Finetuned Language Models Are Zero-Shot Learners

J Wei*, M Bosma*, V Y. Zhao*, K Guu*, A Wei Yu, B Lester, N Du, A M. Dai, Q V. Le
http://arxiv.org/abs/2109.01652

Presented by: Gül Sena Altıntaş
February 14, 2025

Motivation

Landscape so far (until early 2022)

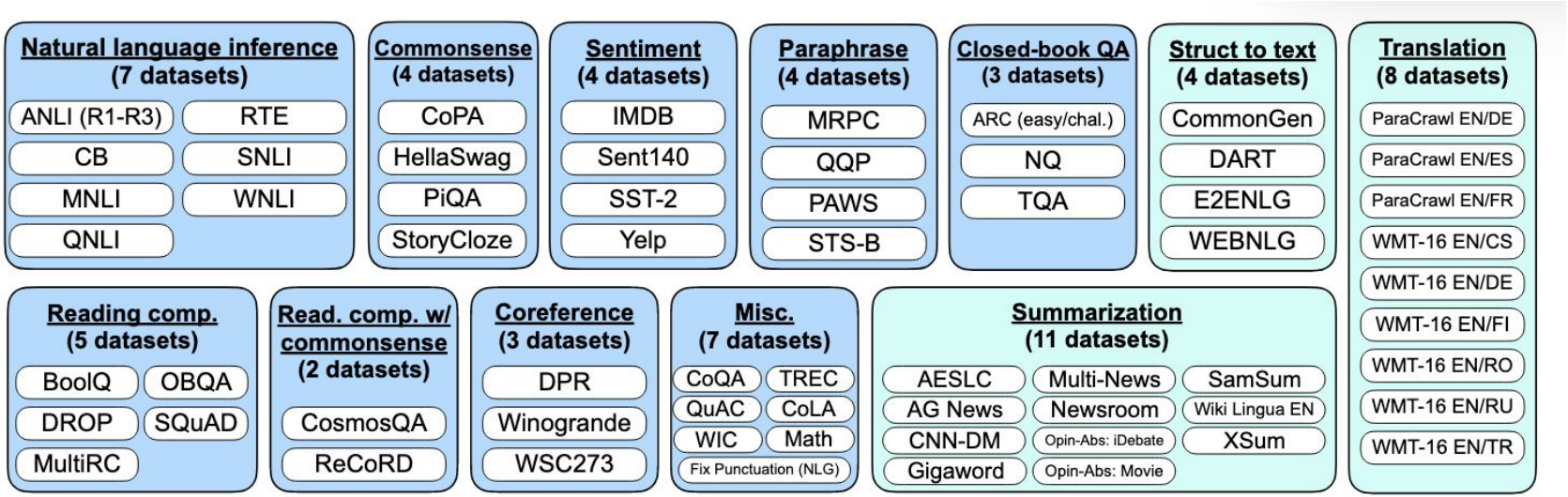
- Large language models excel at few-shot learning (Brown et al., 2020)
- Vanilla zero-shot performance still suffers, and we need to employ alternative prompting techniques
- Main problem: Zero-shot prompts often don't match the pretraining format

Key research question:
Can we improve zero-shot performance on *unseen tasks* by fine-tuning models to follow natural language *instructions*?

