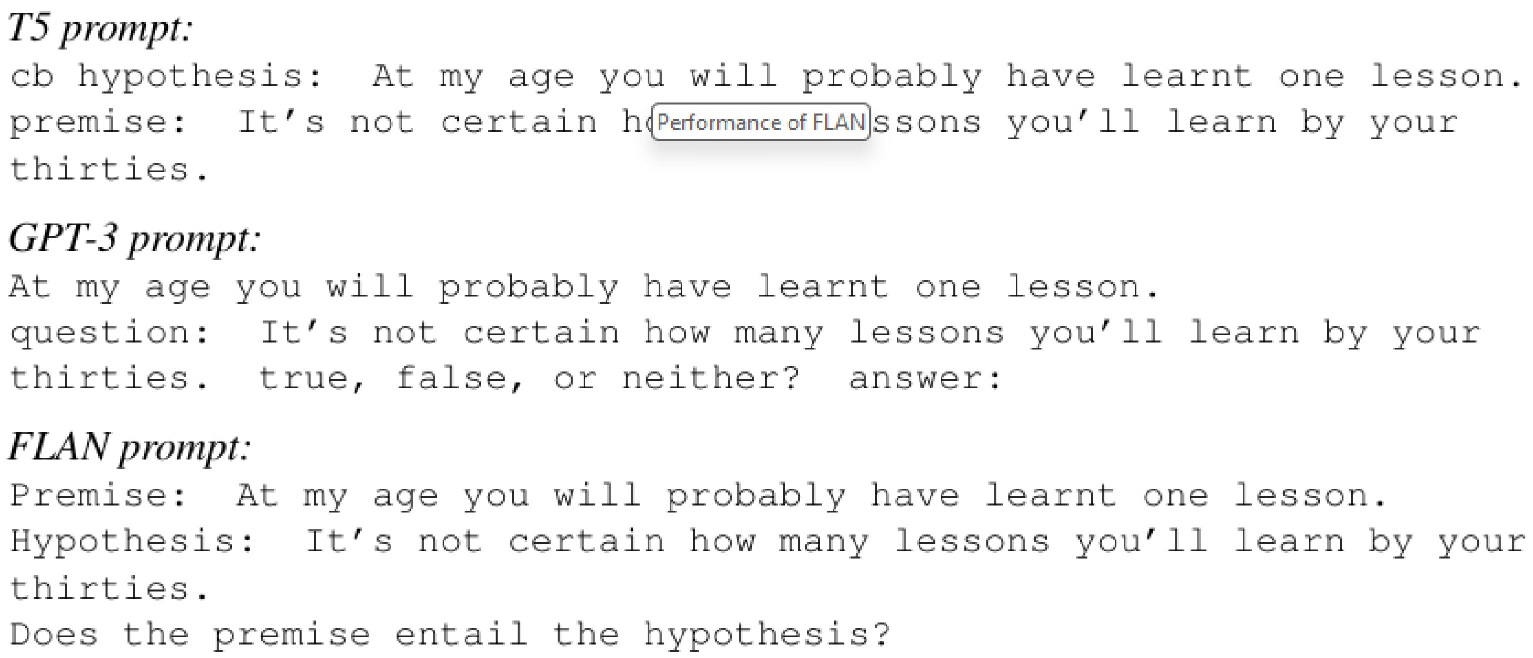
Classification with Options:

- **Classification tasks** used **rank classification approach**
  - Only two outputs ("yes/no") to make the model aware of valid choices
- FLAN improves by explicitly **listing options**
- Makes the model aware of which choices are desired when responding to classification tasks
- **Example:** For NLI, it appends "yes/no/it is not possible to tell."

Training Details

- **Model:** LaMDA-PT, a 137B parameter decoder-only (dense left-to-right) transformer
- **LaMDA-PT** only has **language model pretraining (c.f. LaMDA** finetuned for **dialog)**
- **Pretraining:**
  - 2.49T BPE tokens from web documents, dialog data, and Wikipedia
  - 32k Vocabulary using the SentencePiece library
  - 10% pretraining data was non-English
- **To balance** the different size of datasets: **limit the # of training examples per Dataset to 30k**
- Follows examples-proportional mixing scheme with a mixing rate maximum of 3k
- **Finetuning:** 30k gradient steps, batch size of 8,192 tokens, Adafactor optimizer with learning rate 3e-5
- **Time:** ~60 hours on TPUv3 with 128 cores

Instructions comparison



T5 prompt:

```
cb hypothesis:  At my age you will probably have learnt one lesson.
premise:  It’s not certain h Performance of FLAN ssons you’ll learn by your
thirties.
```

GPT-3 prompt:

```
At my age you will probably have learnt one lesson.
question:  It’s not certain how many lessons you’ll learn by your
thirties.  true, false, or neither?  answer:
```

