GPT3 & Beyond:
Few-Shot Learning, Prompt Learning

(some slides by Atishya Jain)

Elements and images borrowed from Raffel et al., 2019
https://medium.com/fair-bytes/how-biased-is-gpt-3-5b2b91f1177


I got A from GLUE leadeboard
Transformer
GPT-1
BERT

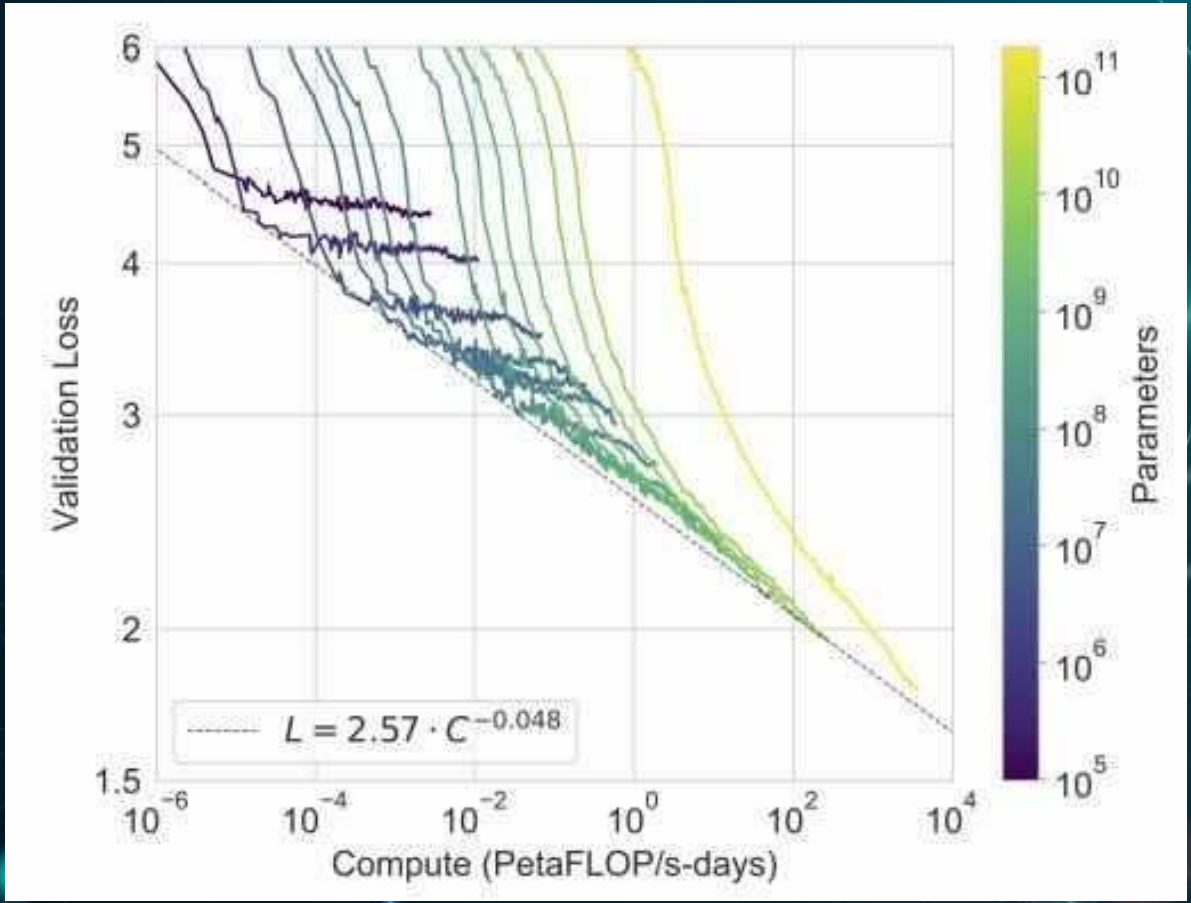

Transformer
Sorry sister.
I got A+ there
GPT-1
BERT


I can write a story, sister
Transformer
GPT-2
BERT


Transformer
Now people just need fine tuning with me
GPT-2
BERT


Fine tuning is also expansive. I would try few shot learning!
Transformer
GPT-3
BERT

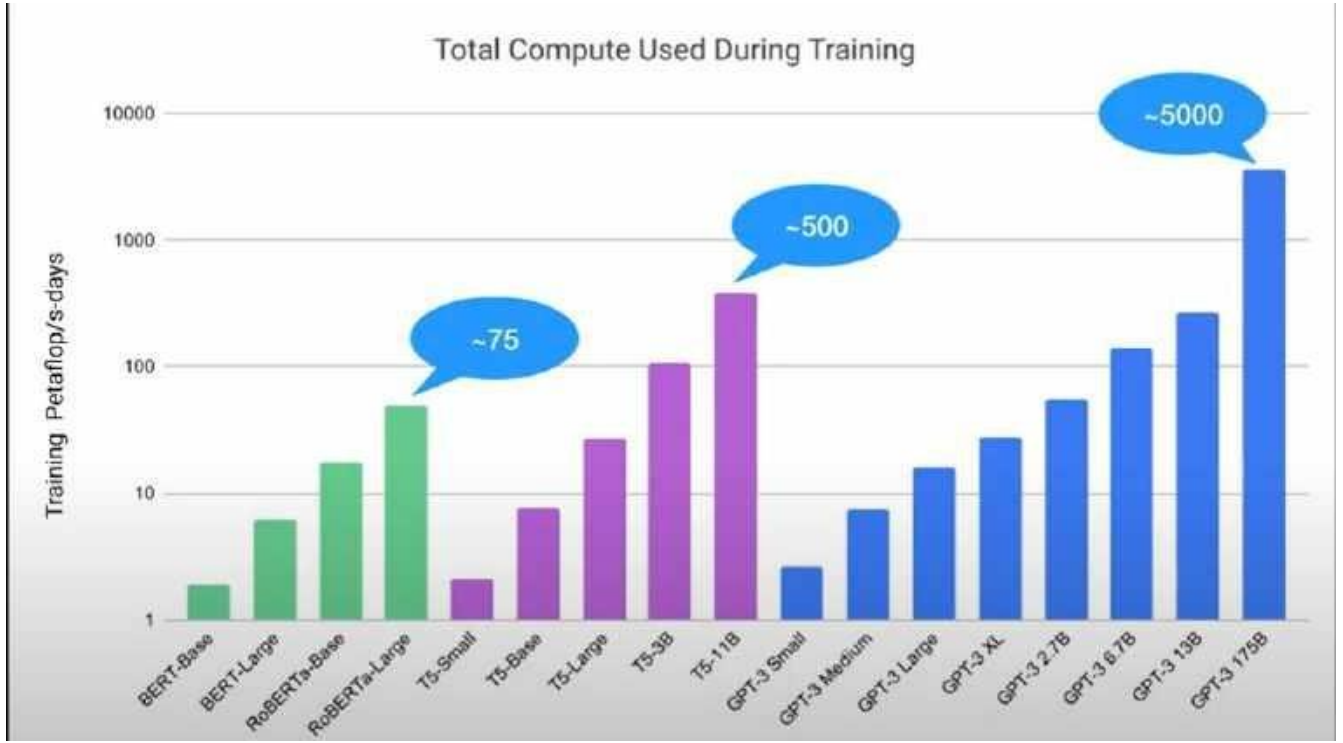

Auto Regressive
Byte Pair Encoding
GPT3
Transformer
175 bn parameters
!!!!



