Prompting and in-context learning

THE UNIVERSITY OF UTAH

Credits: Daniel Kashabi, Colin Raffel, a lot of this material is based on the ACL 2022 tutorial on Zero- and Few-Shot NLP with Pretrained Language Models by Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, Sameer Singh

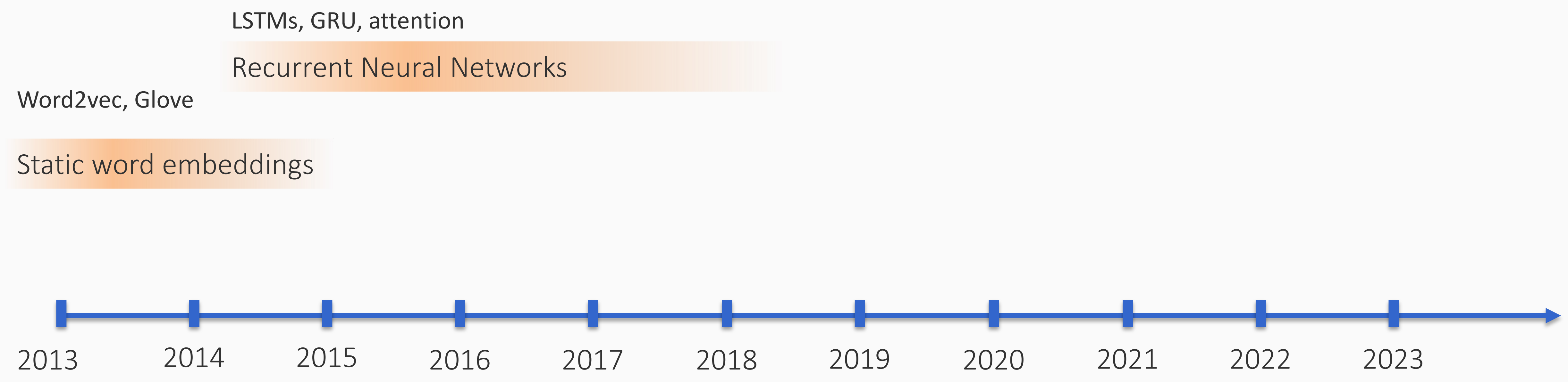
Where are we?

Word2vec, Glove

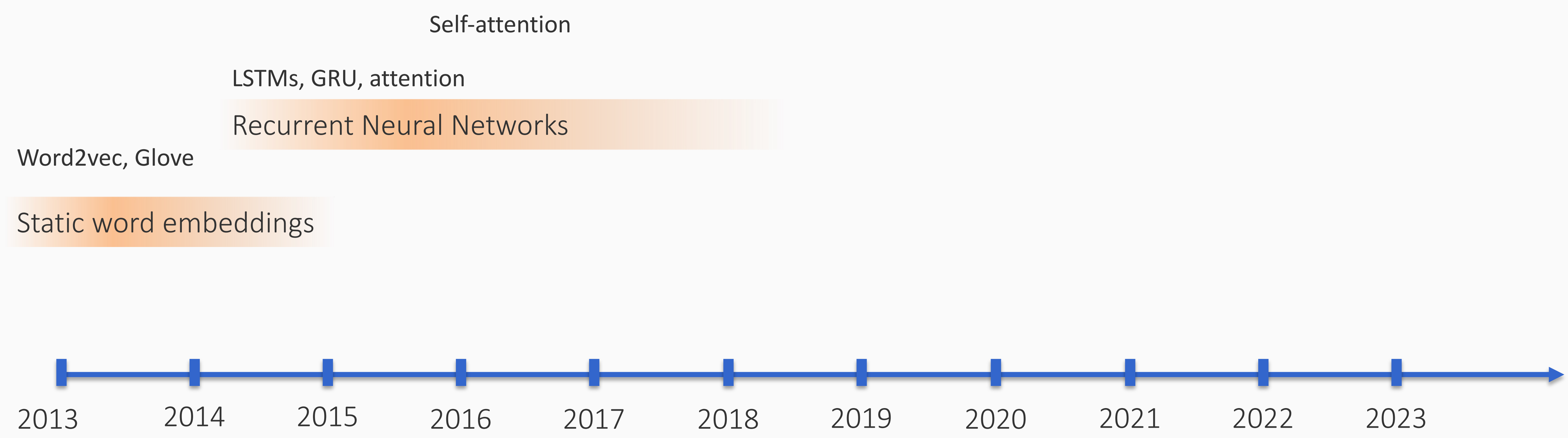
Static word embeddings

2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

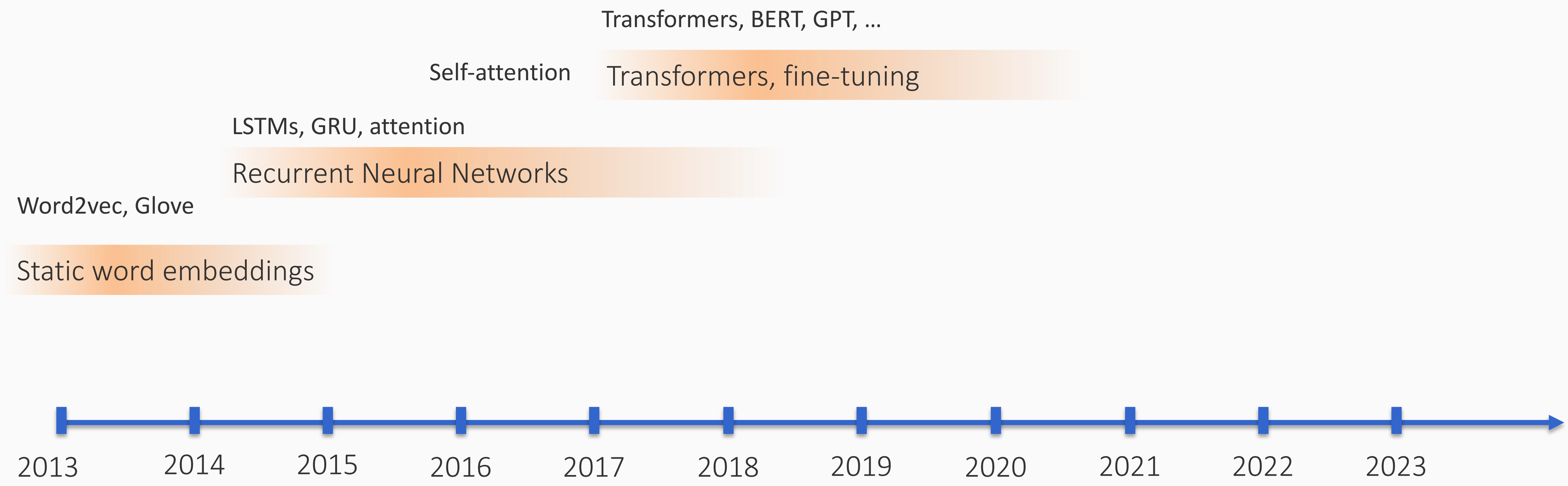
Where are we?


LSTMs, GRU, attention
Recurrent Neural Networks
Word2vec, Glove
Static word embeddings
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

Where are we?


Self-attention
LSTMs, GRU, attention
Recurrent Neural Networks
Word2vec, Glove
Static word embeddings
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

Where are we?


Transformers, BERT, GPT, …
Self-attention
Transformers, fine-tuning
LSTMs, GRU, attention
Recurrent Neural Networks
Word2vec, Glove
Static word embeddings
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

The BERT-flavored model: A recipe

1. Start with a pre-trained transformer
2. Collect a dataset for your task
3. Fine-tune the pre-trained transformer for your model