FLAN prompt:

```
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?
```

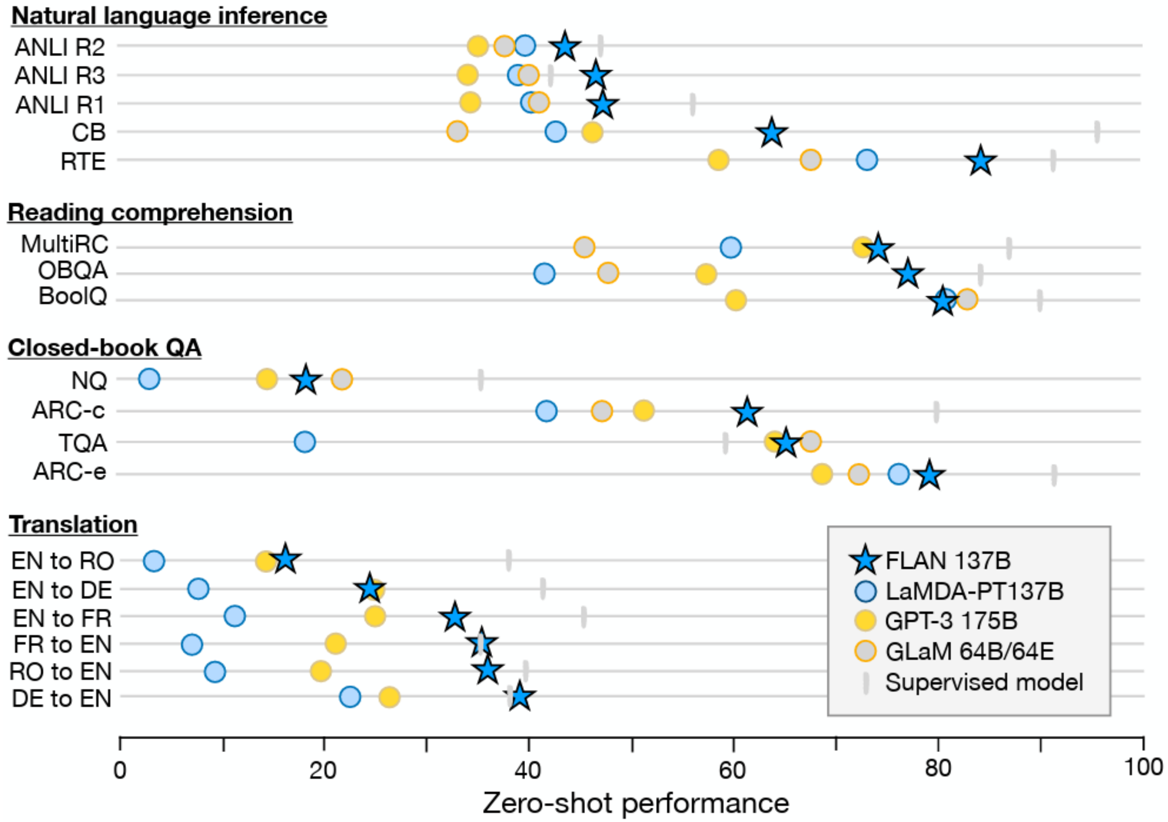
Performance of FLAN

FLAN outperforms GPT-3 on 20 out of 25 datasets

Tasks: NLI, reading comprehension, closed-book QA, translation

Key datasets: ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, StoryCloze

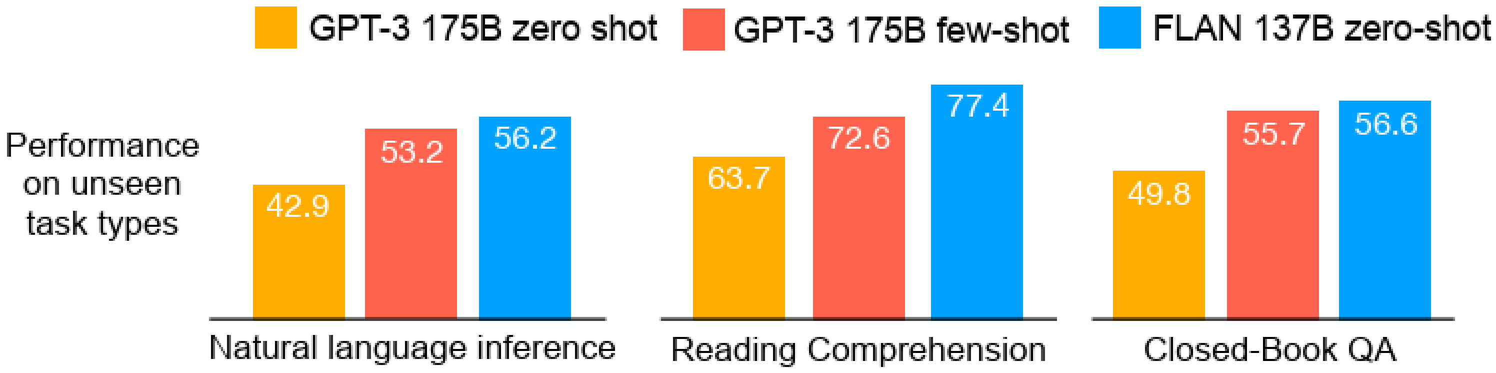
- **Effective** on tasks **naturally verbalized as instructions** (e.g., NLI, QA, translation, struct-to-text)



- **Less effective** on tasks directly formulated as language modeling, where **instructions would be largely redundant** (e.g., commonsense reasoning and coreference resolution tasks that are formatted as finishing an incomplete sentence or paragraph)

