Zero-shot learning

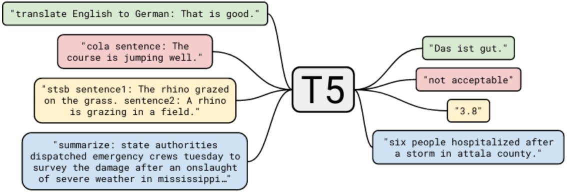
Karol Kaczmarek

Adam Mickiewicz University
Poznań

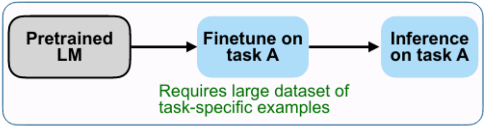
Applica.ai
Warsaw

2021

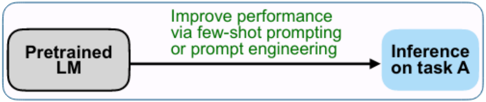
Pretrain-finetune

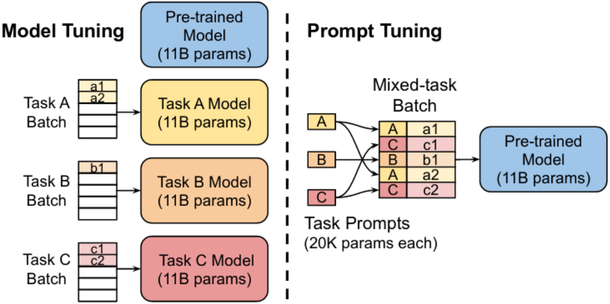


- Standard procedure: pretrain-finetune: BERT [1], RoBERTa [2], T5 [3], ...
  - pretrain on huge text or use pretrained model
  - finetune and evaluate on desired task

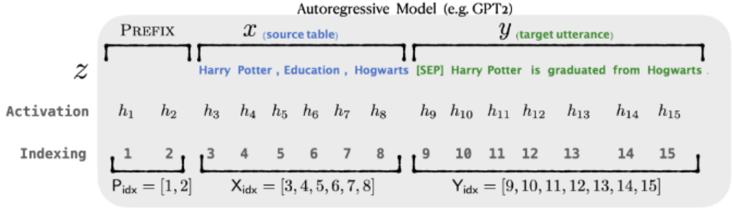


Prompting [4]


Pretrained LM
Improve performance via few-shot prompting or prompt engineering
Inference on task A


Model Tuning
Pre-trained Model (11B params)
Task A Batch
a1
a2
Task A Model (11B params)
Task B Batch
b1
Task B Model (11B params)
Task C Batch
c1
c2
Task C Model (11B params)
Prompt Tuning
Mixed-task Batch
A a1
C c1
B b1
A a2
C c2
A
B
C
Task Prompts (20K params each)
Pre-trained Model (11B params)

Prompting [5]



Summarization Example

Article: Scientists at University College London discovered people tend to think that their hands are wider and their fingers are shorter than they truly are.They say the confusion may lie in the way the brain receives information from different parts of the body.Distorted perception may dominate in some people, leading to body image problems ... [ignoring 308 words] could be very motivating for people with eating disorders to know that there was a biological explanation for their experiences, rather than feeling it was their fault."

Summary: The brain naturally distorts body image – a finding which could explain eating disorders like anorexia, say experts.

Prompting [5]

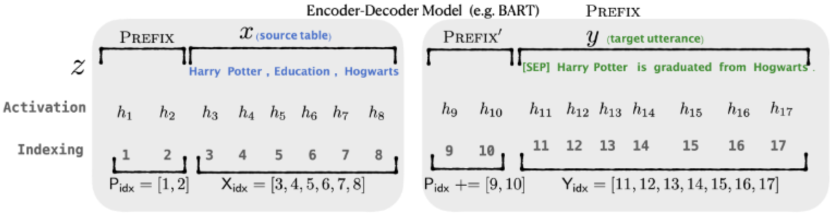
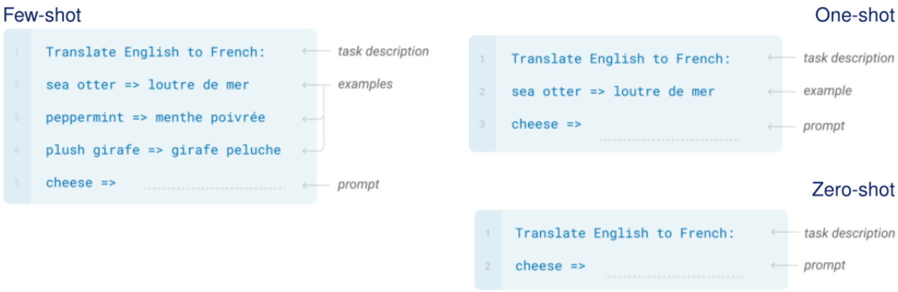


