

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Google Research

<https://github.com/google-research/flan>

kimminhyun@comcom.ai

Quick recap - zero-shot ~ N-shot

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

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

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

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

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1 sea otter => loutre de mer ← example #1
↓
gradient update
↓
1 peppermint => menthe poivrée ← example #2
↓
gradient update
↓
...
↓
1 plush giraffe => girafe peluche ← example #N
↓
gradient update
↓
1 cheese => ..... ← prompt
```

Abstract

- instruction tuning: finetuning language models on a collection of tasks described via instructions
- FLAN: instruction-tuned model
- Better than 175B GPT-3 on 19 of 25 tasks

Introduction

- 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**
- *"Is the sentiment of this movie review positive or negative?"*
- *"Translate 'how are you' into Chinese."*
- instruction tuning—finetuning the model on a mixture of more than **60 NLP tasks** expressed via natural language instructions.
- **Finetuned LAnguage Net: FLAN**

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.

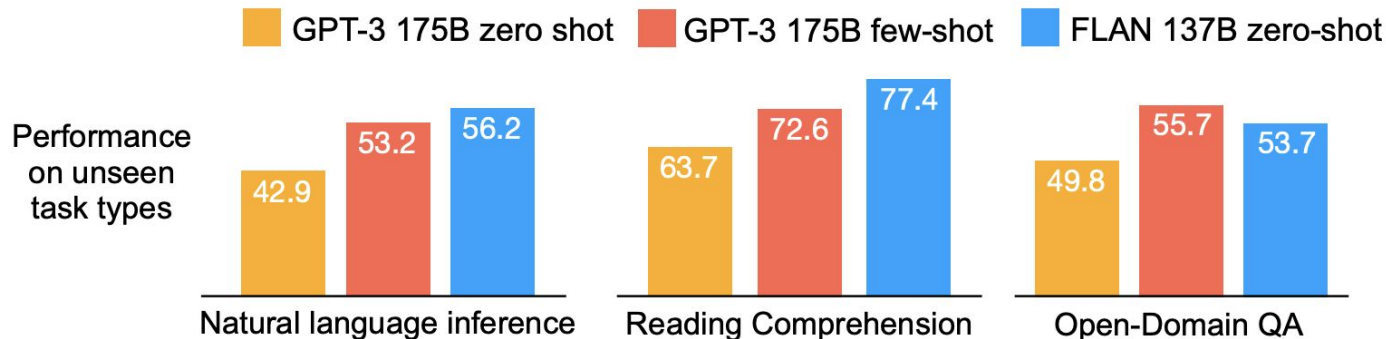
Does the premise entail the hypothesis?

OPTIONS:

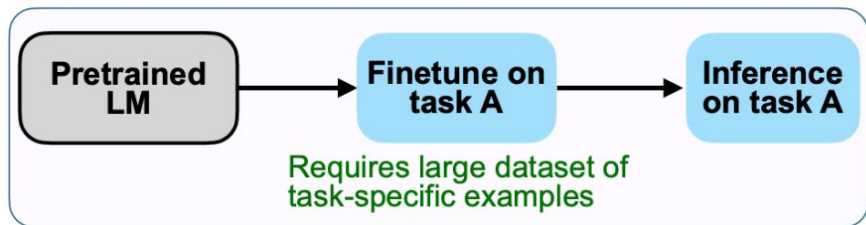
-yes -it is not possible to tell -no

FLAN Response

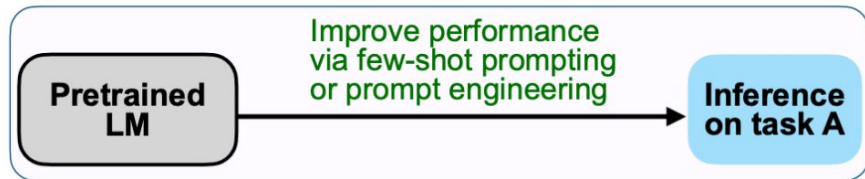
It is not possible to tell



(A) Pretrain–finetune



(B) Prompting



(C) Instruction tuning

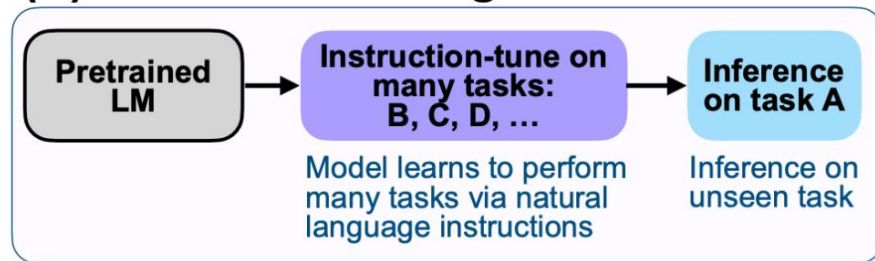


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

FLAN: INSTRUCTION TUNING IMPROVES ZERO-SHOT LEARNING

- To evaluate the model's performance on unseen tasks, we group tasks into clusters by task type and hold out each task cluster for evaluation while instruction tuning on all remaining clusters.