Performance of FLAN

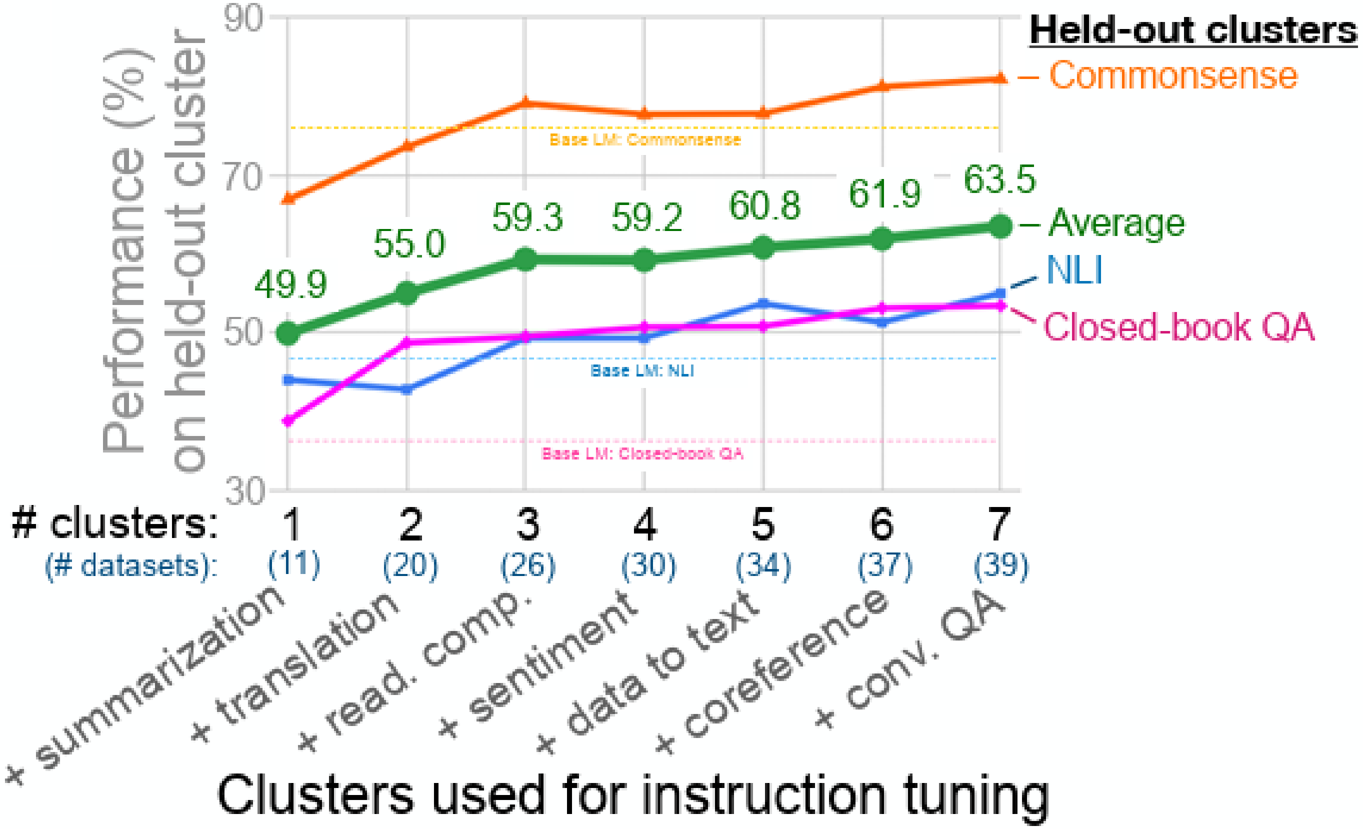
- Improved performance than **zero-shot 175 GBT-3** on 20 of 25 datasets
- Outperforms **few-shot GPT-3** on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA and StoryCloze
- Ablation studies reveal the key to the success of instruction tuning
  - Number of finetuning datasets
  - model scale
  - natural language instructions



Performance of FLAN

Number of Instruction Tuning Clusters:

- NLI, closed-book QA, and commonsense reasoning as **evaluation clusters**
- Performance does not appear to saturate
- Further improvement is possible with even more clusters added to instruction tuning



Which instruction tuning cluster **contributes** the most to each evaluation cluster **???**

THE UNIVERSITY OF ARIZONA

Performance of FLAN

Scaling Laws:

- NLI, closed-book QA, and commonsense reasoning as evaluation clusters
- Evaluate the effect of instruction tuning on models of size 422M, 2B, 8B, 68B, and 137B parameters

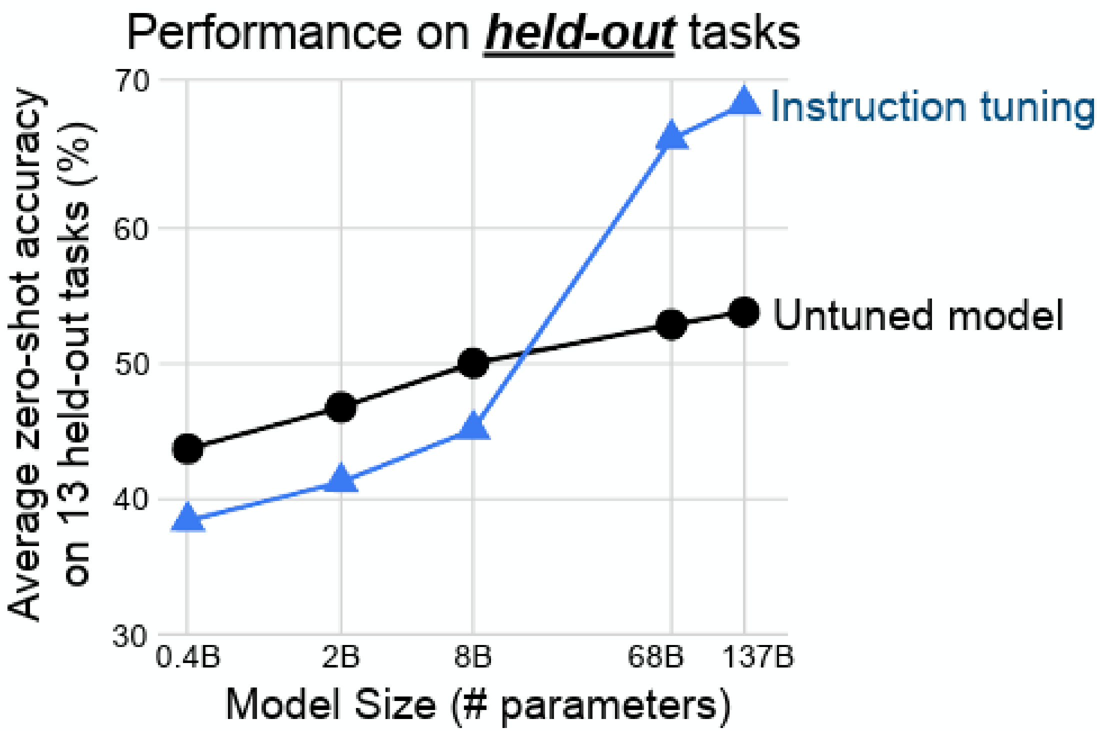
Small-scale models:

- **Learning 40 tasks** used during instruction tuning-

  Fills the entire **model capacity**
- Causing these models to perform worse on new tasks

Larger scale models:

- Instruction tuning fills up some model capacity but also **teaches how to follow instructions**
- Allowing them to **generalize to new tasks with the remaining capacity**



Performance of FLAN

Role of Instructions:

- Performance gains come entirely from multi-task fine-tuning and the model could perform just as well without instructions

Two finetuning setups without instructions

No Template: Model given only inputs and outputs without instructions

Input: "The dog runs." **Output:** "Le chien court."

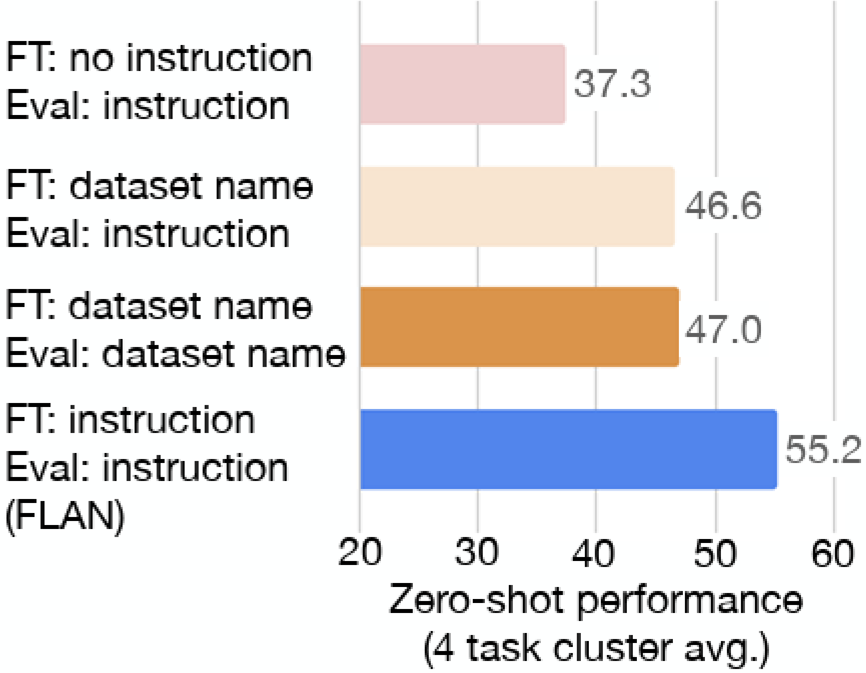
Dataset Name: Model given task and dataset name as input

Each input is prepended with the **name of** the **task** and **dataset** (e.g., for translation to French, the input would be

"[Translation: WMT'14 to French] The dog runs.")

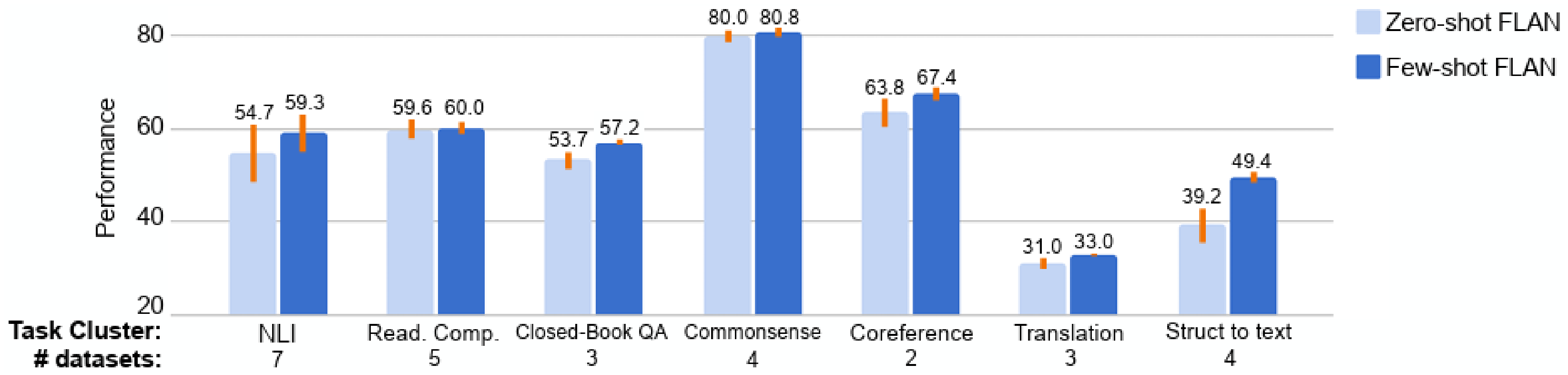
FLAN's finetuning procedure: Used natural instructions

(e.g., "Please translate this sentence to French: 'The dog runs.")