Table-to-text Example

Table: name[Clowns] customer-rating[1 out of 5] eatType[coffee shop] food[Chinese] area[riverside] near[Clare Hall]

Textual Description: Clowns is a coffee shop in the riverside area near Clare Hall that has a rating 1 out of 5 . They serve Chinese food .

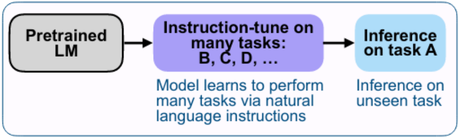
Prompting - GPT-3 [6]



FLAN

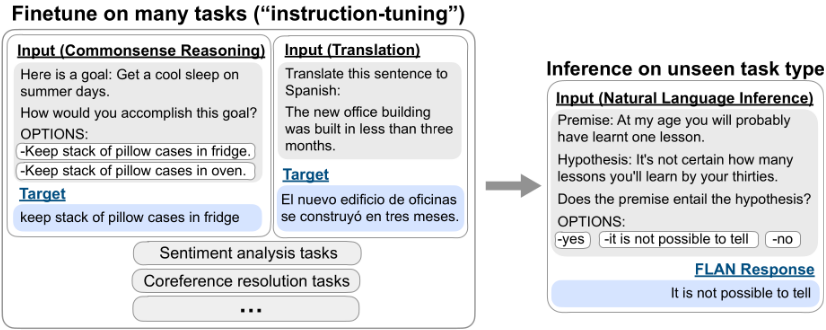
- September 2021, Google
- FLAN (**F**inetuned **LA**nguage **N**et) [7]
- 137B parameter pretrained language model (like GPT-3)
- Improving zero-shot learning on over 60 NLP tasks
- **Instruction tuning** - verbalize NLP tasks by natural language instruction templates

Instruction training



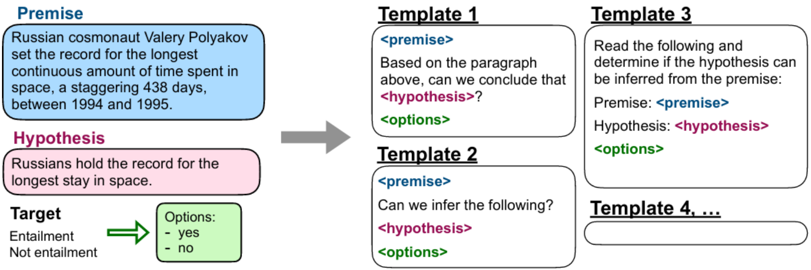
- teach language model to perform tasks described via **instructions**, it will **learn to follow instructions** and do so even for unseen tasks
- group tasks into **clusters by task type** and hold out each task cluster for evaluation while instruction tuning on all remaining clusters
- Intuition - NLP tasks can be described:
  - "Is the sentiment of this movie review positive or negative?"
  - "Translate ‘how are you’ into Chinese."

Instruction training



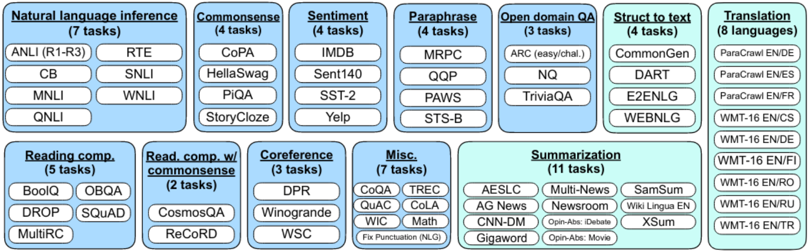
- (!!!) include **OPTIONS** to makes the model aware of which choices are desired when responding

