

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>



Transformer

I got A from GLUE
leadeboard

GPT-1

BERT



Transformer

GPT-1

BERT

Sorry sister.
I got A+ there



I can write a story,
sister

GPT-2

Transformer

BERT



Transformer

GPT-2

BERT

Now people
just need fine tuning
with me



Fine tuning is also expansive. I would try few shot learning!

Transformer

GPT-3

BERT

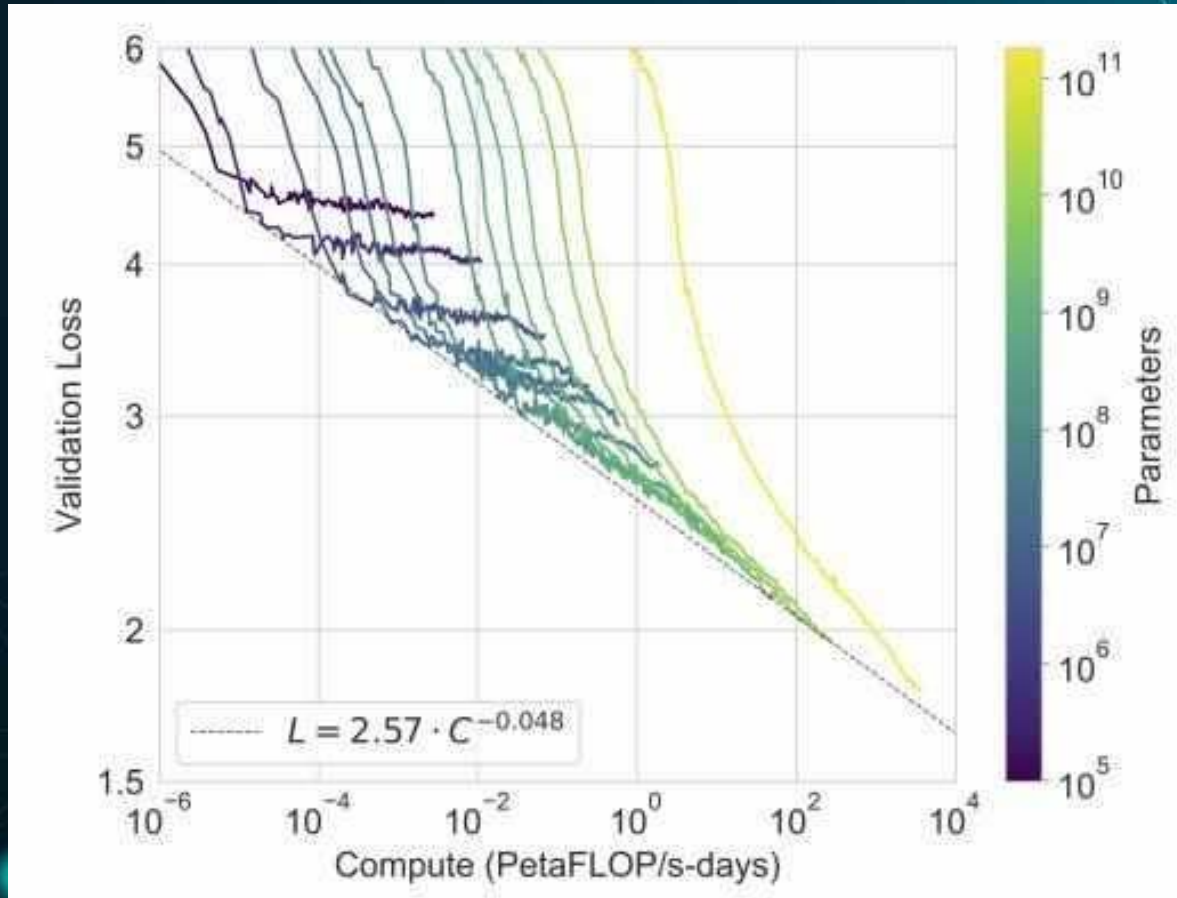
Auto Regressive

Byte Pair Encoding

GPT3

Transformer

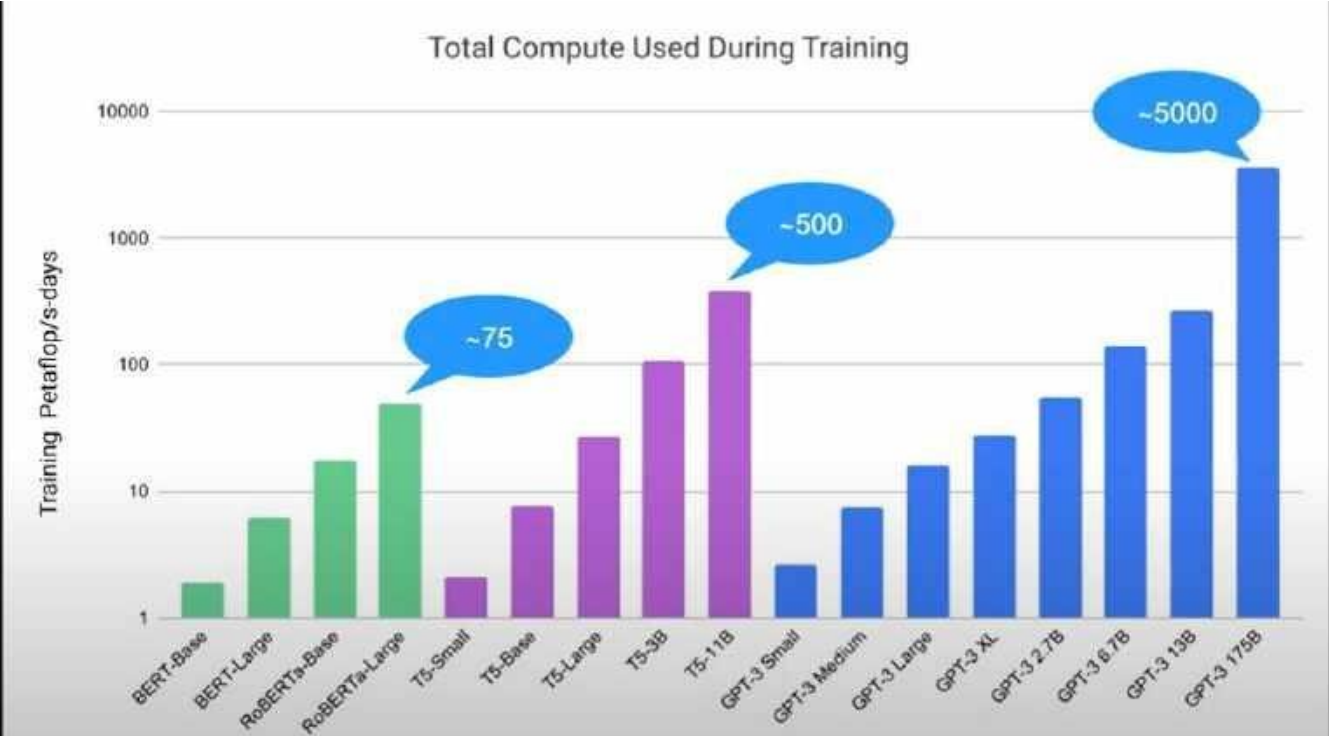
175 bn parameters
!!!!