Performance of FLAN

Instruction with Few-Shot Exemplars:

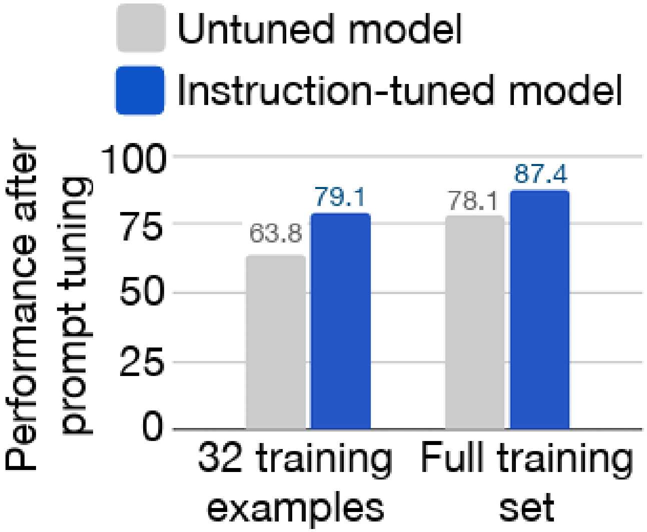


- Few-shot exemplars improve help the model better understands the output format
- Especially **effective** for **tasks with large/complex output spaces**

  (e.g., translation, struct-to-text and closed-book QA)
- For all task clusters, **standard deviation** among templates is **lower for few-shot FLAN**, indicating **reduced sensitivity to prompt engineering**

Performance of FLAN

Instruction Tuning Facilitates Prompt Tuning:



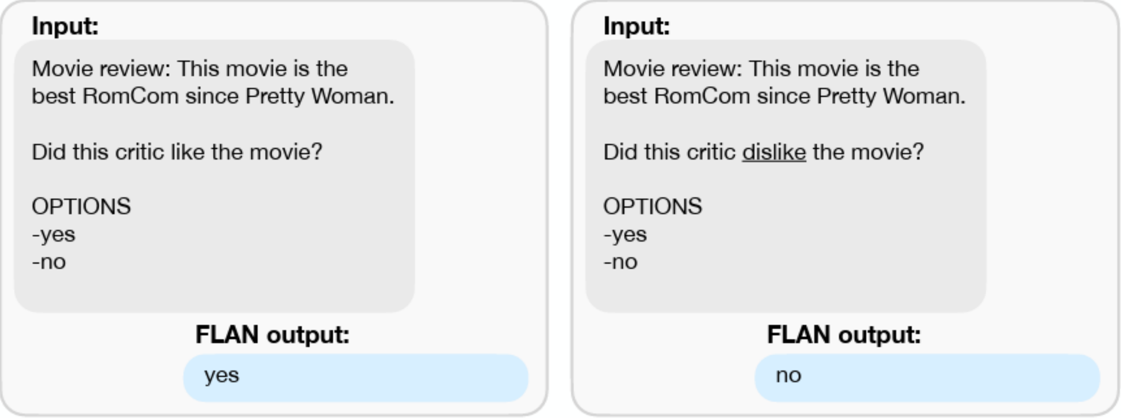
- Outperforms **LaMDA-PT** when performing inference using soft prompts

  (Prepended continuous variables optimized via prompt tuning)

- Train with continuous prompts for each of the **SuperGLUE tasks**

- In **low-resource settings**, FLAN even achieves more than 10% improvement

FLAN Output Examples



Input:

Movie review: This movie is the best RomCom since Pretty Woman.

Did this critic like the movie?

OPTIONS
-yes
-no

FLAN output:

yes

Input:

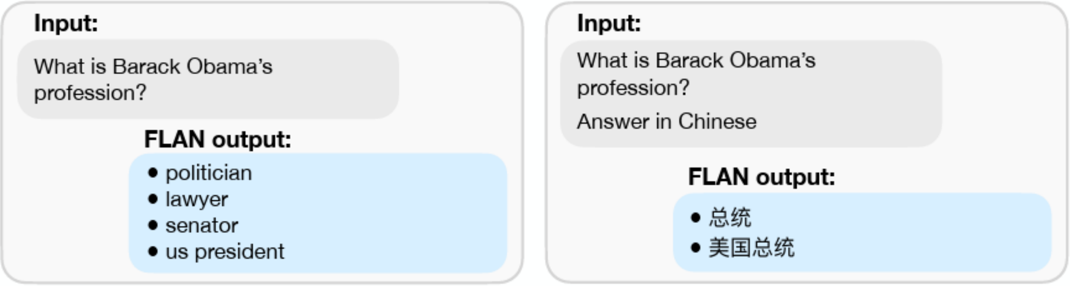
Movie review: This movie is the best RomCom since Pretty Woman.

Did this critic dislike the movie?

OPTIONS
-yes
-no

FLAN output:

no



Input:

What is Barack Obama's profession?

FLAN output:

- politician
- lawyer
- senator
- us president

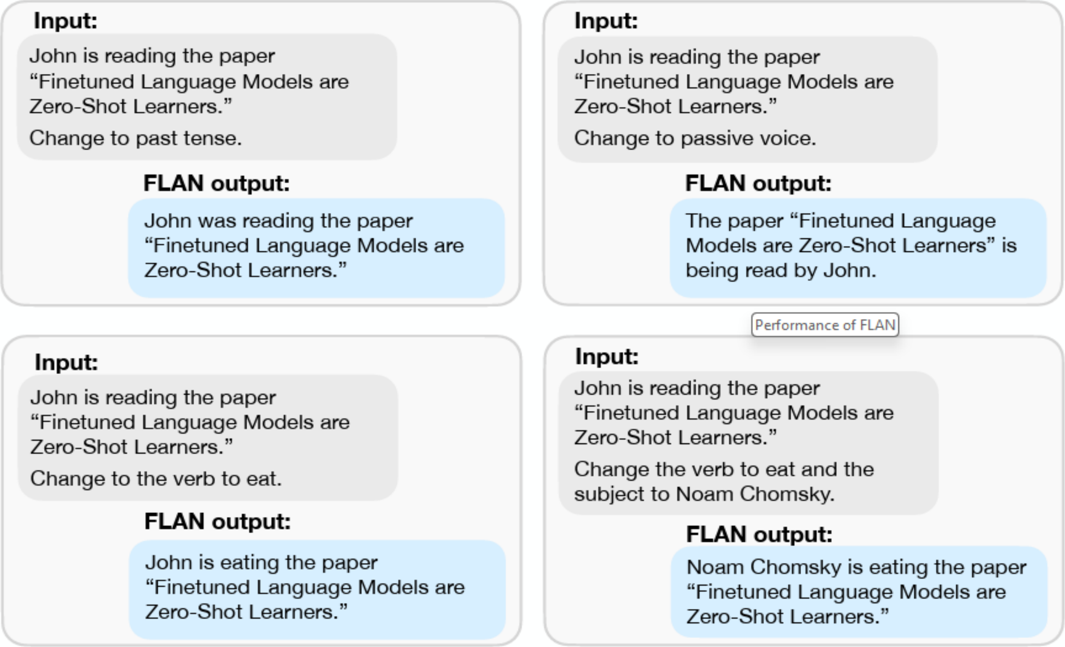
Input:

What is Barack Obama's profession?

Answer in Chinese

FLAN output:

- 总统
- 美国总统



Input:

John is reading the paper "Finetuned Language Models are Zero-Shot Learners."

Change to past tense.

FLAN output:

John was reading the paper "Finetuned Language Models are Zero-Shot Learners."

Input:

John is reading the paper "Finetuned Language Models are Zero-Shot Learners."

Change to passive voice.

FLAN output:

The paper "Finetuned Language Models are Zero-Shot Learners" is being read by John.

Performance of FLAN

Input:

John is reading the paper "Finetuned Language Models are Zero-Shot Learners."

Change to the verb to eat.

FLAN output:

John is eating the paper "Finetuned Language Models are Zero-Shot Learners."

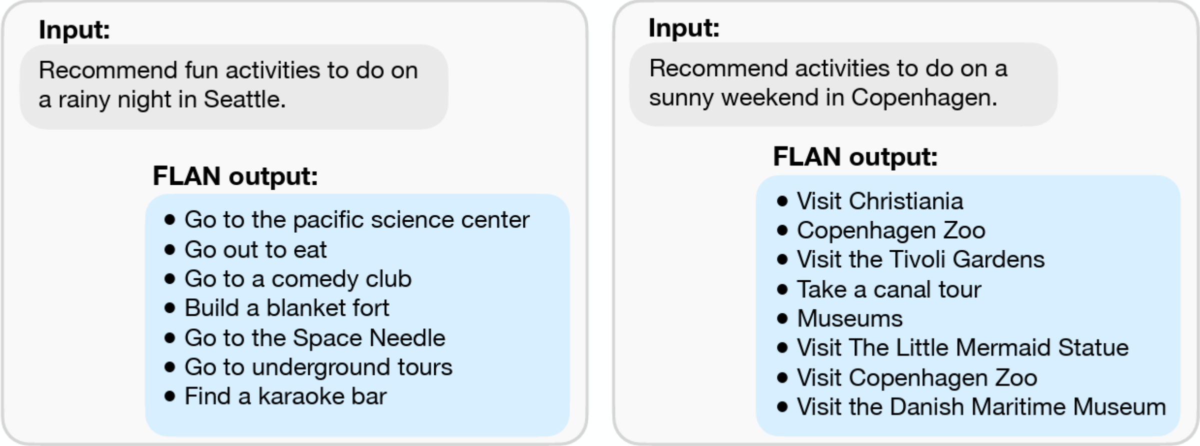
Input:

John is reading the paper "Finetuned Language Models are Zero-Shot Learners."

Change the verb to eat and the subject to Noam Chomsky.

FLAN output:

Noam Chomsky is eating the paper "Finetuned Language Models are Zero-Shot Learners."



Input:

Recommend fun activities to do on a rainy night in Seattle.

FLAN output:

- Go to the pacific science center
- Go out to eat
- Go to a comedy club
- Build a blanket fort
- Go to the Space Needle
- Go to underground tours
- Find a karaoke bar

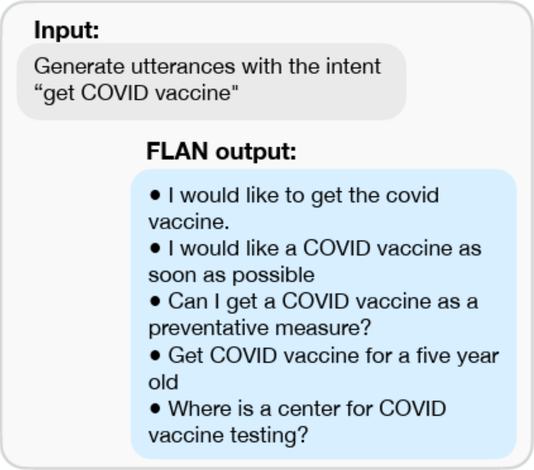
Input:

Recommend activities to do on a sunny weekend in Copenhagen.

