Language Models are Few-Shot Learners

OpenAI

- GPT1.0
  - 12层单向transformer
  - Finetune
  - 不如bert

- Gpt2.0
  - 参数15亿
  - 输入加入任务描述
  - 没有放出来商用 closeAI

GPT3.0

- 不需finetune
- 单向Transformer
- 基本延续GPT2.0
- 参数1750亿
  - 700G硬盘
- 训练花费1200万美元
- 31位作者
- 付费商用 提供接口 waitinglist
- 72页
- Geoffrey Hinton：鉴于GPT3在未来的惊人前景，可以得出结论：生命、宇宙和万物的答案，就只是4.398万亿个参数而已

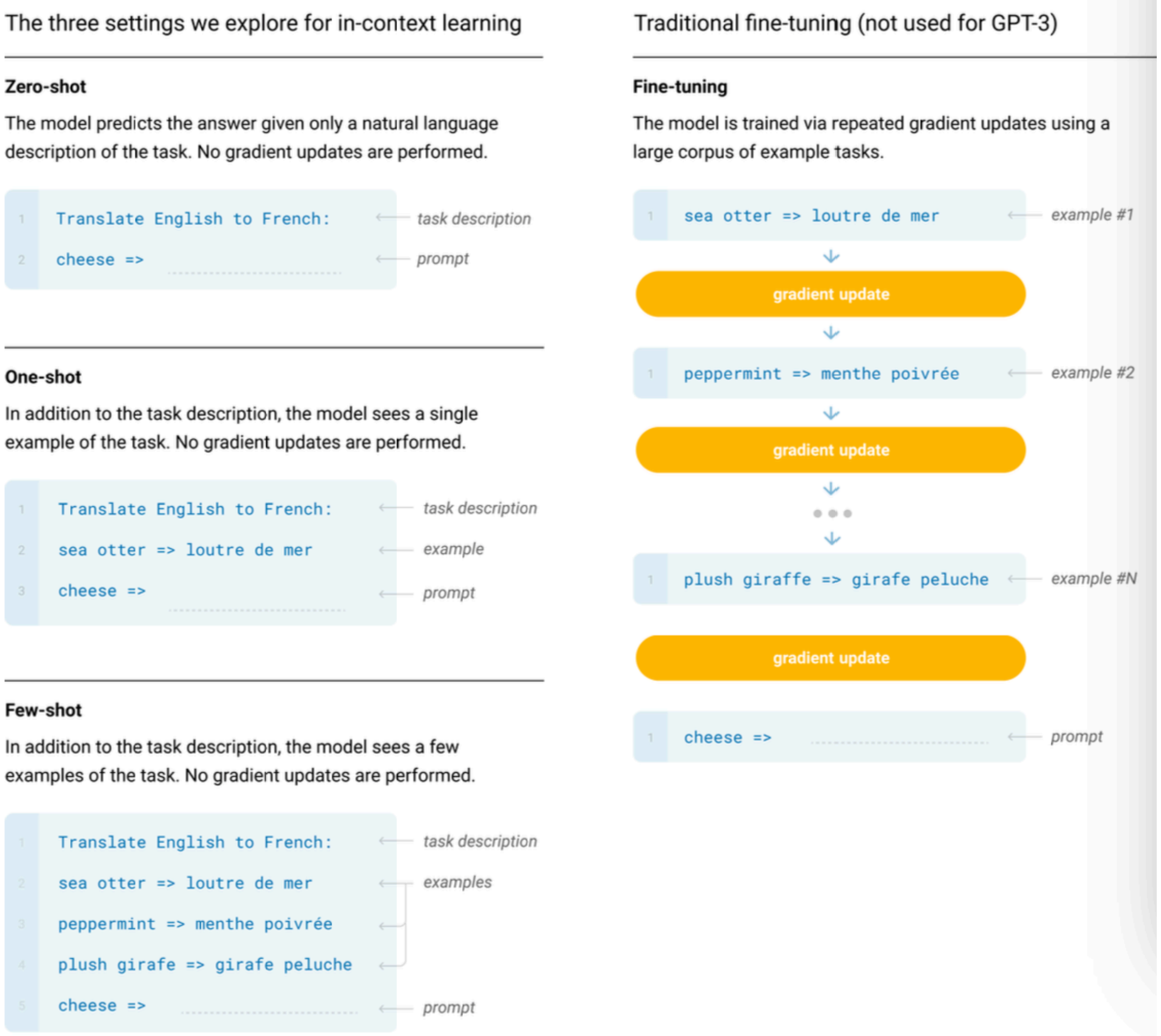

Microsoft Azure
Did you train GPT-3?
OpenAI
Yes
Microsoft Azure
What did it cost?
OpenAI
Everything
知乎 @Mario

- While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples.

