Foundations of Natural Language Processing

Lecture 20: Scaling, Prompting and Few-Shot Learning

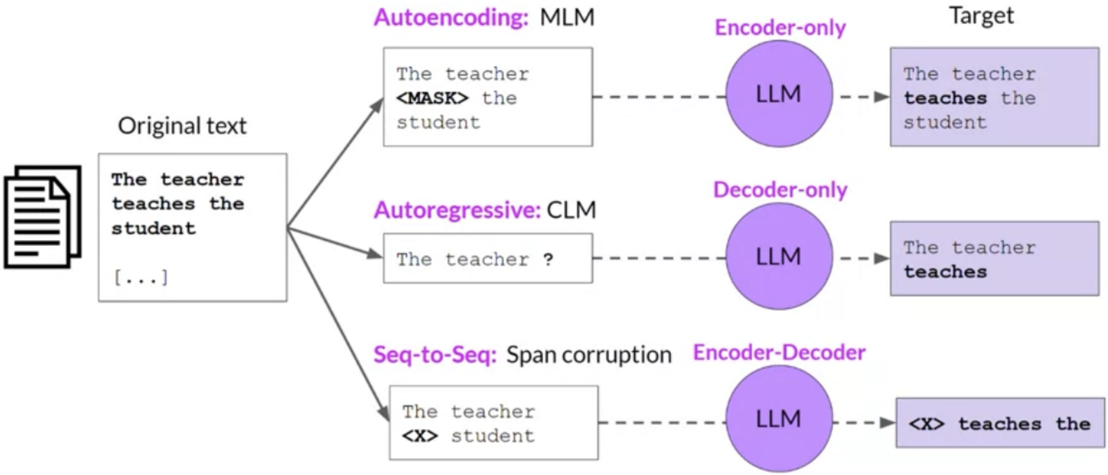
Mirella Lapata
School of Informatics
University of Edinburgh
`mlap@inf.ed.ac.uk`

Edinburgh
University of Edinburgh
Natural Language Processing
NLP

Recap

- Pretraining architectures: BERT, T5, GPT
- Encoder-only, Encoder-Decoder, Decoder-only architectures
- **General trend:** as models grow *bigger* their *performance improves*.
- **General trend:** most NLP tasks can be *reformatted as generation tasks*.
- **Today:** GPT-3, scaling, and prompting.

Model architectures and pretraining objectives

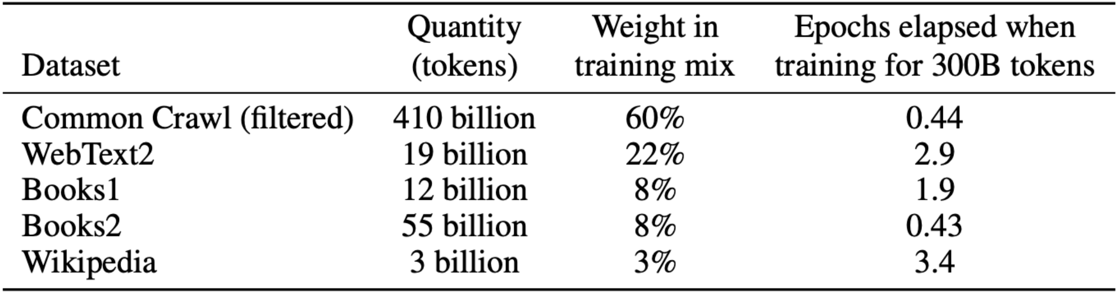


MLM: Masked Language Model, CLM: Causal Language Model

GPT3: A Very Large Language Model

- More layers and parameters, bigger dataset, longer training
- Larger embedding/hidden dimension, larger context window
- **175B parameter model**: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to **cost roughly $10M**
- GPT-4 may be **mixture of experts** combining several similar-sized models