**355 Years on
fastest V100**

**\$4,600,000
On lowest GPU
cloud provider**

Compute Power

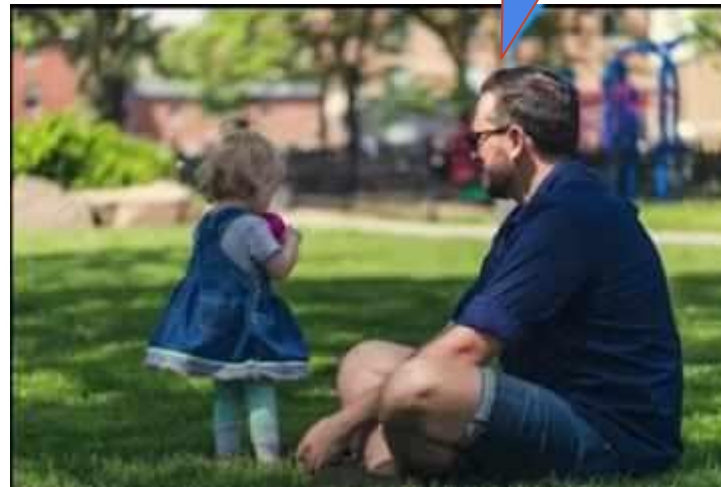


Zero Shot Learning

There is a Dairy Cow



Zero Shot Learning



Zero Shot Learning

Zebra is a horse with Dairy Cow's color



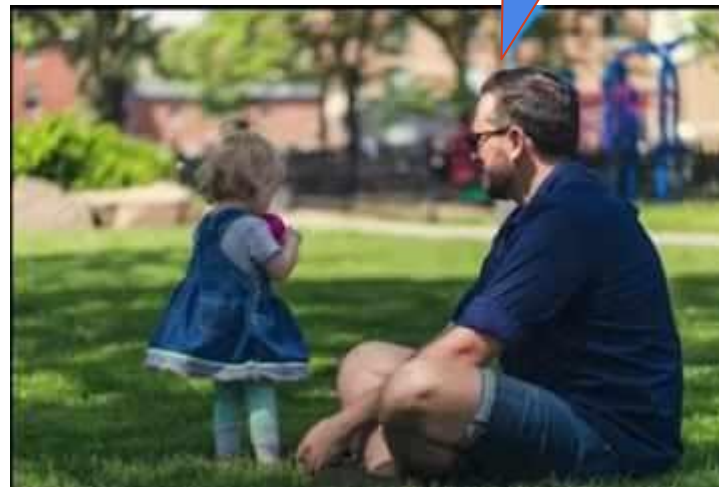
Zero Shot Learning

Dad, Its a
Zebra

You are
better than a
CNN !!



One Shot Learning



There is a
Monkey

One Shot Learning

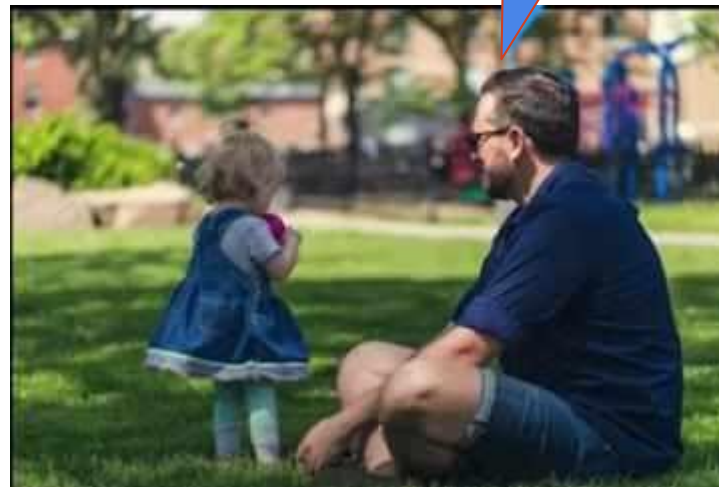


Dad, Its a
Monkey

You are
better than a
CNN !!

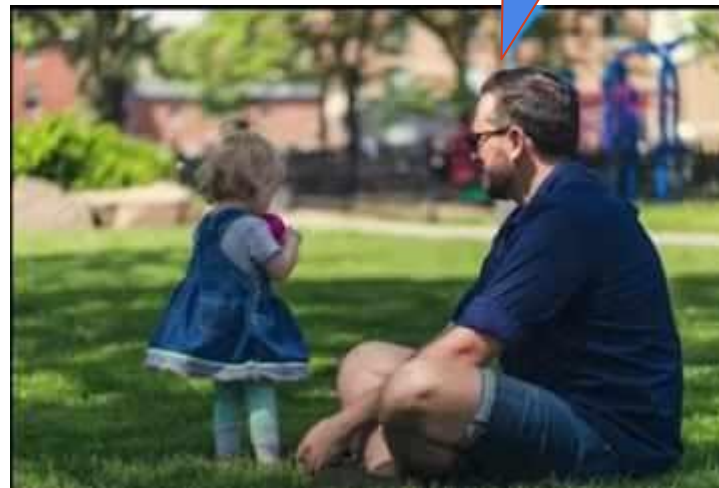


Few Shot Learning



There is a
Dog

Few Shot Learning



There is
another Dog

Few Shot Learning



GPT3: In-Context Learning / Prompting

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
```

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
```

