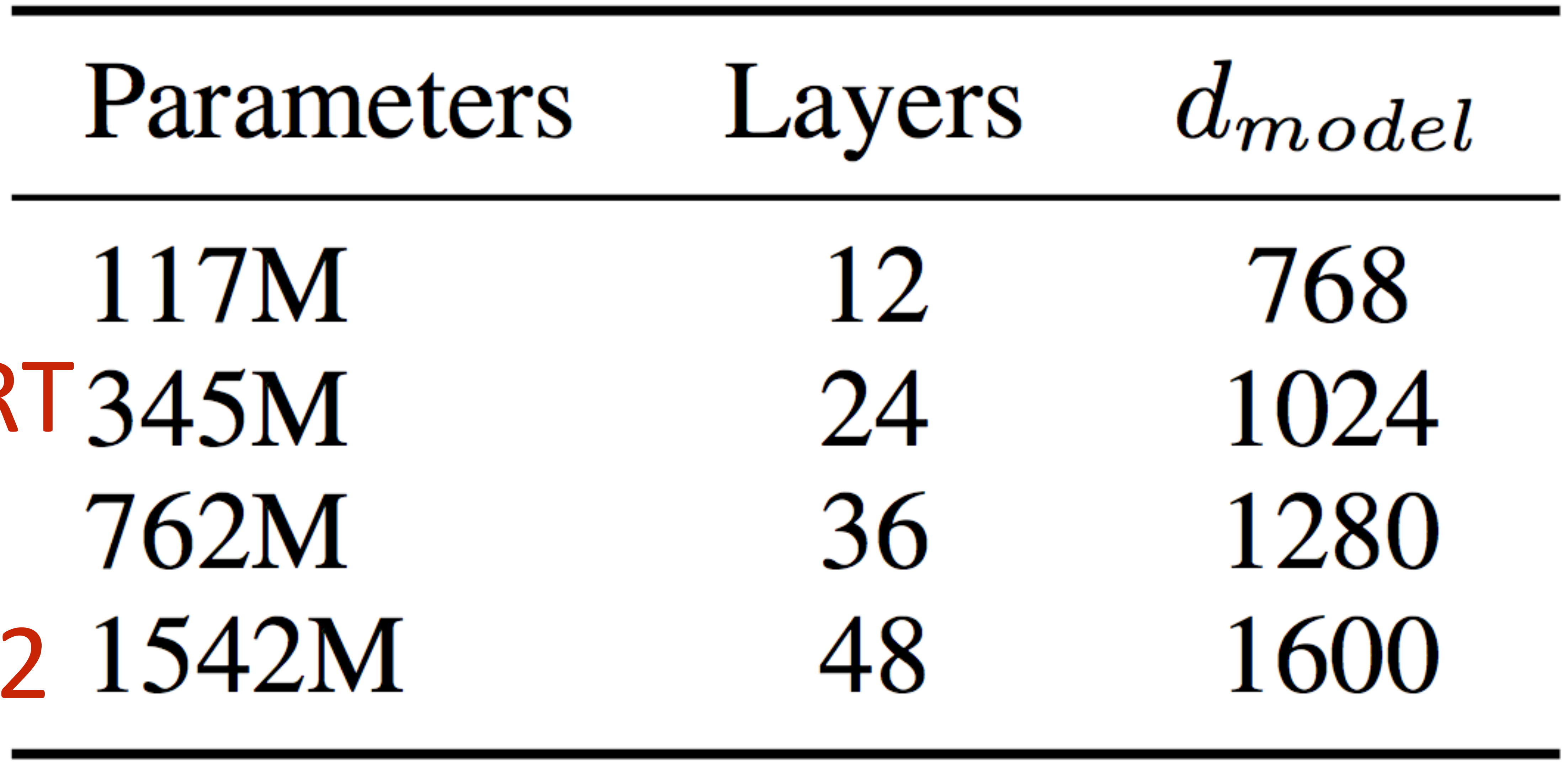
GPT

- GPT models: all very large Transformer language models, left-to-right language models, trained on raw text
- GPT1: came out before BERT, we'll skip it
- GPT2 was trained on 40GB of text:



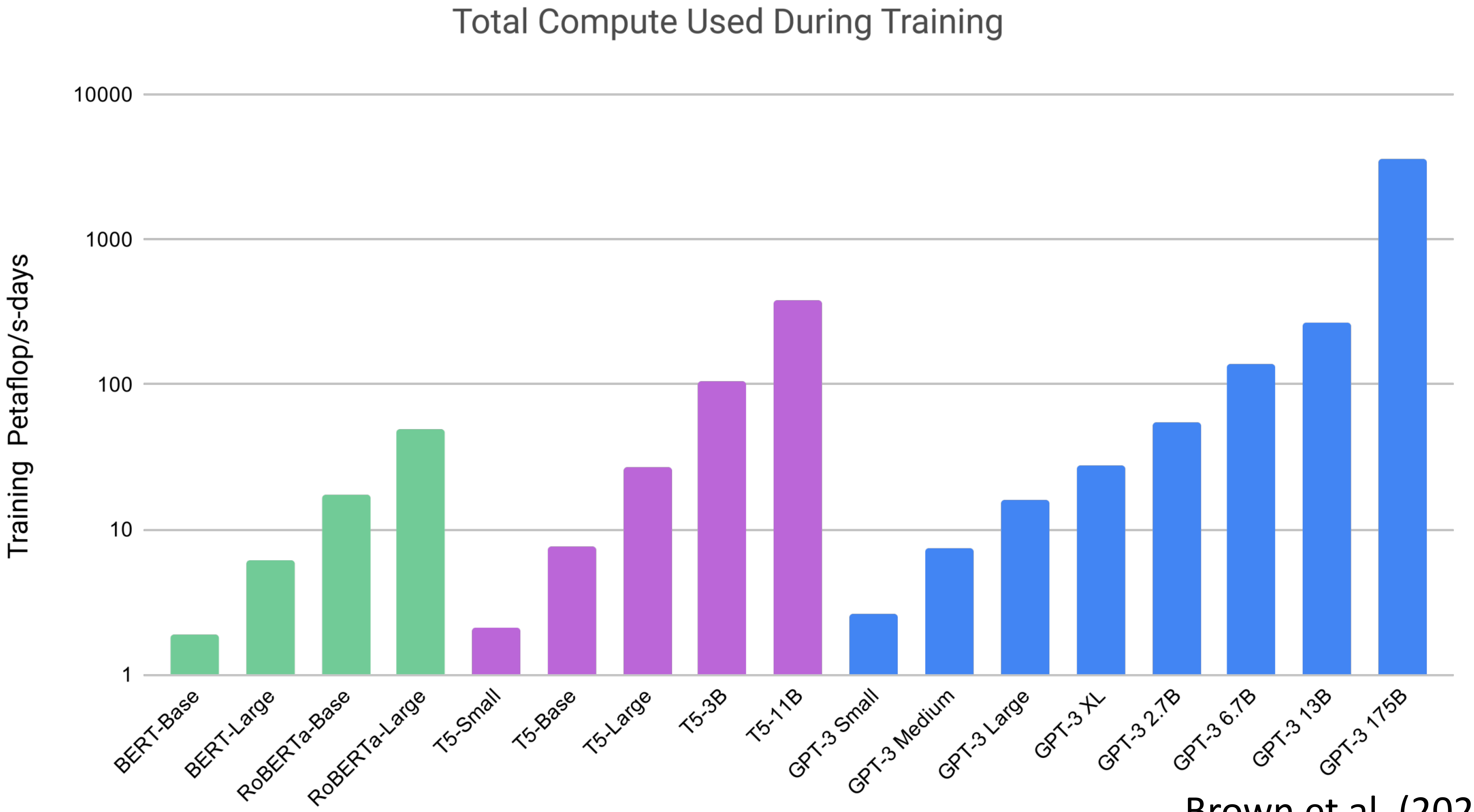
	Parameters	Layers	d_{model}
	117M	12	768
approximate size of BERT	345M	24	1024
	762M	36	1280
GPT-2	1542M	48	1600