-

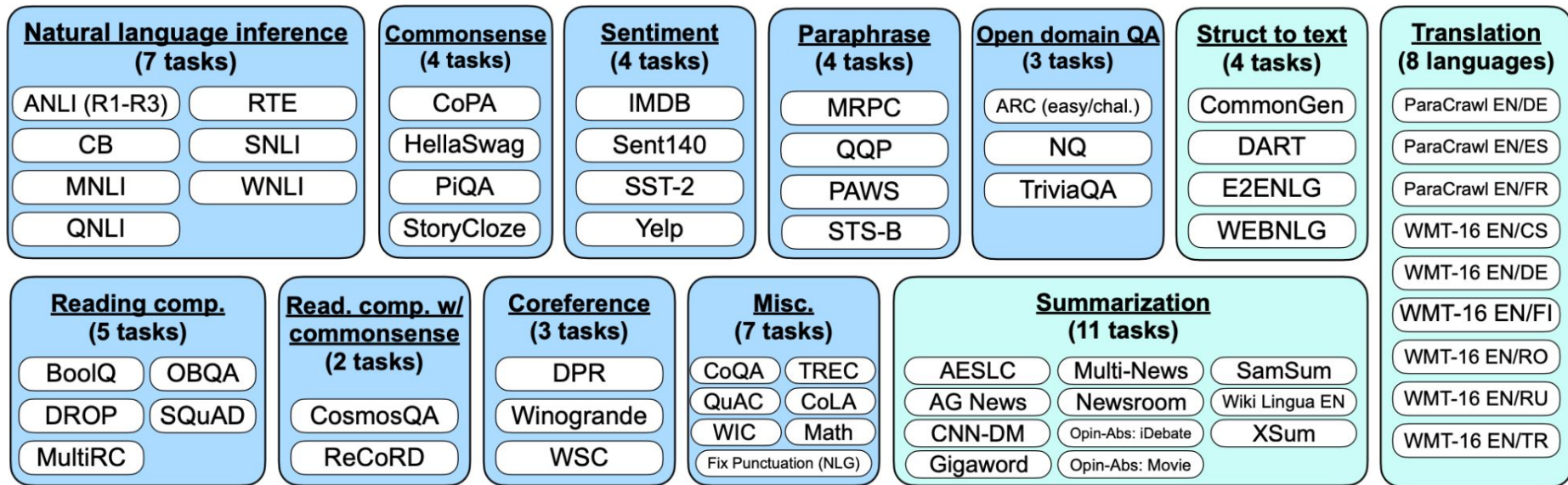


Figure 3: Task clusters used in this paper (NLU tasks in blue; NLG tasks in teal).