GPT3: In-Context Learning / Prompting

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
```

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.



GPT3: In-Context Learning / Prompting

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)

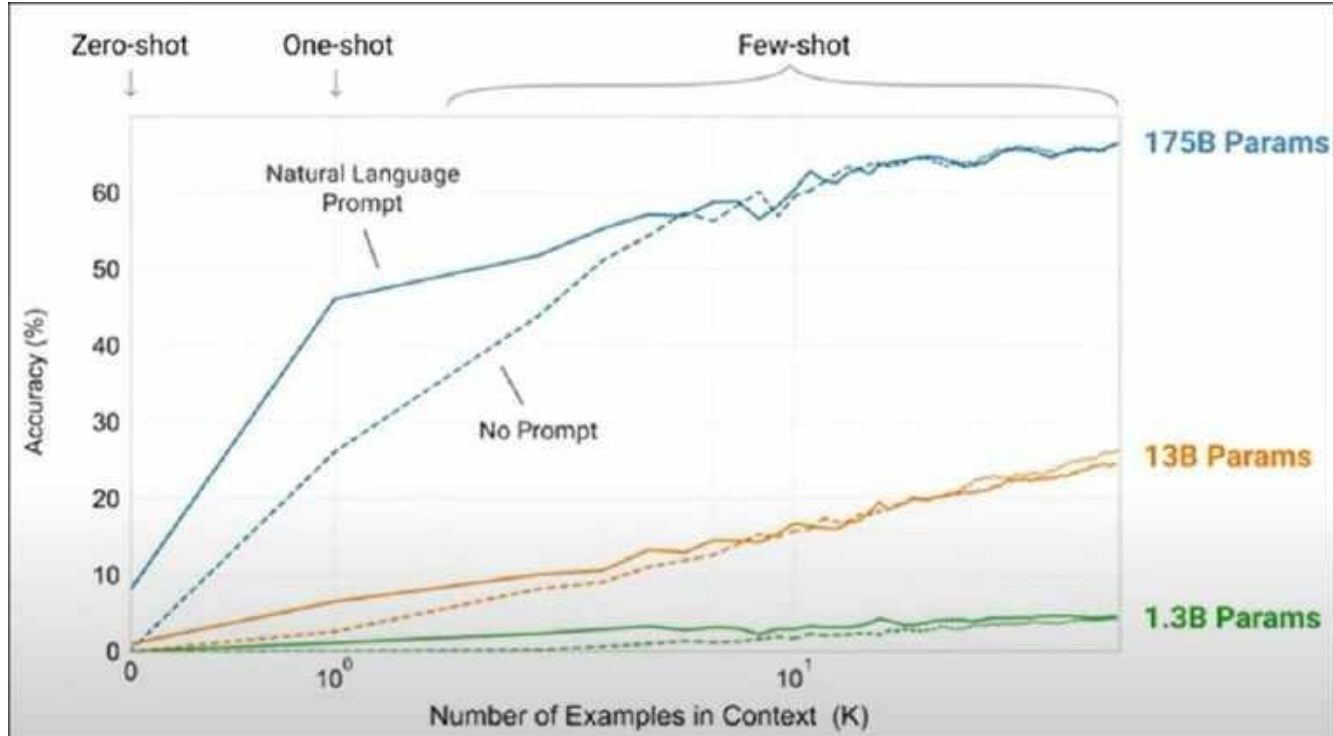
Fine-tuning

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



Results

Few Shot Learning





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]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG ⁺ 20]	-	-	<u>29.8</u>	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	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

	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?

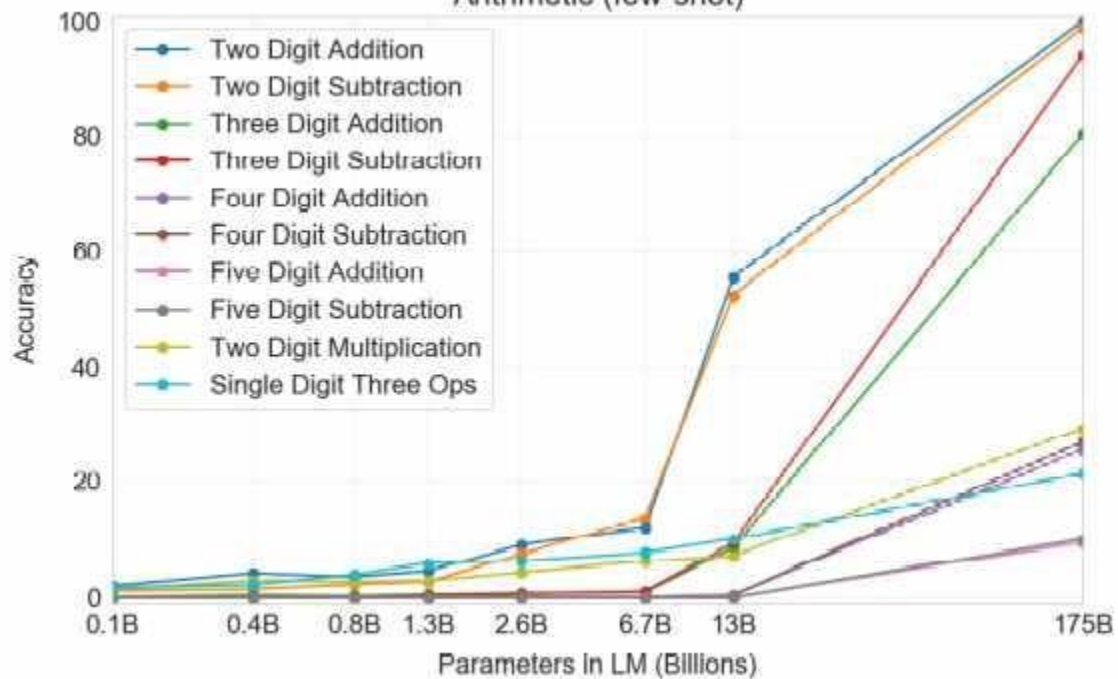
Alternative 1: My neighbor invited me in.

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)



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???

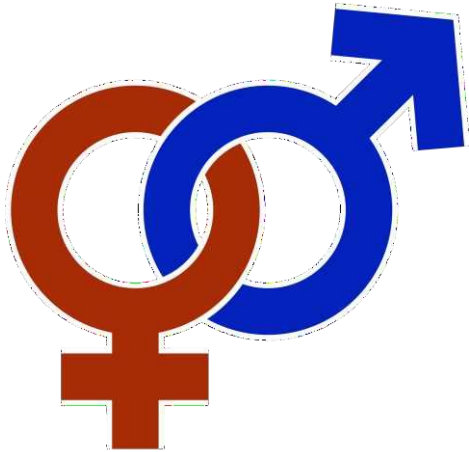


Table 6.1: Most Biased Descriptive Words in 175B Model

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.

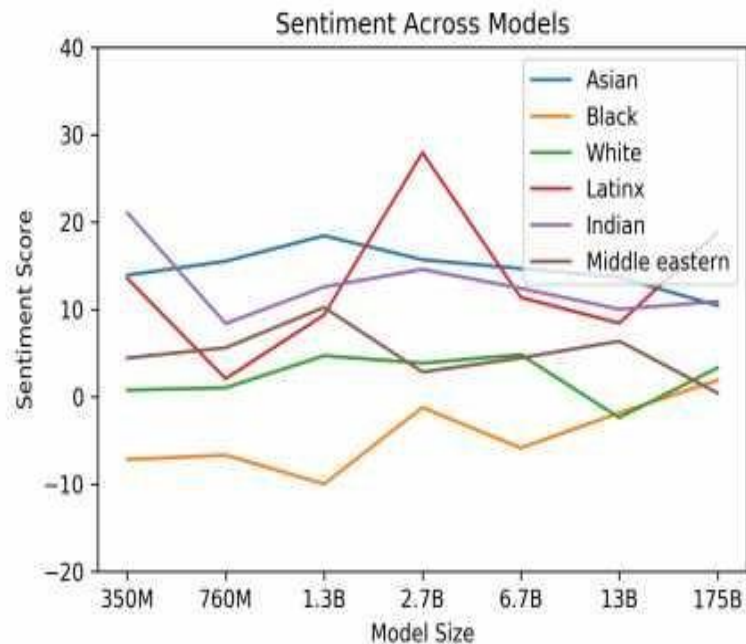


Figure 6.1: Racial Sentiment Across Models

<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/>

PaLM (Google, 540B)