Datasets used to train GPT3

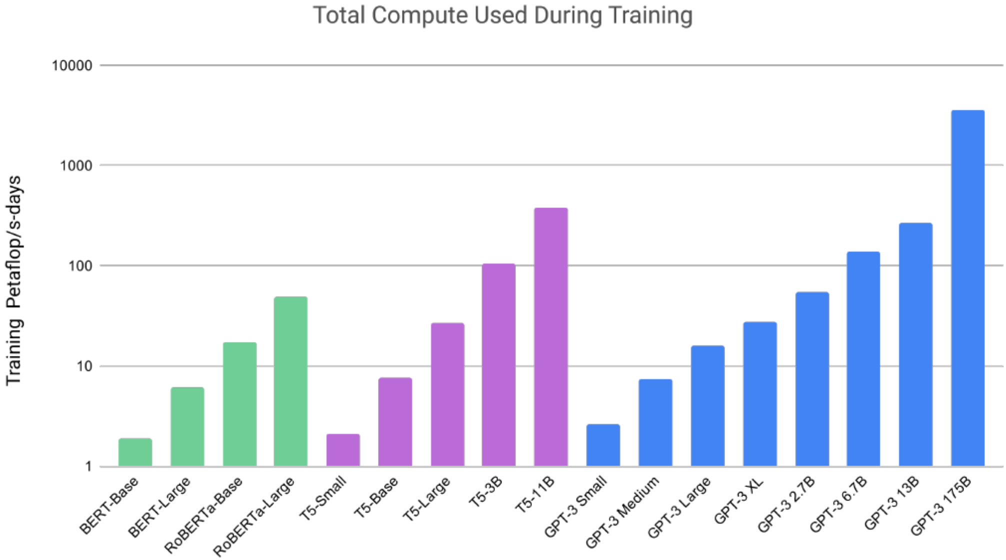


Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Weight in training mix: proportion of examples during training that are drawn from a given dataset. Note it is not a proportion of the size of the dataset.

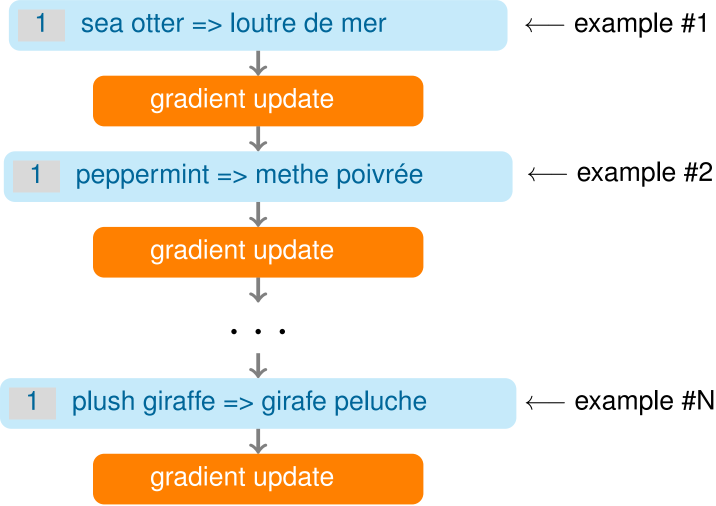
Tried to remove dev/test set of downstream tasks from training (decontamination). However a bug in the code missed out on some overlaps. Too expensive to retrain!!!!!!

Compute used to train GPT3



A petaflop/s-day (pfs-day) consists of performing 1,015 neural net operations per second for one day.

