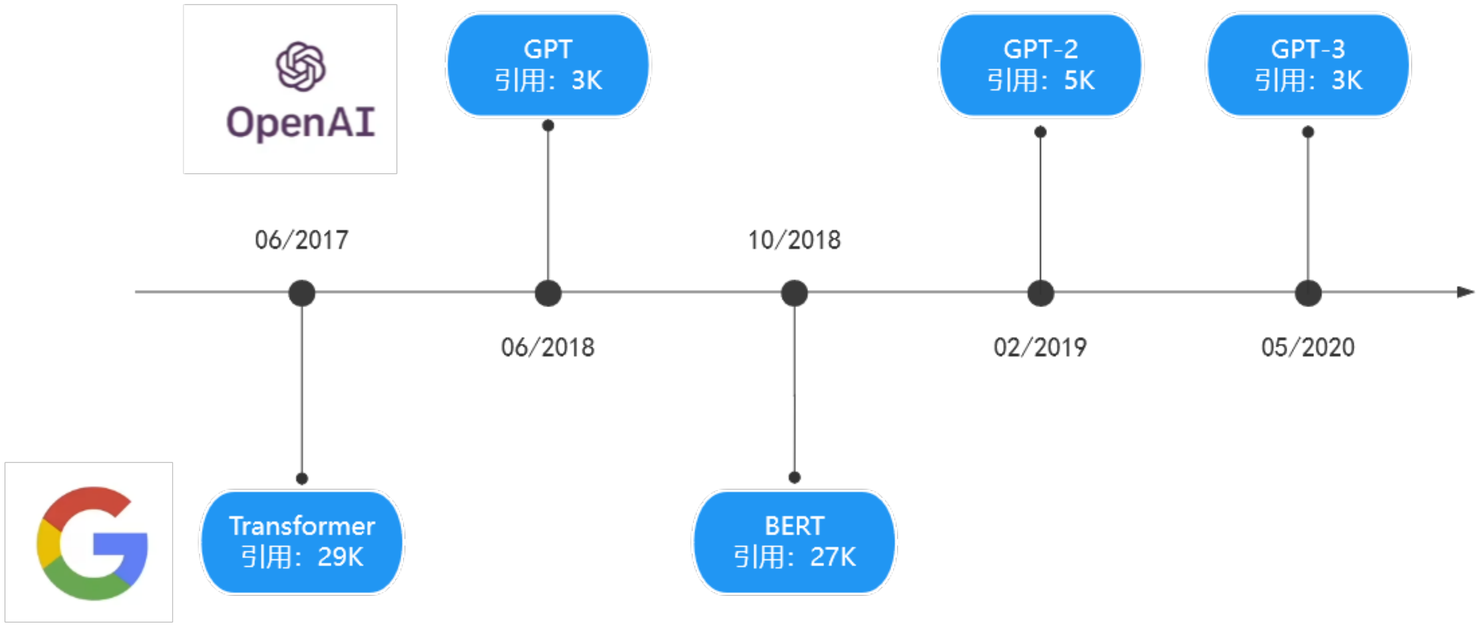
基于Decoder的预训练模型

Decoder-only Models

杜威 52265901025@stu.ecnu.edu.cn

目录

- GPT
- GPT-2
- GPT-3
- PaLM
- OPT



BERT与GPT的区别

- 基础架构不同，BERT采用Encoder结构，GPT采用Decoder结构。
- 预训练任务不同，GPT有两个预训练任务分别针对有标签的数据和无标签的数据：

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$
$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

- 可迁移的下游任务不同，GPT的上限高，但是难度更大。

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

GPT-2

- GPT性能不如BERT

- 怎么办？好像选错路了，但是又不能投敌，只能硬着头继续做。

换赛道

- 更大的数据集，更大的模型
  - GPT-2达到了1.542B的参数量，GPT是117M，大概是十倍
  - 在BERT训练数据的基础上，额外收集了40GB数据，800w文本
- 换赛道
  - fine-tune不行，每次还得重新train
  - Zero-shot，哥们给你整个不用train的
  - 这样的话就算性能不如BERT，但是可以在零样本的条件达到一个和BERT差不多的水平