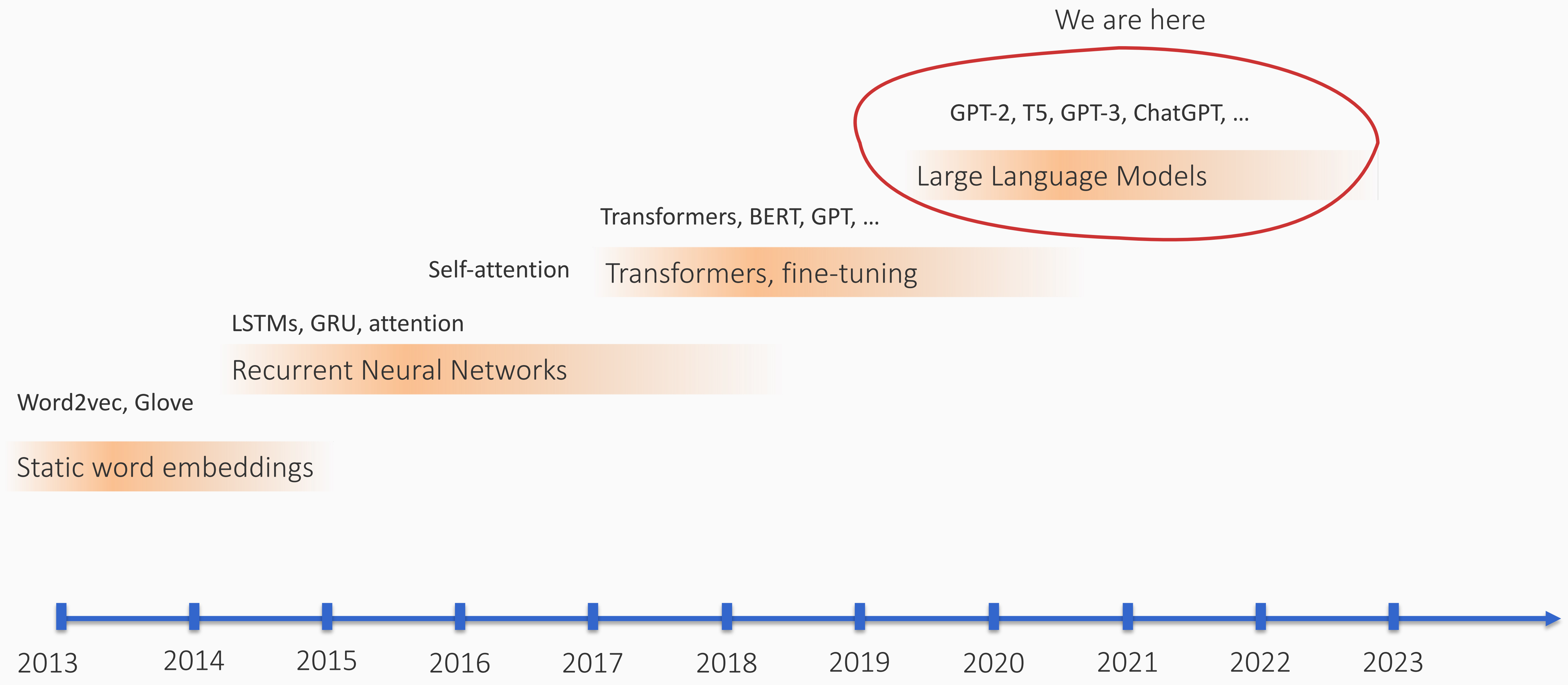
Does this recipe always work?

“I have an extremely large collection of clean labeled data”

- No one

[Colin Raffel]

Where are we?


We are here
GPT-2, T5, GPT-3, ChatGPT, …
Large Language Models
Transformers, BERT, GPT, …
Self-attention
Transformers, fine-tuning
LSTMs, GRU, attention
Recurrent Neural Networks
Word2vec, Glove
Static word embeddings
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

This lecture

- Zero- and few-shot prediction
- Prompting language models a.k.a. in-context learning
- Does prompting work?

This lecture

- Zero- and few-shot prediction

- Prompting language models a.k.a. in-context learning

- Does prompting work?

We tend to be really good at zero-shot predictions

The movie was a two hour masterclass

Positive?

Negative?

We tend to be really good at zero-shot predictions

The movie was a two hour masterclass

Positive? ✓

Negative?

We tend to be really good at zero-shot predictions

The movie was a two hour masterclass

Positive? ✓

Negative?

This movie doesn't scrape the bottom of the barrel. This movie isn't the bottom of the barrel. This movie isn't below the bottom of the barrel. This movie doesn't deserve to be mentioned in the same sentence with barrels.

(from Roger Ebert's review of "Freddie Got Fingered")

Positive?

Negative?

We tend to be really good at zero-shot predictions

The movie was a two hour masterclass

Positive? ✓

Negative?

This movie doesn't scrape the bottom of the barrel. This movie isn't the bottom of the barrel. This movie isn't below the bottom of the barrel. This movie doesn't deserve to be mentioned in the same sentence with barrels.

Positive?

Negative? ✓

(from Roger Ebert's hilariously negative review of "Freddie Got Fingered")

We tend to be really good at zero-shot predictions

The movie was a two hour masterclass

Positive? ✓

Negative?

This movie doesn't scrape the bottom of the barrel. This movie isn't the bottom of the barrel. This movie isn't below the bottom of the barrel. This movie doesn't deserve to be mentioned in the same sentence with barrels.

Positive?

Negative? ✓

(from Roger Ebert's hilariously negative review of "Freddie Got Fingered")

How were you able to predict the label?
Did you have have access to a labeled sentiment dataset? Did you train yourself on the data?

What can you do without large (or any) training datasets?

Answer: *Few- or Zero-shot learning*

Learning to perform a task with minimal task description, and perhaps a small number of examples (~10)

What can you do without large (or any) training datasets?

Answer: *Few- or Zero-shot learning*

Learning to perform a task with minimal task description, and perhaps a small number of examples (~10)

Why is this interesting?

What can you do without large (or any) training datasets?

Answer: *Few- or Zero-shot learning*

Learning to perform a task with minimal task description, and perhaps a small number of examples (~10)

Why is this interesting?

Practically useful

- Labeling data costs money and takes expertise
- Fine-tuning models is computationally expensive

What can you do without large (or any) training datasets?

Answer: *Few- or Zero-shot learning*

Learning to perform a task with minimal task description, and perhaps a small number of examples (~10)

Why is this interesting?

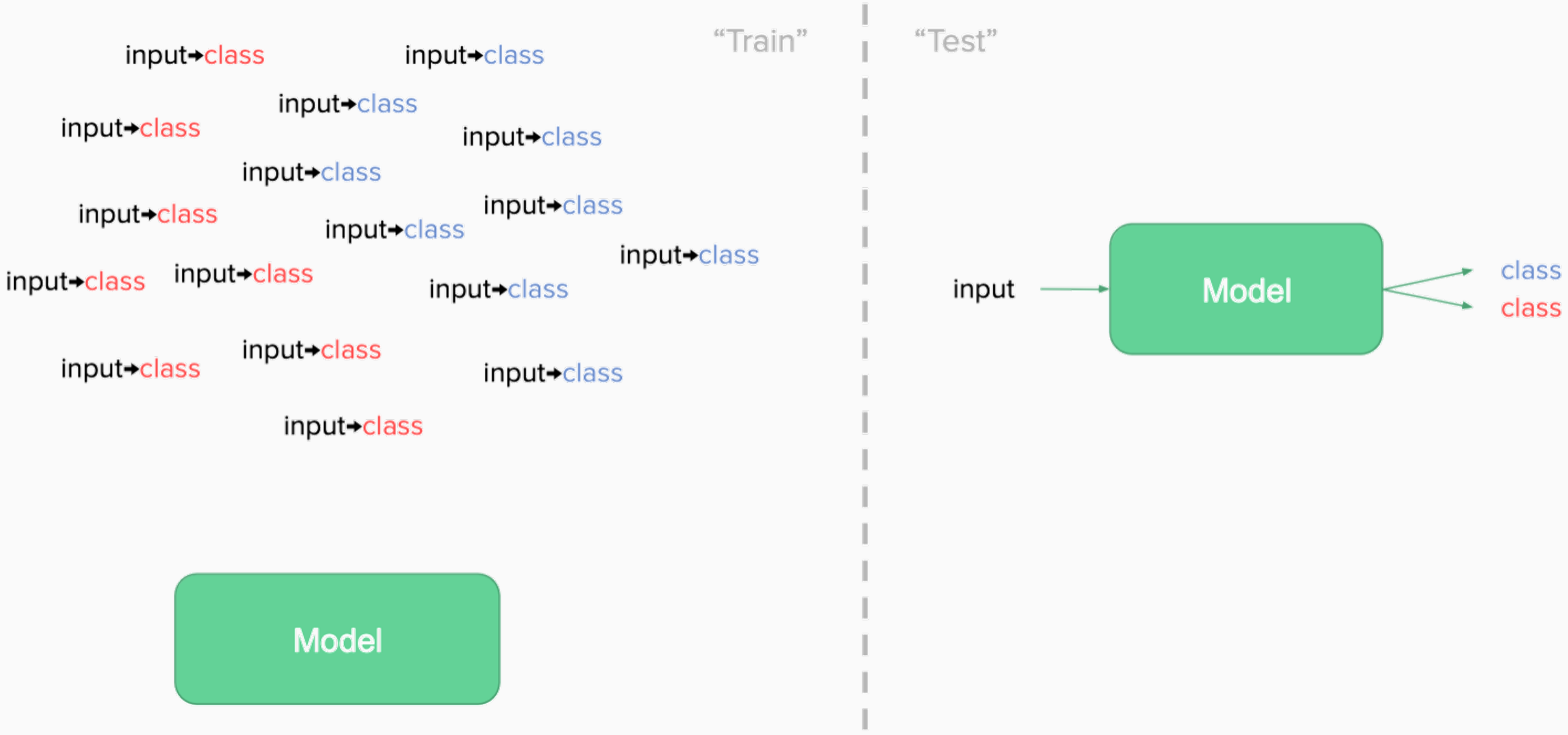
Practically useful

- Labeling data costs money and takes expertise
- Fine-tuning models is computationally expensive