Revisiting Finetuning



1 sea otter => loutre de mer ⟵ example #1

gradient update

1 peppermint => methe poivrée ⟵ example #2

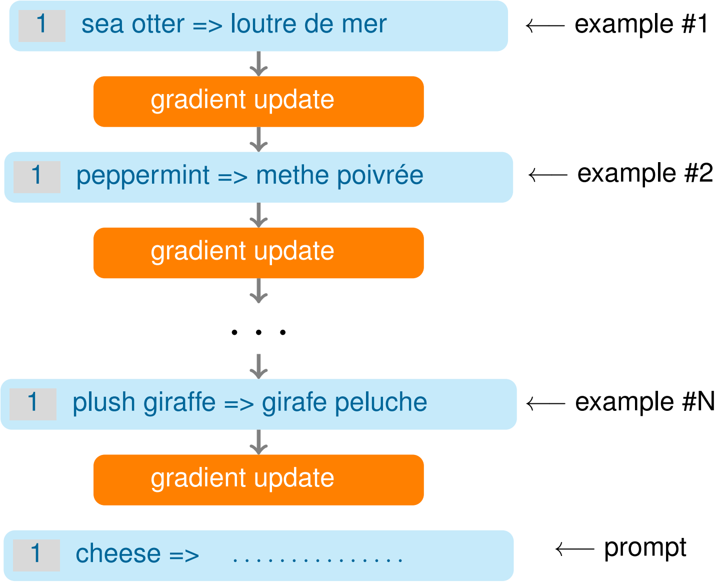
gradient update

• • •

1 plush giraffe => girafe peluche ⟵ example #N

gradient update

Revisiting Finetuning



1 sea otter => loutre de mer ⟵ example #1

gradient update

1 peppermint => methe poivrée ⟵ example #2

gradient update

. . .

1 plush giraffe => girafe peluche ⟵ example #N

gradient update

1 cheese => . . . . ⟵ prompt

In-context learning

- Finetuning is the “normal way” of doing learning in models like GPT-2
- Finetuning requires **computing the gradient** and applying a **parameter update** on **every example**
- This is **super-expensive** for 175B parameters

In-context learning

During unsupervised pretraining, a language model develops a set of skills and pattern recognition abilities. It then uses these abilities at inference time to solve a new task (*without any change to its weights*) by being fed a prompt with examples of that task. The model performs the task *only with forward passes* at test time.

In-context learning

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
```

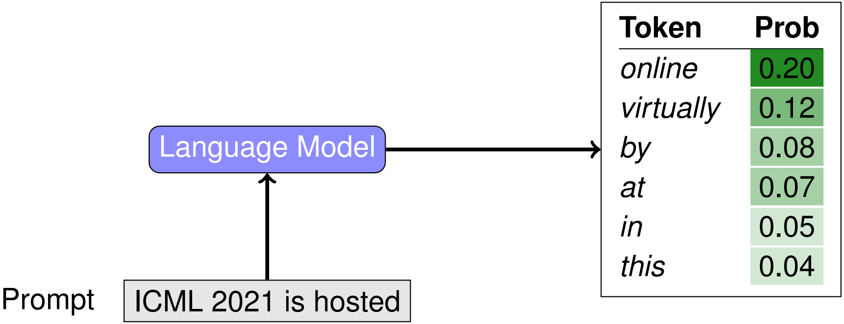
In-context learning

Few-shot

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

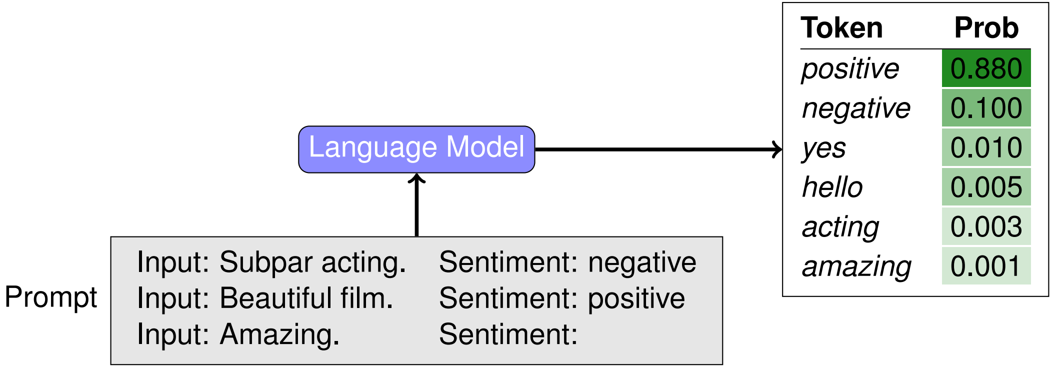
```
Translate English to French:              ⟵ task description

