文本大模型

- 得唠唠这几个事（故事背景是我很穷，电费都付不起那种）
  - llama：Meta开源语言模型（我们能负担得起下游任务了）
  - LoRA：给你模型你也得能训练的动才行（咱们也能微调下游任务了）
  - Self-Instruct：下游任务得规矩一些，输入和输出都得有一个标准格式
  - PEFT：Parameter-Efficient Fine-Tuning（将上面三个大哥整合到一起）

llama

- 流行的原因
  - 让NLP走进千万家（模型参数量，7B - 65B，我单卡也能跑了）
  - 13B模型的效果超越GPT3（130亿吊打1750亿的故事）
  - 仅使用公开数据集进行训练（光走正道都很强，那整点捷径不更狠了）
  - 算法上没啥特别的，完全开源，提供预训练模型（都可以来玩了）

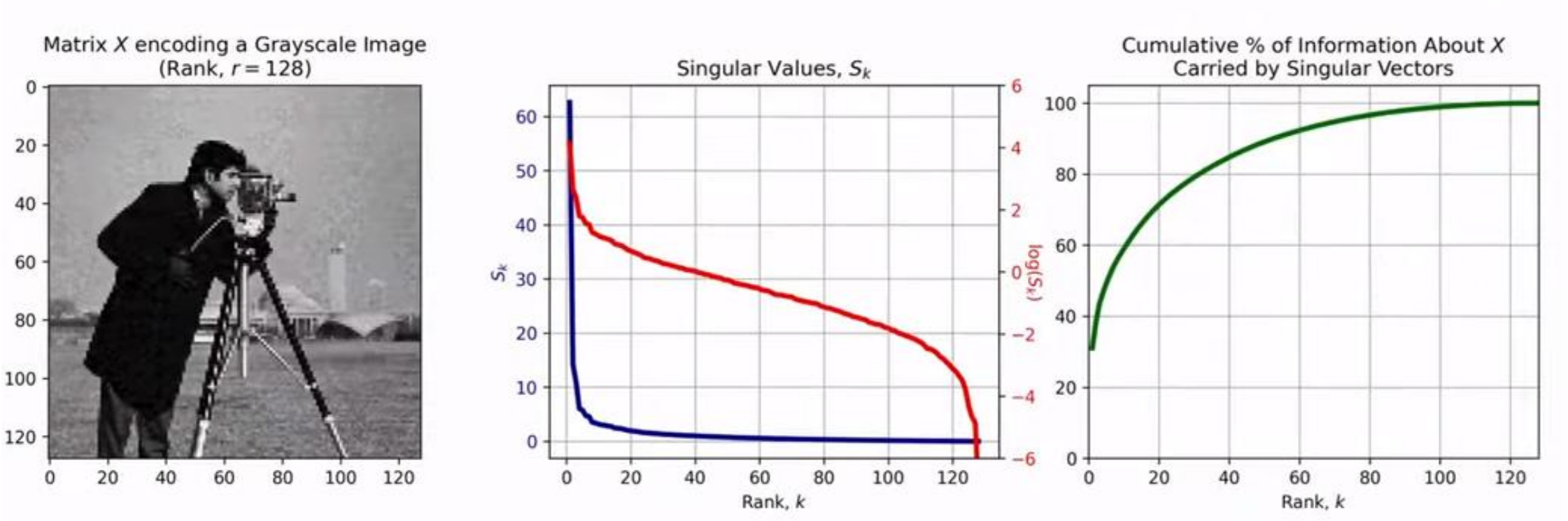
LoRA

- NLP下游任务可能遇到的问题
  - 微调不仅需要数据，更需要算力（微调个llama也得吃我几个卡）
  - 针对不同下游任务都要重新训练模型是不是忒麻烦了点
  - 基础语言模型已经很强了，就不能让他额外注意下我的任务非得重头来吗
  - 我手里就几个卡，等微调完了都过年了，新的大模型又出了还得重来