数据集获取

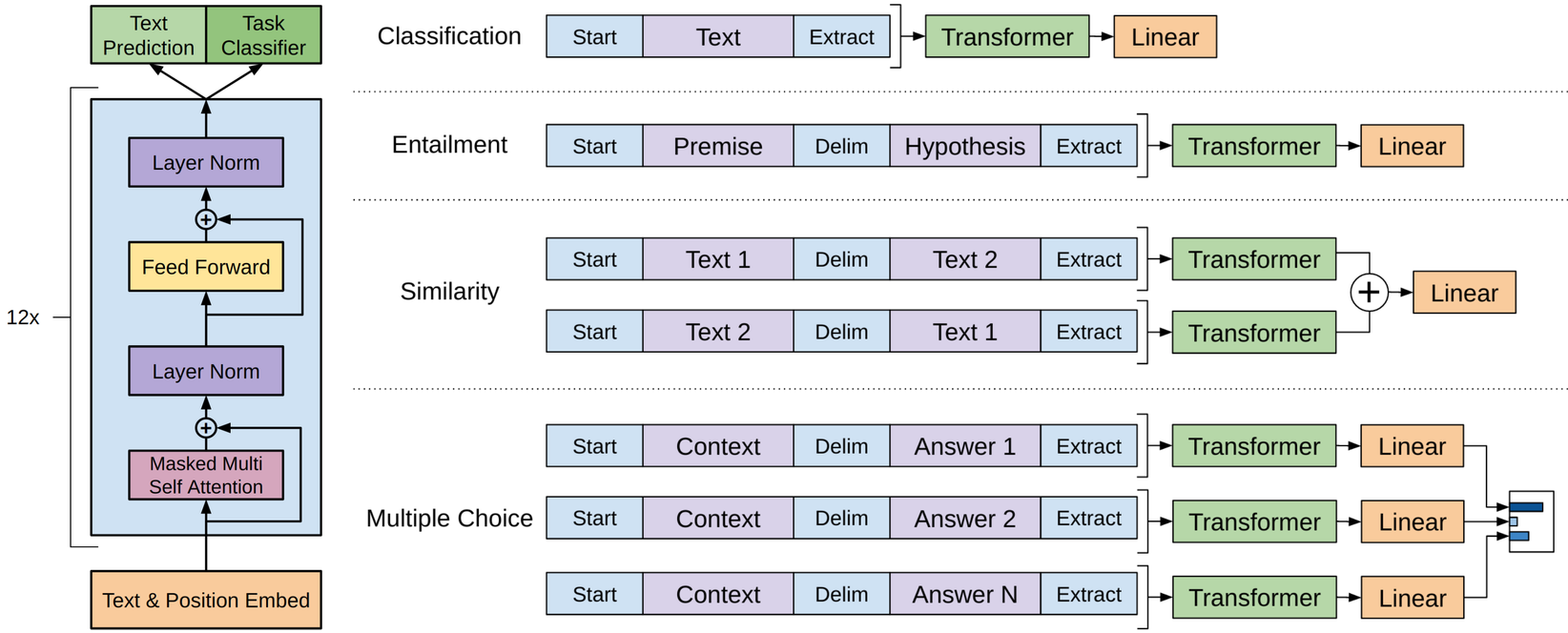


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

数据集获取

- 与1不一样的地方，GPT-2会对下游任务进行构造，加入开始符，结束符，分隔符，这些符号是之前的预训练模型没见过的，是通过fine-tune环节去认识这些符号。所以输入的形式必须由模型见过的文本组成（prompt）

sequence of symbols. For example, a translation training example can be written as the sequence (translate to french, english text, french text). Likewise, a reading comprehension training example can be written as (answer the question, document, question, answer). McCann et al. (2018) demon-

数据集获取

- 爬去Reddit上点赞较高的帖子，保证帖子中的文本质量，这样的话很可能模型在训练数据集中就已经见过prompt了

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose**,” which translates as, ”**Lie lie and something will always remain.**”
“I hate the word ‘**perfume**,”’ Burr says. ‘It’s somewhat better in French: ‘**parfum**.’
If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