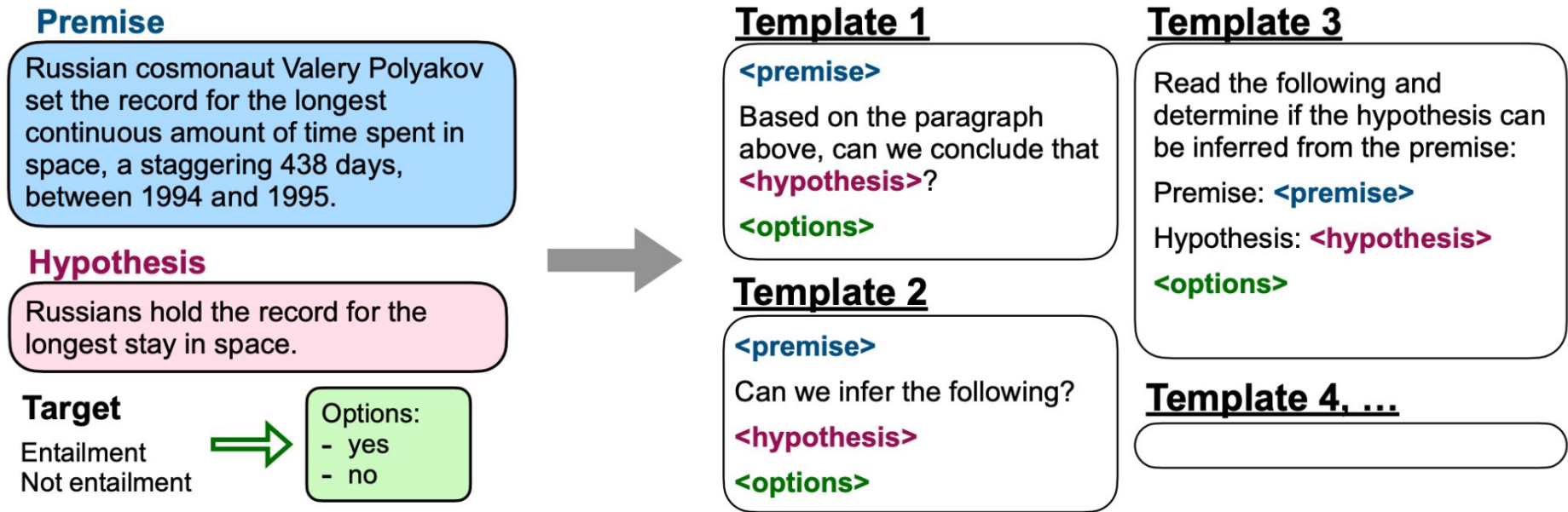


Figure 4: Multiple instruction templates describing a natural language inference task.

CLASSIFICATION WITH OPTIONS

append the token OPTIONS to the end of a classification task along with a list of the output classes for that task. This makes the model aware of which choices are desired when responding to classification tasks.

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

TRAINING DETAILS

- Model architecture and pretraining

dense left-to-right, decoder-only transformer language model of 137B parameters

pretrained on a collection of web documents (including those with computer code), dialog data, and Wikipedia

2.81T BPE tokens with a vocabulary of 32K tokens using the SentencePiece library

10% of the pretraining data was non-Engl

not as clean as the GPT-3 training set and also has a mixture of dialog and code -> *Base LM*

TRAINING DETAILS

- Instruction tuning procedure

Translation: ten million training examples

CommitmentBank: 250

examples-proportional mixing scheme: ([code](#))

$$w_i = \min(n_i, 3000) / \sum_{j \in \mathcal{D}} \min(n_j, 3000),$$

30,000 gradient updates at a batch size of 8,192 / Adafactor Optimizer / learning rate 3e-5

Input sequence length: 1024 / target sequence length: 256

Packing: multiple training examples into a single sequence, separating inputs from targets using a special end-of-sequence token

Result: Natural Language Inference

“Does <premise> mean that <hypothesis>?”

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction	The man is sleeping.
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.