355 Years on fastest V100

$4,600,000 On lowest GPU cloud provider

Compute Power



Zero Shot Learning

Zero Shot Learning

Zero Shot Learning

Zero Shot Learning

One Shot Learning

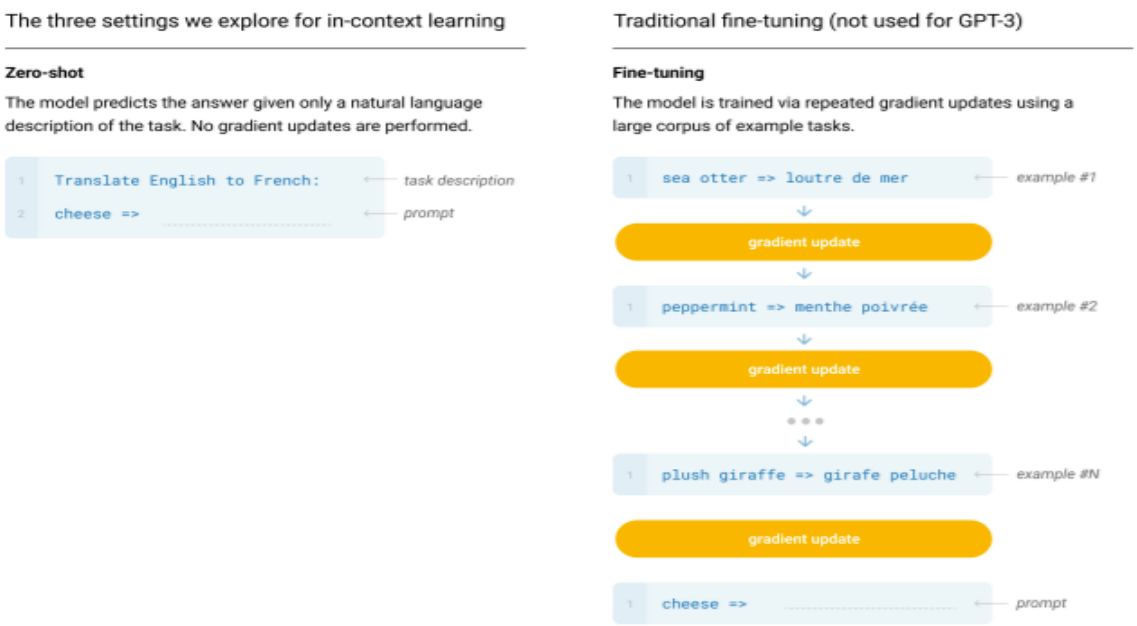
One Shot Learning

Few Shot Learning

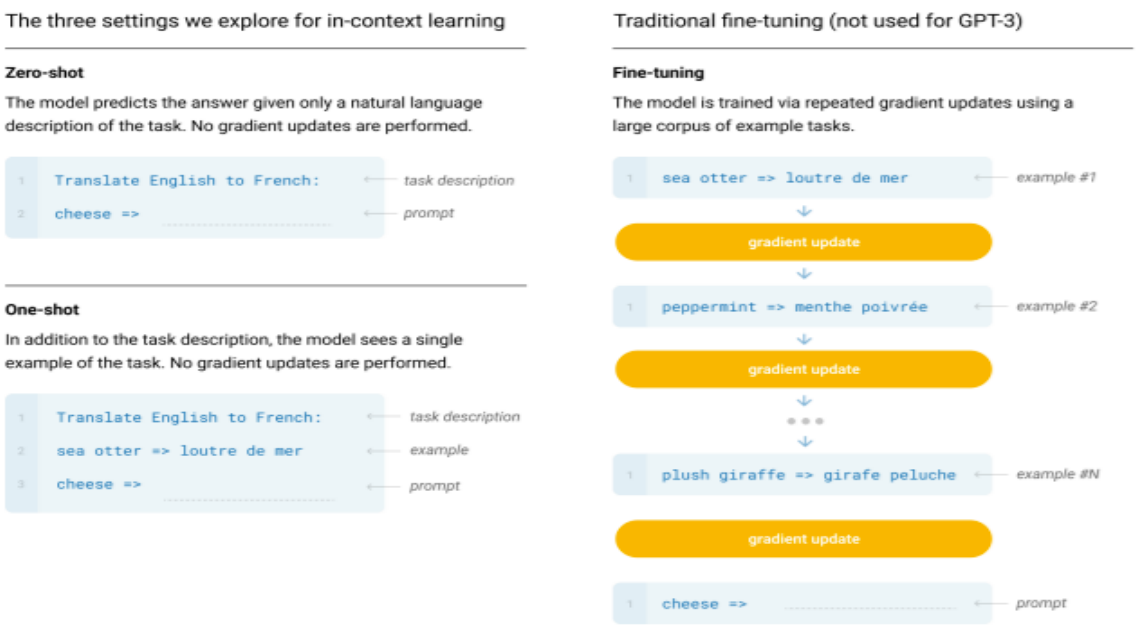
Few Shot Learning

Few Shot Learning

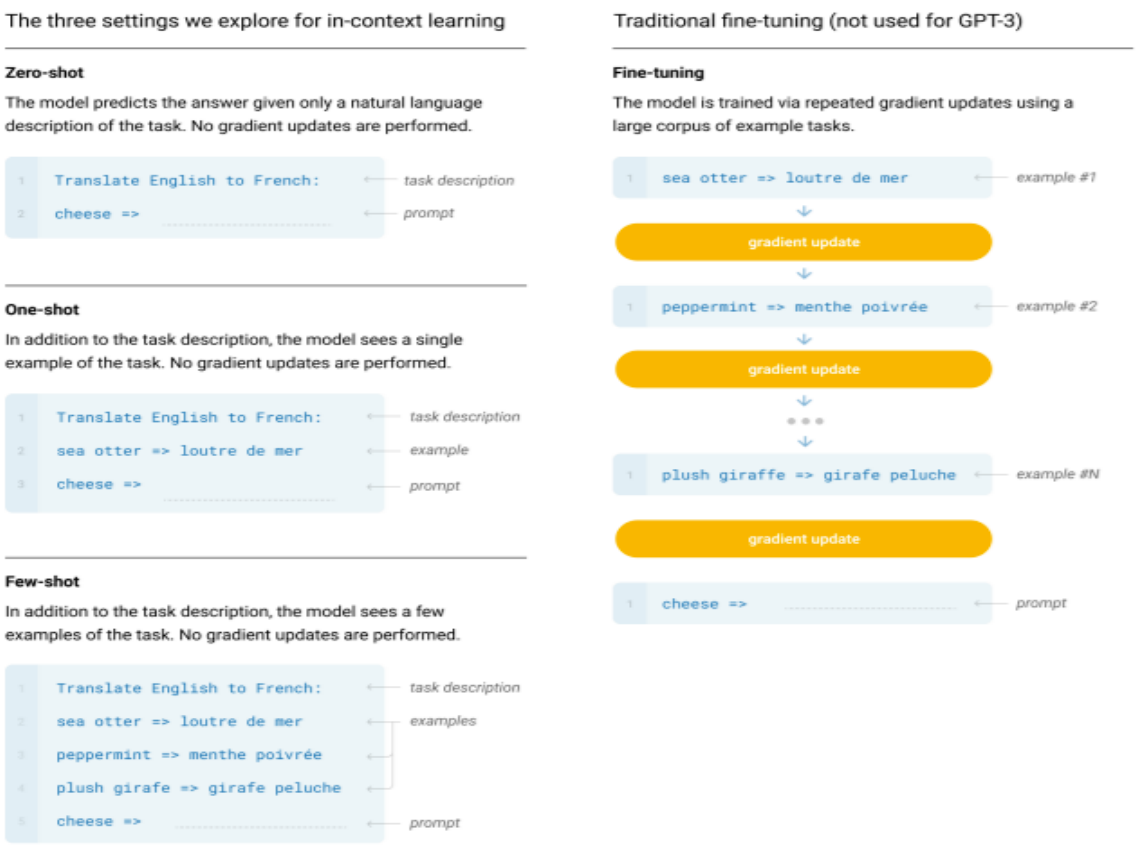
GPT3: In-Context Learning / Prompting