LoRA

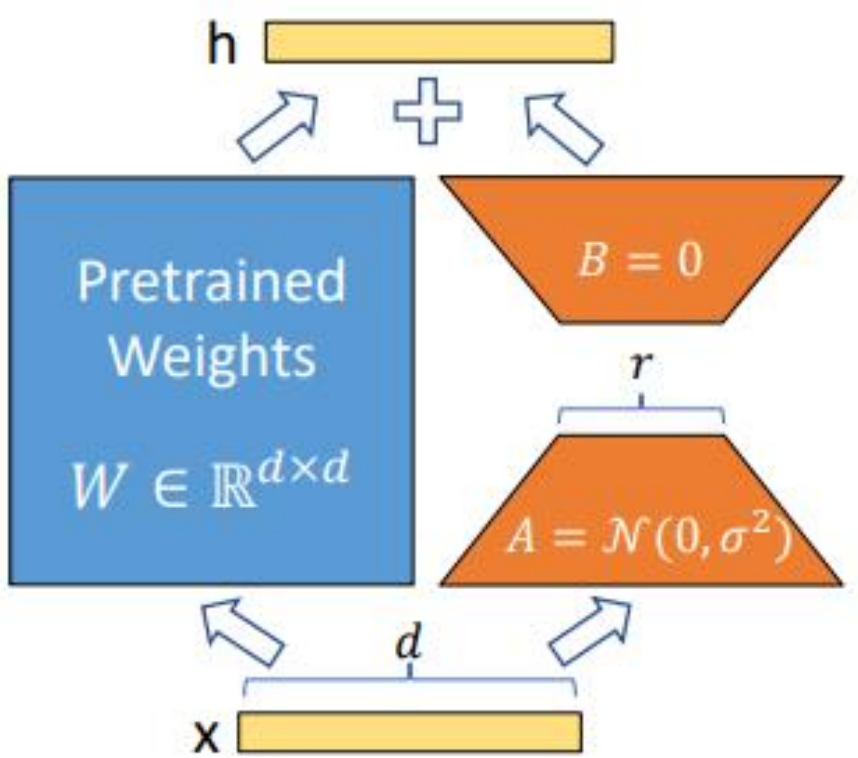
- 计算机的世界里能不能以偏概全呢

- 我们的世界里大部分的财富被少数人所拥有，NLP的世界也是如此



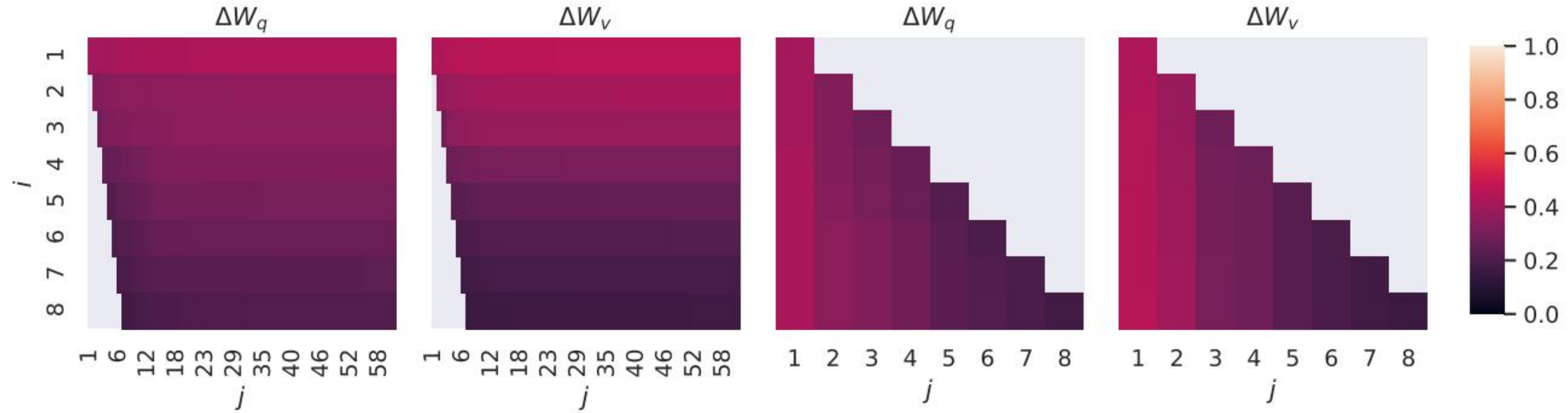
LoRA

- 解决方案
  - 训练好的模型你别动，给我稳住
  - 额外往里面加参数，训练加的即可
  - 结果：稳住的+我加的=更新后的
  - 如果我加的很少，那需要训练的也就少了