Cluster templates



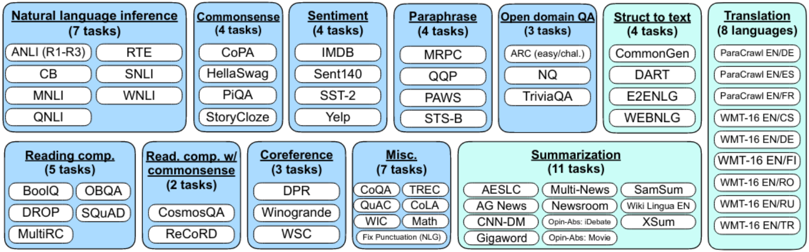
- manually compose **10 unique templates** that describe the task using natural language instructions (10 templates per task)
- include up to **3 templates** that "turned the task around" (generate negative movie review for sentiment classification)
- **randomly selected** instruction template for tasks in pretraining

Task clusters



- aggregate **62 text datasets** (including both language understanding and language generation tasks) into a single mixture
- each dataset is **categorized into one of twelve** task clusters (given cluster are of the same task type)
- evaluate on cluster that **were not** during instruction tuning

Task clusters



- some datasets have more than ten million training examples (translation) - limit the number of training examples per dataset to **30000**
- other datasets have few training examples (CommitmentBank only has **250**) - use **examples-proportional mixing** scheme (from T5 - probability of example sampling) to prevent datasets from being marginalized

Task clusters

- **Natural language inference (NLI)** concerns how two sentences relate, typically asking, given a first sentence, whether a second sentence is true, false, or possibly true
- **Reading comprehension** tests the ability to answer a question when given a passage that contains the answer
- **Open-domain QA** asks models to answer questions about the world without specific access to information that contains the answer
- **Commonsense reasoning** evaluates the ability to perform physical or scientific reasoning with an element of common sense
- **Coreference resolution** tests the ability to identify expressions of the same entity in some given text
- **Translation** is the task of translating text from one language into a different language

Architecture

- **left-to-right**, **decoder-only** transformer language model of **137B parameters**
  - model from Google publication - generate computer programs in a programming language (program synthesis)
- pretrained on a collection of web documents (including those with **computer code**), **dialog data**, and **Wikipedia** - dataset **is not as clean** as the GPT-3 training set and also **has a mixture of dialog and code**
- used models:
  - **Base LM** - pretrained model that was used for program synthesis
  - **FLAN** - instruction-tuned version of **Base LM**

Score

	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

[a] T5-11B, [b] BERT-large, ▲ improvement over few-shot GPT-3, ↑ improvement only over zero-shot GPT-3

- using the same prompts as GPT-3 for **zero** and **few-shot** Base LM results
- NLI examples are unlikely to have appeared naturally in an training set
- FLAN question: "Does <premise> mean that <hypothesis>?"

Score

	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

[a] T5-11B, ▲ improvement over few-shot GPT-3, ↑ improvement only over zero-shot GPT-3

- **reading comprehension** is where models are asked to answer a question about a provided passage

Score

	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

[a] T5-11B, [b] BERT-large, ▲ improvement over few-shot GPT-3, ↑ improvement only over zero-shot GPT-3

- **Commonsense reasoning** evaluates the ability to perform physical or scientific reasoning with an element of common sense
- **Coreference resolution** tests the ability to identify expressions of the same entity in some given text
- Model fails when instructions are not crucial for describing task

Score

	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						
- average template	32.0 ↑6.8 std=2.0	35.6 ↑14.4 std=1.5	24.2 std=2.7	39.4 ↑12.2 std=0.6	16.9 ↑2.8 std=1.4	36.1 ↑16.2 std=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

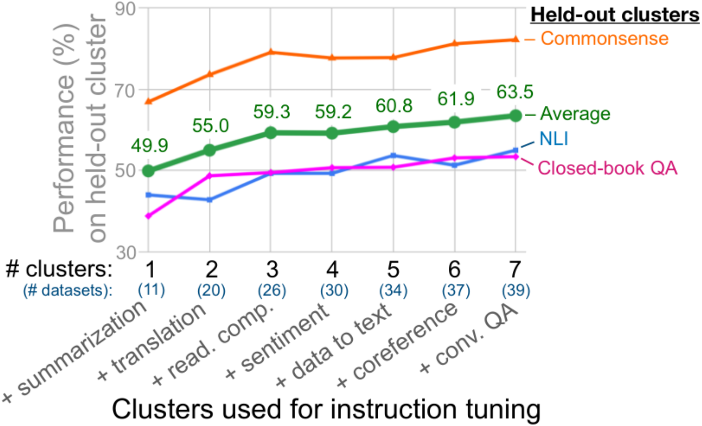
▲ improvement over few-shot GPT-3, ↑ improvement only over zero-shot GPT-3