FLAN output:

- Visit Christiania
- Copenhagen Zoo
- Visit the Tivoli Gardens
- Take a canal tour
- Museums
- Visit The Little Mermaid Statue
- Visit Copenhagen Zoo
- Visit the Danish Maritime Museum

FLAN Output Examples

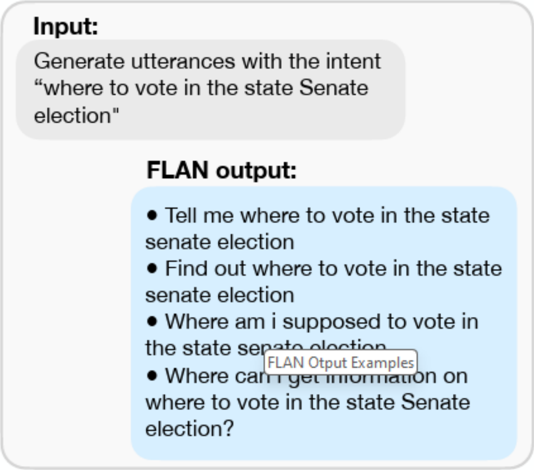


Input:

Generate utterances with the intent "get COVID vaccine"

FLAN output:

- I would like to get the covid vaccine.
- I would like a COVID vaccine as soon as possible
- Can I get a COVID vaccine as a preventative measure?
- Get COVID vaccine for a five year old
- Where is a center for COVID vaccine testing?

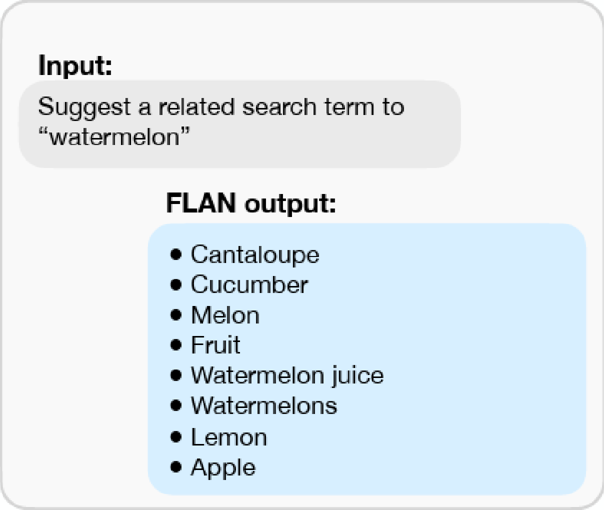


Input:

Generate utterances with the intent "where to vote in the state Senate election"

FLAN output:

- Tell me where to vote in the state senate election
- Find out where to vote in the state senate election
- Where am i supposed to vote in the state senate election
- Where can I get information on where to vote in the state Senate election?

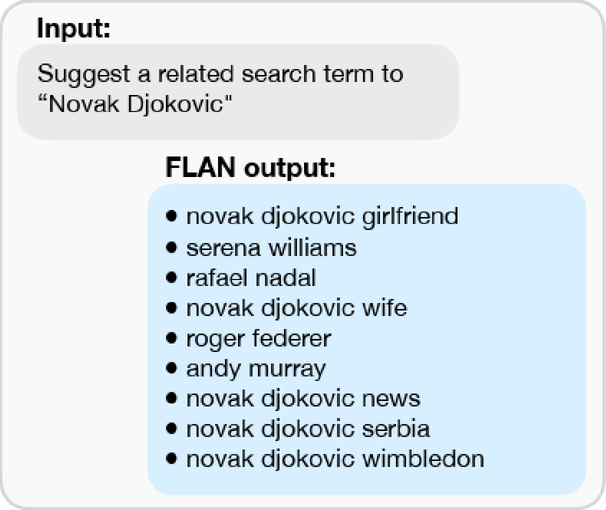


Input:

Suggest a related search term to "watermelon"

FLAN output:

- Cantaloupe
- Cucumber
- Melon
- Fruit
- Watermelon juice
- Watermelons
- Lemon
- Apple

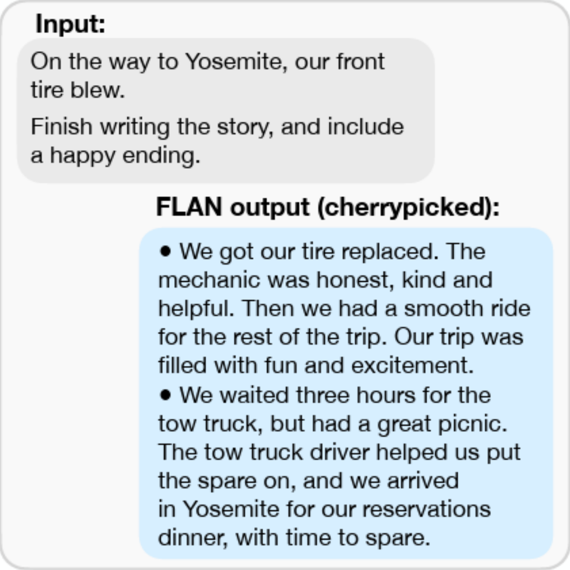


Input:

Suggest a related search term to "Novak Djokovic"

FLAN output:

- novak djokovic girlfriend
- serena williams
- rafael nadal
- novak djokovic wife
- roger federer
- andy murray
- novak djokovic news
- novak djokovic serbia
- novak djokovic wimbledon



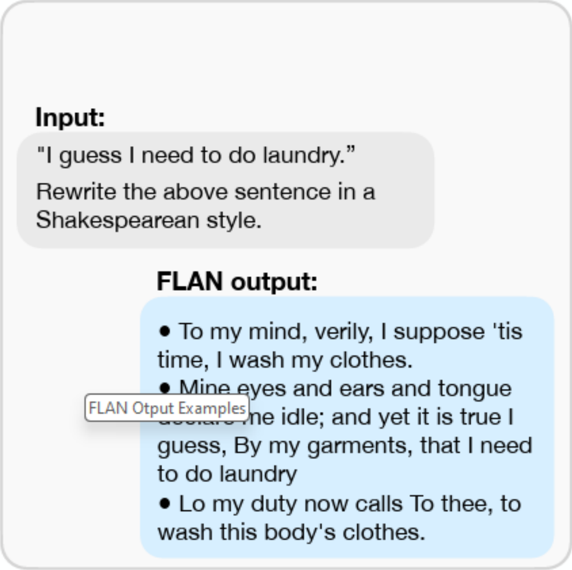
Input:

On the way to Yosemite, our front tire blew.

