# Title：

Language Models are Few-Shot Learners

GPT：在子任务上提供少量样本就能完成

GPT-2：在子任务上具有zero-shot的能力，但是效果不是很好

GPT-3：解决GPT-2的有效性问题，给一些样本但不多

# Abstract：

训练了一个自回归的GPT-3模型，有175B的可学习参数，比那些非稀疏的模型大10倍。GPT-3在做子任务的时候不做任何权重更新和微调。在很多NLP的任务上取得了很好的成绩。

GPT-3能够生成一些人类很难区分的新闻类文章

# Introduction

最近都用预训练模型+微调去做一些任务，这当然是有问题的，因为需要和任务相关的数据集和微调。

【PROBLEM】：

1.大的数据集需要标号

2.样本没有出现在预训练数据集中时泛化性不见得比小模型要好，如果在微调上表现好可能是过拟合了训练数据

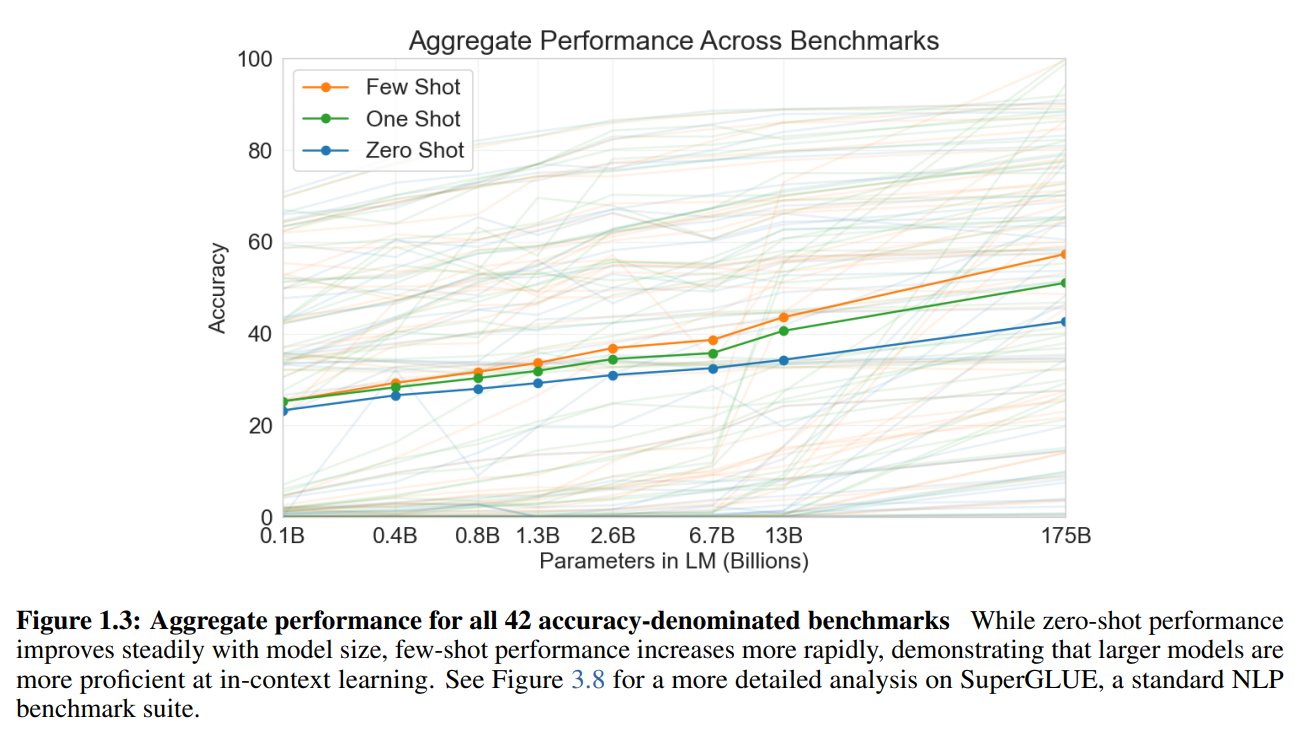
3.人类不需要一个很大的数据集来做任务

评估GPT-3从以下三个方面：

1.few-shot learning：在上下文只提供10-100的样例