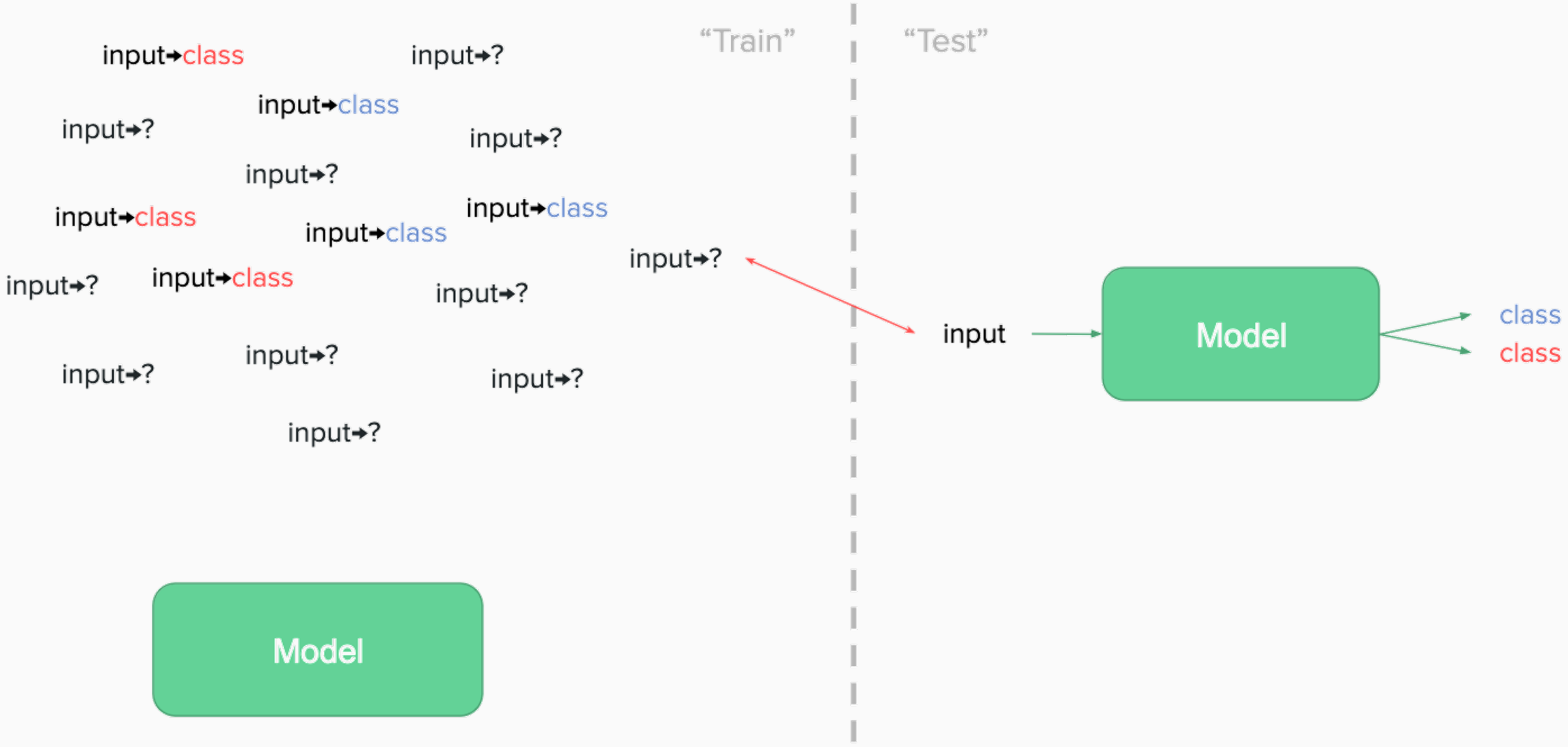
Scientifically interesting

- Generalizing correctly from a small number of examples is a good test of intelligence
- Provides insights into what models encode