- Finetune缺点
  - 过分依赖领域数据集
  - 数据少 过拟合
- **对于所有任务，应用GPT-3无需进行任何梯度更新或微调，而仅通过与模型的文本交互指定任务和少量演示即可**

Approach

- Fine-Tuning(FT)
  - 可以没必要
- Few-Shot(FS)
  - Giving K examples of context and completion,
- One-shot(1S)
  - Giving 1 examples of context and completion,
- Zero-shot(0s)
  - Giving 0 examples of context and completion,


The three settings we explore for 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:
cheese =>
task description
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:
sea otter => loutre de mer
cheese =>
task description
example
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:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
task description
examples
prompt
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.
sea otter => loutre de mer
example #1
gradient update
peppermint => menthe poivrée
example #2
gradient update
plush giraffe => girafe peluche
example #N
gradient update
cheese =>
prompt

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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Dateset

- Common Crawl dataset
  - A Trillion words

- We use filtered versions of Common Crawl.

- 基本思路
  - 文档级别、数据集之间进行模糊重复消除 防止冗余
  - 加入一些已知高质量语料库

- 45TB数据
- 但是不幸的是去重也不完美，还是看到了一些下游任务的数据，但是由于训练成本，就没停止，后来证明没影响

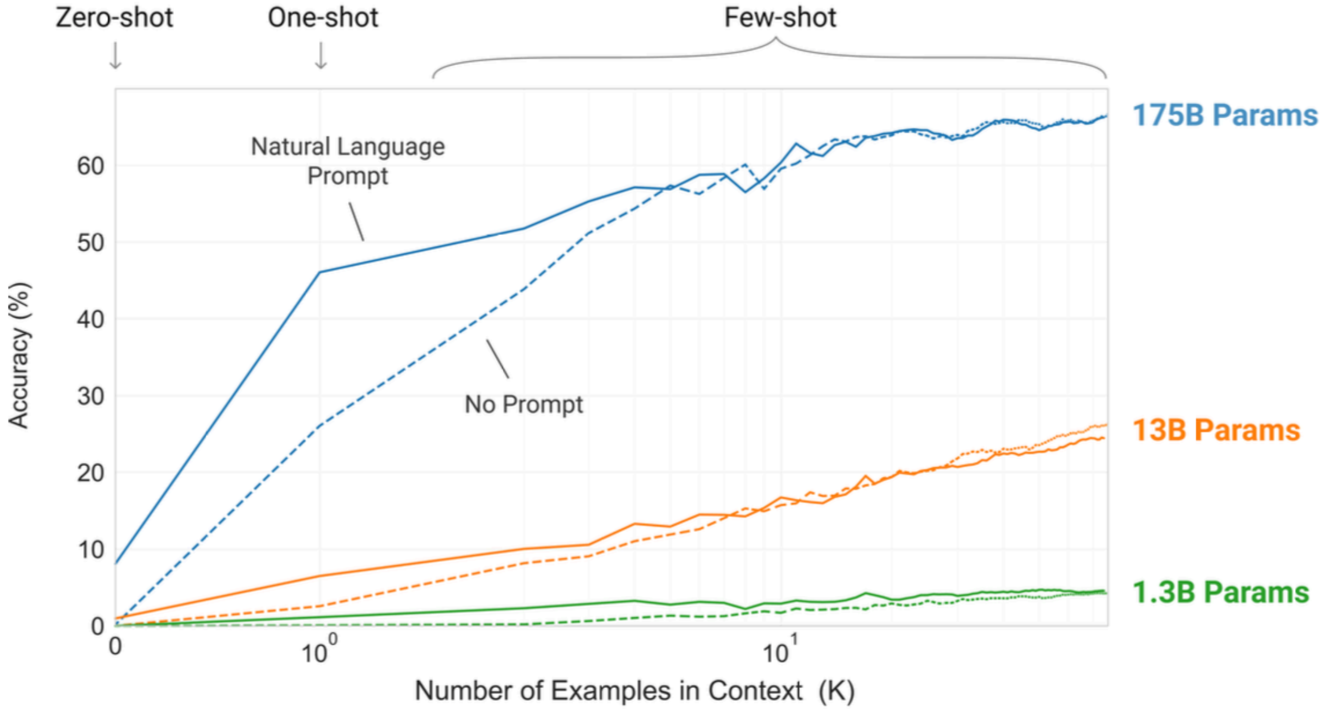


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

LAMBADA

- Test the modeling of long-range dependencies in text
- predict the last word of sentences
- The HellaSwag dataset involves picking the best ending to a story or set of instructions.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Table 3.2: Performance on cloze and completion tasks. GPT-3 significantly improves SOTA on LAMBADA while achieving respectable performance on two difficult completion prediction datasets. [a][Tur20] [b][RWC+19] [c][LDL19] [d][LCH+20]

