

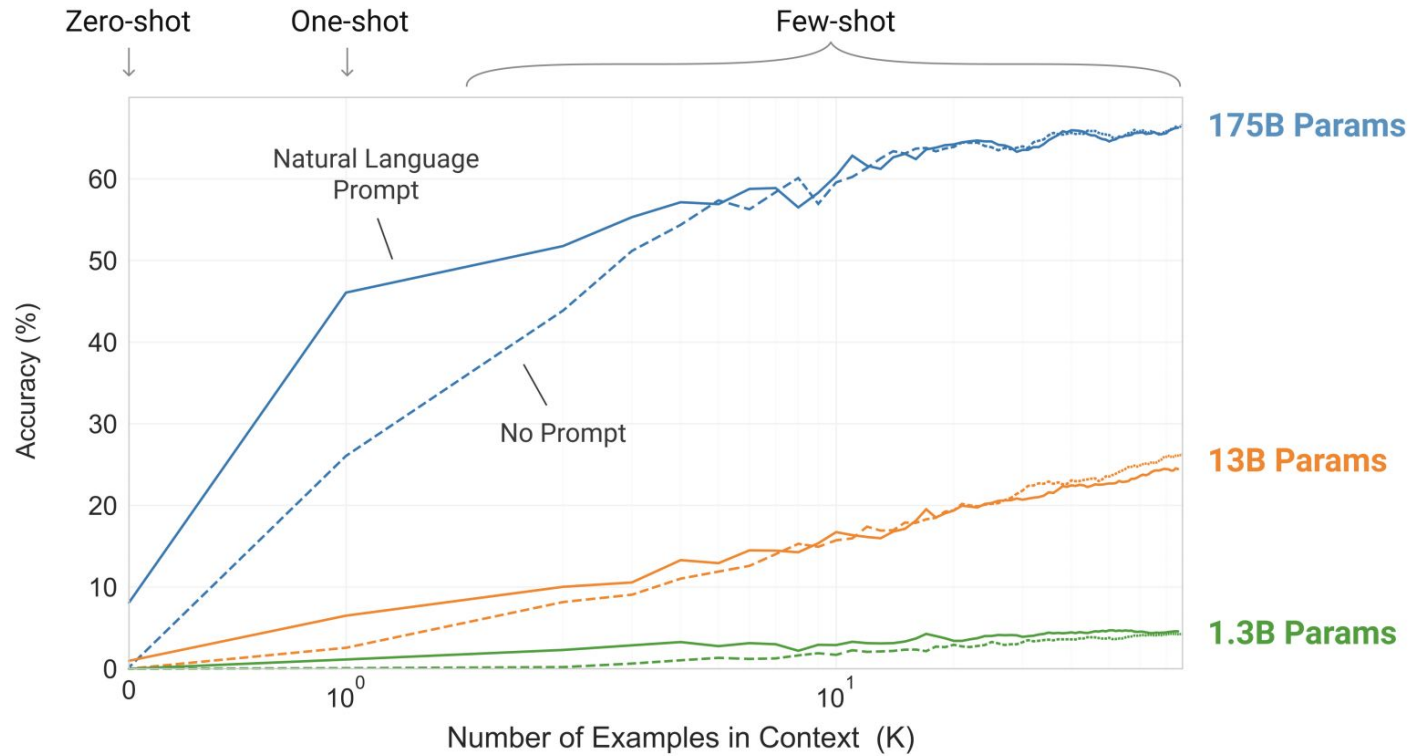


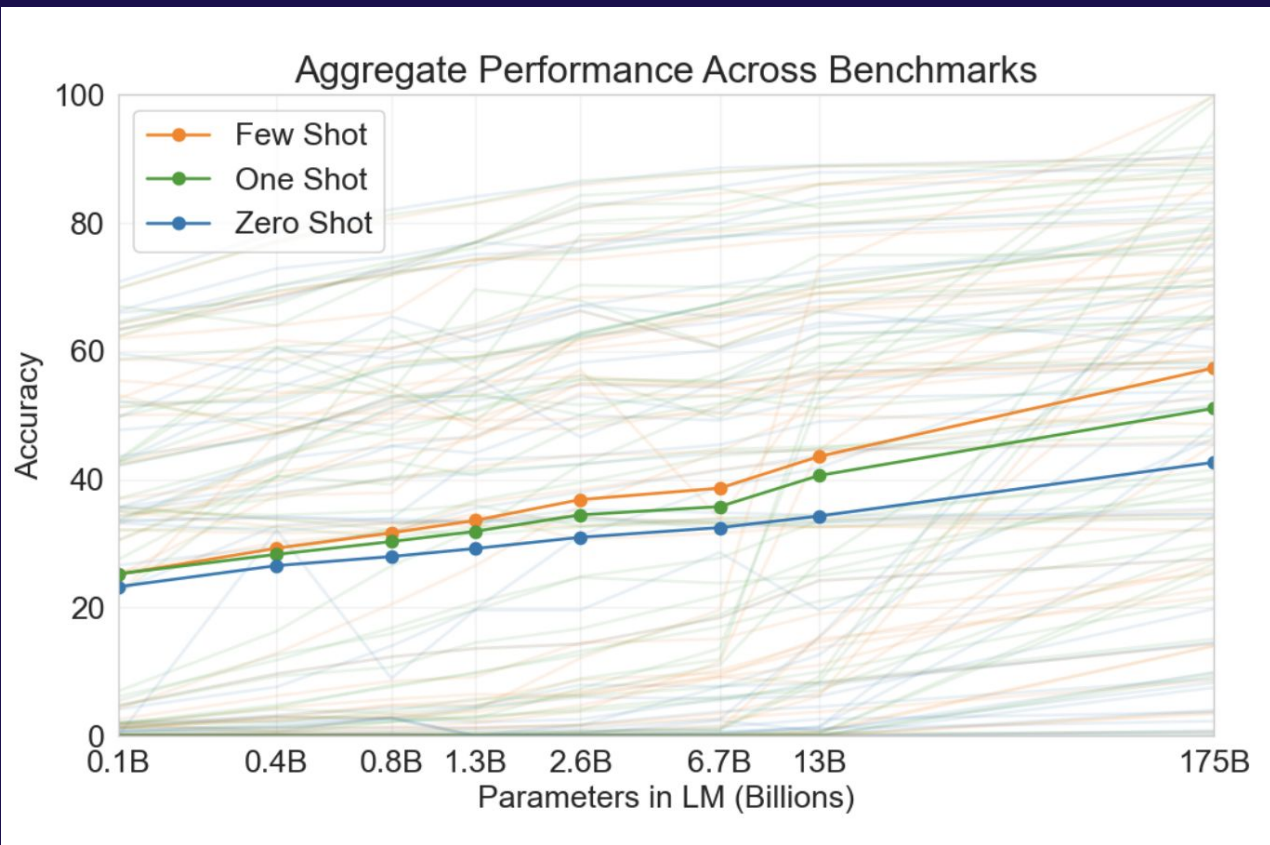
OpenAI

Language Models are Few-Shot Learners

Ben Mann

2020.07.24





Zero shot

1 Translate English to French:

← *task description*

2 cheese =>

← *prompt*

.....

Few shot

The diagram illustrates a few-shot prompt structure for a translation task. It consists of five numbered lines within a light blue rounded rectangle. Line 1 is the task description. Lines 2, 3, and 4 are examples of the translation. Line 5 is the prompt for the model to complete. 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.