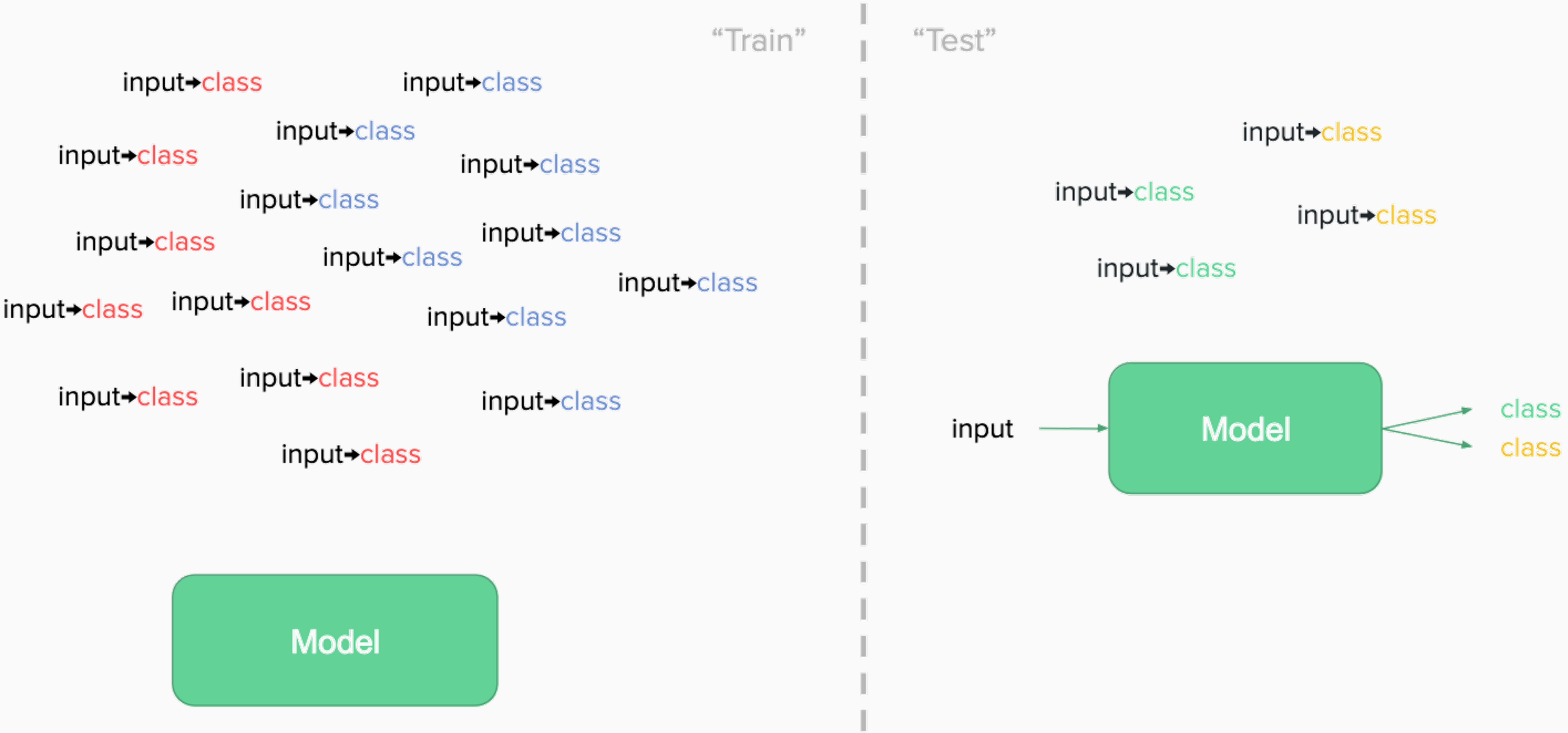
Supervised learning



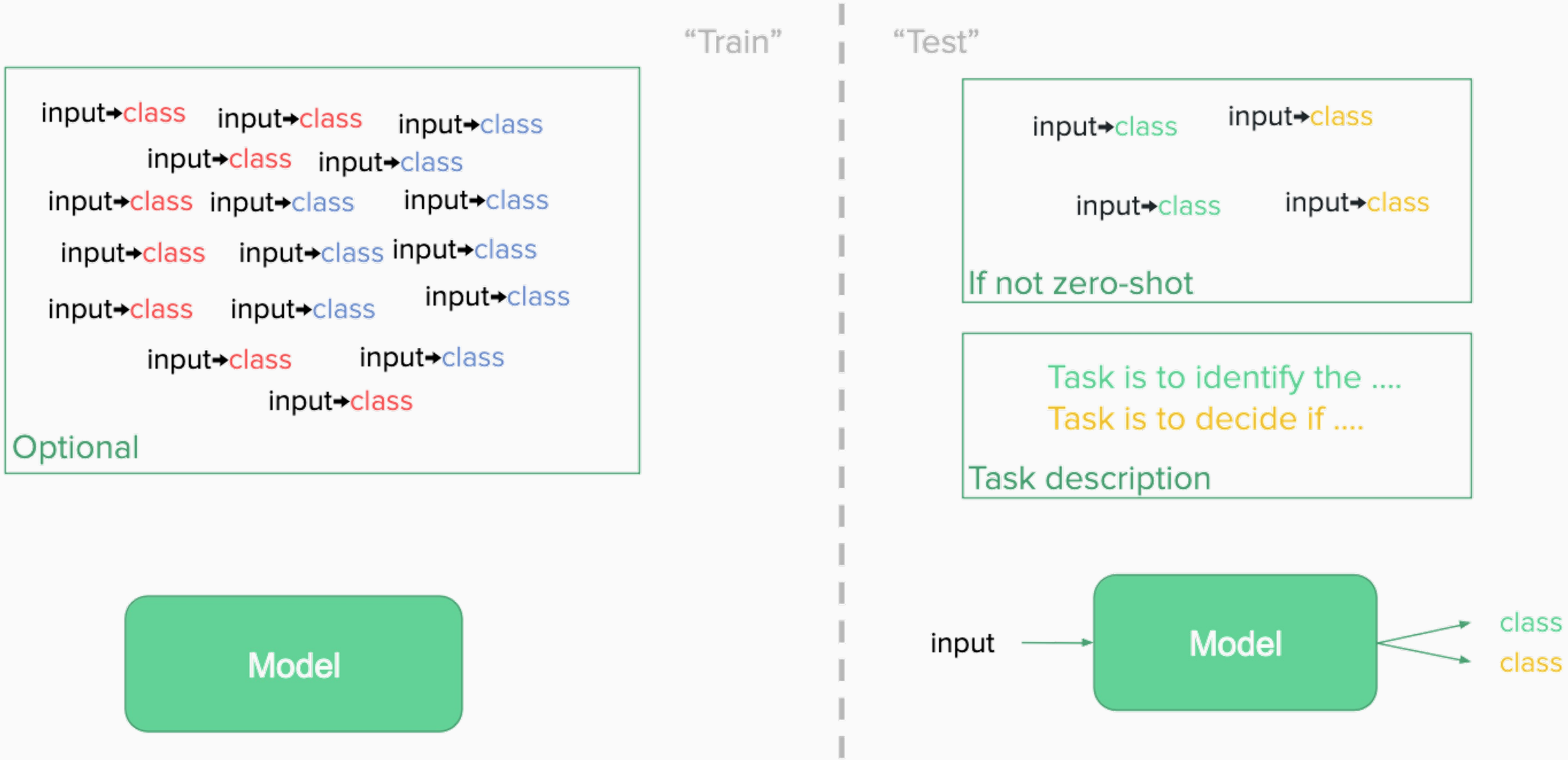
Semi-supervised learning



Few-shot learning (before LLMs)



Modern few-shot learning



This lecture

- Zero- and few-shot prediction
- Prompting language models a.k.a. in-context learning
- Does prompting work?

Zero- and Few-shot predictions with a language model

Zero-shot

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

```
Translate English to French:
cheese =>
```

task description

prompt

From: Brown et al. 2020. “Language Models are Few-Shot Learners”

Zero- and Few-shot predictions with a language model

Zero-shot

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

One-shot

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

From: Brown et al. 2020. “Language Models are Few-Shot Learners”

Zero- and Few-shot predictions with a language model

Zero-shot

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

```
Translate English to French:        task description
cheese =>                           prompt
```

One-shot

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

```
Translate English to French:        task description
sea otter => loutre de mer          example
cheese =>                           prompt
```

Few-shot

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

```
Translate English to French:        task description
sea otter => loutre de mer          examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           prompt
```

From: Brown et al. 2020. “Language Models are Few-Shot Learners”

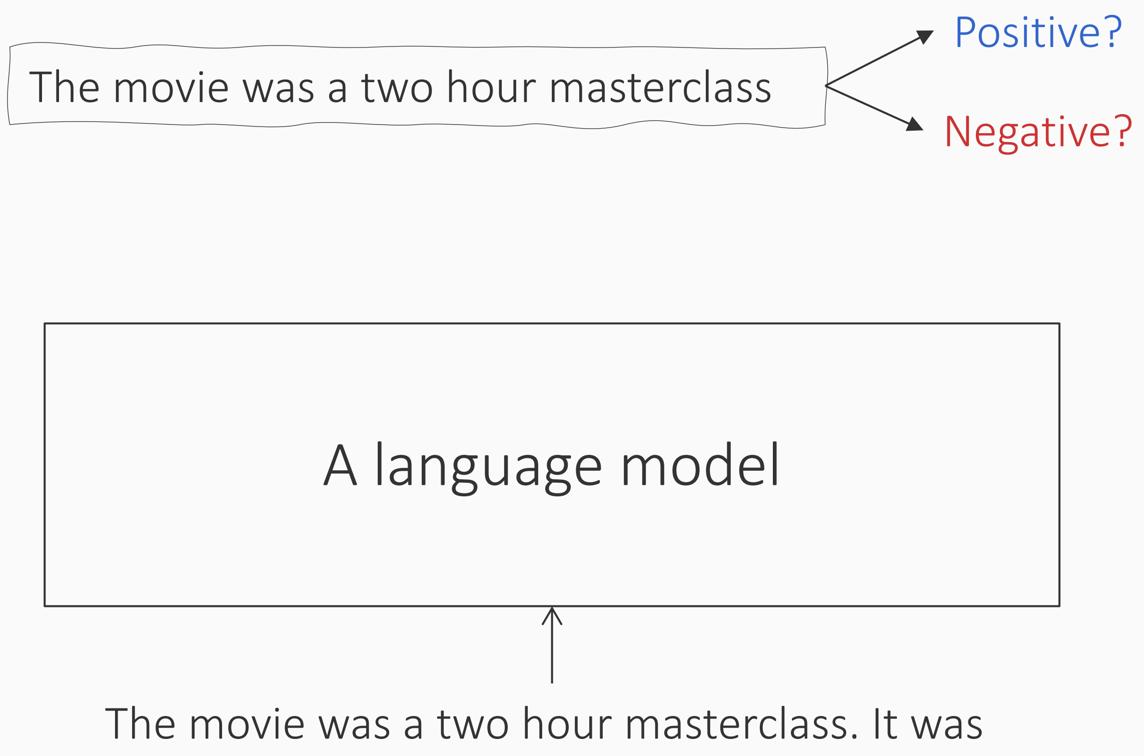
A concrete example: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

A concrete example: Sentiment classification



A concrete example: Sentiment classification

The movie was a two hour masterclass

Positive?

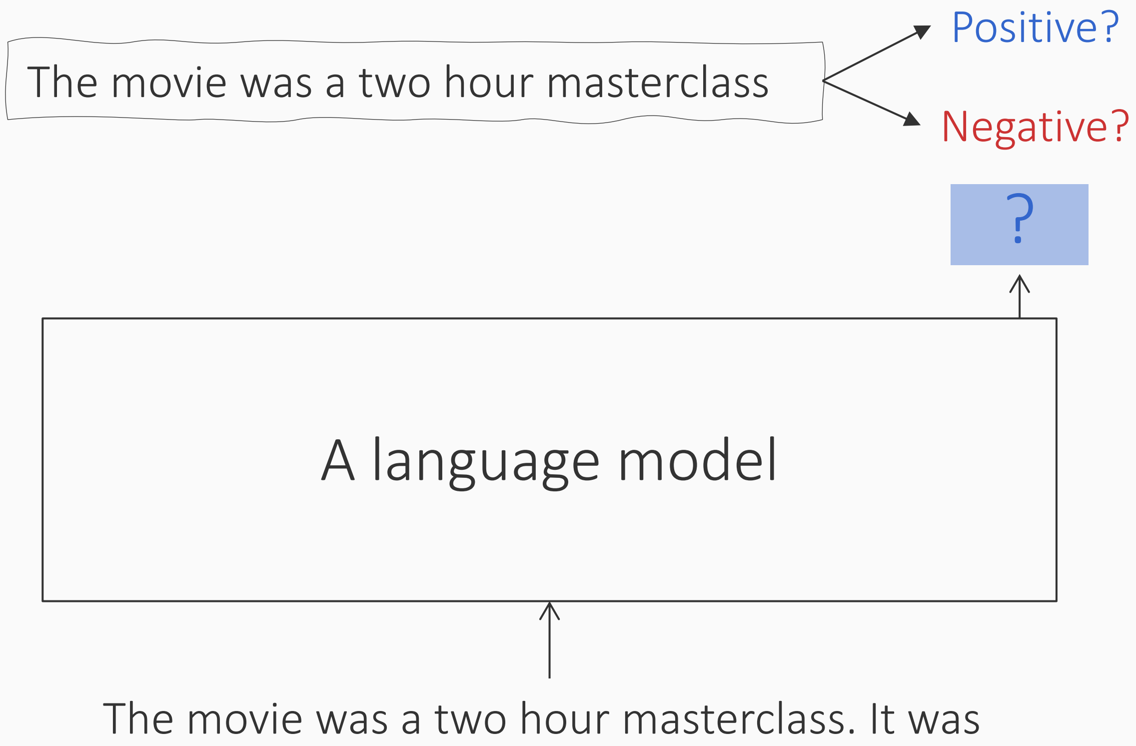
Negative?