Result: Natural Language Inference













	NATURAL LANGUAGE INFERENCE				
	ANLI-R1	ANLI-R2	ANLI-R3	CB	RTE
	acc.	acc.	acc.	acc.	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

Table 1: Results on natural language inference. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. The triangle  indicates improvement over few-shot GPT-3. The up-arrow  indicates improvement only over zero-shot GPT-3. ^aT5-11B, ^bBERT-large.

READING COMPREHENSION & OPEN-DOMAIN QA

	READING COMPREHENSION			OPEN-DOMAIN QA			
	BoolQ acc.	MultiRC F1	OBQA acc.	ARC-e acc.	ARC-c acc.	NQ EM	TriviaQA EM
Supervised model	91.2 ^a	88.2 ^a	85.4 ^a	92.6 ^a	81.1 ^a	36.6 ^a	60.5 ^a
Base LM 137B zero-shot	81.0	60.0	41.8	76.4	42.0	3.2	18.4
· few-shot	79.7	59.6	50.6	80.9	49.4	22.1	55.1
GPT-3 175B zero-shot	60.5	72.9	57.6	68.8	51.4	14.6	64.3
· few-shot	77.5	74.8	65.4	70.1	51.5	29.9	71.2
FLAN 137B zero-shot							
- no prompt engineering	80.2▲2.7 stdev=3.1	74.5↑2.4 stdev=3.7	77.4▲12.0 stdev=1.3	79.5▲8.6 stdev=0.8	61.7▲10.2 stdev=1.4	18.6▲4.0 stdev=2.7	55.0 stdev=2.3
- best dev template	82.9▲5.4	77.5▲2.7	78.4▲13.0	79.6▲8.7	63.1▲11.6	20.7▲6.1	56.7

Table 2: Results on reading comprehension and open-domain question answering. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3. ^aT5-11B.

TriviaQA dataset

(1) has relatively complex, compositional questions,

(2) has considerable syntactic and lexical variability between questions and corresponding answer-evidence sentences

(3) requires more cross sentence reasoning to find answers.

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?

Answer: Fitness

Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Figure 1: Question-answer pairs with sample excerpts from evidence documents from TriviaQA exhibiting lexical and syntactic variability, and requiring reasoning from multiple sentences.

COMMONSENSE REASONING & COREFERENCE RESOLUTION

- FLAN will be marginal when **instructions are not crucial** for describing the given task. Moreover, we note a further limitation with FLAN for these five language modeling tasks, including **options actually hurts performance**, and so the reported results are for rank classification without options.
- ReCORD dataset:
<https://sheng-z.github.io/ReCoRD-explorer/examples/002.html>
-

	COMMONSENSE REASONING					COREFERENCE	
	CoPA acc.	HellaSwag acc.	PiQA acc.	StoryCloze acc.	ReCoRD acc.	WSC273 acc.	Winogrande acc.
Supervised model	94.8 ^a	47.3 ^b	66.8 ^b	89.2 ^b	93.4 ^a	72.2 ^b	93.8 ^a
Base LM 137B zero-shot	90.0	57.0	80.3	79.5	87.8	81.0	68.3
· few-shot	89.0	58.8	80.2	83.7	87.6	61.5	68.4
GPT-3 175B zero-shot	91.0	78.9	81.0	83.2	90.2	88.3	70.2
· few-shot	92.0	79.3	82.3	87.7	89.0	88.6	77.7
FLAN 137B zero-shot							
- no prompt engineering	90.6 stdev=2.0	56.4 stdev=0.5	80.9 stdev=0.8	92.2▲4.5 stdev=1.3	67.8 stdev=3.0	80.8 stdev=3.7	67.3 stdev=2.5
- best dev template	91.0	56.7	80.5	93.4▲5.7	72.5	-	71.2↑1.0

Table 3: Results (accuracy in %) for commonsense reasoning and coreference resolution. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. ^aT5-11B, ^bBERT-large. The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3.

Translation

Similar to GPT-3, FLAN shows

- strong X-> English
- Weak English -> X
- FLAN uses an English sentencepiece tokenizer and the majority of pretraining data is English.