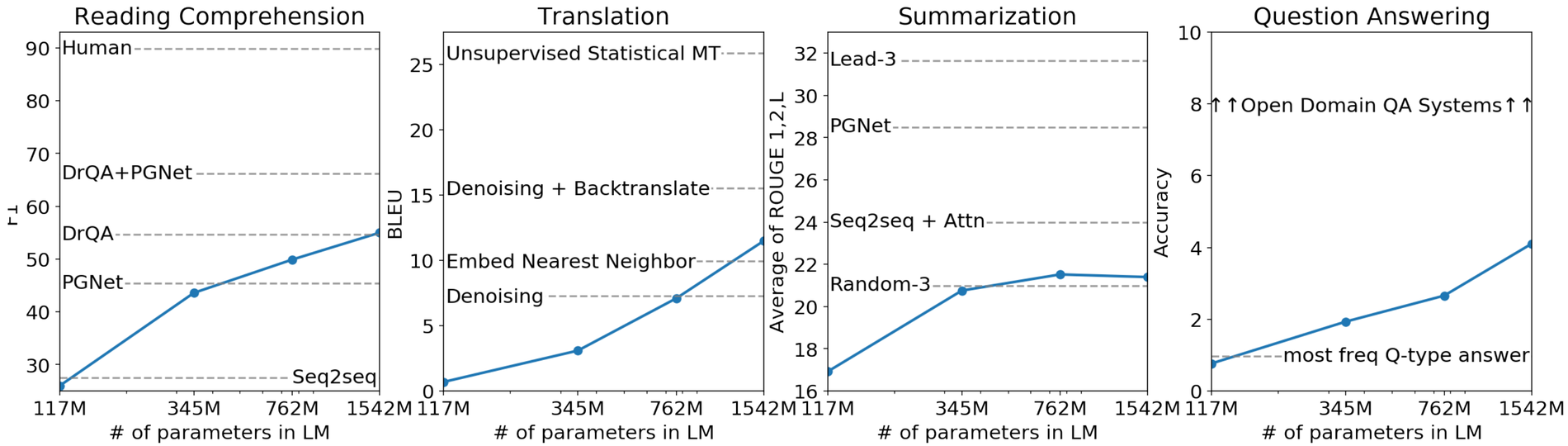
Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

零样本性能

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

零样本性能


Reading Comprehension
Translation
Summarization
Question Answering
Human
DrQA+PGNet
DrQA
PGNet
Seq2seq
Unsupervised Statistical MT
Denoising + Backtranslate
Embed Nearest Neighbor
Denoising
Lead-3
PGNet
Seq2seq + Attn
Random-3
↑ ↑ Open Domain QA Systems ↑ ↑
most freq Q-type answer
BLEU
Average of ROUGE 1,2,L
Accuracy
117M
345M
762M
1542M
# of parameters in LM

Language Models are Few-Shot Learners

Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Jared Kaplan† **Prafulla Dhariwal** **Arvind Neelakantan** **Pranav Shyam** **Girish Sastry**

Amanda Askell **Sandhini Agarwal** **Ariel Herbert-Voss** **Gretchen Krueger** **Tom Henighan**

Rewon Child **Aditya Ramesh** **Daniel M. Ziegler** **Jeffrey Wu** **Clemens Winter**

Christopher Hesse **Mark Chen** **Eric Sigler** **Mateusz Litwin** **Scott Gray**

Benjamin Chess **Jack Clark** **Christopher Berner**

Sam McCandlish **Alec Radford** **Ilya Sutskever** **Dario Amodei**

OpenAI

GPT-3

- 更大的模型，更大的数据集。

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

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.