A language model

The movie was a two hour masterclass. It was

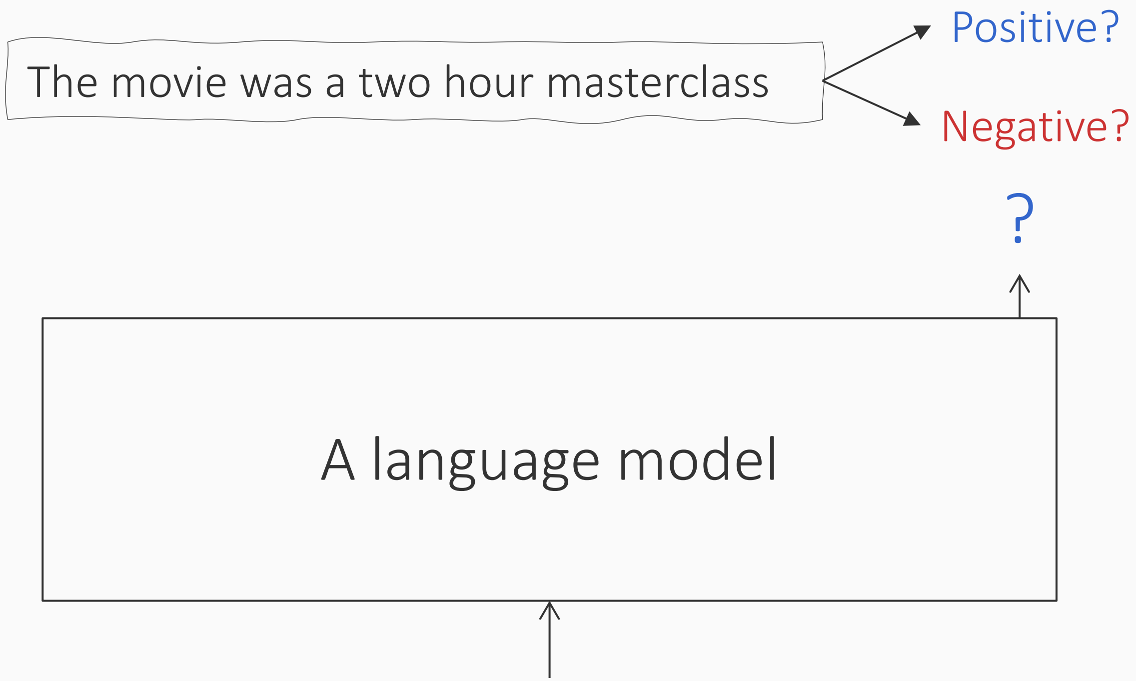
This part was not part of the original input. We add it to make the language model behave in ways that we would like.

A concrete example: Sentiment classification



The language model will generate a distribution over the next word

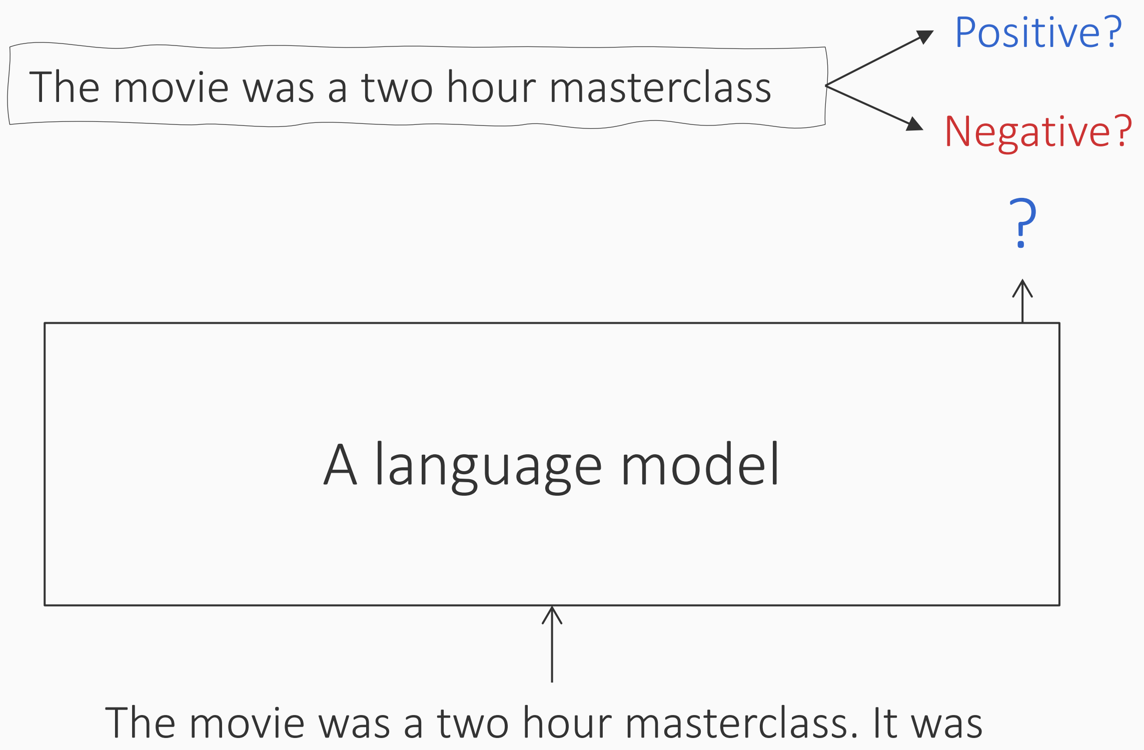
A concrete example: Sentiment classification



The movie was a two hour masterclass. It was

$$p_1 = P(great \mid \text{The movie was a two hour masterclass. It was})$$

A concrete example: Sentiment classification



$$p_1 = P(great \mid \text{The movie was a two hour masterclass. It was})$$
$$p_2 = P(terrible \mid \text{The movie was a two hour masterclass. It was})$$

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

?

A language model

The movie was a two hour masterclass. It was

Suppose we have these two examples

An effortlessly accomplished and richly resonant work. It was great!
A mostly tired retread of several other mob tales. It was terrible!

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

?

A language model

The movie was a two hour masterclass. It was

Suppose we have these two examples

E1= An effortlessly accomplished and richly resonant work. It was great!

E2=A mostly tired retread of several other mob tales. It was terrible!

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

?

A language model

Suppose we have these two examples

E1= An effortlessly accomplished and richly resonant work. It was great!

E2=A mostly tired retread of several other mob tales. It was terrible!

E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was

$$p_1 = P(great \mid \text{E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was})$$

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

?

A language model

E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was

Suppose we have these two examples

E1= An effortlessly accomplished and richly resonant work. It was great!

E2=A mostly tired retread of several other mob tales. It was terrible!

$$p_1 = P(great \mid \text{E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was})$$

$$p_2 = P(terrible \mid \text{E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was})$$

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

?

A language model

Suppose we have these two examples

E1= An effortlessly accomplished and richly resonant work. It was great!

E2=A mostly tired retread of several other mob tales. It was terrible!

E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was

$p_1 = P(great$ | E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was)

$p_2 = P(terrible$ | E1 [SEP] E2 [SEP] The movie was a two hour masterclass. It was)

If $p_1 > p_2$ then label = **Positive** otherwise label = **Negative**

In the few-shot setting: Sentiment classification

The movie was a two hour masterclass

Positive?

Negative?

rk. It was great!

was terrible!

Why might this idea have any hope of working?

E1 [SEP] E2 [SE

$p_1 =$

$p_2 =$

)

If $p_1 > p_2$ then label = Positive otherwise label = Negative

This has been the subject of much research discussion

- Demonstrations do not teach a new task; instead, it is about locating an already-learned task during pretraining (Reynolds & McDonell, 2021)
- LMs do not exactly understand the meaning of their prompt (Webson & Pavlick, 2021)
- Demonstrations are about providing a latent concept so that LM generates coherent next tokens (Xie et al. 2022)
- In-context learning performance is highly correlated with term frequencies during pretraining (Razeghi et al. 2022)
- LMs do not need input-label mapping in demonstrations, instead, it uses the specification of the input & label distribution separately (Min et al. 2022)
- Data properties lead to the emergence of few-shot learning (burstiness, long-tailedness, many-to-one or one-to-many mappings, a Zipfian distribution) (Chan et al. 2022)

Prompting language models: Some terminology

An effortlessly accomplished and richly resonant work. It was great!
A mostly tired retread of several other mob tales. It was terrible!
A movie was a two hour masterclass. It was _______!

Prompting language models: Some terminology

Prompt: A conditioning text coming before the test input

Demonstrations: A special instance of prompt which is a concatenation of the k-shot training data (in in-context learning, prompt==demonstrations)

An effortlessly accomplished and richly resonant work. It was great!
A mostly tired retread of several other mob tales. It was terrible!
A movie was a two hour masterclass. It was ________!