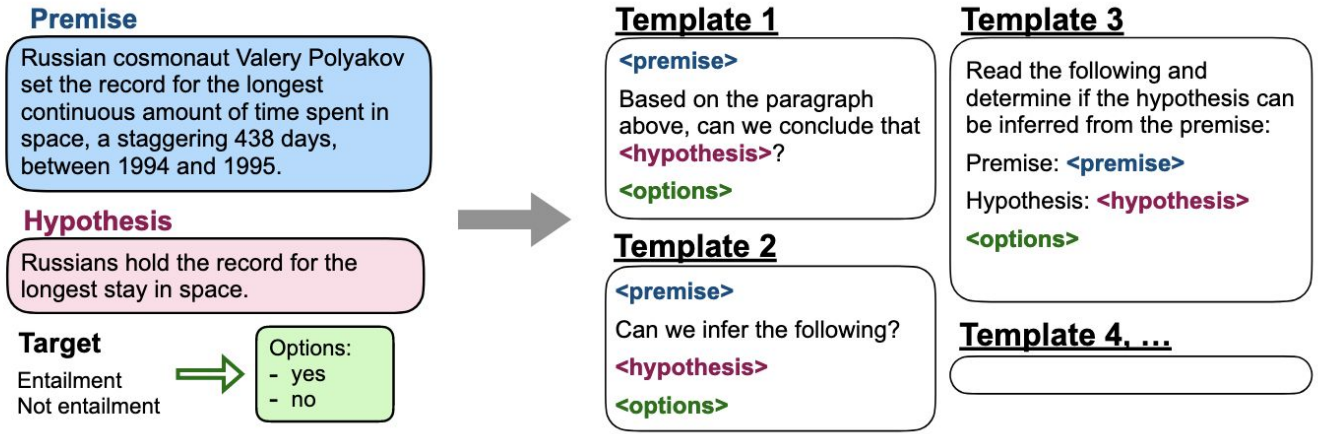
Instruction Tuning (IF). Also see P3/T0(Sanh et al., 2021), xP3 (Muennighoff et al., 2022), Super Natural Instructions (Wang et al., 2022f), LIMA (Zhou et al., 2023a), Dolly (Conover et al., 2023a)

FLAN (Finetuned Language Net) - The Recipe

Base Model: 137B parameter LaMDA-PT, pre-trained on web documents, dialog data, and Wikipedia



Instruction Templates



- Natural language instructions
- 10 unique templates per dataset
- When evaluating on a task type, no tasks from that cluster are used in training

Results

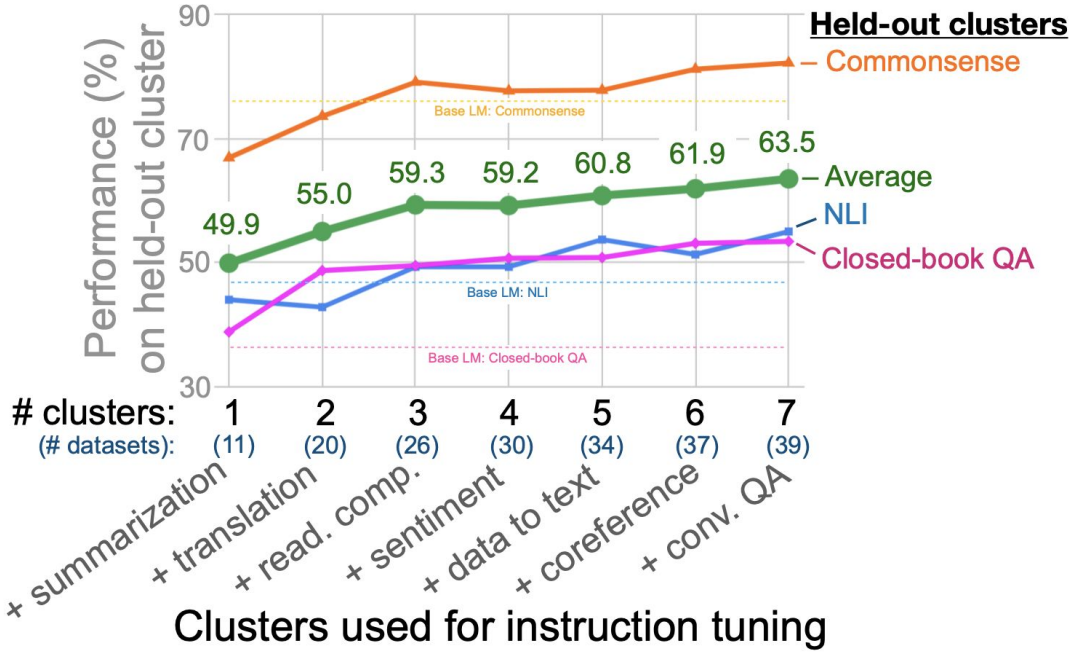
Zero-shot performance of FLAN compared to other approaches

Take-aways:

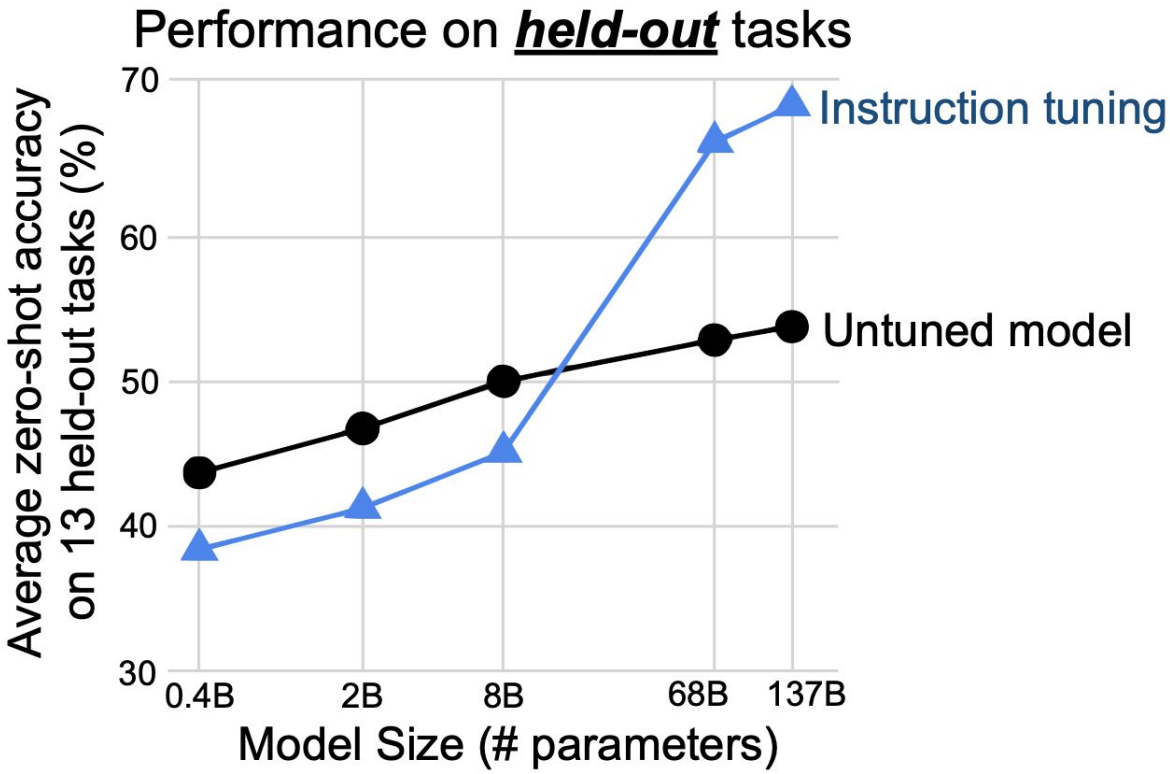
- Strong gains in NLI and Reading Comprehension
- Particularly effective for tasks that don't naturally occur in pre-training data, e.g. NLI
- Competitive in translation despite English-centric training, but underperforms few-shot GPT-3
- Less effective when task matches original LM objective
-

Zero-shot FLAN outperforms the base model and GPT-3

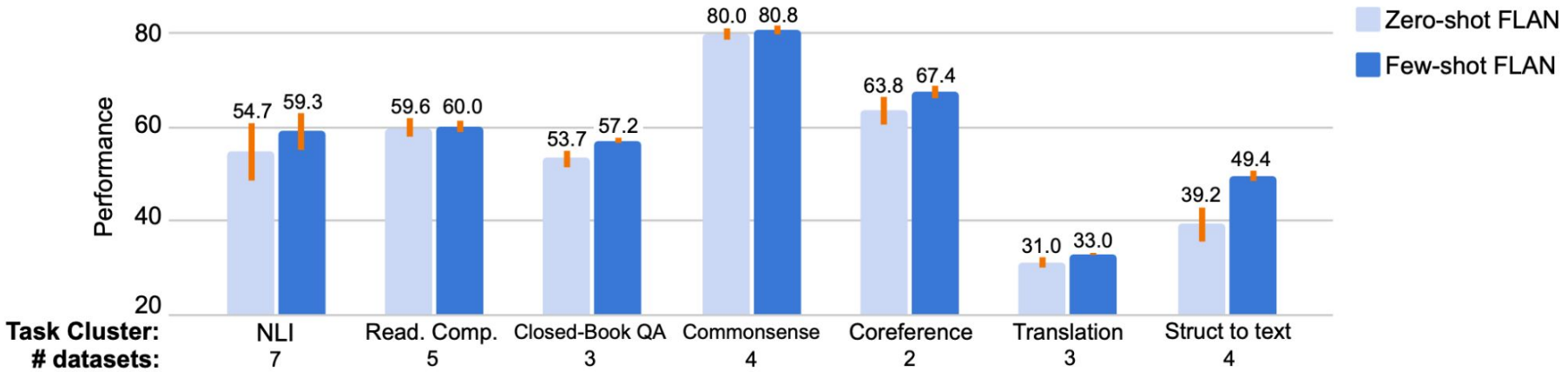
More tasks in training = Better performance on held-out (and new) tasks