sea otter => loutre de mer                ⟵ example

peppermint = > menthe poivreée            ⟵ example

plush giraffe ==> peluche girafe          ⟵ example

cheese ==> . . . . . . . . .              ⟵ prompt
```

Neural Language Models



Token	**Prob**
online	0.20
virtually	0.12
by	0.08
at	0.07
in	0.05
this	0.04

Few-shot Learning with LMs: Sentiment Analysis



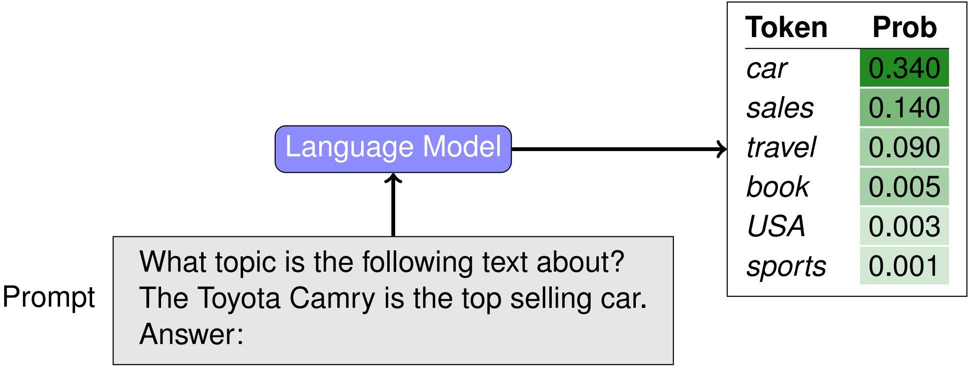
Prompt

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: Amazing. Sentiment:

Language Model

Token	Prob
positive	0.880
negative	0.100
yes	0.010
hello	0.005
acting	0.003
amazing	0.001

Few-shot Learning with LMs: Topic Classification



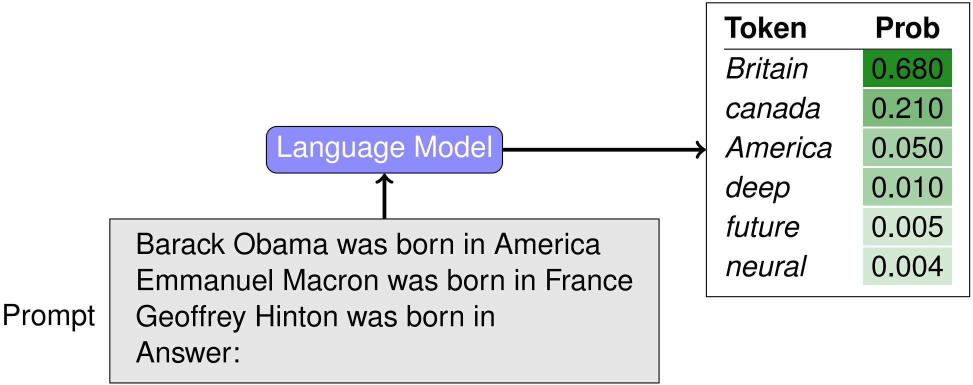
Language Model

Prompt

What topic is the following text about?
The Toyota Camry is the top selling car.
Answer:

Token	Prob
car	0.340
sales	0.140
travel	0.090
book	0.005
USA	0.003
sports	0.001

Few-shot Learning with LMs: Knowledge Base Completion