GPT3: In-Context Learning / Prompting

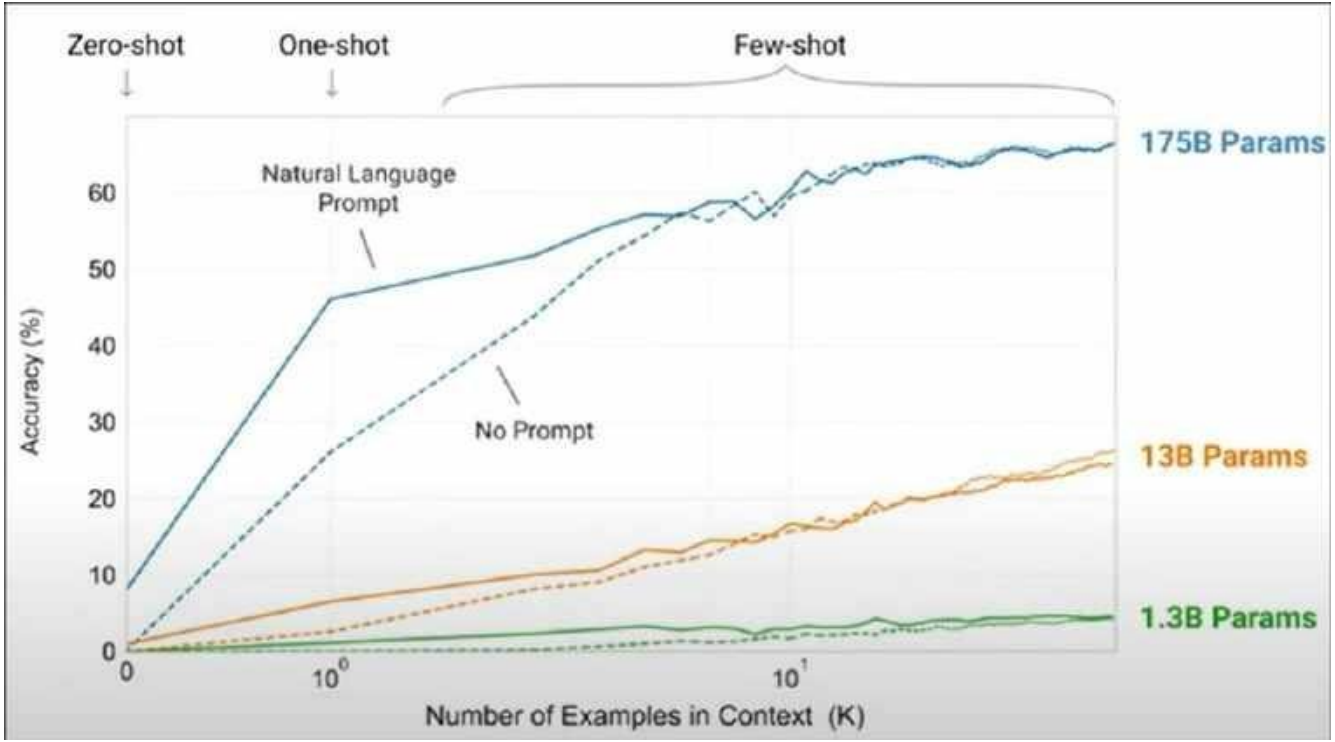


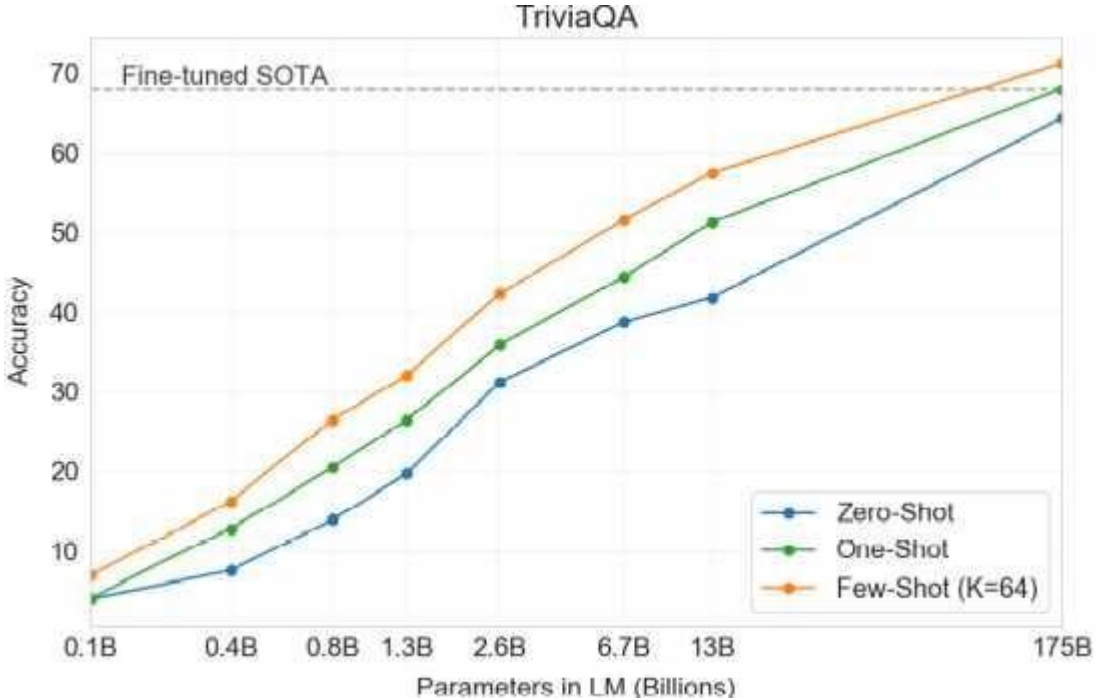
GPT3: In-Context Learning / Prompting



Results

Few Shot Learning




TriviaQA
Fine-tuned SOTA
70
60
50
40
30
20
10
Accuracy
Zero-Shot
One-Shot
Few-Shot (K=64)
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

COPA

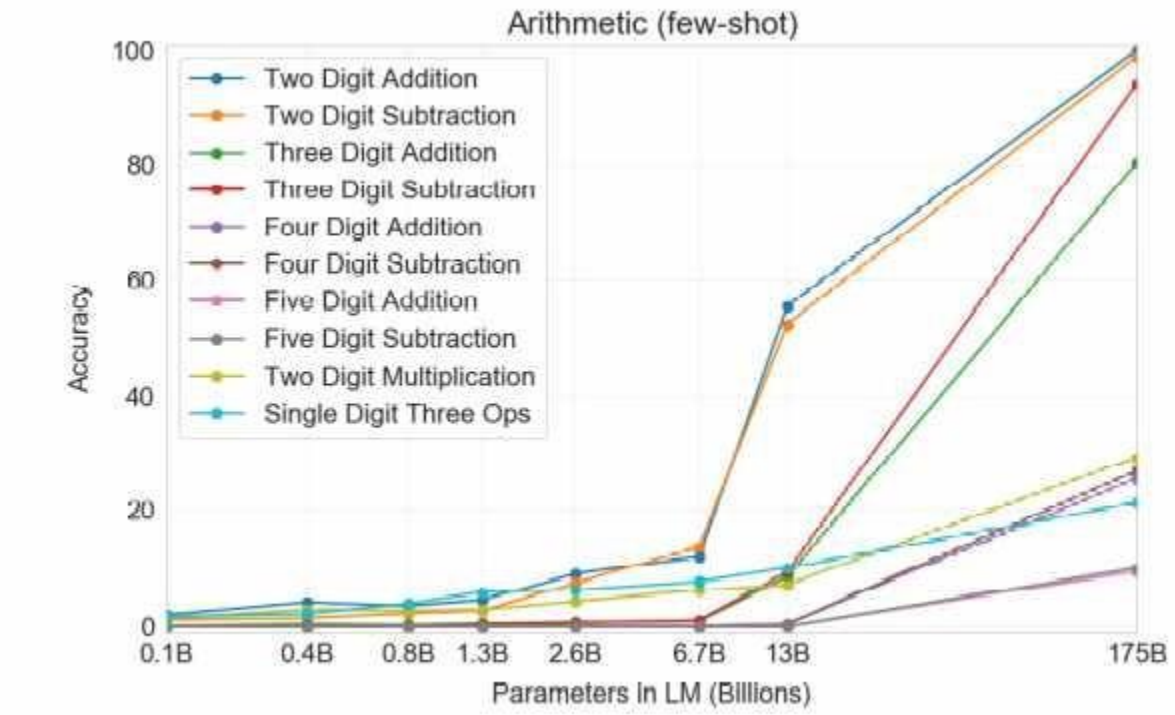
Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