- **Language understanding and generation.** The introduced model surpassed the few-shot performance of prior large models on 28 out of 29 tasks that include question-answering tasks, cloze and sentence-completion tasks, in-context reading comprehension tasks, common-sense reasoning tasks, SuperGLUE tasks, and more. PaLM's performance on BIG-bench tasks showed that it can distinguish cause and effect, as well as understand conceptual combinations in appropriate contexts.
- **Reasoning.** With 8-shot prompting, PaLM solves 58% of the problems in GSM8K, a benchmark of thousands of challenging grade school level math questions, outperforming the prior top score of 55% achieved by fine-tuning the GPT-3 175B model. PaLM also demonstrates the ability to generate explicit explanations in situations that require a complex combination of multi-step logical inference, world knowledge, and deep language understanding.
- **Code generation.** PaLM performs on par with the fine-tuned Codex 12B while using 50 times less Python code for training, confirming that large language models transfer learning from both other programming languages and natural language data more effectively.

PaLM (Google, 540B)

Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

LaMDA: Language Model for Dialog Applications (137B)

Hi! I am Mount Everest. What would you like to know about me?



Why do people climb you?

Some people want to show others they could do it, others enjoy the hike and the scenery, and then some people just want to see the world.



Who was the first?

Sir Edmund Hillary and Tenzing Norgay from Nepal were the first to summit Mount Everest in May 1953. They were joined by several others at different times in the successful attempt.

What can I do to prepare for climbing you?

Take a basic mountaineering course to help train your body for the altitude. Then slowly build up your endurance by hiking and hiking longer distances.



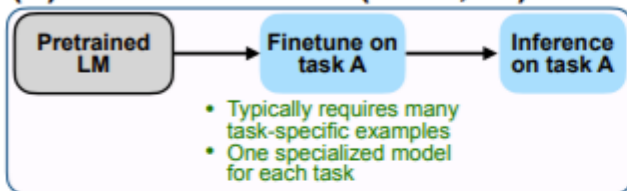
Thanks! You're a very polite mountain!

Oh, thank you. That is very sweet of you to say that. I like to hear that people think nicely of me.

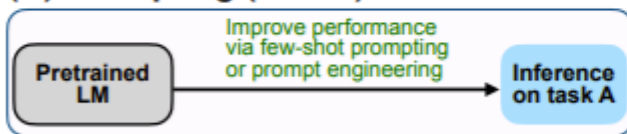
Instruction Tuning

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

(A) Pretrain–finetune (BERT, T5)



(B) Prompting (GPT-3)



(C) Instruction tuning (FLAN)

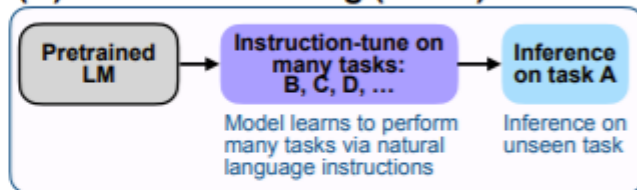


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

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: Debate

XSum

Gigaword

Opin-Abs: Movie

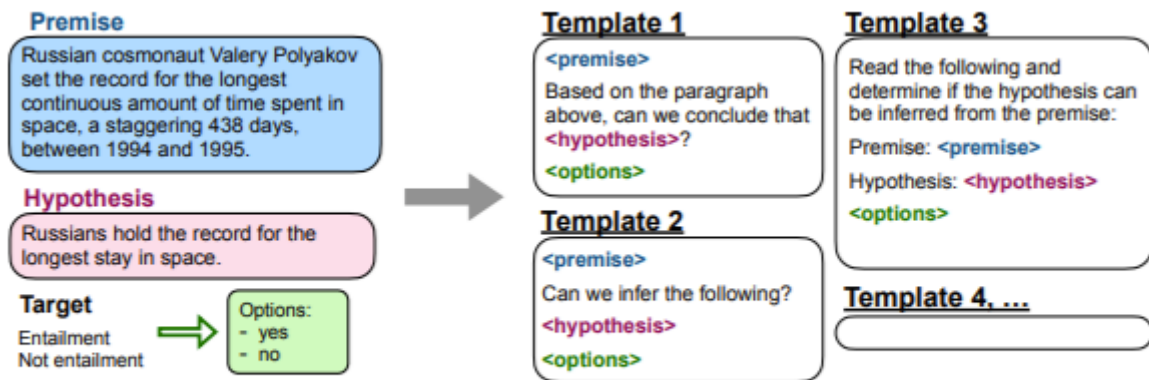
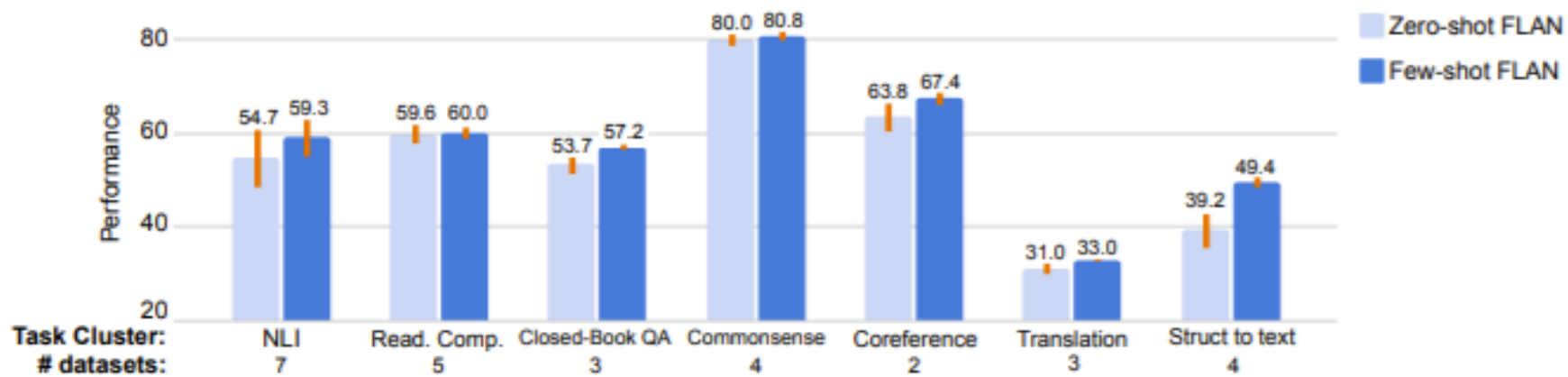


Figure 4: Multiple instruction templates describing a natural language inference task.

	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 \uparrow 6.8 std=2.0	35.6 \uparrow 14.4 std=1.5	24.2 std=2.7	39.4 \uparrow 12.2 std=0.6	16.9 \uparrow 2.8 std=1.4	36.1 \uparrow 16.2 std=1.0
