

Intro to GPT & X-shot learning

GPT & Benchmarks

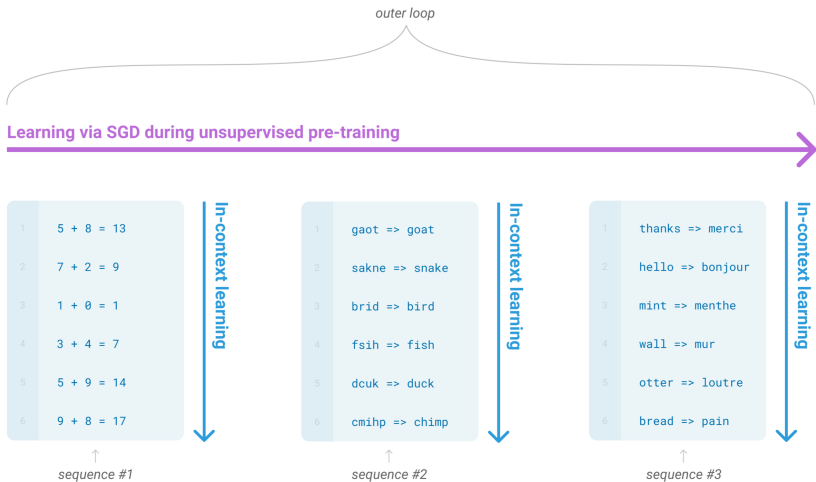
Learning goals

- Recap GPT and the ideas behind standard language modelling
- Understand the difference between fine-tuning and X-shot learning

GPT

- Like BERT, GPT is a language model.
- But not MLM, but a conventional language model: it predicts the next word (or subword).
- Like BERT, GPT is trained on a huge corpus, actually an even huger corpus.
- Like BERT, GPT is a transformer architecture.
- Difference 1: GPT is a **single model** that aims to solve **all tasks**.
 - It can also switch back and forth between tasks and solve tasks within tasks, another human capability that is important in practice. **“fluidity”**
- Difference 2: GPT leverages **task descriptions**.
- Difference 3: GPT is effective at **few-shot learning**.

GPT: IN-CONTEXT LEARNING



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.



ZERO-SHOT (NO GRADIENT UPDATE)

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 (NO GRADIENT UPDATE)

One-shot

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

1	<code>Translate English to French:</code>	← <i>task description</i>
2	<code>sea otter => loutre de mer</code>	← <i>example</i>
3	<code>cheese =></code>	← <i>prompt</i>

.....

FEW-SHOT (NO GRADIENT UPDATE)

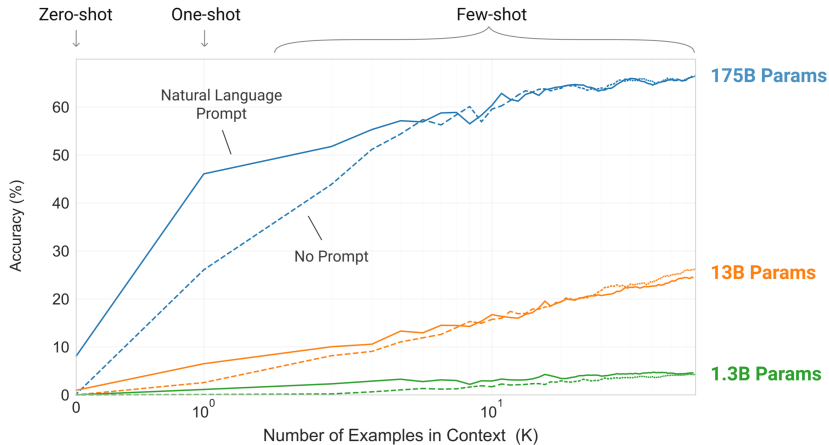
Few-shot

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

The diagram shows a prompt structure for a few-shot learning task. It consists of five lines of text, each preceded by a number in a light blue box. The first line is the task description. The next three lines are examples of the task. The fifth line is the prompt to be completed. Arrows on the right point from labels to the corresponding lines: 'task description' points to line 1, 'examples' points to lines 2, 3, and 4, and 'prompt' points to line 5.

```
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
```

GPT: EFFECTIVE IN-CONTEXT LEARNING



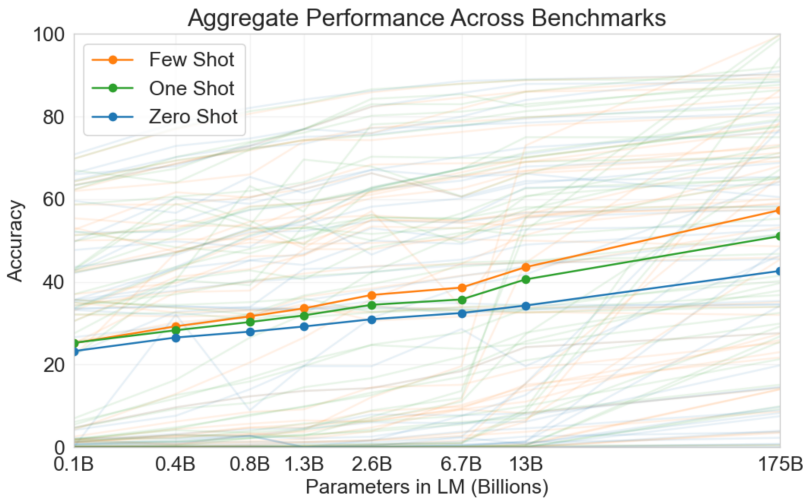
ARCHITECTURE

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

TRAINING CORPUS

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

X-SHOT COMPARISON AND EFFECT OF LARGER CORPORA



LOSS AS A FUNCTION OF COMPUTE