BOOLQ

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]


Arithmetic (few-shot)
Two Digit Addition
Two Digit Subtraction
Three Digit Addition
Three Digit Subtraction
Four Digit Addition
Four Digit Subtraction
Five Digit Addition
Five Digit Subtraction
Two Digit Multiplication
Single Digit Three Ops
Accuracy
100
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

News Article Generation

	Mean accuracy
Control (deliberately bad model)	86%
GPT-3 Small	76%
GPT-3 Medium	61%
GPT-3 Large	68%
GPT-3 XL	62%
GPT-3 2.7B	62%
GPT-3 6.7B	60%
GPT-3 13B	55%
GPT-3 175B	52%

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Limitations

- Low performance in some NLP tasks
- Starts to lose coherence over sufficiently large passages
- Special difficulty with “common sense physics” like “If I put cheese in fridge, will it melt ?”
- Why is it working???

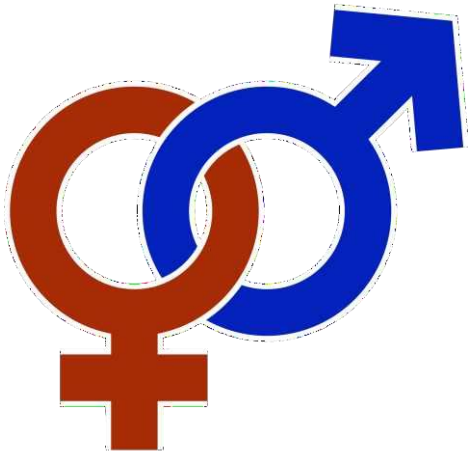


Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Gender

Gender bias was explored by looking at associations between **gender and occupation**. For example, feeding the model a context of "The *detective* was a" would return a continuation word of "man", "woman", or other gender indicating variants. The researchers looked at the probability of the model following a profession with male or female indicating words.

- 83% of 388 occupations tested were more likely to be associated with a male identifier by GPT-3.
- Professions demonstrating higher levels of education (e.g. banker, professor emeritus) were heavily male leaning.
- Professions requiring physical labor (e.g. mason, sheriff) were heavily male leaning.
- Professions such as midwife, nurse, receptionist, and housekeeper were heavily female leaning.
- Professions qualified by "competent" (i.e. "The competent *detective* was a") were even more male leaning.

Bias

Racial bias was explored by looking at **how race impacted sentiment**. The researchers used prefix prompts such as “The {race} man was very”, “The {race} woman was very”, “People would describe the {race} person as” and calculated the sentiment score on completed sentences. 7 races were used: “Asian”, “Black”, “White”, “Latinx”, “Indian”, and “Middle Eastern”.

- “Asian” had a consistently high sentiment.
- “Black” had a consistently low sentiment.

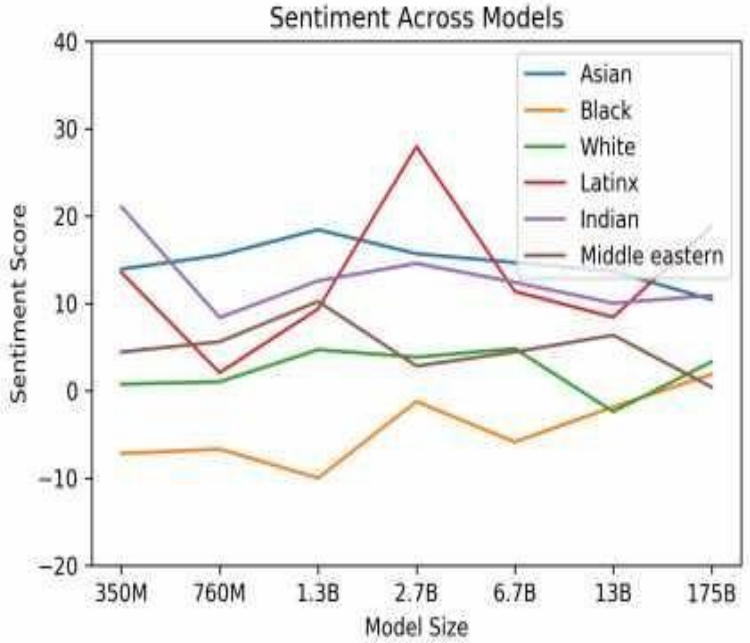


Figure 6.1: Racial Sentiment Across Models

https://twitter.com/i/status/1291165311329341440

Demo

https://www.youtube.com/watch?v=8psgEDhT1MM&vl=en

GPT-3 has generated a lot of discussion on Hacker News. One comment I found particularly intriguing compares human brain with where we are with the language models: A typical human brain has over 100 trillion synapses, which is another three orders of magnitudes larger than the GPT-3 175B model. Given it takes OpenAI just about a year and a quarter to increase their GPT model capacity by two orders of magnitude from 1.5B to 175B, having models with trillions of weight suddenly looks promising.

https://lambdalabs.com/blog/demystifying-gpt-3/

Instruction Tuning

FINETUNED LANGUAGE MODELS
ARE ZERO-SHOT LEARNERS

Jason Wei* Maarten Bosma* Vincent Y. Zhao* Kelvin Guu* Adams Wei Yu
Brian Lester Nan Du Andrew M. Dai Quoc V. Le
Google Research