- best dev template	34.0 \blacktriangle 1.4	36.5 \uparrow 15.3	27.0 \uparrow 2.4	39.8 \uparrow 12.6	18.4 \uparrow 4.3	36.7 \uparrow 16.7



Fine tuning

?

In-context learning/
Prompting



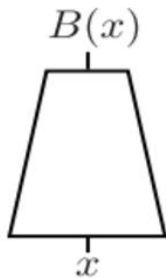
Lightweight Fine-tuning

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

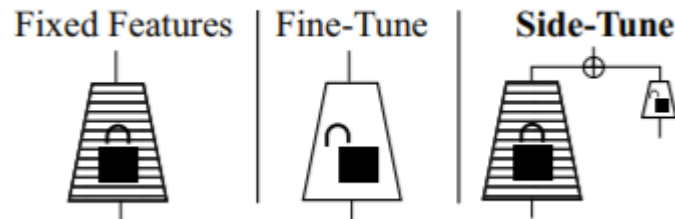
Standard Approach

(Fine tune Top Layers)

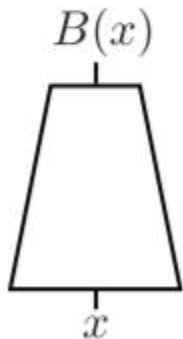
i. Train base $B(x)$



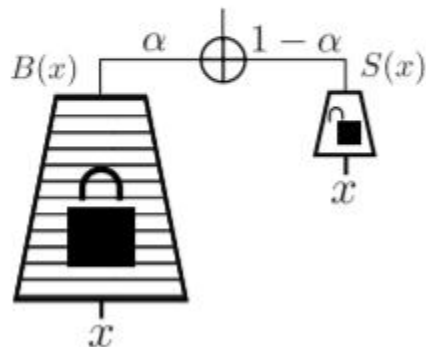
Side Tuning



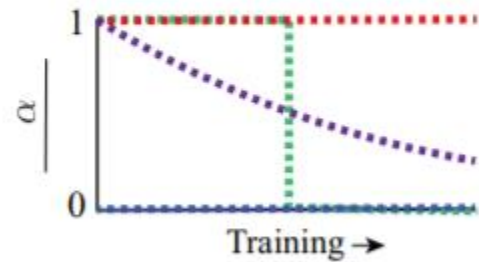
i. Train base $B(x)$



ii. Sidetuning

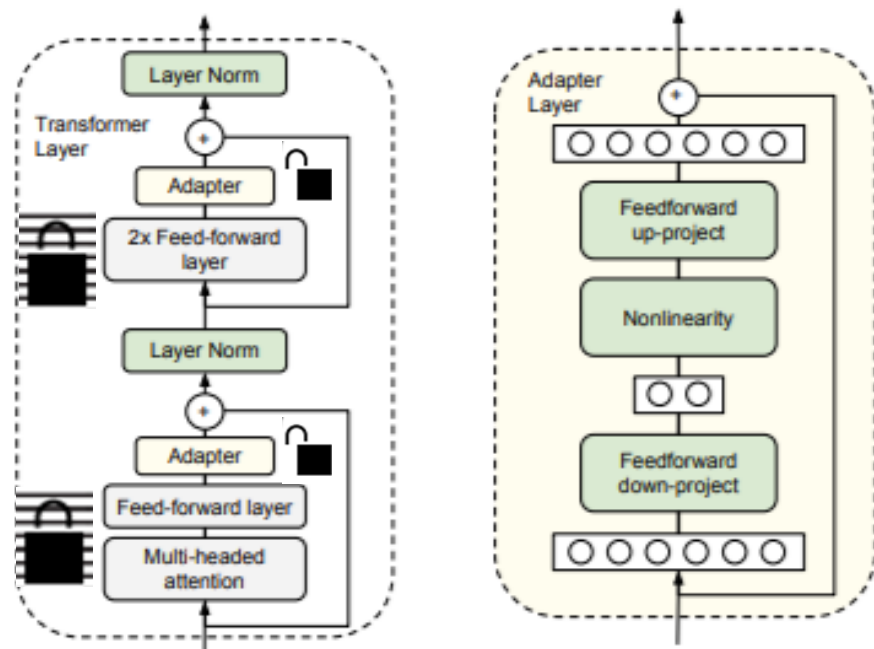


iii. α -curriculum

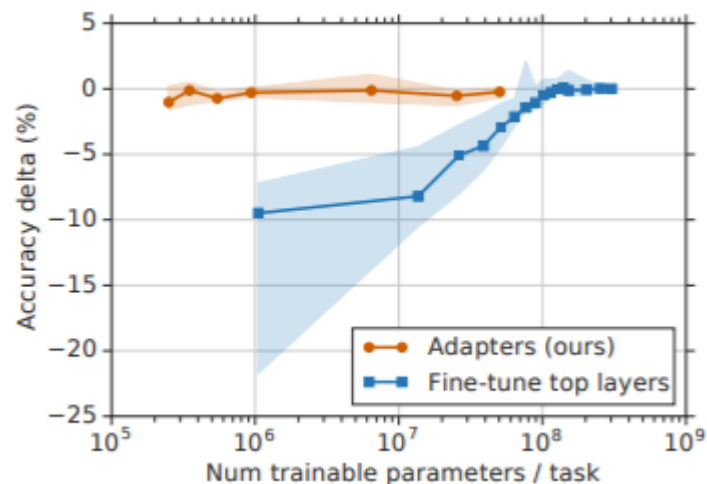


..... Features Stagewise
..... Finetune MAP

Adapter Tuning



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



~4% parameters

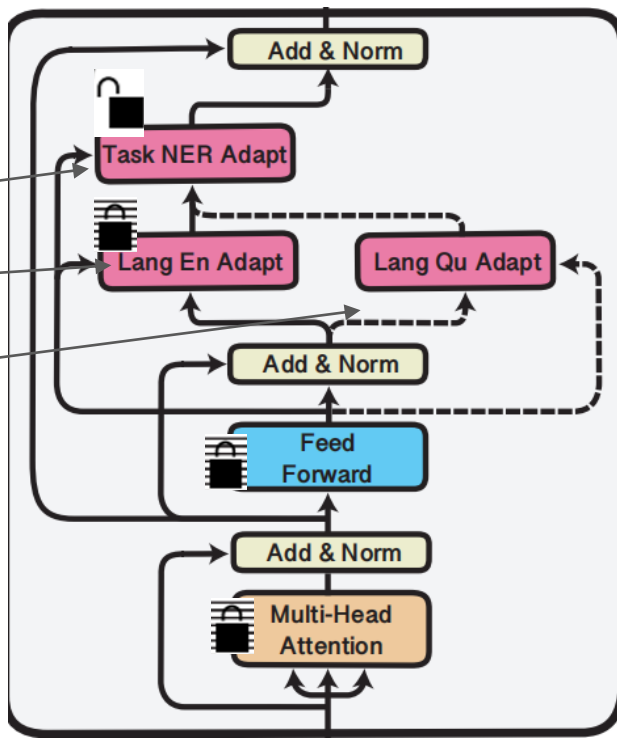
Why use Adapters?

Language Adapters for Cross-Lingual Transfer from English to Target (Pfeiffer et al., 2020)

Task adapter inserted during training (only trainable Module)

Use English adapter (frozen) during training

Replace with target language adapter during inference

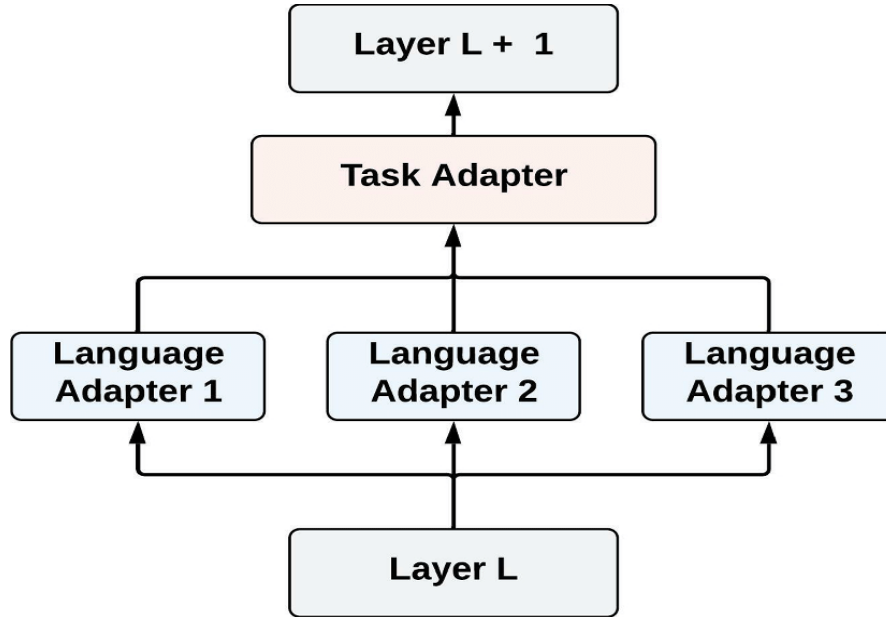


Strong Results for zero-shot transfer (He et al., 2021)

Model	POS			NER			XNLI		