Prompting language models: Some terminology

Prompt: A conditioning text coming before the test input

Demonstrations: A special instance of prompt which is a concatenation of the k-shot training data (in in-context learning, prompt==demonstrations)

Pattern: A function that maps an input to the text (a.k.a. template)

An effortlessly accomplished and richly resonant work. It was great!
A mostly tired retread of several other mob tales. It was terrible!
A movie was a two hour masterclass. It was ________!

Prompting language models: Some terminology

Prompt: A conditioning text coming before the test input

Demonstrations: A special instance of prompt which is a concatenation of the k-shot training data (in in-context learning, prompt==demonstrations)

Pattern: A function that maps an input to the text (a.k.a. template)

Verbalizer: A function that maps a label to the text (a.k.a. label words)

An effortlessly accomplished and richly resonant work. It was great!
A mostly tired retread of several other mob tales. It was terrible!
A movie was a two hour masterclass. It was ________!

Examples of patterns and verbalizers

An effortlessly accomplished and richly resonant work.	It was great!
A mostly tired retread of several other mob tales.	It was terrible!
A three-hour cinema master class.	It was great!

Pattern: f(<x>) = <x>
Verbalizer: v("positive") = "It was great!", f("negative") = "It was terrible!"

Examples of patterns and verbalizers

An effortlessly accomplished and richly resonant work.	It was great!
A mostly tired retread of several other mob tales.	It was terrible!
A three-hour cinema master class.	It was great!

Pattern: f(<x>) = <x>
Verbalizer: v("positive") = "It was great!", f("negative") = "It was terrible!"

Review: An effortlessly accomplished and richly resonant work.	Sentiment: positive
Review: A mostly tired retread of several other mob tales.	Sentiment: negative
Review: A three-hour cinema master class.	Sentiment: positive

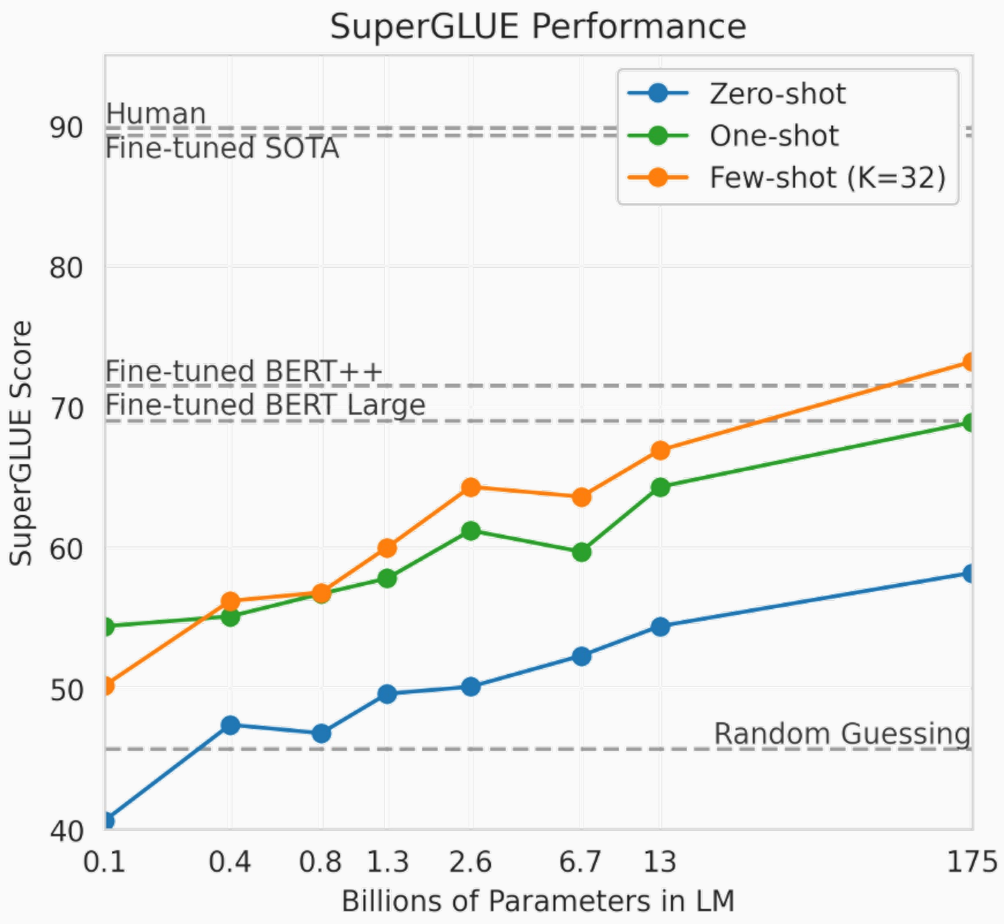
Pattern: f(<x>) = "Review: <x>"
Verbalizer: v(<x>) = "Sentiment: <x>"

Practical notes

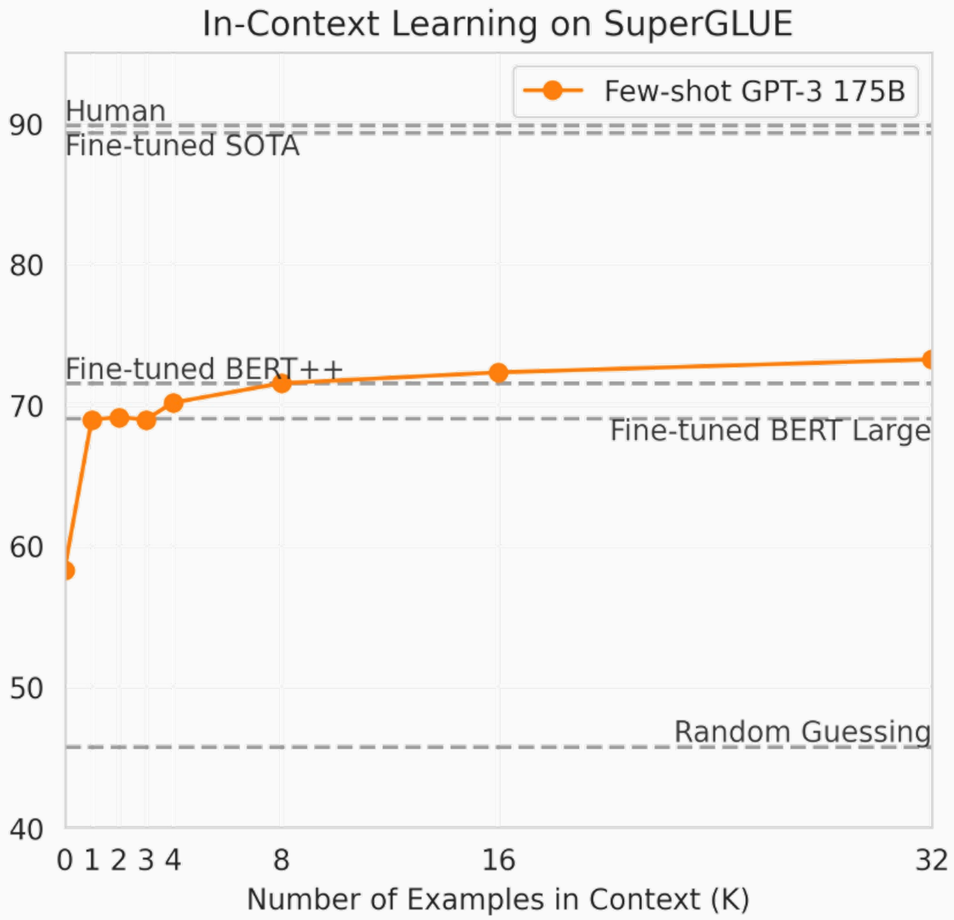
- There are many different possible patterns/verbalizers even for the same task.
- In practice, it is better to use patterns/verbalizers that makes the sequence closer to language modeling, i.e. closer to the text that the model might have seen during pretraining.
- It turns out there is huge variance in performance based on the choice of patterns/verbalizers (more in the next slide).
- You should not choose patterns/verbalizers based on the test data

This lecture

- Zero- and few-shot prediction
- Prompting language models a.k.a. in-context learning
- Does prompting work?