- GPT-2 was by far the largest model trained when it came out in March 2019
- Could generate several fluent and coherent sentences back-to-back, which was not seen in smaller models or LSTMs

Radford et al. (2019)

GPT-3

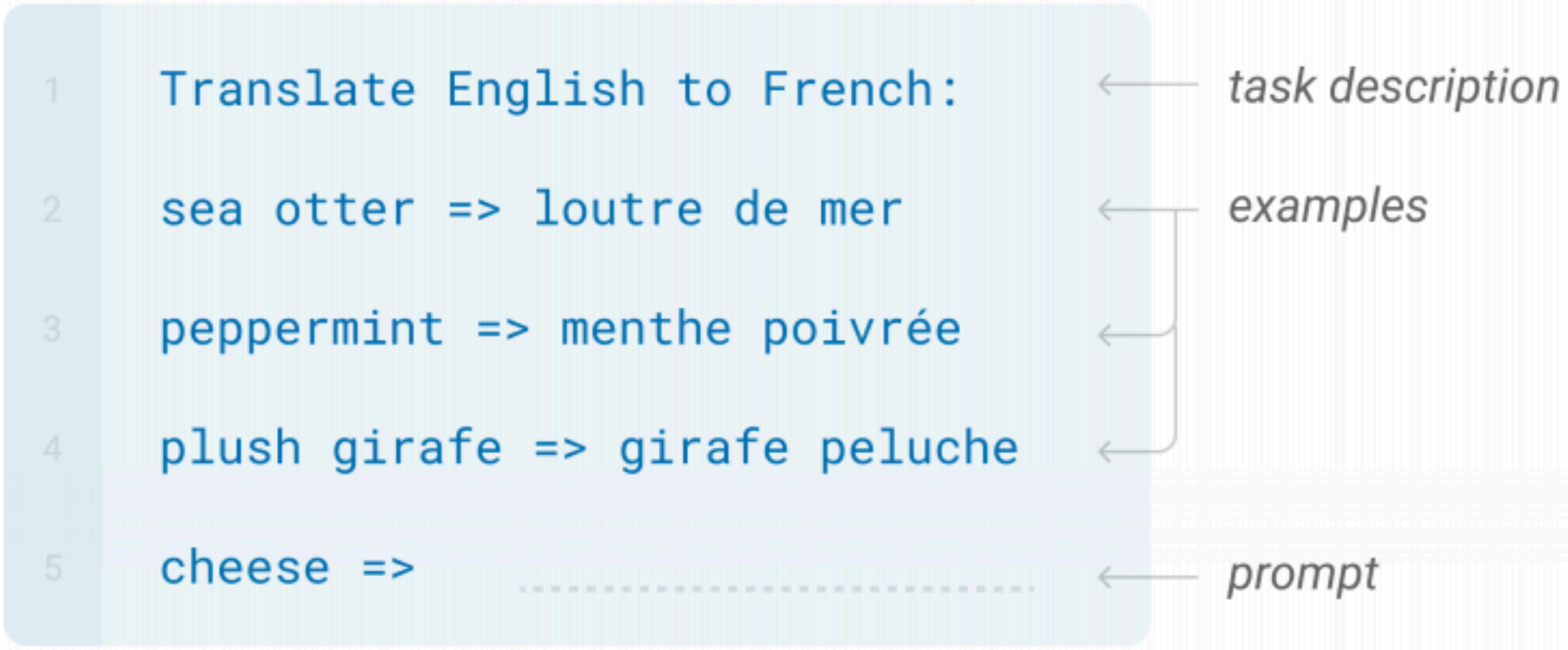
- Released in mid-2020
- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors



Brown et al. (2020)

In-context Learning

- GPT-3 proposes an alternative to fine-tuning: **in-context learning.** Just uses the off-the-shelf model, no gradient updates.
- Key concept: an LM should be able to continue an observed pattern

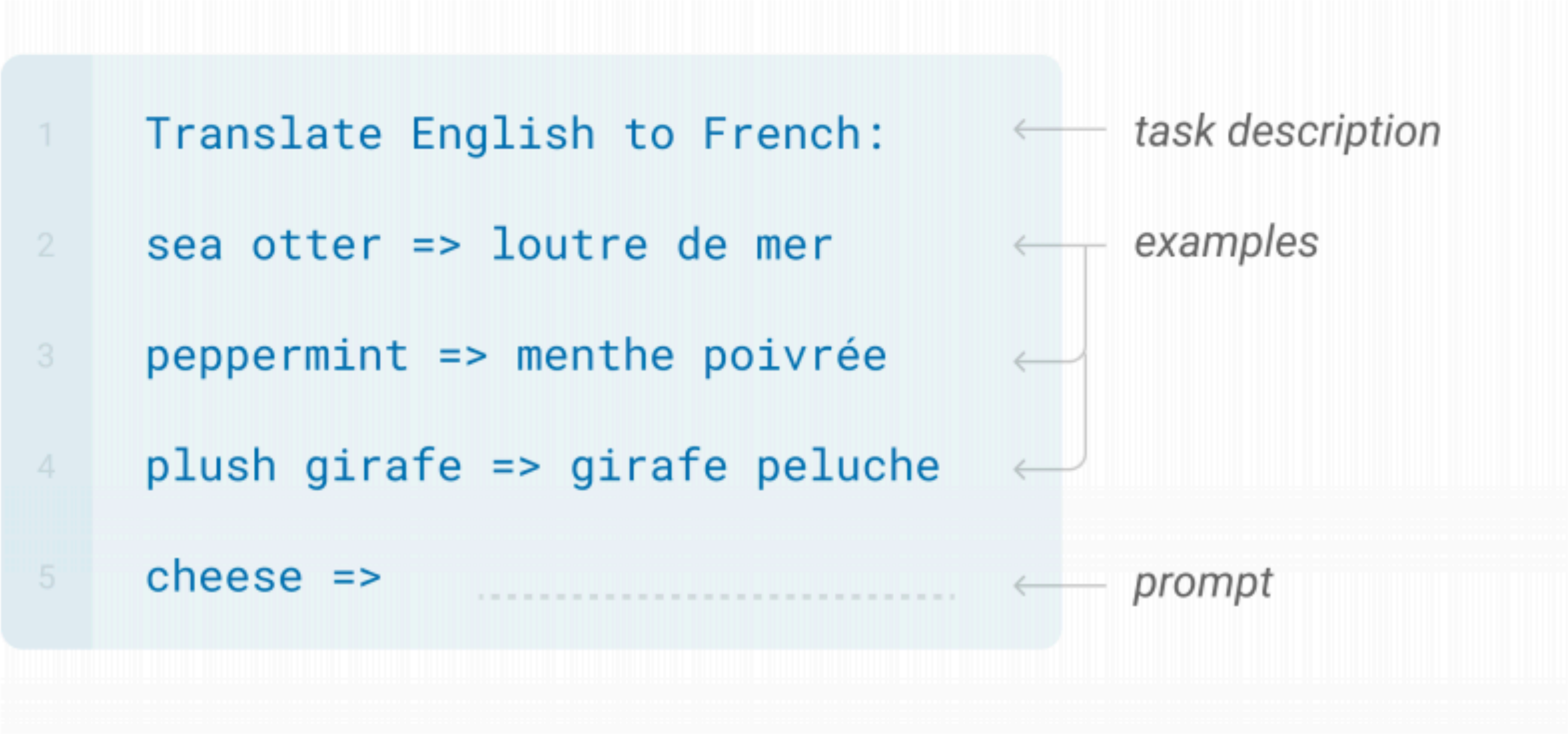


```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese => ..............................
```