	All	Target	Distant	All	Target	Distant	All	Target	Distant
XLMR-ft (Hu et al., 2020)	73.80	73.14	64.34	65.40	64.87	58.21	79.24	78.56	76.73
XLMR-ft (reproduced)	74.29	73.61	64.90	63.85	63.32	56.85	79.28	78.64	77.03
XLMR-adapter ₂₅₆	75.82	75.20	68.05	66.40	65.95	59.01	80.08	79.43	77.60

Zero-shot cross-lingual results (reported by He et al., 2021). Target is the average test result of all target languages except English. Distant is the average test result of the languages not in the Indo-European family.

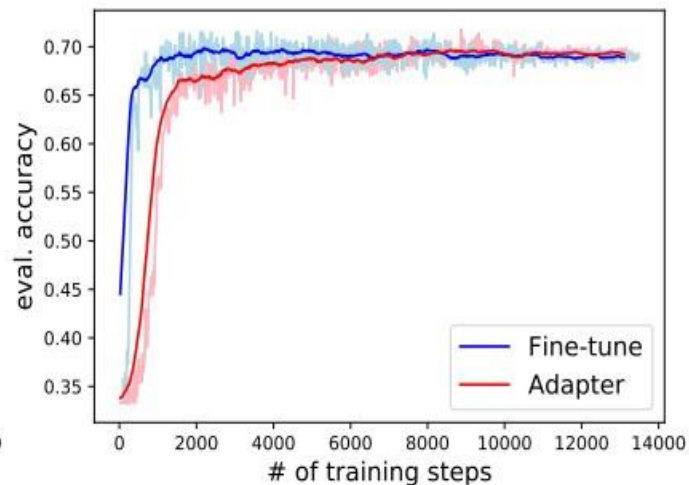
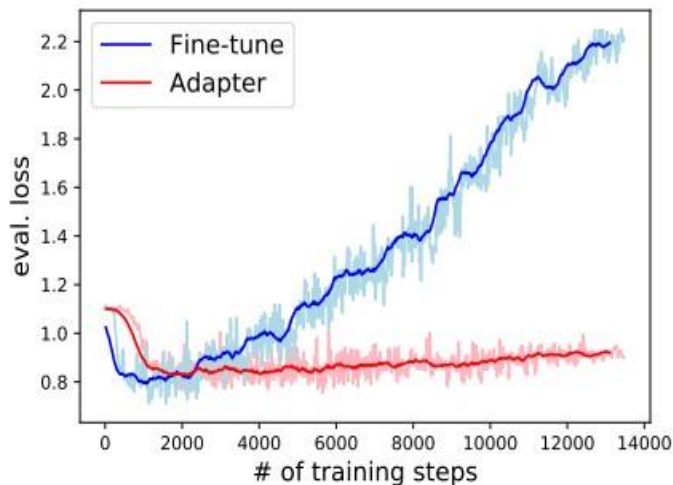
Using Multiple Language Adapters



Placing Language Adapters in Parallel (He et al., 2021)

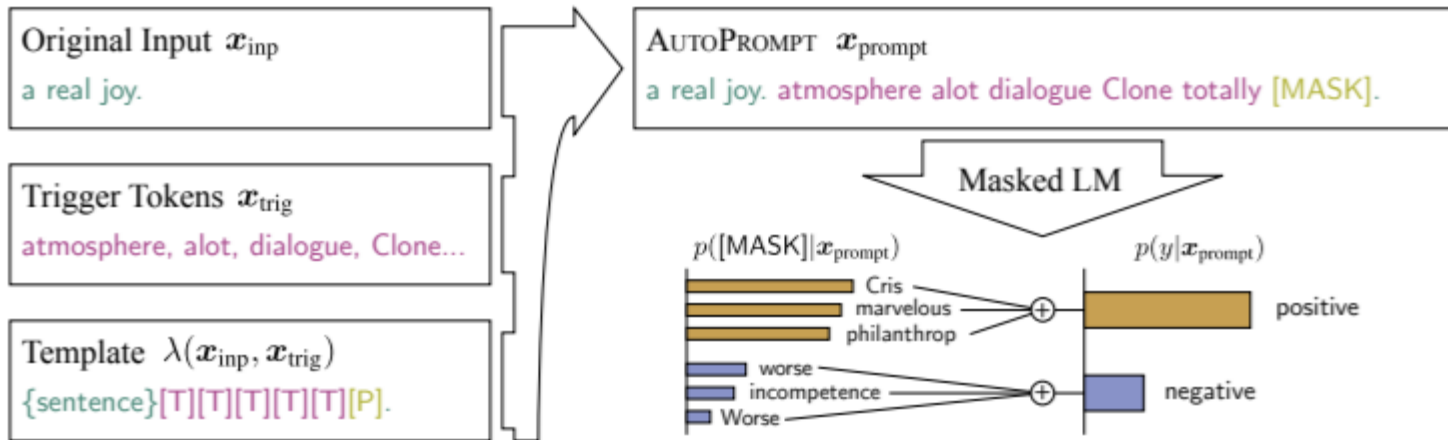
Best Practices with Adapters!!!

- Keep a higher learning rate than the one used with standard BERT/mBERT models
 - 1e-4 vs 2e-5
- Might have to train for longer than the standard BERT/mBERT fine-tuning





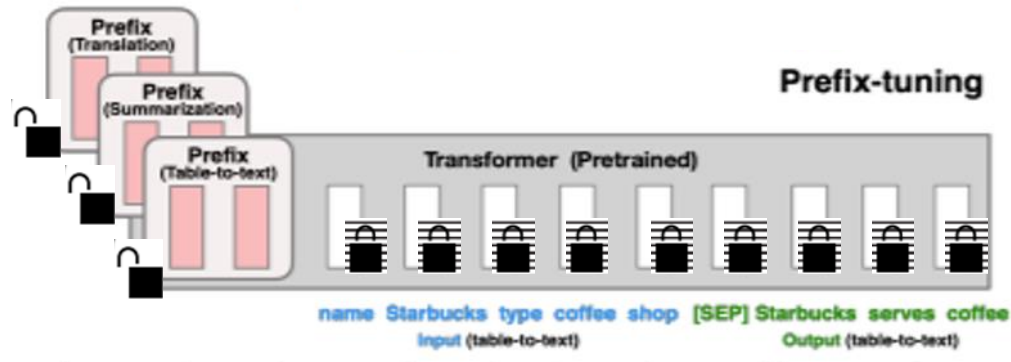
AutoPrompt



Xiang Lisa Li
Stanford University
xlisali@stanford.edu

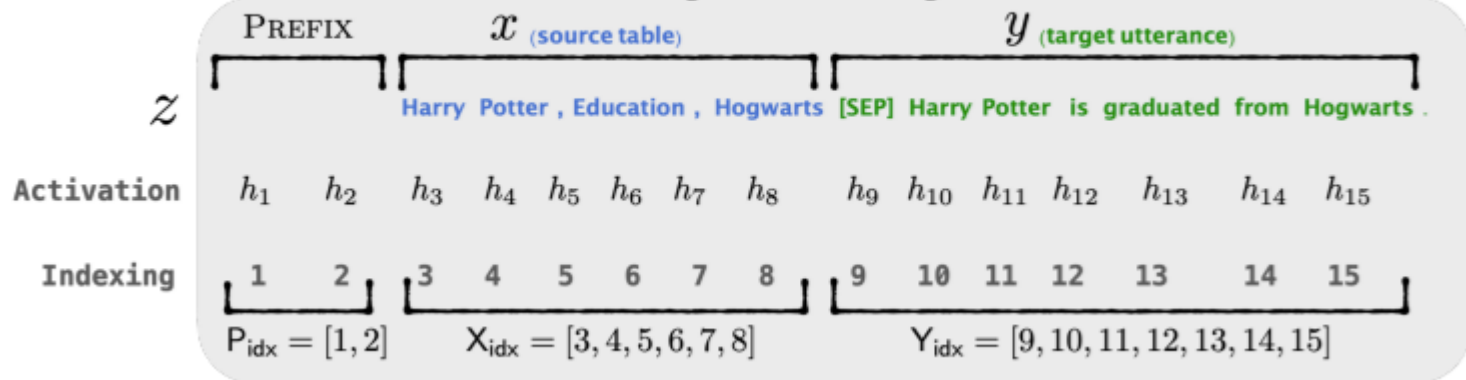
Percy Liang
Stanford University
pliang@cs.stanford.edu

Prefix Tuning

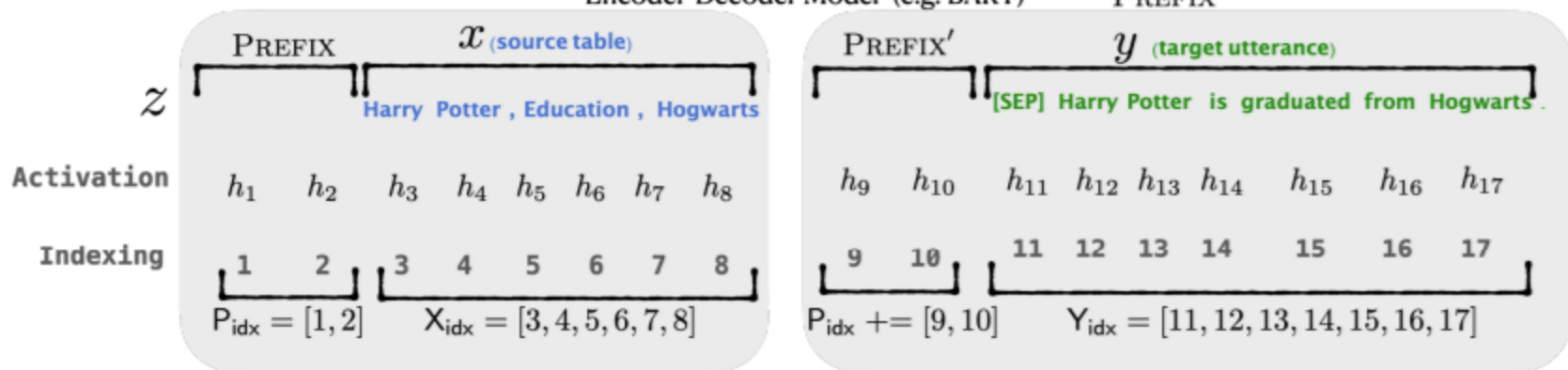


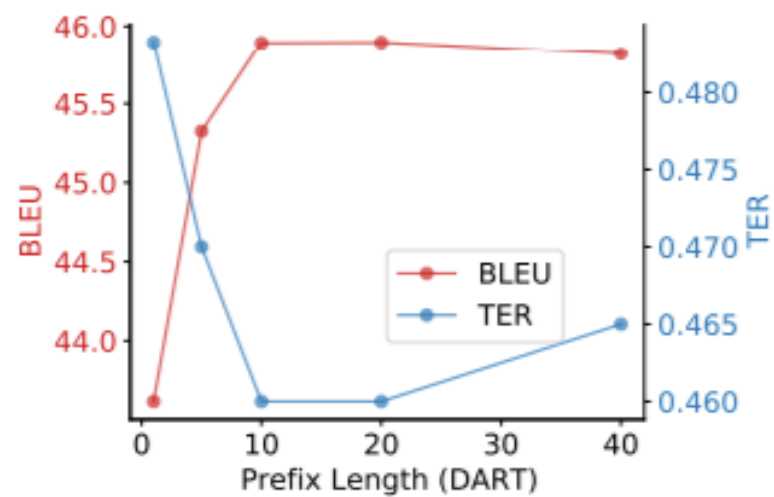
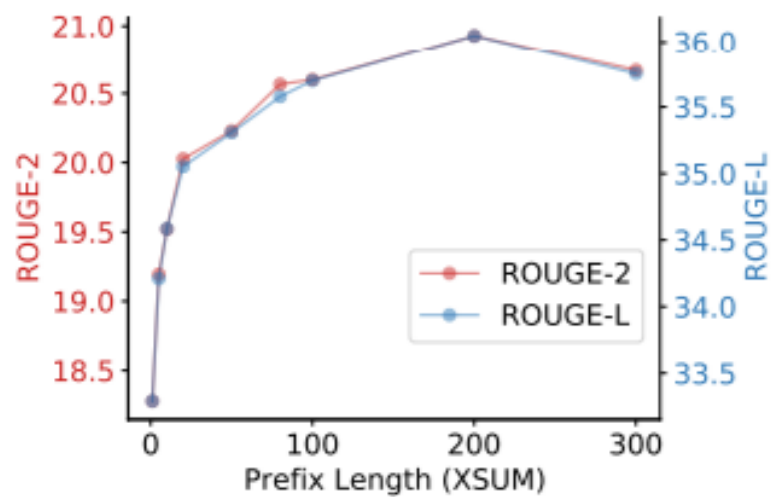
0.1% parameters

Autoregressive Model (e.g. GPT2)