GPT-3

- GPT-2创新性有余，有效性不足，作者希望GPT-3能够拥有更强的性能。于是将重点转移到few-shot

- 对标人类学习过程，实际上我们在上课的时候并不需要有很多的训练数据，都是几个例题，那么GPT-3应该要能够做类似的事情。

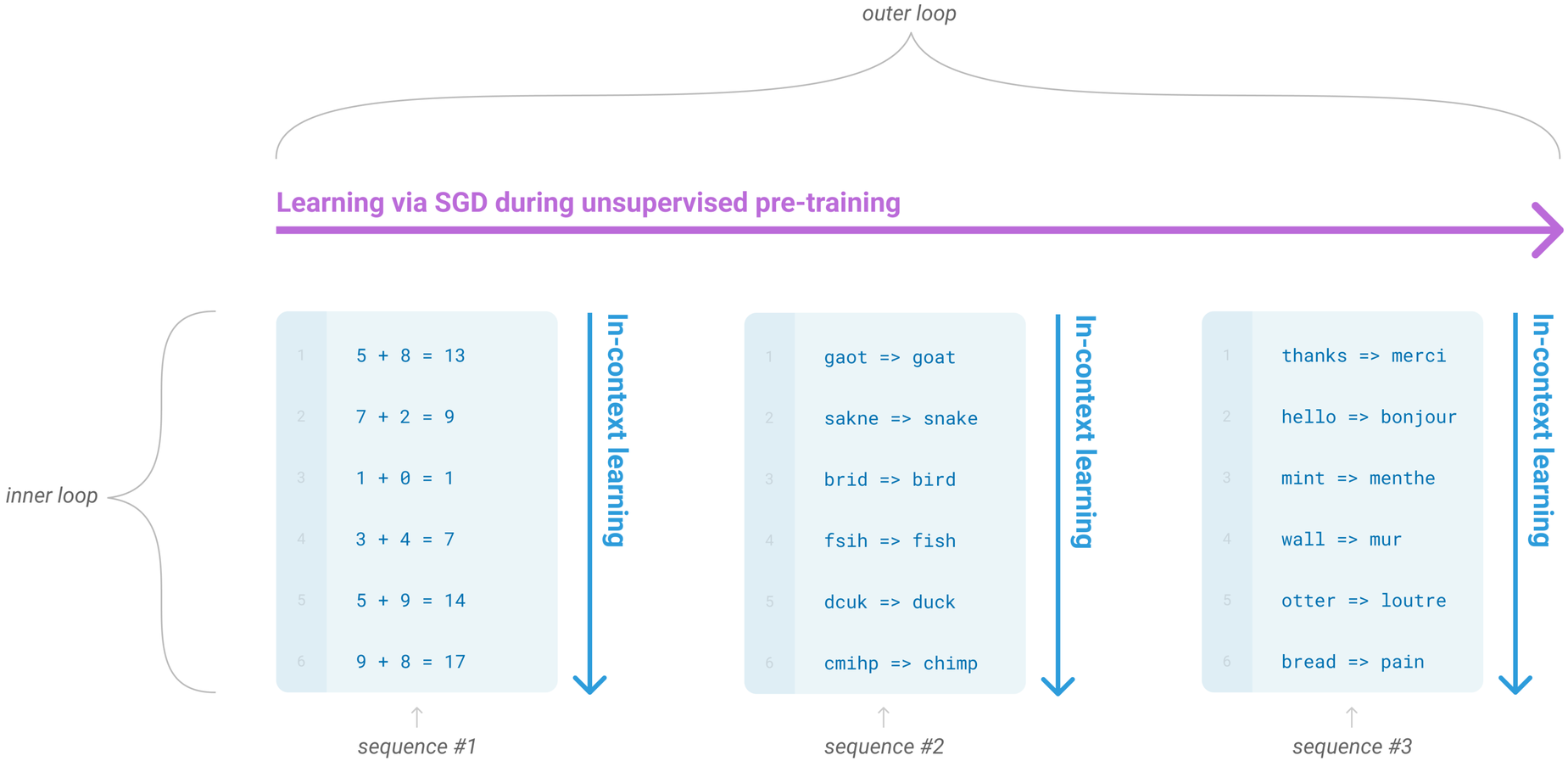
few-shot方法

- 随着模型的膨胀，就算是few-shot，如果要调整模型参数也是一个不小的开销。于是GPT-3放弃了微调，使用了一种新的few-shot方法。“Meta learning”，“in-context learning”（Facebook：？）

- 对于以往的fine-tune方法，GPT-3提出
  - 还是需要重新训练模型，需要有标号的数据集。
  - 可能不是模型训练的好，而是fine-tune的时候拟合的好。