Instruction tuning yields big gains on large models



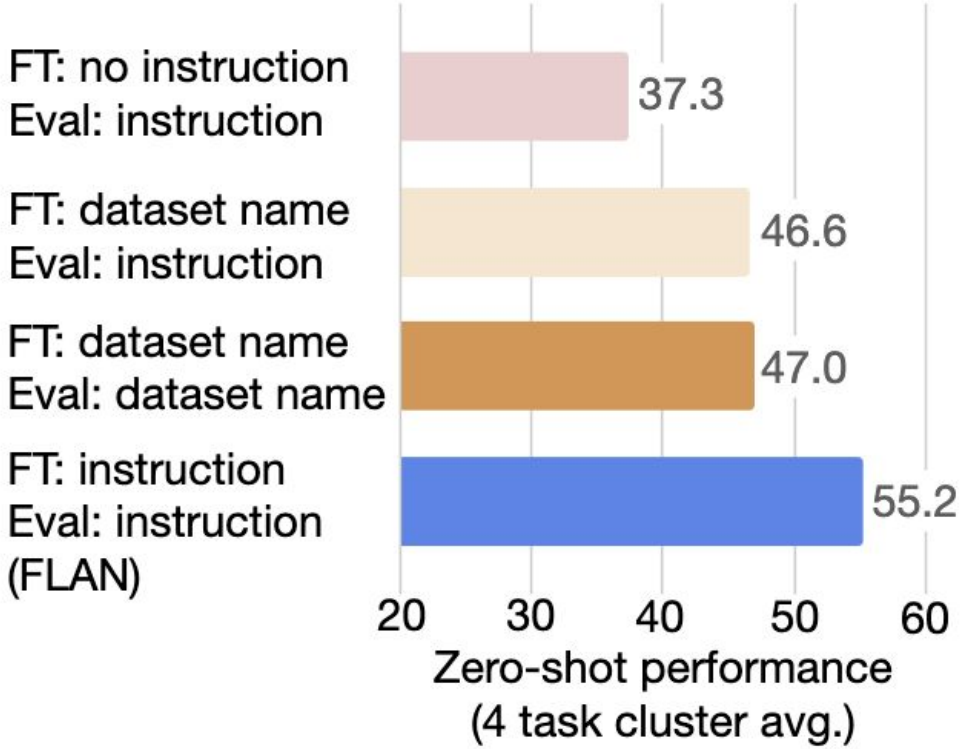
Adding few-shot exemplars further improves FLAN



Code

- https://drive.google.com/file/d/1rcKU8xnp0VHTVzM_LSNLk-xMESUSxgD5/view?usp=sharing
  - Scaling performance FLAN-T5
  - Zero-shot vs. Few-shot performance

Instruction fine-tuning is actually important



No Instruction: Input-Output
Dataset Name: *[Translation: WMT'14 to French] The dog runs.*
FLAN: Please translate this sentence to French: 'The dog runs.'

Training with instructions is important for **zero-shot** performance.

Limitations & Summary

- Only works at large scale (>68B parameters)
- Not effective when task matches language modeling
- Instructions were simple, single-sentence (vs modern complex prompts)
- English-centric performance
- No alignment considerations

Impact & Evolution

What FLAN Showed:

- Instruction tuning works (for zero-shot performance)
- More tasks → better cross-task generalization

How Field Evolved:

- Complex instructions
- RLHF & alignment
- More scaling in models & tasks