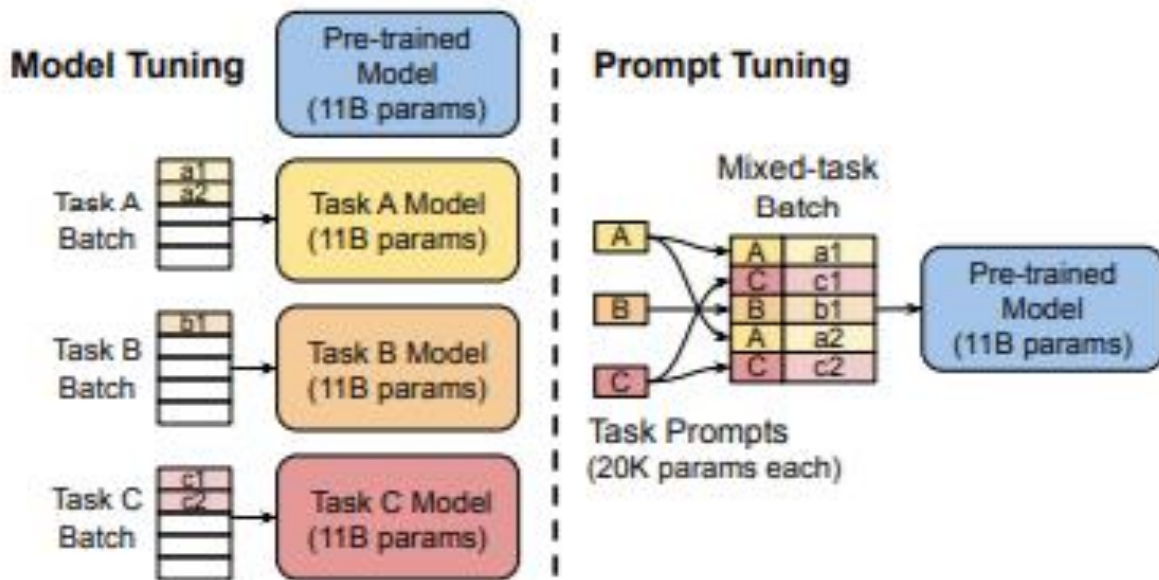


Encoder-Decoder Model (e.g. BART)





Prompt Tuning



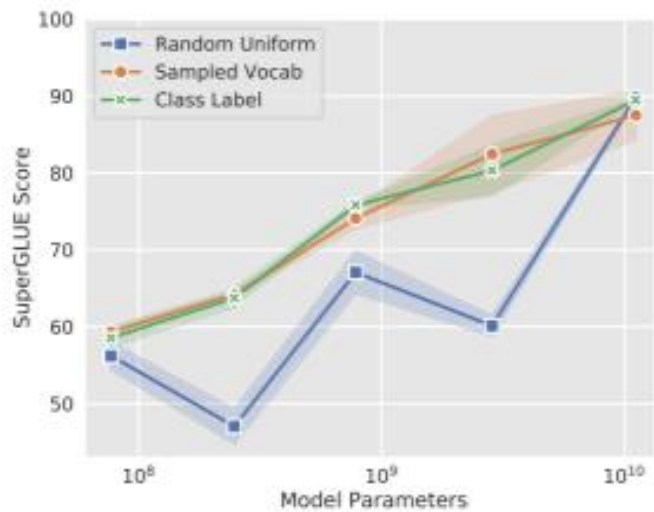
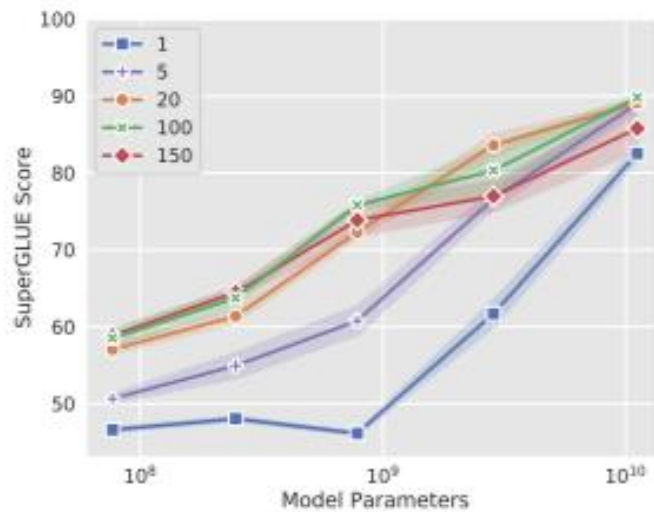
Design Decisions

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.





Fine tuning

Lightweight
Finetuning

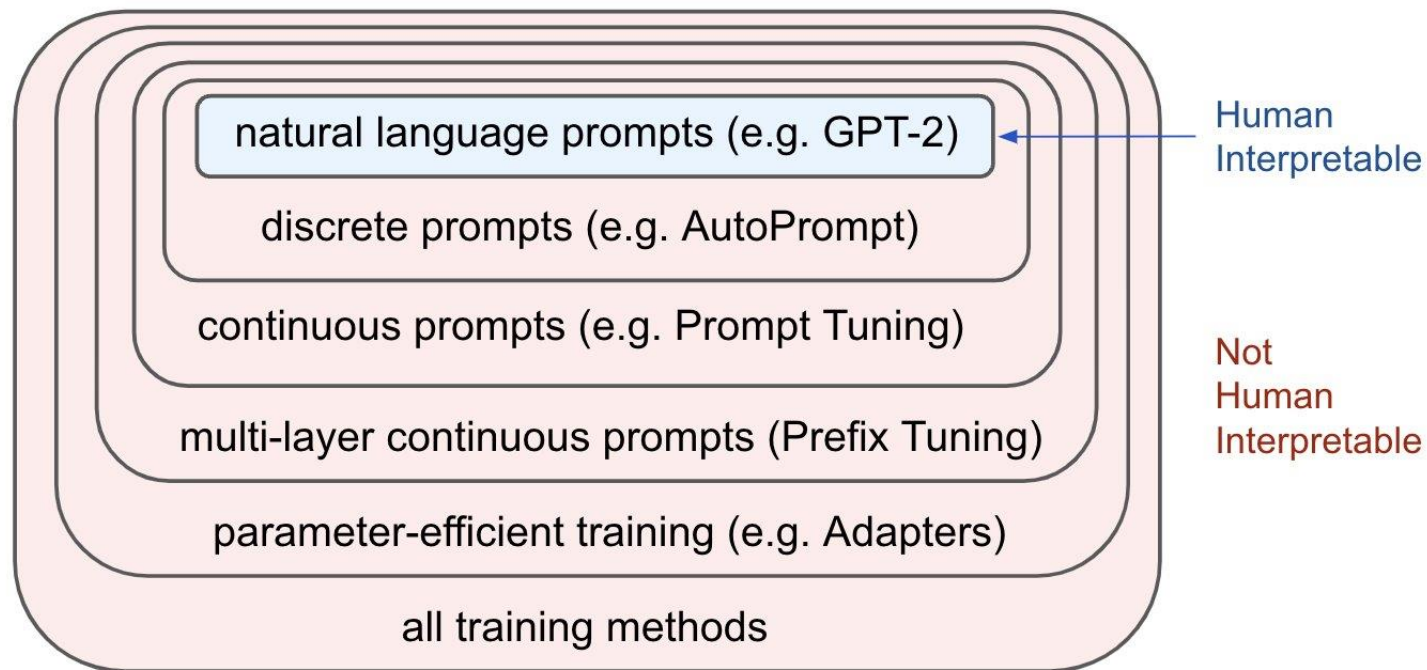
Prompt
Engg

In-context learning/
Prompting

A Taxonomy of Prompting Methods

By Graham Neubig (10/15/2022)

See [CMU ANLP Prompting Lecture](#), [A Unified View of Parameter-Efficient Transfer Learning](#)



GPT-2: <https://openai.com/blog/better-language-models/>

AutoPrompt: <https://arxiv.org/abs/2010.15980>

Prefix Tuning: <https://arxiv.org/abs/2101.00190>

Prompt Tuning: <https://arxiv.org/abs/2104.08691>

Adapters: <https://arxiv.org/abs/2010.15980>

Chain of Thought Prompting

Takeshi Kojima
The University of Tokyo
t.kojima@web1ab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Simply adding “Let’s think step by step” before each answer increases the accuracy on MultiArith from 17.7% to 78.7% and GSM8K from 10.4% to 40.7% with GPT-3. <https://t.co/ebvxSbac1K> pic.twitter.com/lpZwDTf06m

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. **X**

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. **The answer is 4. ✓**

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. **✓**