- This procedure depends heavily on the examples you pick as well as the prompt (“*Translate English to French*”)

Brown et al. (2020)

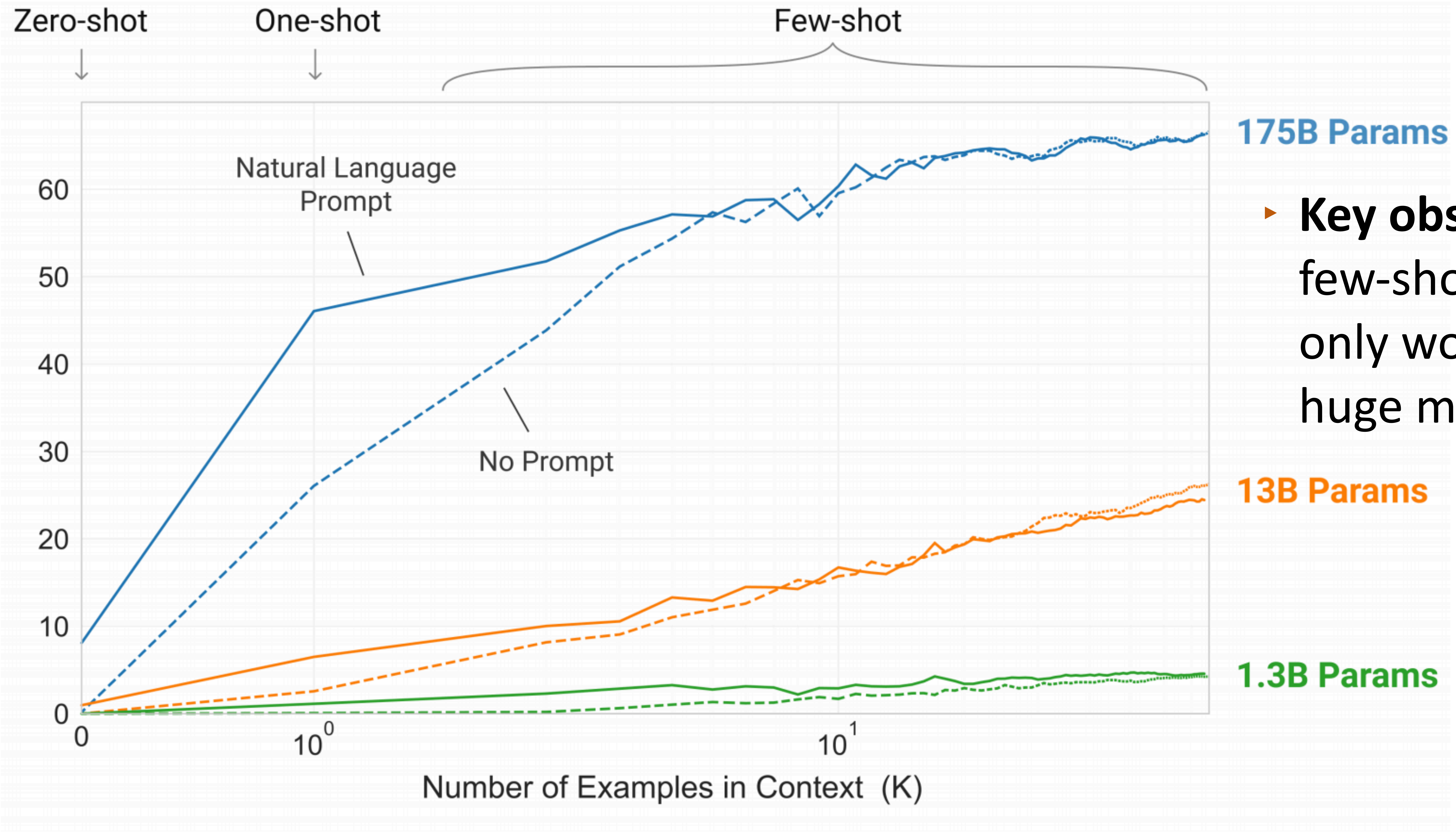
In-context Learning



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

Brown et al. (2020)