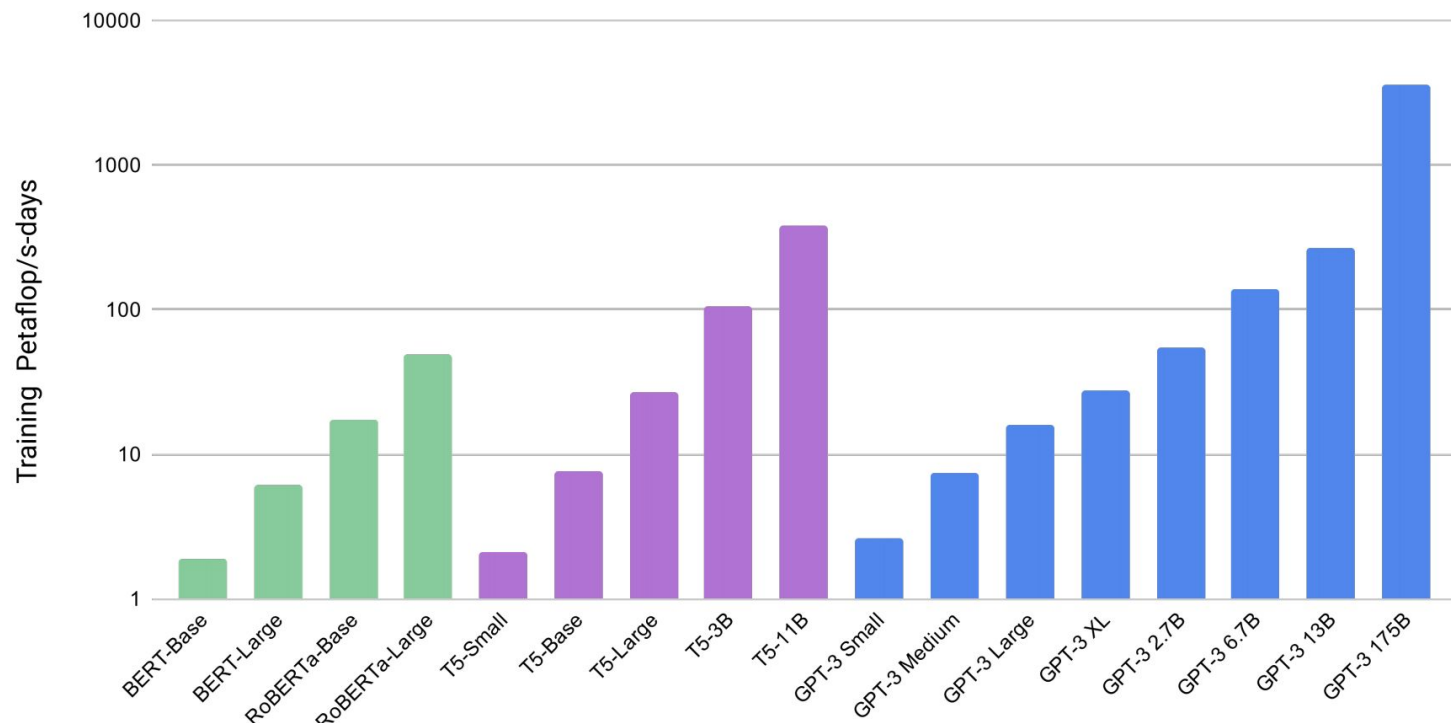
```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Finetuning



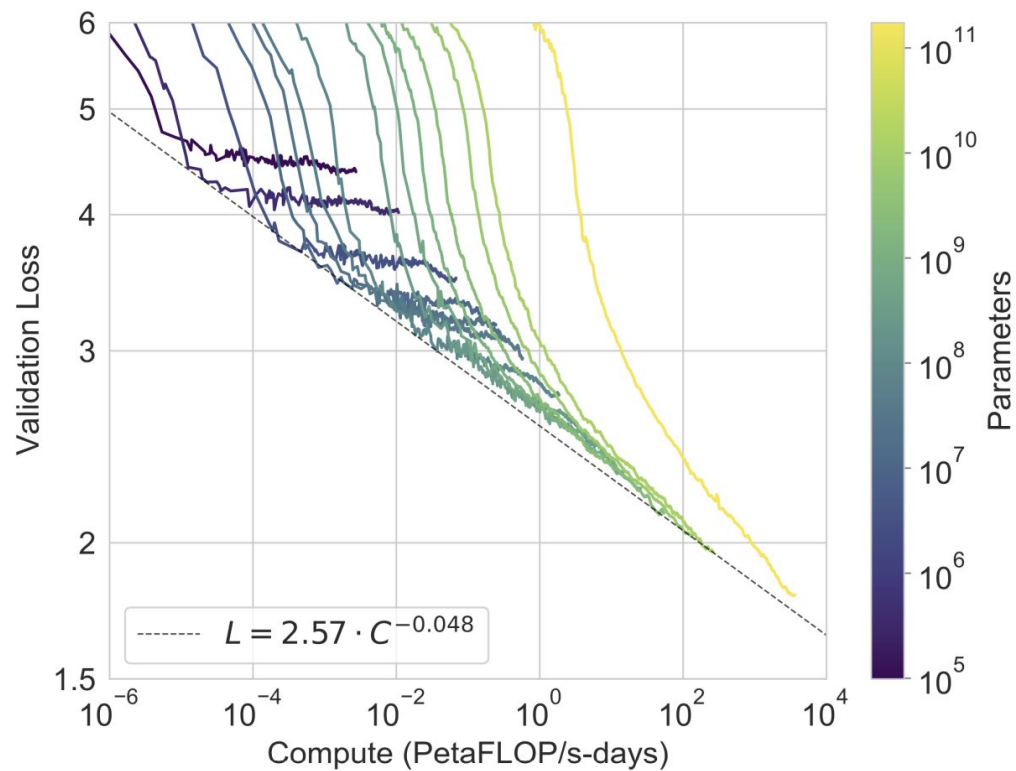
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	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}

Total Compute Used During Training

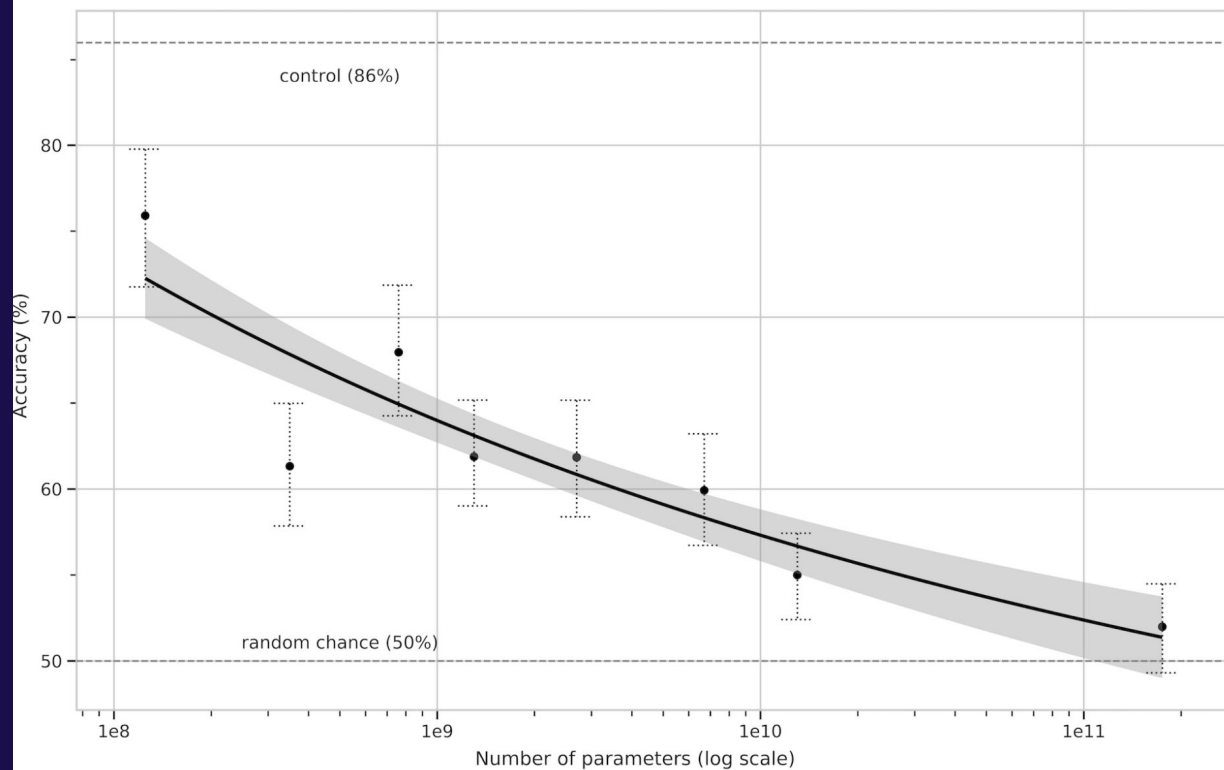


Datasets

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

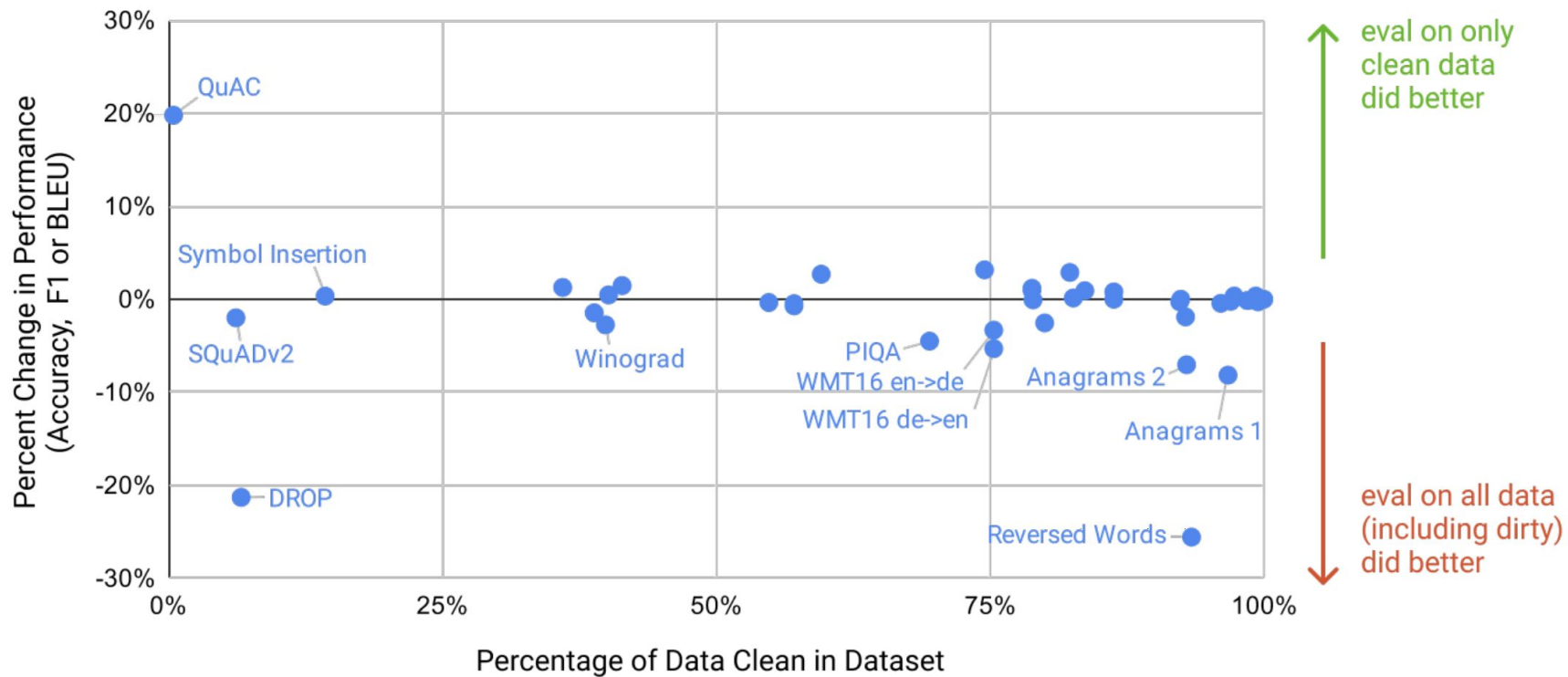


Human ability to detect model generated news articles



Measuring and Preventing Memorization Of Benchmarks

- Train/test overlap
- Removal bug
- Clean vs dirty evals
- What does it mean to have a clean training set?



Limitations

- Perf/FLOP still low
- Poor multilingual performance
- Query entailment doesn't perform
- Didn't publish fine-tuning results
- "common sense"
- Reliability and bias
- Dataset quality
 - Extraction
 - Filtering
- Cost & deployability

Poor English input: I eated the purple berries.

Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I'd appreciate it.

Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.

Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I'd be more than happy to work with you in another project.

Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.

Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.

Good English output: The patient died.

Thank You

Visit openai.com for more information.

FOLLOW @OPENAI ON TWITTER