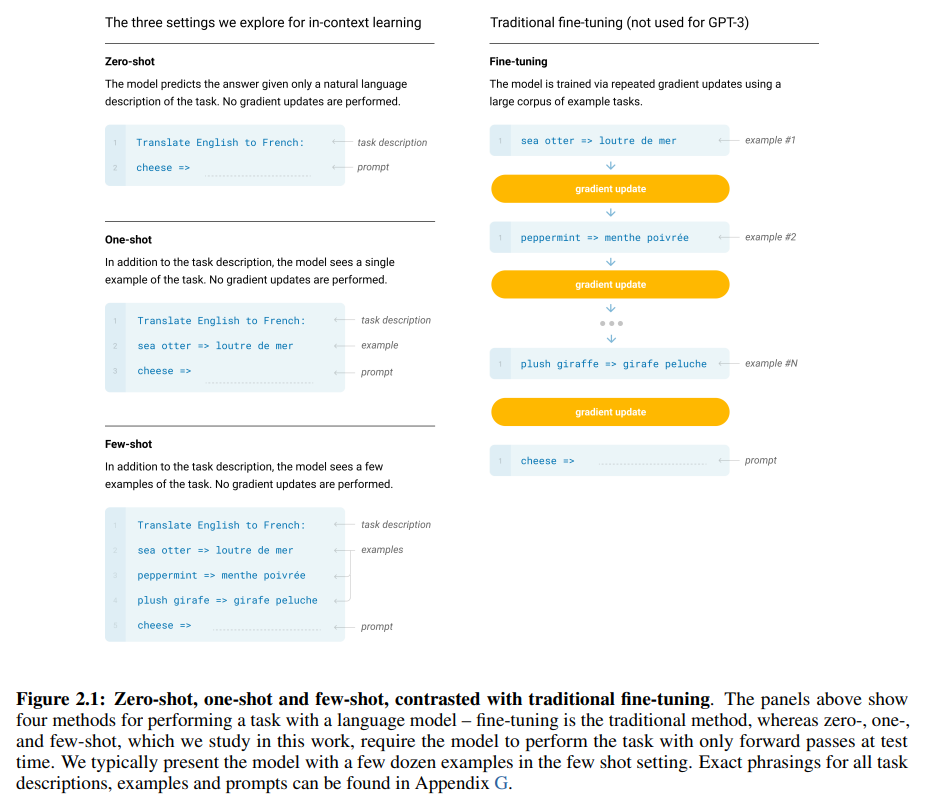
2.one-shot learning：只给一个样本

3.zero-shot learning：不给样本



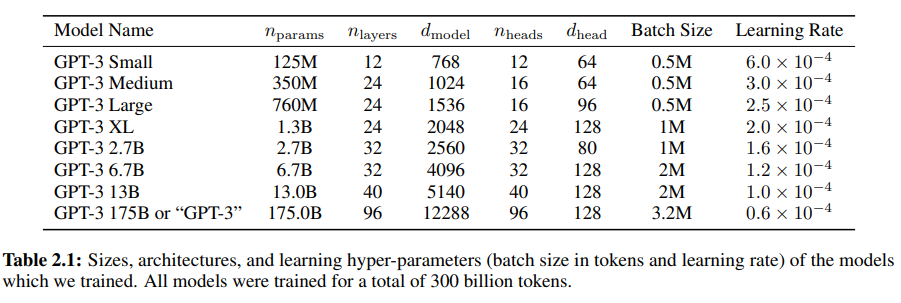
在三个设定下模型的区别

# Approach

上图讲述了fune-tunning，few-shot，one-shot，zero-shot分别怎么做

## Model and Architectures

GPT-3和GPT-2的模型是一样的，但是把Sparse Transformer的工作拿了过来。设计了8个不同大小的模型



# Results

当训练数据指数增长，验证精度线性增加，符合power-law分布

# Limitations

GPT-3在文本生成上有些弱，如果让它写一个很长的文本，则可能出现循环现象。

结构和算法的局限性，不能像BERT一样用双向的视角。

GPT-3是很均匀的预测下一个词，没有告诉哪个词是否重要，只用文本训练，没见过视频、音频和图片长什么样子。

样本有效性不够。

不知道是否是真的学习到了样本还是在大量的数据里记住了样本

GPT-3和很多深度学习模型一样是无法解释的