- GPT-3的新方法希望能够避免上述的问题

In-Context-Learning



In-Context-Learning

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.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

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

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

Few-shot

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

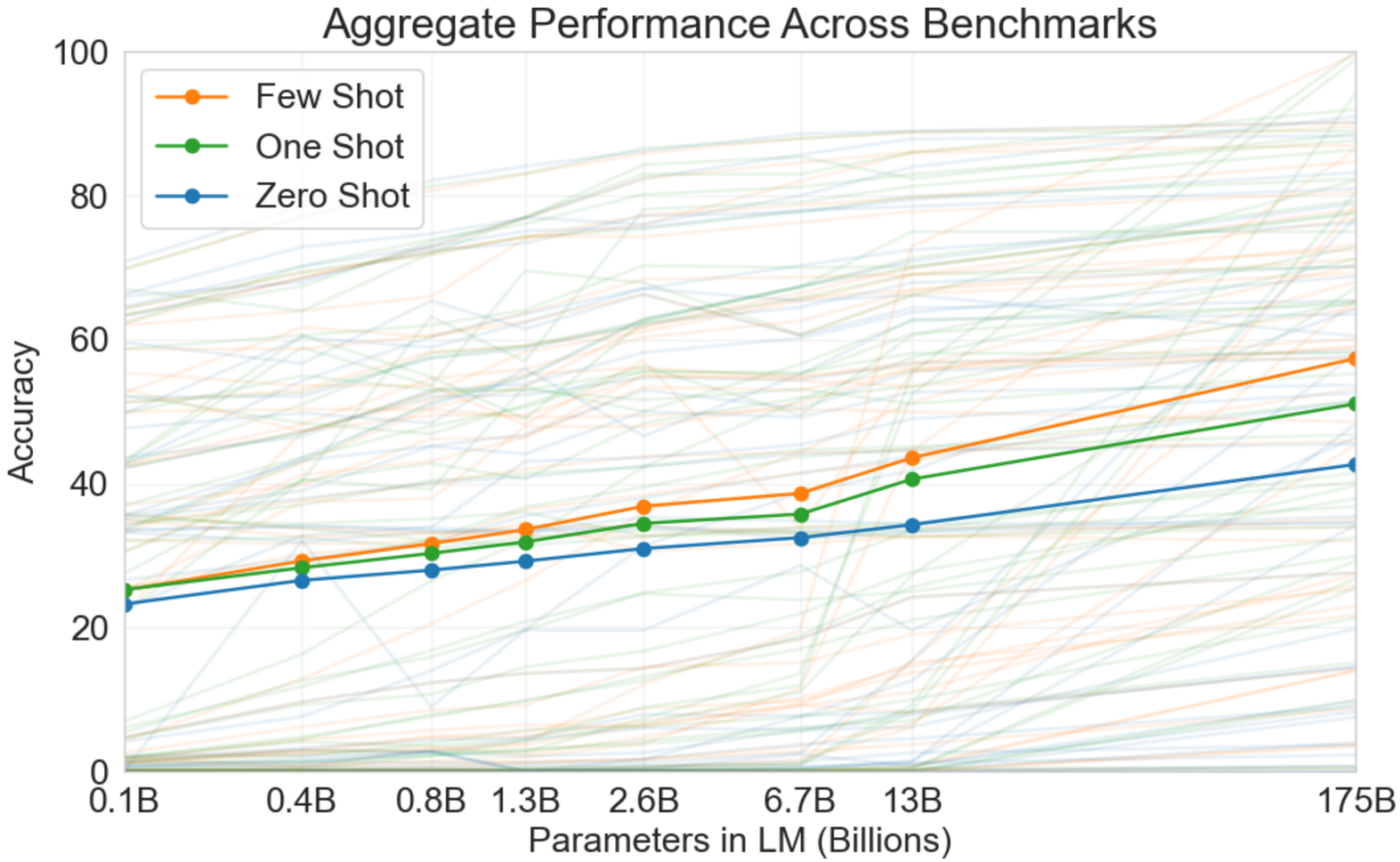
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

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.