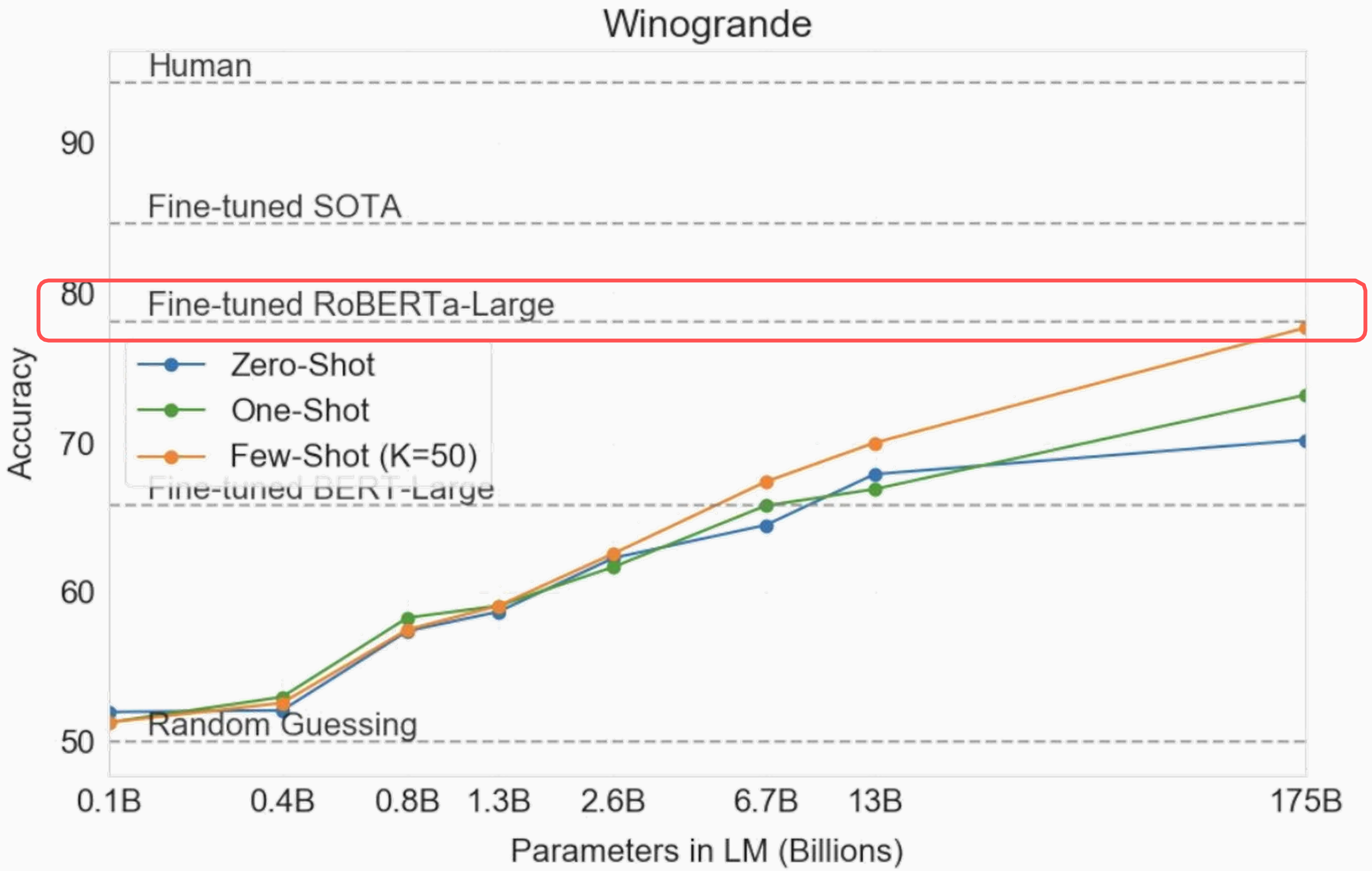


From: Brown et al. 2020. “Language Models are Few-Shot Learners”



From: Brown et al. 2020. “Language Models are Few-Shot Learners”

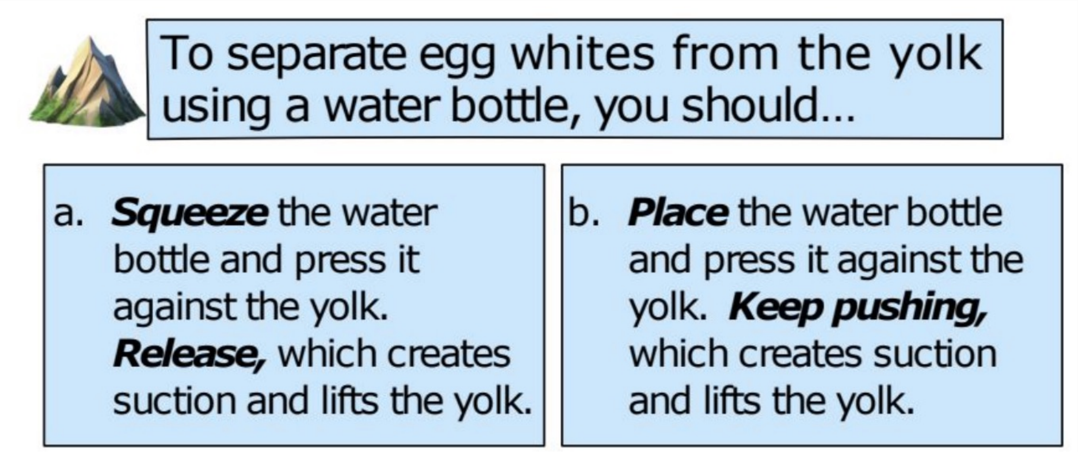
In-context learning results

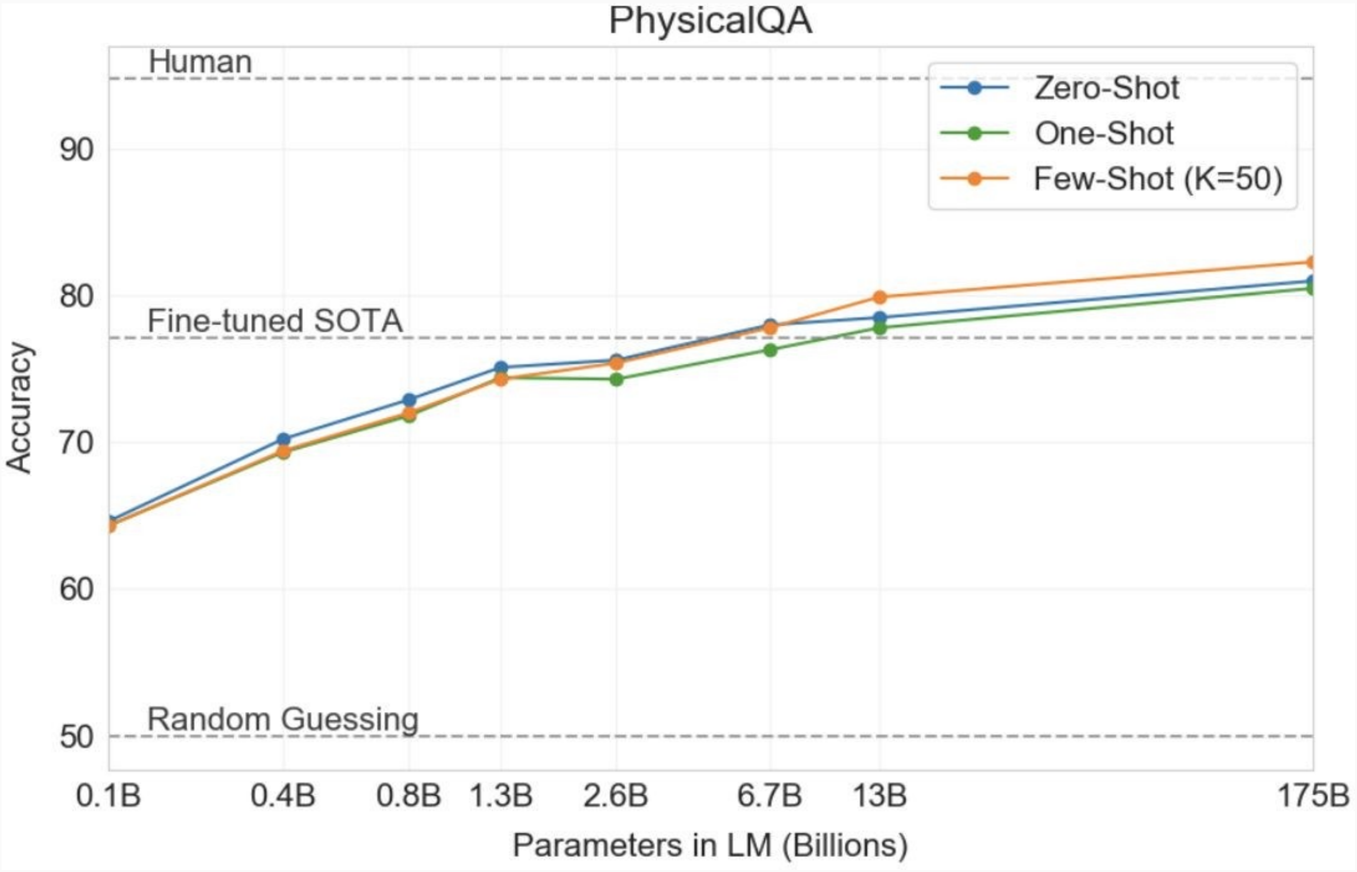


✓	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *less* time to get ready for school.	**Robert** / Samuel
	Robert woke up at 9:00am while Samuel woke up at 6:00am, so **he** had *more* time to get ready for school.	Robert / **Samuel**

From: Brown et al. 2020. "Language Models are Few-Shot Learners"