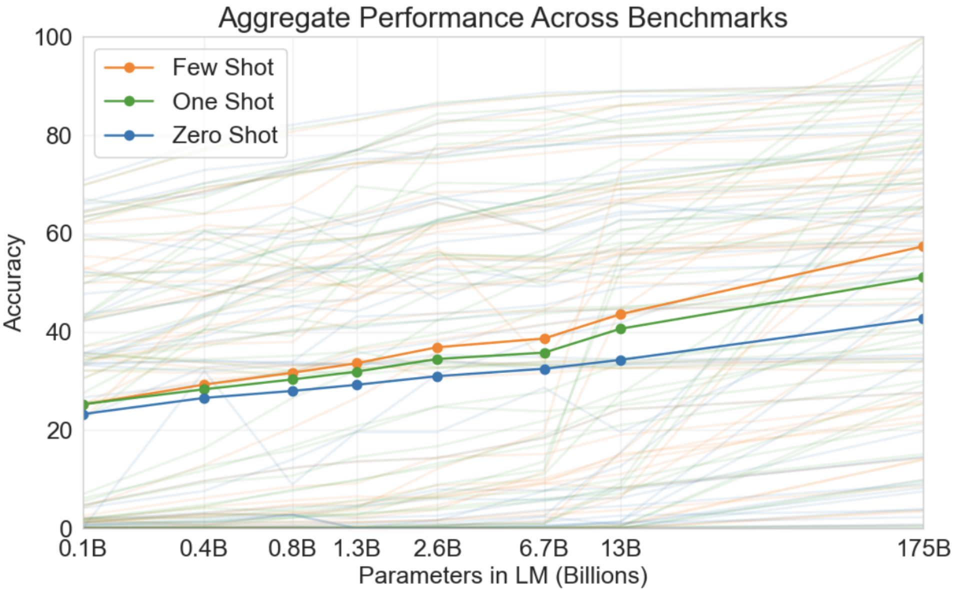
Prompt

Barack Obama was born in America
Emmanuel Macron was born in France
Geoffrey Hinton was born in
Answer:

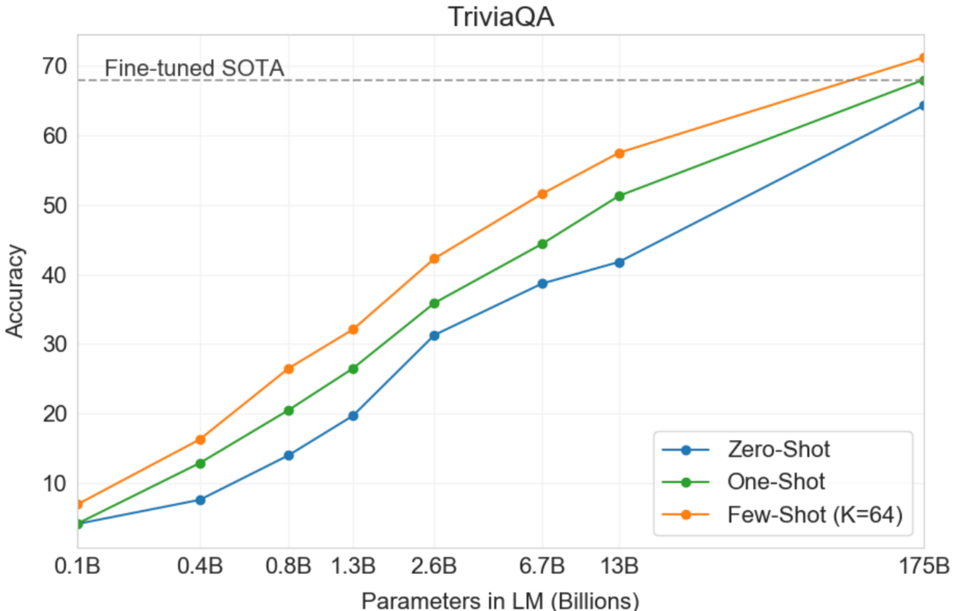
Language Model

Token	Prob
Britain	0.680
canada	0.210
America	0.050
deep	0.010
future	0.005
neural	0.004

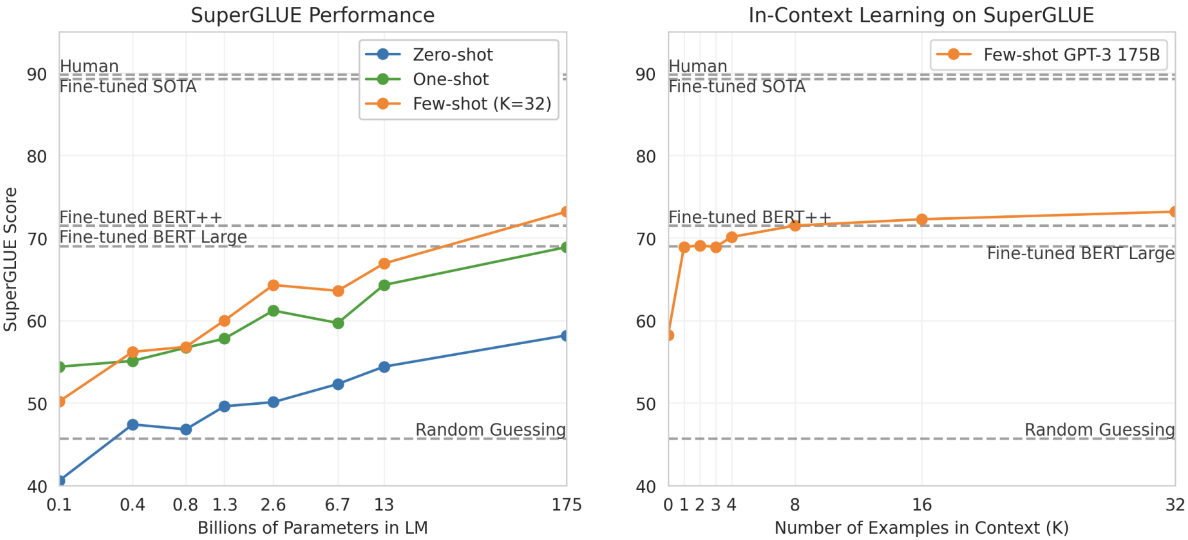
GPT-3 Aggregate Performance across benchmarks



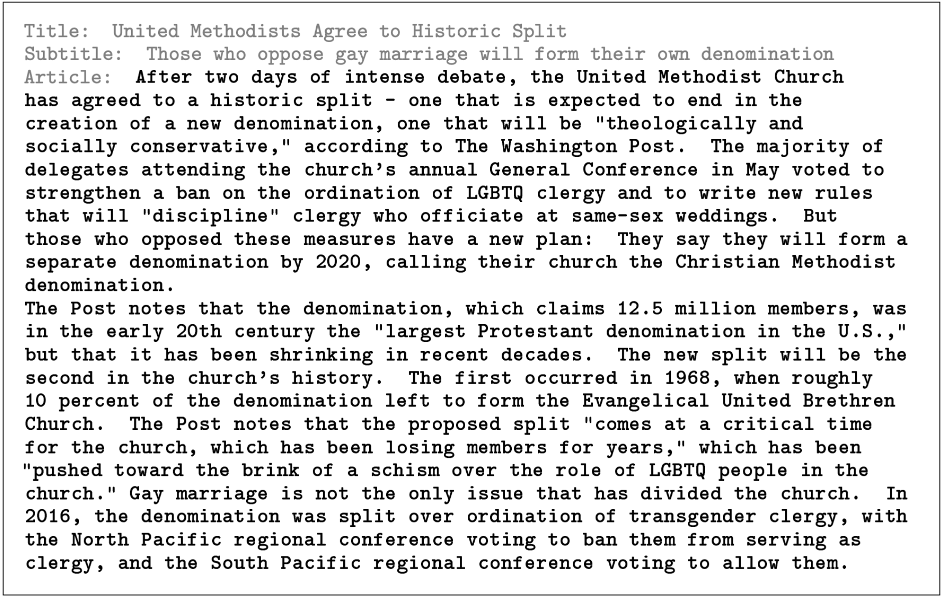
GPT-3 Trivia QA Results



GPT-3 SuperGlue Results



Article generated by GPT-3



Title: United Methodists Agree to Historic Split

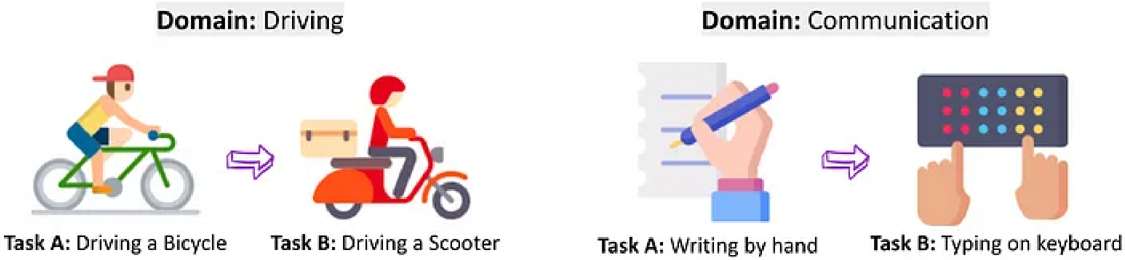
Subtitle: Those who oppose gay marriage will form their own denomination

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.

Why do we care about few-shot learning?

- **Labeling data is costly**, it often requires expertise (e.g., medical, legal, financial domains) and can be complex (e.g., parsing).
- **Finetuning is expensive** to train (time, memory).
- You want to do best with what you have, also there can be emergent, **time-sensitive scenarios** where no data is available.
- Potential **test for intelligent behavior** (humans transfer skills across related tasks).


Domain: Driving
Task A: Driving a Bicycle
Task B: Driving a Scooter
Domain: Communication
Task A: Writing by hand
Task B: Typing on keyboard

Components of the prompt: Prompt Format

The **prompt format** is a template which consists of placeholders for the training and test example(s) and possibly a natural language description of the task.

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

Input: Amazing. Sentiment:

Components of the prompt: Prompt Format

The **prompt format** is a template which consists of placeholders for the training and test example(s) and possibly a natural language description of the task.