Result


Aggregate Performance Across Benchmarks
Few Shot
One Shot
Zero Shot
Accuracy
100
80
60
40
20
0
0.1B
0.4B
0.8B
1.3B
2.6B
6.7B
13B
175B
Parameters in LM (Billions)

GPT-3 APP

- GPT-3 Demo：https://towardsdatascience.com/gpt-3-demos-use-cases-implications-77f86e540dc1

- Copilot：https://github.com/features/copilot/

GPT-3 APP

- GPT-3 Demo：https://towardsdatascience.com/gpt-3-demos-use-cases-implications-77f86e540dc1

- Copilot：https://github.com/features/copilot/

OPT: Open Pre-trained Transformer Language Models

Susan Zhang[*], **Stephen Roller**[*], **Naman Goyal**[*],
Mikel Artetxe, **Moya Chen**, **Shuohui Chen**, **Christopher Dewan**, **Mona Diab**, **Xian Li**,
Xi Victoria Lin, **Todor Mihaylov**, **Myle Ott**[†], **Sam Shleifer**[†], **Kurt Shuster**, **Daniel Simig**,
Punit Singh Koura, **Anjali Sridhar**, **Tianlu Wang**, **Luke Zettlemoyer**

Meta AI

{susanz,roller,naman}@fb.com

OPT

- Open Source
- Meta AI表示，最低只需要16块英伟达V100 GPU，就能训练并部署OPT-175B模型。
- 参数相同，效果更好，更环保

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* **Sharan Narang*** **Jacob Devlin***
Maarten Bosma **Gaurav Mishra** **Adam Roberts** **Paul Barham**
Hyung Won Chung **Charles Sutton** **Sebastian Gehrmann** **Parker Schuh** **Kensen Shi**
Sasha Tsvyashchenko **Joshua Maynez** **Abhishek Rao†** **Parker Barnes** **Yi Tay**
Noam Shazeer‡ **Vinodkumar Prabhakaran** **Emily Reif** **Nan Du** **Ben Hutchinson**
Reiner Pope **James Bradbury** **Jacob Austin** **Michael Isard** **Guy Gur-Ari**
Pengcheng Yin **Toju Duke** **Anselm Levskaya** **Sanjay Ghemawat** **Sunipa Dev**
Henryk Michalewski **Xavier Garcia** **Vedant Misra** **Kevin Robinson** **Liam Fedus**
Denny Zhou **Daphne Ippolito** **David Luan‡** **Hyeontaek Lim** **Barret Zoph**
Alexander Spiridonov **Ryan Sepassi** **David Dohan** **Shivani Agrawal** **Mark Omernick**
Andrew M. Dai **Thanumalayan Sankaranarayana Pillai** **Marie Pellat** **Aitor Lewkowycz**
Erica Moreira **Rewon Child** **Oleksandr Polozov†** **Katherine Lee** **Zongwei Zhou**
Xuezhi Wang **Brennan Saeta** **Mark Diaz** **Orhan Firat** **Michele Catasta†** **Jason Wei**
Kathy Meier-Hellstern **Douglas Eck** **Jeff Dean** **Slav Petrov** **Noah Fiedel**