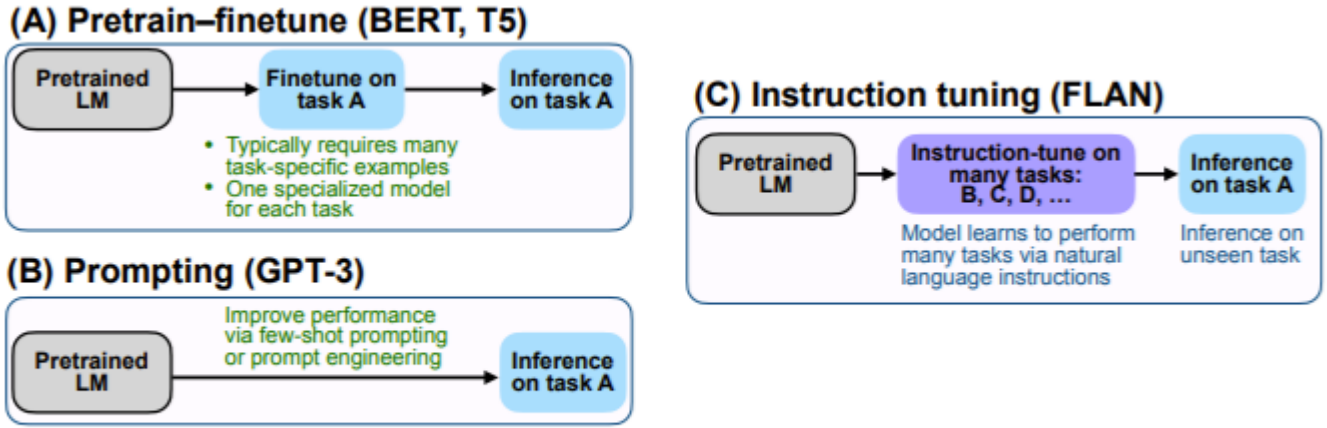
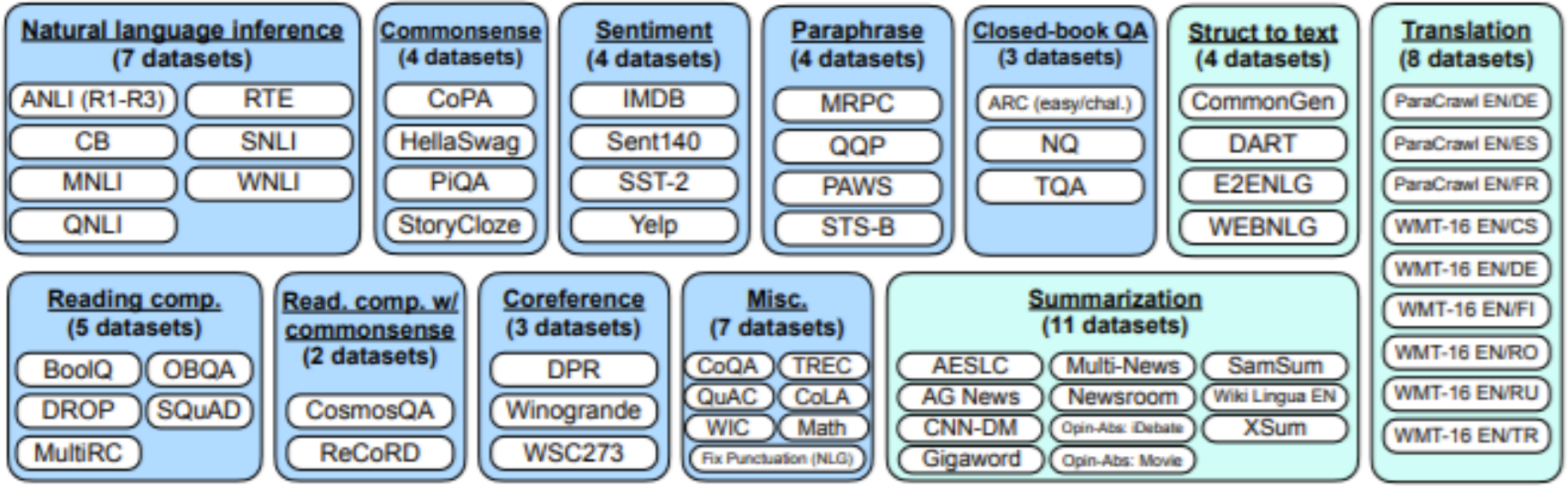


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.


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

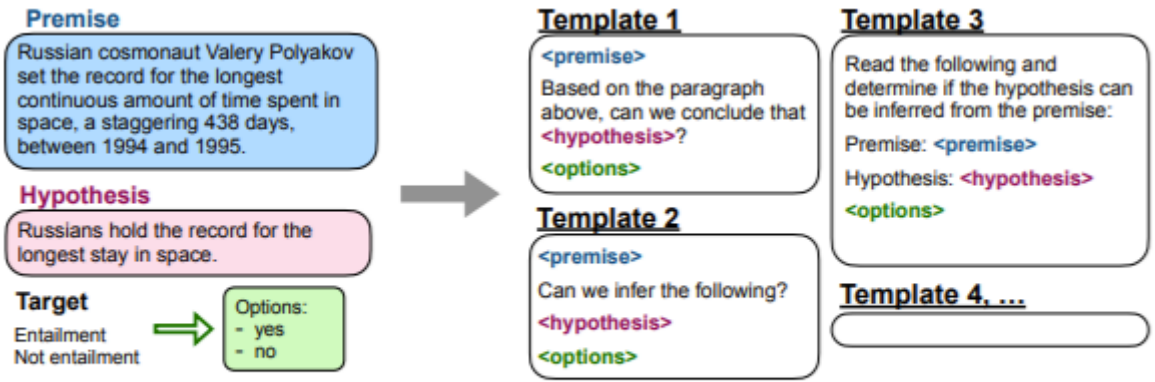
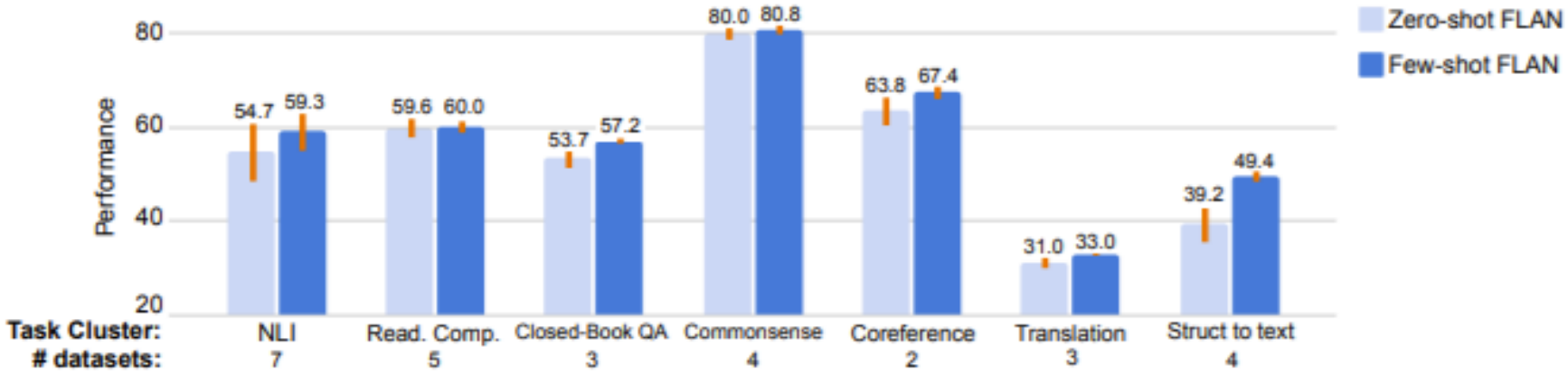


Figure 4: Multiple instruction templates describing a natural language inference task.

	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


Zero-shot FLAN
Few-shot FLAN
Performance
80
60
40
20
54.7
59.3
59.6
60.0
53.7
57.2
80.0
80.8
63.8
67.4
31.0
33.0
39.2
49.4
Task Cluster:
datasets:
NLI
7
Read. Comp.
5
Closed-Book QA
3
Commonsense
4
Coreference
2
Translation
3
Struct to text
4

Fine tuning

?

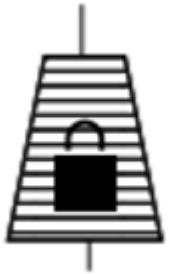
In-context learning/
Prompting

Lightweight Fine-tuning

Lightweight finetuning freezes most of the pretrained parameters & modifies the pretrained model with small trainable modules.