TRANSLATION

	TRANSLATION					
	French		German		Romanian	
	En→Fr BLEU	Fr→En BLEU	En→De BLEU	De→En BLEU	En→Ro BLEU	Ro→En BLEU
Supervised model	45.6 ^c	35.0 ^d	41.2 ^e	38.6 ^f	38.5 ^g	39.9 ^g
Base LM 137B zero-shot	11.2	7.2	7.7	20.8	3.5	9.7
· few-shot	31.5	34.7	26.7	36.8	22.9	37.5
GPT-3 175B zero-shot	25.2	21.2	24.6	27.2	14.1	19.9
· few-shot	32.6	39.2	29.7	40.6	21.0	39.5
FLAN 137B zero-shot						
- no prompt engineering	32.0 ^{↑6.8} stdev=2.0	35.6 ^{↑14.4} stdev=1.5	24.2 stdev=2.7	39.4 ^{↑12.2} stdev=0.6	16.9 ^{↑2.8} stdev=1.4	36.1 ^{↑16.2} stdev=1.0
- best dev template	34.0 ^{▲1.4}	36.5 ^{↑15.3}	27.0 ^{↑2.4}	39.8 ^{↑12.6}	18.4 ^{↑4.3}	36.7 ^{↑16.7}

Table 4: Translation results (BLEU) for WMT’14 En/Fr and WMT’16 En/De and En/Ro. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. ^cEdunov et al. (2018), ^dDurrani et al. (2014), ^eWang et al. (2019b), ^fSennrich et al. (2016), ^gLiu et al. (2020). The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3.

NUMBER OF INSTRUCTION TUNING CLUSTERS

- performance does not appear to saturate, implying that performance may further improve with even more clusters added to instruction tuning
- Contributions are not clear
- Minimal contribution = sentiment

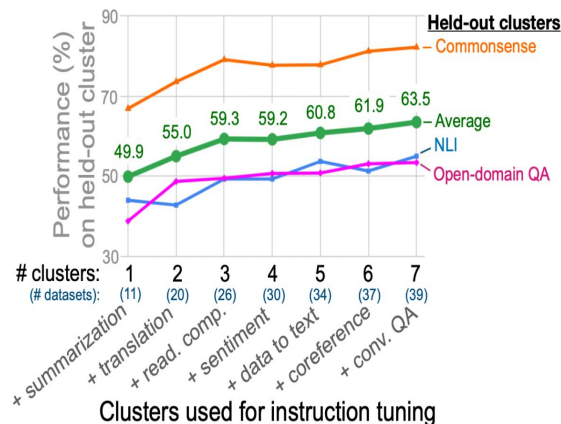


Figure 5: Adding additional task clusters to instruction tuning improves zero-shot performance on held-out task clusters. The evaluation tasks are the following. Commonsense: CoPA, HellaSwag, PiQA, and StoryCloze. NLI: ANLI R1-R3, QNLI, RTE, SNLI, and WNLI. Open-domain QA: ARC easy, ARC challenge, Natural Questions, and TriviaQA.