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG+20]	-	-	<u>29.8</u>	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>

Table 3.4: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM. We report BLEU scores on the WMT'14 Fr↔En, WMT'16 De↔En, and WMT'16 Ro↔En datasets as measured by `multi-bleu.perl` with XLM's tokenization in order to compare most closely with prior unsupervised NMT work. SacreBLEU[f] [Pos18] results reported in Appendix H. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. [a][EOAG18] [b][DHKH14] [c][WXH+18] [d][oR16] [e][LGG+20] [f][SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

PIQA common sense Reasoning

- To Make a Breakfast Pizza

- "To prepare eggs to top your breakfast pizza, pour five beaten eggs into a pan and gently scramble over low-medium heat. Season with salt and pepper and be careful not to overcook."

Setting	PIQA	ARC (Easy)	ARC (Challenge)	OpenBookQA
Fine-tuned SOTA	79.4	**92.0**[KKS+20]	**78.5**[KKS+20]	**87.2**[KKS+20]
GPT-3 Zero-Shot	**80.5***	68.8	51.4	57.6
GPT-3 One-Shot	**80.5***	71.2	53.2	58.8
GPT-3 Few-Shot	**82.8***	70.1	51.5	65.4

Table 3.6: GPT-3 results on three commonsense reasoning tasks, PIQA, ARC, and OpenBookQA. GPT-3 Few-Shot PIQA result is evaluated on the test server. See Section 4 for details on potential contamination issues on the PIQA test set.

- 2 digit subtraction (2D-) – The model is asked to subtract two integers sampled uniformly from [0, 100); the answer may be negative. Example: “Q: What is 34 minus 53? A: -19”.