```
Input: Subpar acting.    Sentiment: negative

Input: Beautiful film.   Sentiment: positive

Input: Amazing.          Sentiment:
```

```
Sentence: Subpar acting.    Label: negative

Sentence: Beautiful film.   Label: positive

Sentence: Amazing.          Label:
```

Components of the prompt: Prompt Format

Sentence: Subpar acting. Label: bad

Sentence: Beautiful film. Label: good

Sentence: Amazing. Label:

Components of the prompt: Prompt Format

Sentence: Subpar acting. Label: bad

Sentence: Beautiful film. Label: good

Sentence: Amazing. Label:

Q: What’s the sentiment of “ Subpar acting ”?
A: negative

Q: What’s the sentiment of” Beautiful film ”?
A: positive

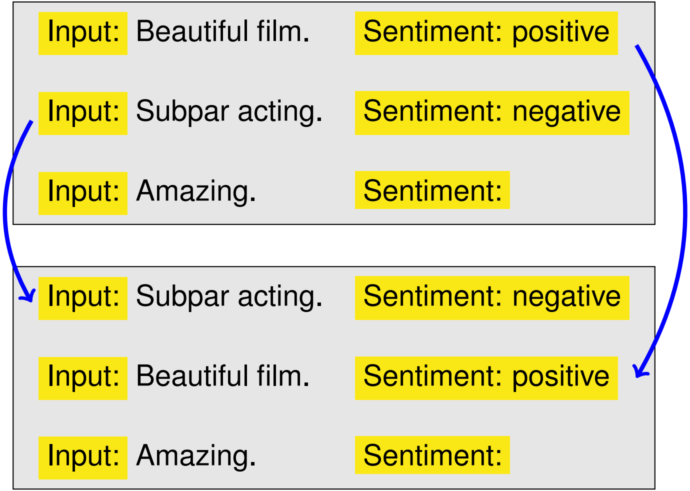
What’s the sentiment of “ Amazing ”?
A:

Components of the prompt: Example Selection

In the few-shot case, the prompt contains training examples to teach the LM how to solve the task at hand. So how do we select these examples? Does it matter?