Google Research

PaLM

- 5400亿参数
- 一个新的AI架构
- 以pathway系统为基础的分布式计算

PaLM

Pathways 在 AI 和模型上的定义是**一个**新的AI架构

（1）能同时执行**众多**（AI）任务

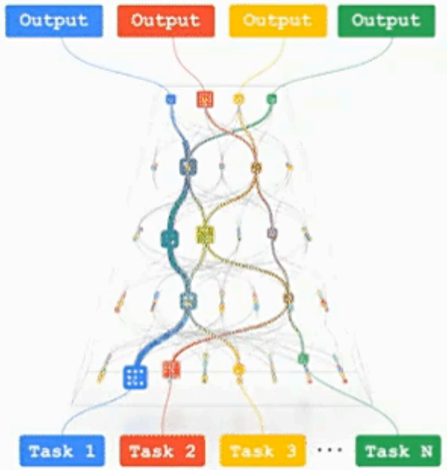
（2）**快速**学习新任务

（3）拥有对（真实）**世界的更好理解**。

pathways 由一堆 paths 组成，每个path都是一个任务输入到输出的处理，它们之间的交错意味着共享共同的权重/结构。

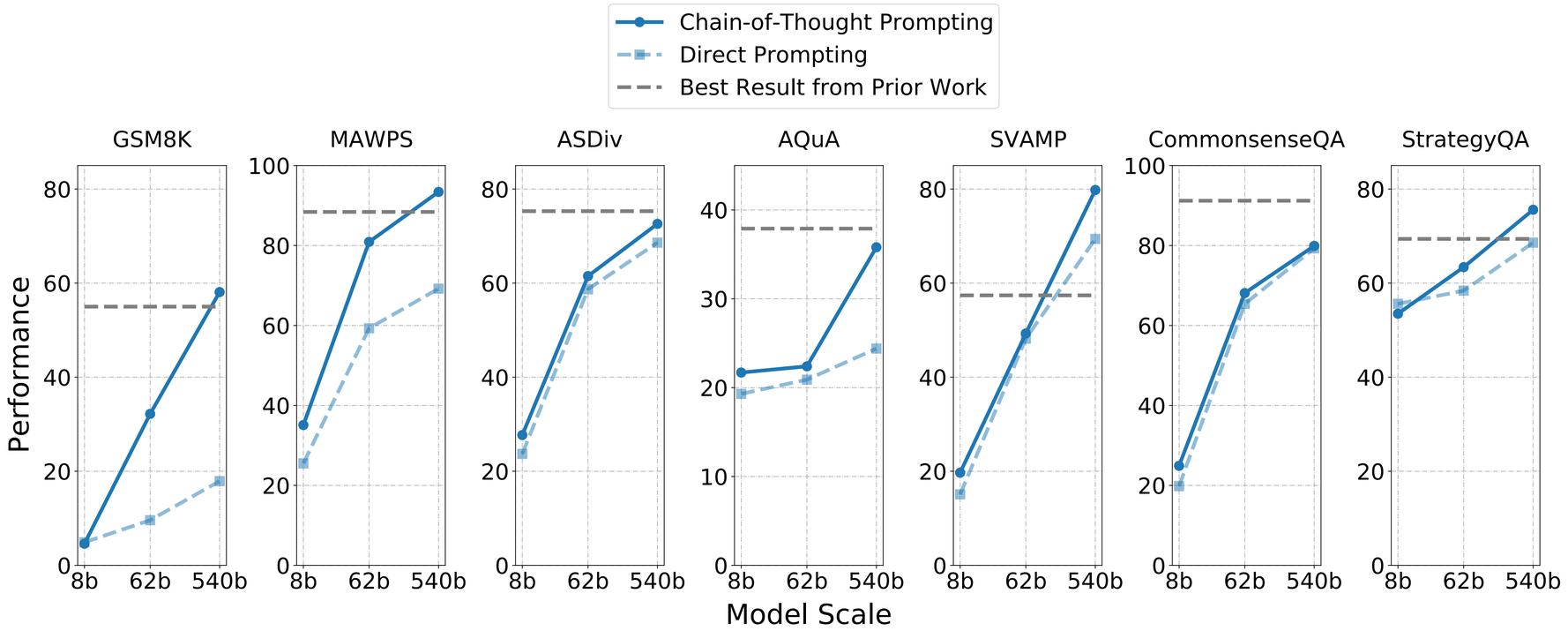
•

PaLM



Pathways: A single model that can generalize across millions of tasks.

PaLM


Chain-of-Thought Prompting
Direct Prompting
Best Result from Prior Work
GSM8K
MAWPS
ASDiv
AQuA
SVAMP
CommonsenseQA
StrategyQA
Performance
Model Scale
8b
62b
540b

总结

模型	参数规模
BERT-base(Google)	110M
BERT-large(Google)	340M
GPT(OpenAI)	117M
GPT-2(OpenAI)	1.542B
GPT-3(OpenAI)	175B
GPT-4(OpenAI)	知情人士透露可能达到100000B or 175B~280B
OPT(Meta AI)	175B
Gopher(DeepMind)	280B
PaLM(Google)	540B
ERNIE(百度)	260B

总结

- 用大规模预训练模型做Teacher Model

- 做数据

END