Arithmetic

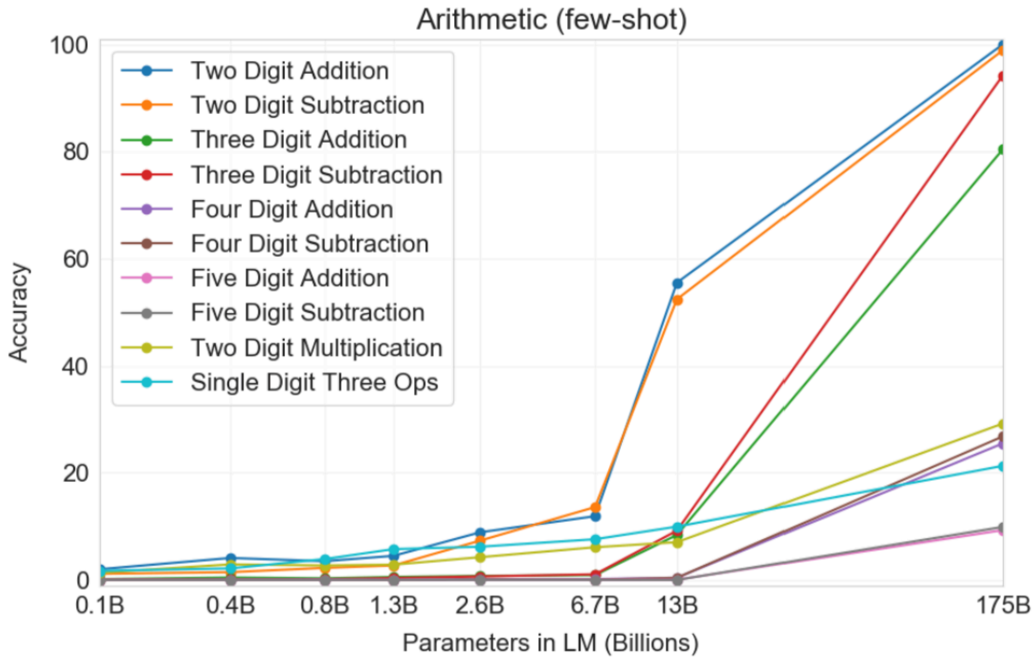


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

SAT Analogies

- audacious is to boldness (胆大 大胆）
- sanctimonious is to hypocrisy, （谦虚 虚伪）
- anonymous is to identity （匿名 身份）

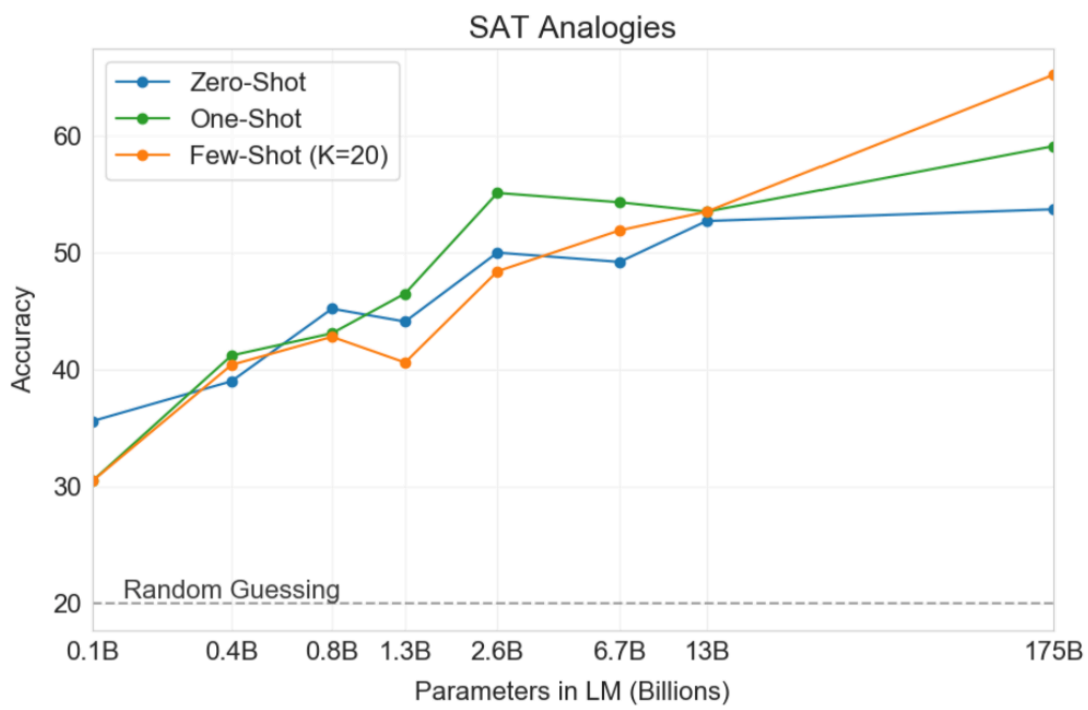


Figure 3.12: Zero-, one-,and few-shot performance on SAT analogy tasks, for different sizes of model. The largest model achieves 65% accuracy in the few-shot setting, and also demonstrates significant gains to in-context learning which are not present in smaller models.

- GPT想证明的事情，像是人类对基于广泛阅读的语境理解能力的极限探索。

- 量变引起质变