Finish writing the story, and include a happy ending.

FLAN output (cherrypicked):

- We got our tire replaced. The mechanic was honest, kind and helpful. Then we had a smooth ride for the rest of the trip. Our trip was filled with fun and excitement.
- We waited three hours for the tow truck, but had a great picnic. The tow truck driver helped us put the spare on, and we arrived in Yosemite for our reservations dinner, with time to spare.



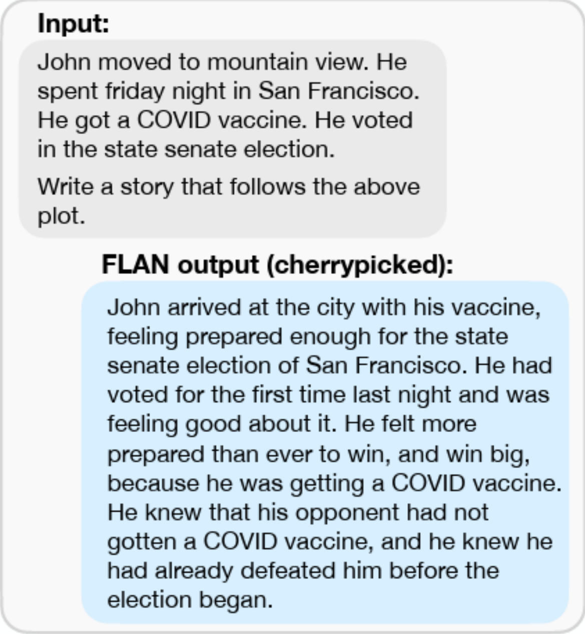
Input:

"I guess I need to do laundry."

Rewrite the above sentence in a Shakespearean style.

FLAN output:

- To my mind, verily, I suppose 'tis time, I wash my clothes.
- Mine eyes and ears and tongue declare me idle; and yet it is true I guess, By my garments, that I need to do laundry
- Lo my duty now calls To thee, to wash this body's clothes.



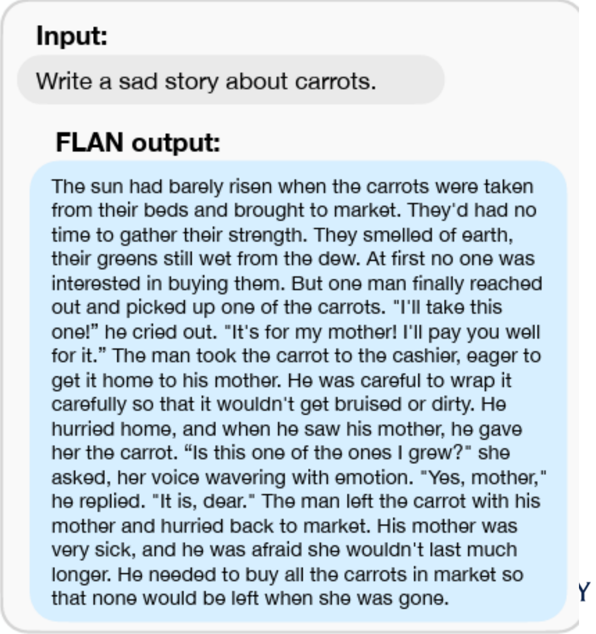
Input:

John moved to mountain view. He spent friday night in San Francisco. He got a COVID vaccine. He voted in the state senate election.

Write a story that follows the above plot.

FLAN output (cherrypicked):

John arrived at the city with his vaccine, feeling prepared enough for the state senate election of San Francisco. He had voted for the first time last night and was feeling good about it. He felt more prepared than ever to win, and win big, because he was getting a COVID vaccine. He knew that his opponent had not gotten a COVID vaccine, and he knew he had already defeated him before the election began.



Input:

Write a sad story about carrots.

FLAN output:

The sun had barely risen when the carrots were taken from their beds and brought to market. They'd had no time to gather their strength. They smelled of earth, their greens still wet from the dew. At first no one was interested in buying them. But one man finally reached out and picked up one of the carrots. "I'll take this one!" he cried out. "It's for my mother! I'll pay you well for it." The man took the carrot to the cashier, eager to get it home to his mother. He was careful to wrap it carefully so that it wouldn't get bruised or dirty. He hurried home, and when he saw his mother, he gave her the carrot. "Is this one of the ones I grew?" she asked, her voice wavering with emotion. "Yes, mother," he replied. "It is, dear." The man left the carrot with his mother and hurried back to market. His mother was very sick, and he was afraid she wouldn't last much longer. He needed to buy all the carrots in market so that none would be left when she was gone.

Conclusions

- Instruction tuning significantly boosts zero-shot performance on unseen tasks
- FLAN 137B outperforms GPT-3 on most tasks, proving the value of task-specific finetuning.
- Model scale is critical: benefits emerge only with large models (e.g., 137B parameters).

Limitations:

- FLAN uses relatively short instructions (typically one sentence), unlike detailed instructions given to human workers
- Individual examples might have appeared in the models' pretraining data, which includes web documents, though in post-hoc analysis showed that data overlap substantially impacted the results
- Scale of FLAN 137B makes it costly to serve