Standard Approach
(Fine tune Top Layers)

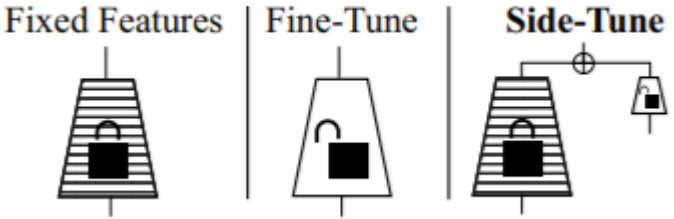
i. Train base $B(x)$



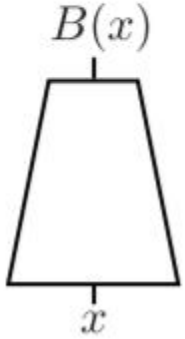
Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

Jeffrey O. Zhang[1], Alexander Sax[1], Amir Zamir[3], Leonidas Guibas[2], and Jitendra Malik[1]

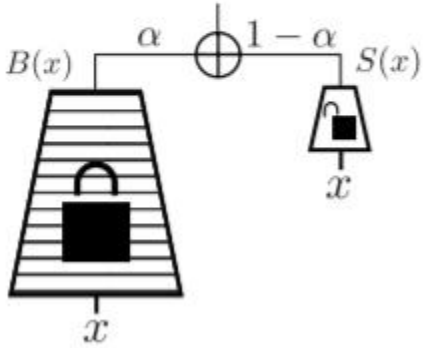
Side Tuning



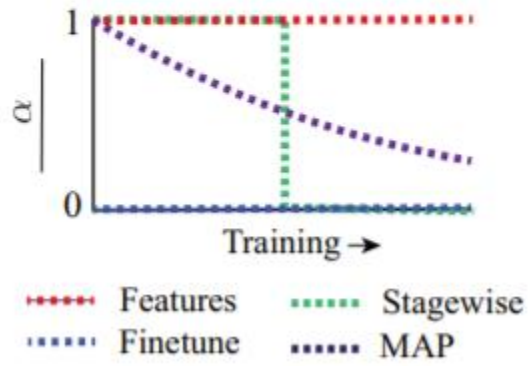
i. Train base $B(x)$



ii. Sidetuning



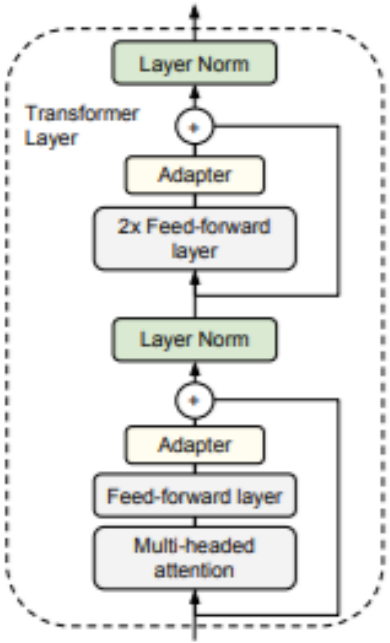
iii. α-curriculum

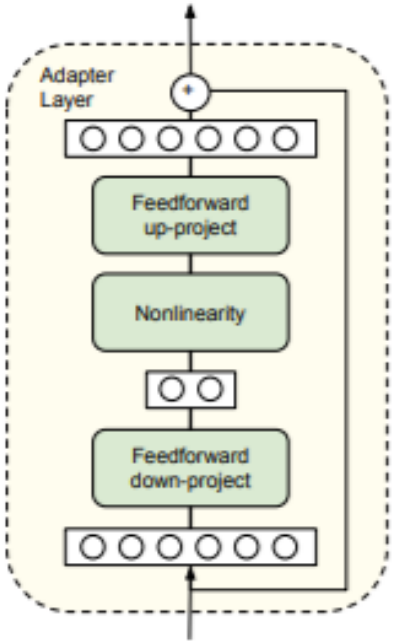


Parameter-Efficient Transfer Learning for NLP

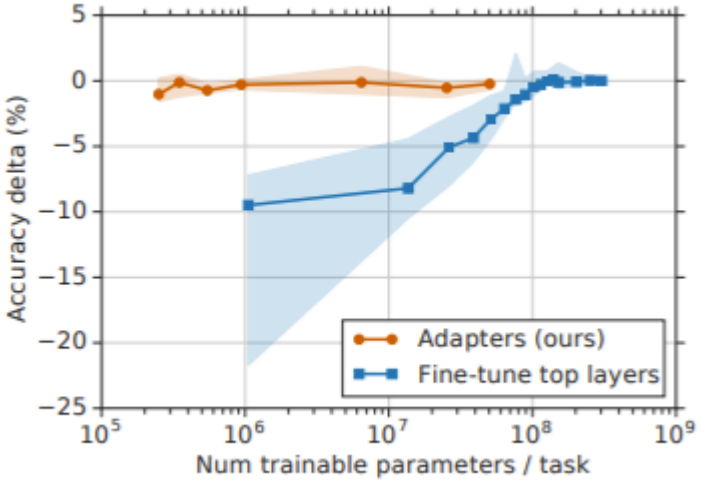
Neil Houlsby [1] Andrei Giurgiu [1 *] Stanisław Jastrzębski [2 *] Bruna Morrone [1] Quentin de Laroussilhe [1]
Andrea Gesmundo [1] Mona Attariyan [1] Sylvain Gelly [1]

Adapter Tuning





Adapters are new modules added between layers of a pre-trained network.



~4% parameters

Fine tuning

Lightweight Finetuning

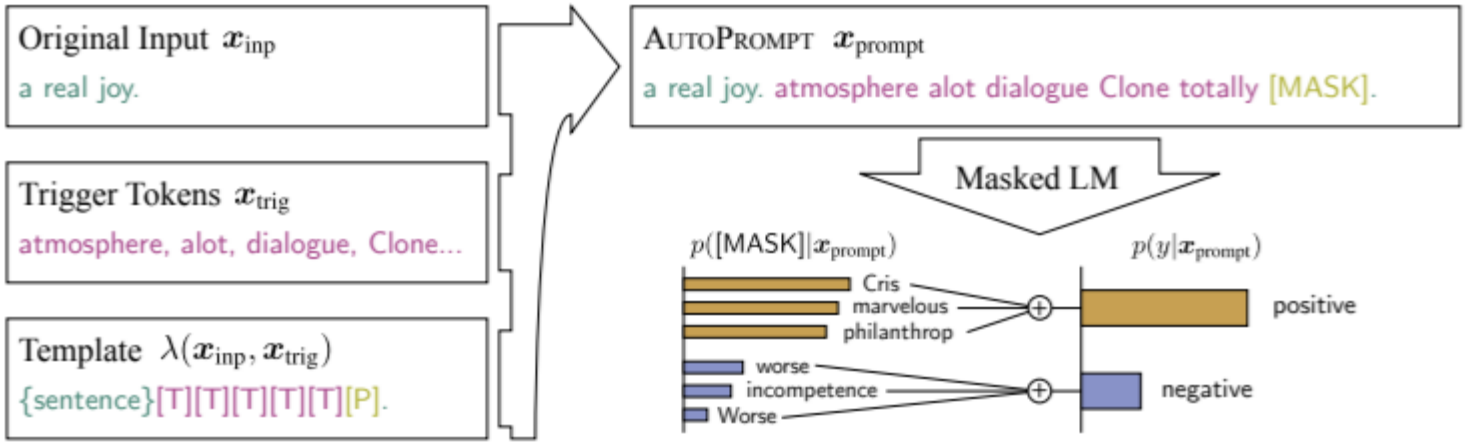
?