- GPT-3: ∼7% of text in other language (∼1,5 fr, ∼1,5 de)
- FLAN: ∼10% of text in other language

Instruction tuning

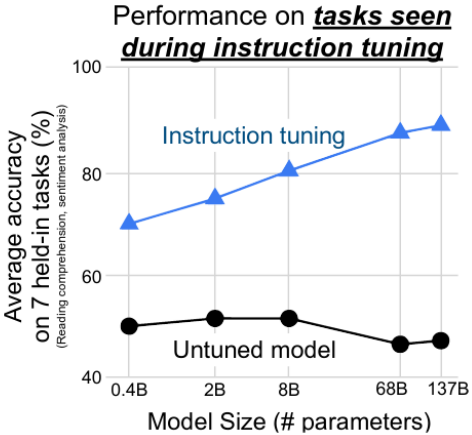
- is **very effective** on tasks that **can be naturally verbalized** as instructions (natural language inference and question answering)
- is **less effective** on tasks that are **directly formulated** as language modeling (commonsense reasoning and coreference resolution)

Clusters used for instruction tuning



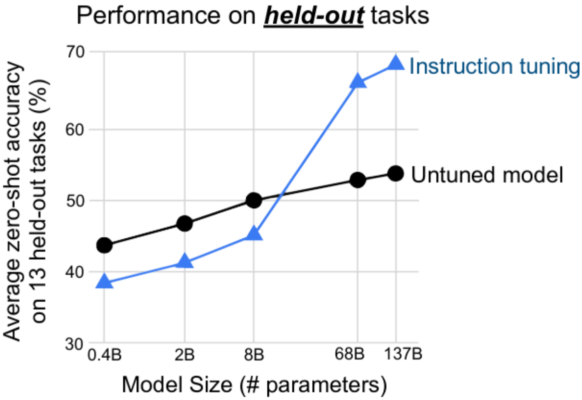
Increasing the number of task clusters improves perform.

Instruction tuning - performance on **seen** tasks



Untuned model - untuned model without instruction tuning

Instruction tuning - performance on **unseen** tasks



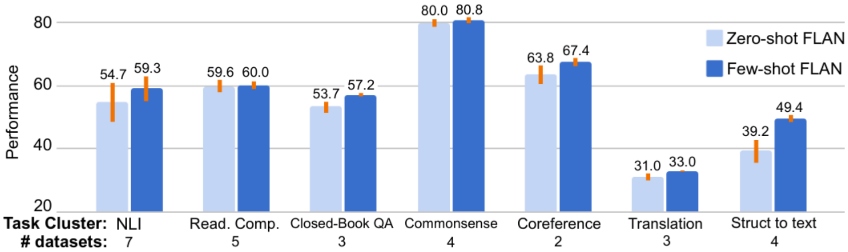
Hurts performance on 8B and smaller models - learning the ∼40 tasks **fills the entire model capacity**, causing these models to perform worse on new tasks

Prompt tuning - SuperGLUE

	Prompt tuning train. examples	BoolQ acc.	CB acc.	CoPA acc.	MultiRC F1	ReCoRD acc.	RTE acc.	WiC acc.	WSC acc.
		PROMPT TUNING ANALYSIS							
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

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

Few-shot



Standard deviation (orange color) among templates is **lower** for few-shot FLAN, indicating reduced sensitivity to prompt engineering.

Environmental consideration

- energy cost and carbon footprint for the pretrained models were **451 MWh** and **26 tCO2e**
- additional instruction tuning gradient-steps for finetuning FLAN is **less than 2%** of the number of pretraining steps

References I

[1] J. Devlin and et al., “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018.

[2] Y. Liu and et al., “Roberta: A robustly optimized bert pretraining approach,” 2019.

[3] C. Raffel and et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2019.

[4] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” 2021.

[5] X. Lisa Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” 2021.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, and et al., “Language models are few-shot learners,” 2020.

[7] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, and et al., “Finetuned language models are zero-shot learners,” 2021.