4.2 SCALING LAWS

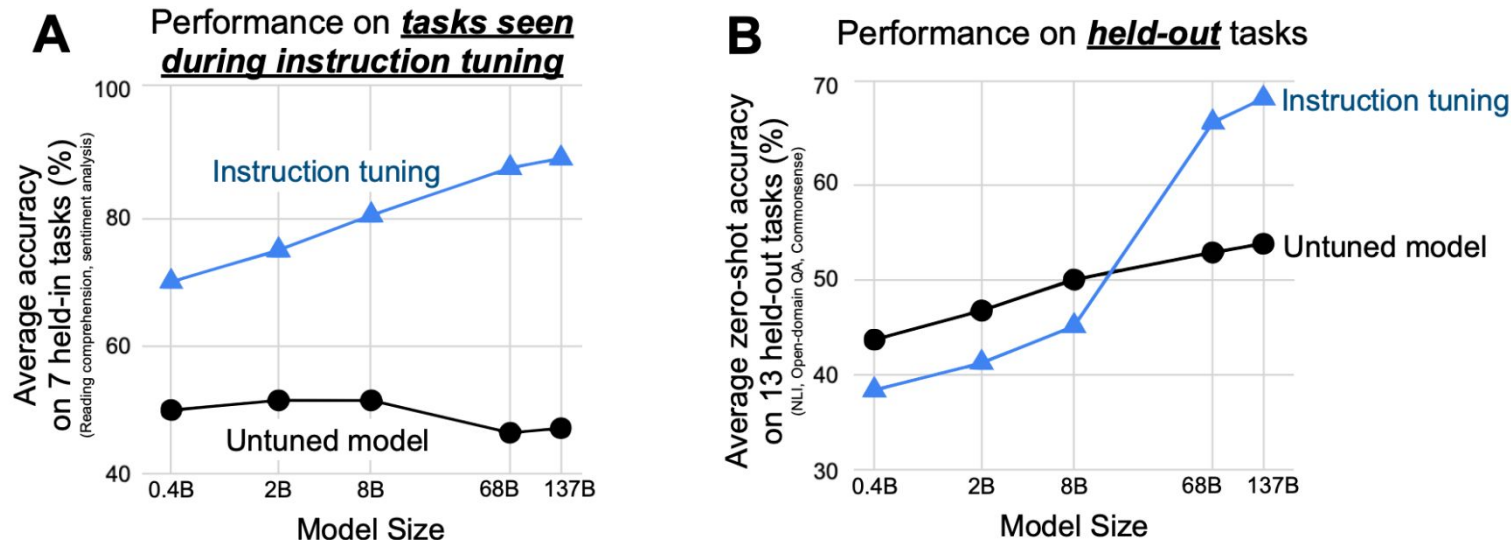


Figure 6: (A) Performance on tasks seen during instruction tuning improves for all model sizes. (B) The effect of instruction tuning on performance on unseen tasks depends on the model scale. Whereas instruction tuning helps large models generalize to new tasks, for small models it actually hurts generalization to unseen tasks, potentially because all model capacity is used to learn the mixture of instruction tuning tasks.

Prompt tuning works better in flan

PROMPT TUNING ANALYSIS									
	Prompt tuning train. examples	BoolQ acc.	CB acc.	CoPA acc.	MultiRC F1	ReCoRD acc.	RTE acc.	WiC acc.	WSC acc.
Base LM	32	55.5	55.4	87.0	65.4	78.0	52.4	51.6	65.4
FLAN		77.5	87.5	91.0	76.8	80.8	83.0	57.8	70.2
Base LM	full dataset	82.8	87.5	90.0	78.6	84.8	82.0	54.9	72.7
FLAN		86.3	98.2	94.0	83.4	85.1	91.7	74.0	86.5

Table 5: FLAN responds better to continuous inputs attained via prompt tuning than Base LM. When prompt tuning on a given dataset, no tasks from the same cluster as that dataset were seen during instruction tuning.

DISCUSSION

- does instruction tuning a language model improve its ability to perform unseen tasks?
- that the benefits of instruction tuning emerge only with sufficient model scale.
- FLAN appears to respond better to prompt tuning than the unmodified base model,
-

Limitation

- there is no accepted method for operationalizing the similarity between two tasks
 - Conservative approach: excluding reading comprehension with commonsense from instruction tuning when evaluating both reading comprehension and commonsense reasoning
- short instructions (typically a single sentence)
- only?? cover ten task clusters
- potentially be used to improve model behavior with respect to bias and fairness
-

ETHICAL CONSIDERATIONS

- labeled datasets such as those we use for finetuning can contain undesirable biases => propagated into zero-shot applications of the model on downstream tasks.
- instruction-tuned models can potentially require less data and expertise to use; such lower barriers
- for most datasets, supervised models such as BERT and T5 still outperform the zero-shot instructions-only model