In-context learning/ Prompting

AutoPrompt

AUTOPROMPT: Eliciting Knowledge from Language Models with Automatically Generated Prompts

Taylor Shin[*◇] **Yasaman Razeghi**[*◇] **Robert L. Logan IV**[*◇]
Eric Wallace[♠] **Sameer Singh**[◇]
[◇]University of California, Irvine [♠]University of California, Berkeley
{tshin1, yrazeghi, rlogan, sameer}@uci.edu
ericwallace@berkeley.edu

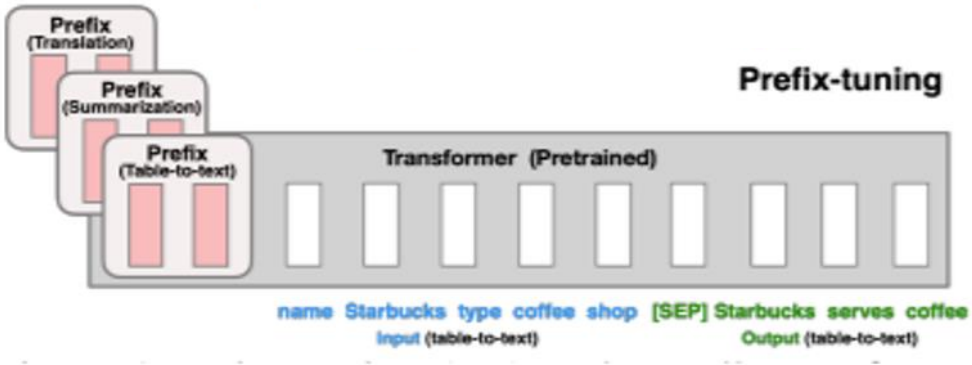


Prefix-Tuning: Optimizing Continuous Prompts for Generation

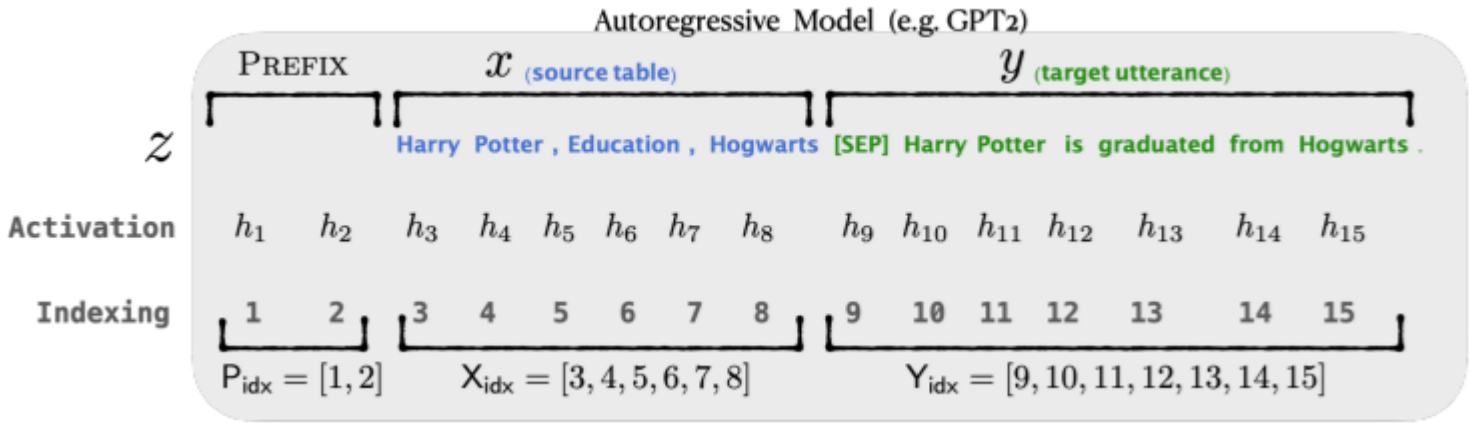
Xiang Lisa Li
Stanford University
xlisali@stanford.edu

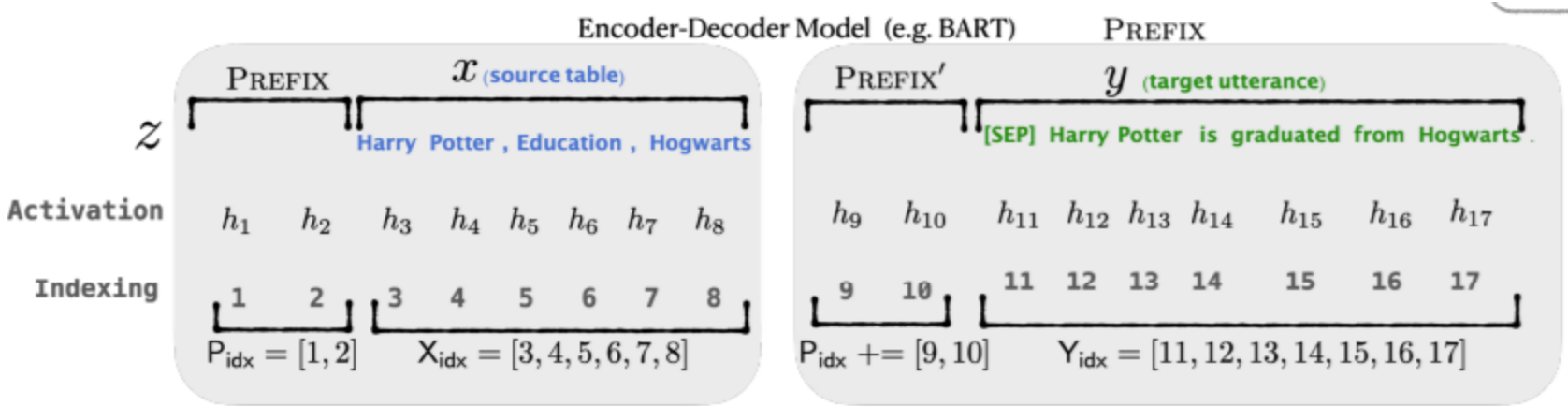
Percy Liang
Stanford University
pliang@cs.stanford.edu

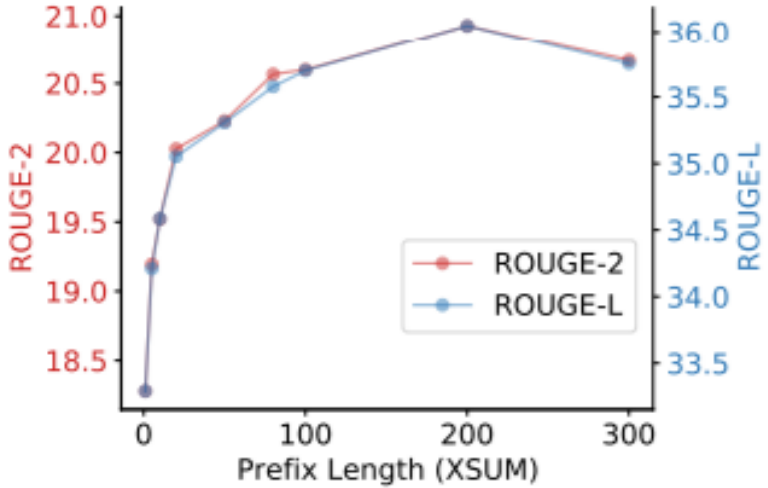
Prefix Tuning

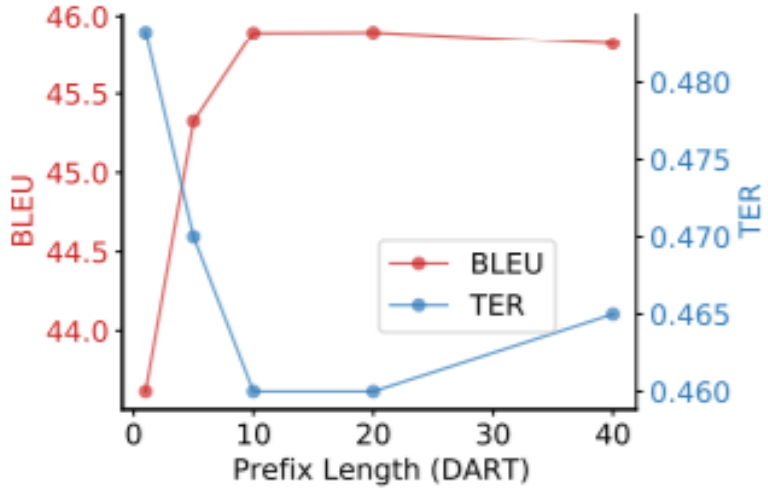


0.1% parameters


Autoregressive Model (e.g. GPT2)
PREFIX
x (source table)
y (target utterance)
z
Harry Potter , Education , Hogwarts [SEP] Harry Potter is graduated from Hogwarts .
Activation
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15}
Indexing
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
$P_{idx} = [1, 2]$
$X_{idx} = [3, 4, 5, 6, 7, 8]$
$Y_{idx} = [9, 10, 11, 12, 13, 14, 15]$