In-context learning results



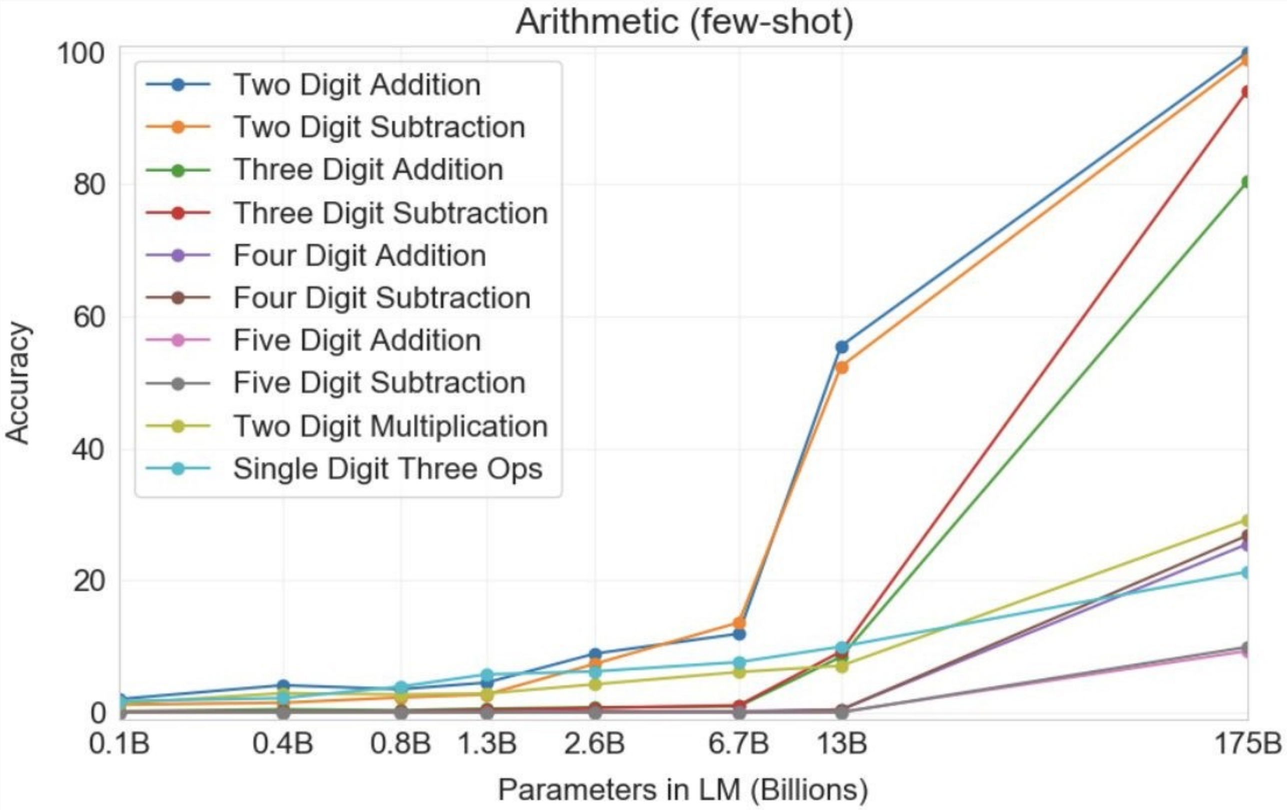


Brown et al. 2020. “Language Models are Few-Shot Learners” + Daniel Kashabi

In-context learning results

Example:

- Q: What is 48 plus 76?
- A: 124

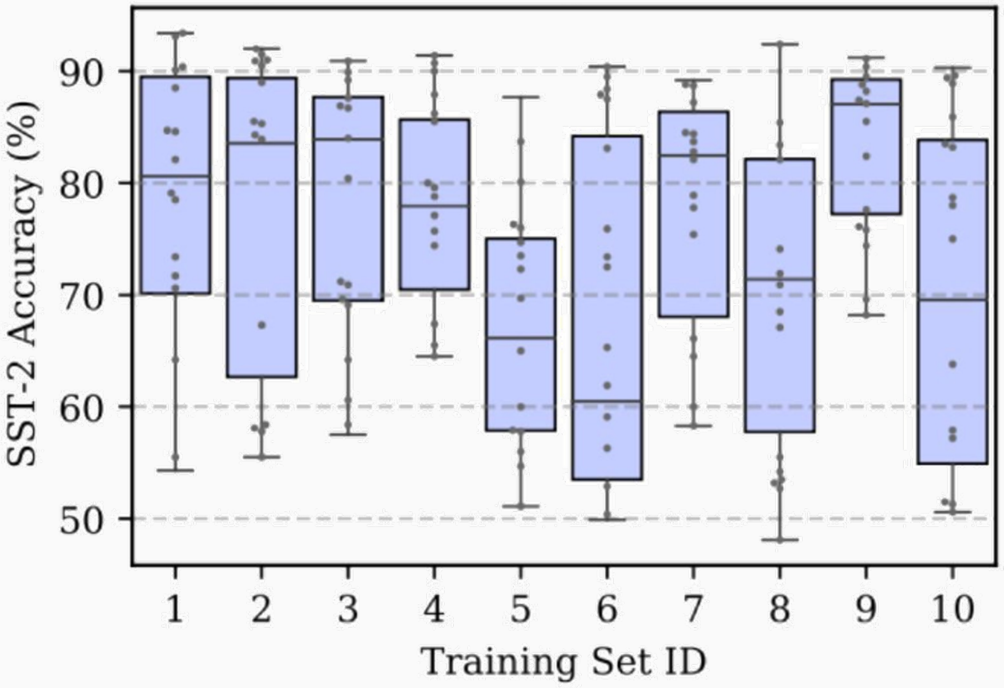


Observations:

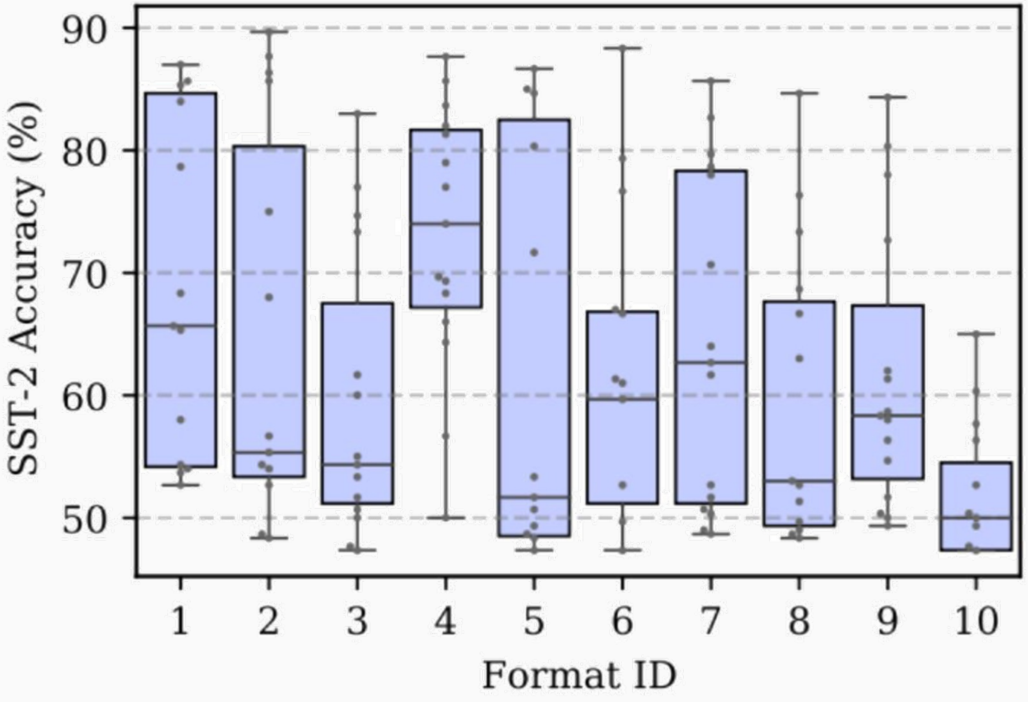
- Scale is important
- Number of digits correlate with their difficulty.
- Multiplication is harder than summation!

Variance across design choices

Across different training sets and permutations



Across different training sets and patterns/verbalizers



Zhao et al. 2021. "Calibrate Before Use: Improving Few-Shot Performance of Language Models"

The Phases of Our Understanding

"Language modeling is a useful subtask for many NLP tasks"
– pre-2018

"Language modeling is a useful supertask for many NLP tasks"
– post-2018

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pengfei Liu
Carnegie Mellon University
pliu3@cs.cmu.edu

Weizhe Yuan
Carnegie Mellon University
weizhey@cs.cmu.edu

Jinlan Fu
National University of Singapore
jinlanjonna@gmail.com

Zhengbao Jiang
Carnegie Mellon University
zhengbaj@cs.cmu.edu

Hiroaki Hayashi
Carnegie Mellon University
hiroakih@cs.cmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

In-Context (Few Shot) Prompting

- Popularized by GPT-3 (but predates that model)
- Perform a task based on a few examples provided in the inference time.
- The model identifies patterns in examples and replicates it