```
Input: Good film.        Sentiment: positive

Input: Don’t watch.      Sentiment: negative

Input: Amazing.          Sentiment:
```

Components of the prompt: Example Permutation



Input: Beautiful film. Sentiment: positive

Input: Subpar acting. Sentiment: negative

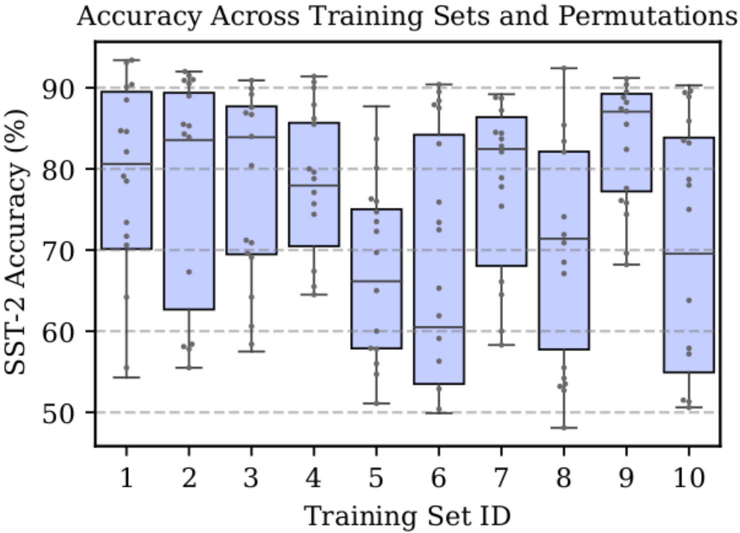
Input: Amazing. Sentiment:

Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

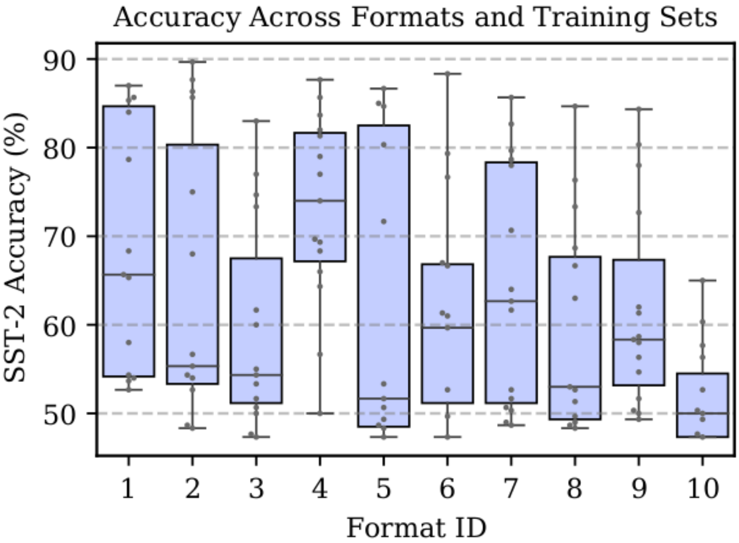
Input: Amazing. Sentiment:

Accuracy is highly sensitive to prompt design



- Prompt format is fixed
- 10 sets, each with 4 examples (4-shot)
- Plot accuracy for all permutations per set
- The box represents the interquartile range (IQR) (middle 50% of data).
- The line inside the box is the median
- Whiskers show the range of most of the data

Accuracy is highly sensitive to prompt design



- Change prompt format
- Keep examples and permutations fixed
- High variance!

Summary so far

- There are many possible ways to *encode in-context examples* for a given task:
  1. (1) Many ways to describe the task and format examples.
  2. (2) Many examples/demonstrations to choose from.
  3. (3) Many ways to order the examples in the prompt.
- There is a *huge variance* in performance depending on the encoding.
- Generally, it is better to use encoding that *makes the sequence closer to language* modeling — closer to what is observed during pretraining.

Where we stand now

- Bigger is better: performance improved across all tasks with scaling
- More shots are usually better: few-shot $>$ one-shot $>$ zero-shot
- No limit in sight of performance being bottlenecked by model size.
- **Next time:** scaling laws for LLMs and instruction tuning.