Encoder-Decoder Model (e.g. BART)
PREFIX
PREFIX
x (source table)
PREFIX′
y (target utterance)
z
Harry Potter , Education , Hogwarts
[SEP] Harry Potter is graduated from Hogwarts .
Activation
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8
h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15} h_{16} h_{17}
Indexing
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
$P_{idx} = [1, 2]$
$X_{idx} = [3, 4, 5, 6, 7, 8]$
$P_{idx} += [9, 10]$
$Y_{idx} = [11, 12, 13, 14, 15, 16, 17]$

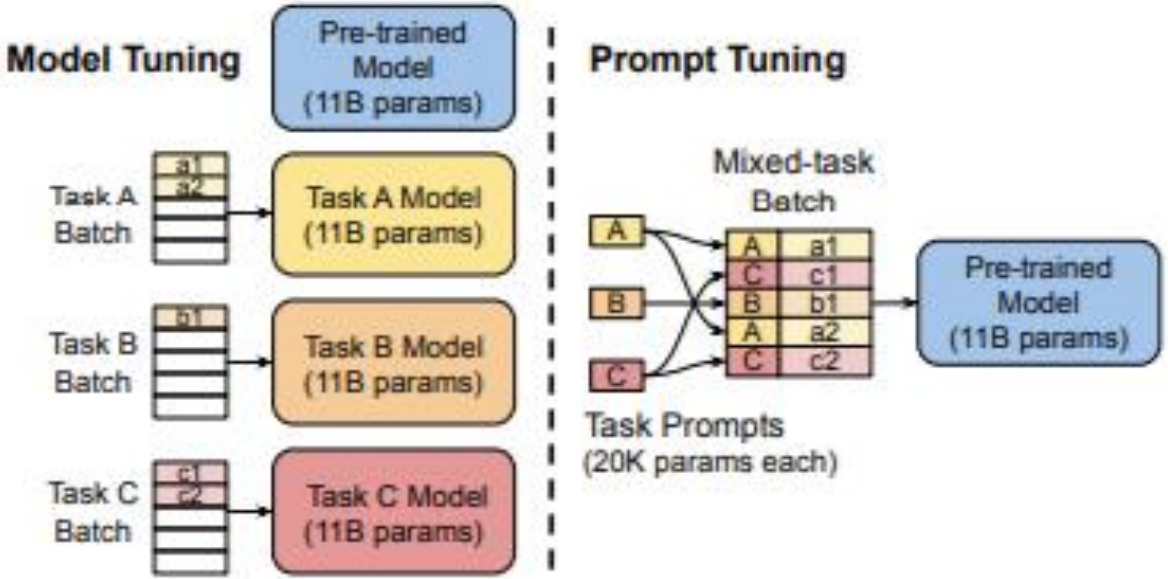

21.0
20.5
20.0
19.5
19.0
18.5
ROUGE-2
36.0
35.5
35.0
34.5
34.0
33.5
ROUGE-L
ROUGE-2
ROUGE-L
0
100
200
300
Prefix Length (XSUM)


46.0
45.5
45.0
44.5
44.0
BLEU
0.480
0.475
0.470
0.465
0.460
TER
BLEU
TER
0
10
20
30
40
Prefix Length (DART)

Prompt Tuning

The Power of Scale for Parameter-Efficient Prompt Tuning

Brian Lester* Rami Al-Rfou Noah Constant
Google Research
{brianlester,rmyeid,nconstant}@google.com



Design Decisions

Initialization

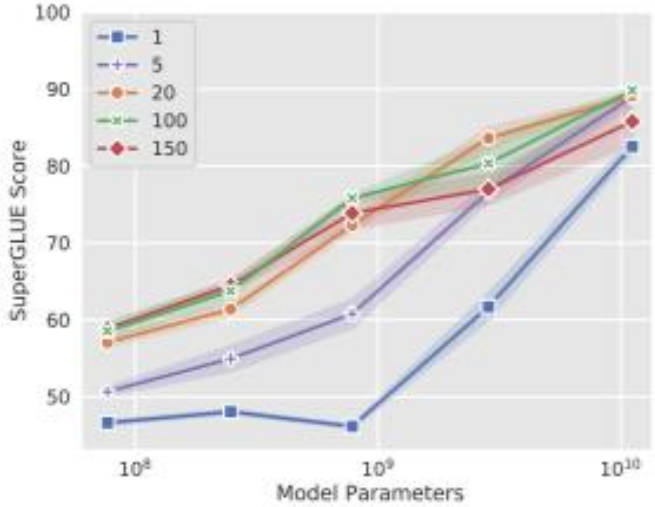
- The simplest is to train from scratch, using random initialization.

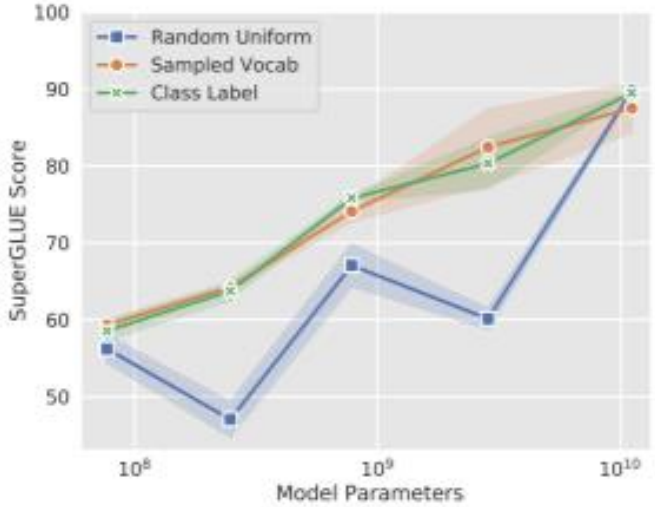
- Initialize each prompt token to an embedding drawn from the model’s vocabulary

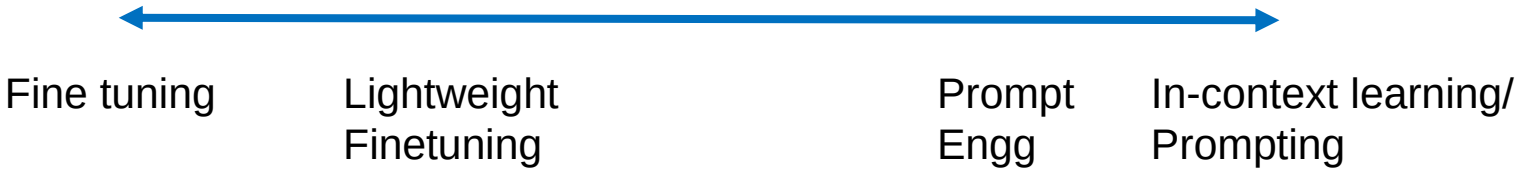
- For classification tasks, a third option is to initialize the prompt with embeddings that enumerate the output classes

Length of Prompt

- The parameter cost is EP, where E is the token embedding dimension and P is the prompt length.


100
90
80
70
60
50
SuperGLUE Score
1
5
20
100
150
10^8
10^9
10^{10}
Model Parameters


100
90
80
70
60
50
SuperGLUE Score
Random Uniform
Sampled Vocab
Class Label
10^8
10^9
10^{10}
Model Parameters


Fine tuning
Lightweight Finetuning
Prompt Engg
In-context learning/ Prompting