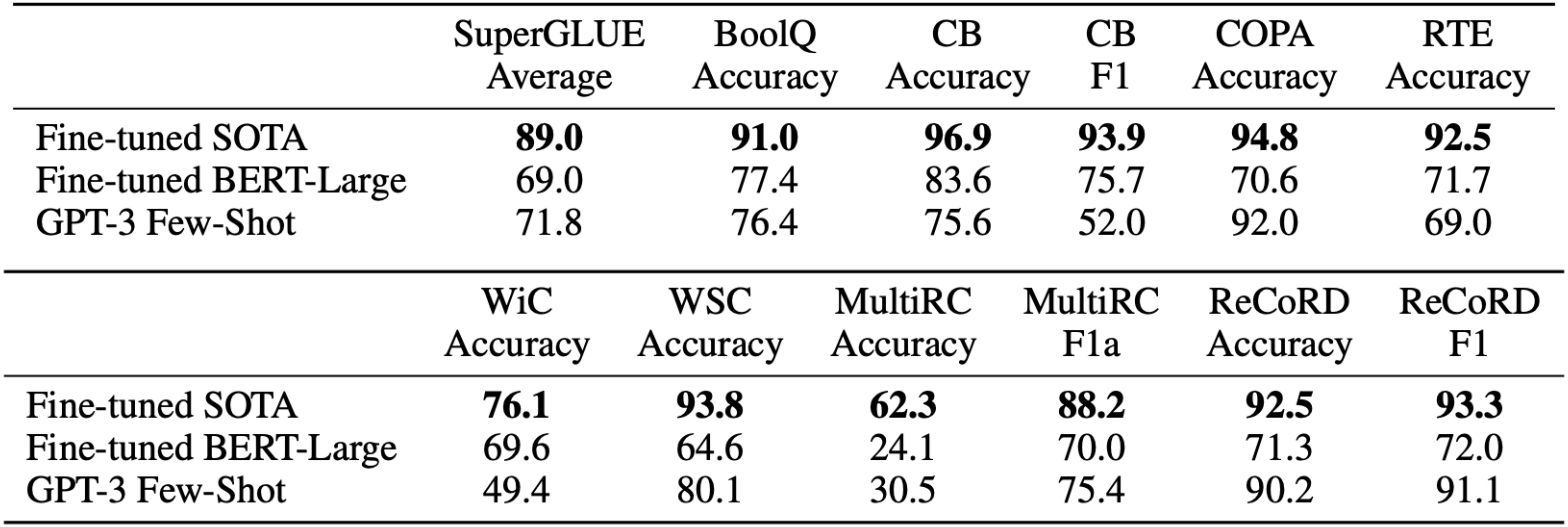
In-context Learning



- **Key observation:** few-shot learning only works with huge models!

Brown et al. (2020)

GPT-3: Results



	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

- Comparison to fine-tuned state-of-the-art models, fine-tuned BERT-Large Note that these models train on much more data, GPT-3 is "few-shot" and **only** uses in-context learning
- Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad
- Results on other datasets are equally mixed — but still strong for a few-shot model!

Other Models

- GPT-3 represents a fundamental paradigm shift in model capabilities

- Other strong large language models (LLMs) that are widely used: LLaMA, PaLM (Google), OPT, BLOOM, Pythia (all open except PaLM)

- Modern models like ChatGPT, GPT-4 use additional reinforcement learning from human feedback or instruction tuning. We will come to these later in the course; they are basically souped-up fine-tuning techniques