LoRA

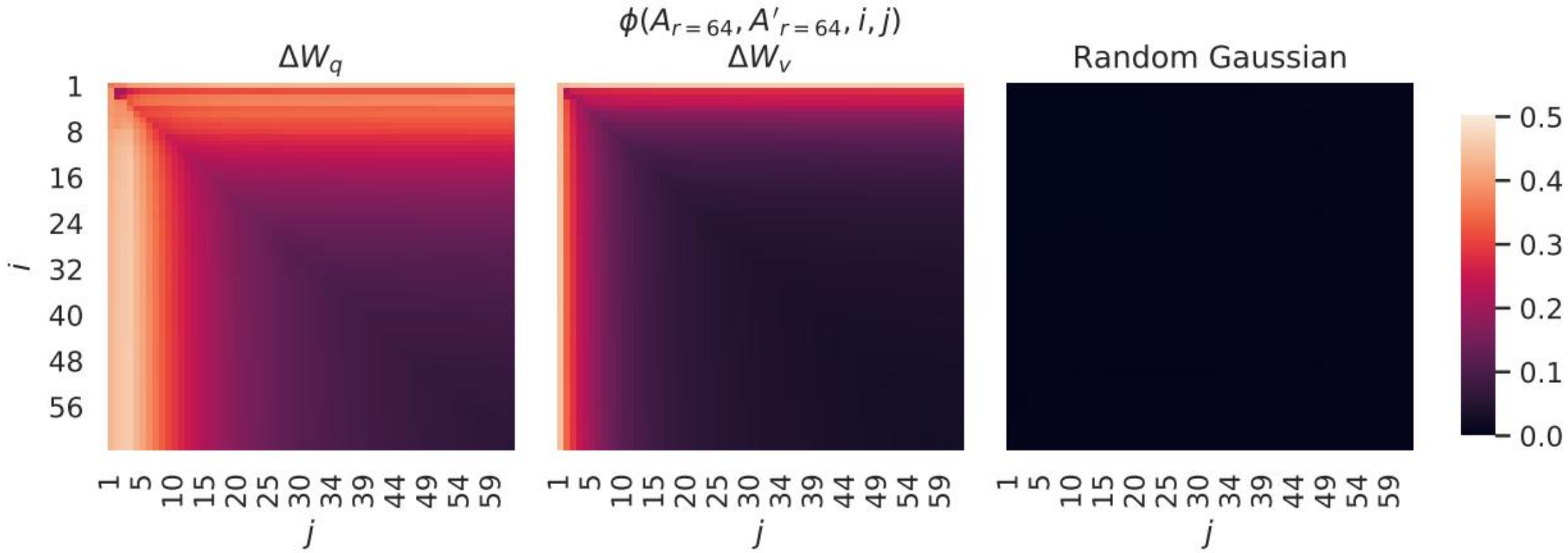
- LORA中为啥能low rank
  - r=8与r=64中，只有top的特征向量之间有关联，其他的基本都没啥关系
  - 其实就是r等于很小的时候就已经有足够多的信息来做下游任务了



LoRA

- Transformer中的Q与V

- 看起来Q需要的rank更多一点而V意思一下感觉就可以了



Self-Instruct

- 下游任务怎么才能做的好
  - 我做一个医疗智能助手，那用户基本都是在提问，XXX的原因是什么?
  - 我做一个售后客服，那用户也在提问，XXX什么时候到?
  - 既然是下游任务了，就要针对这个领域的模板来做一些专门的数据
  - 如何把模板做的好呢？咱们可以在LLM中取取经来自动生成一些标准格式

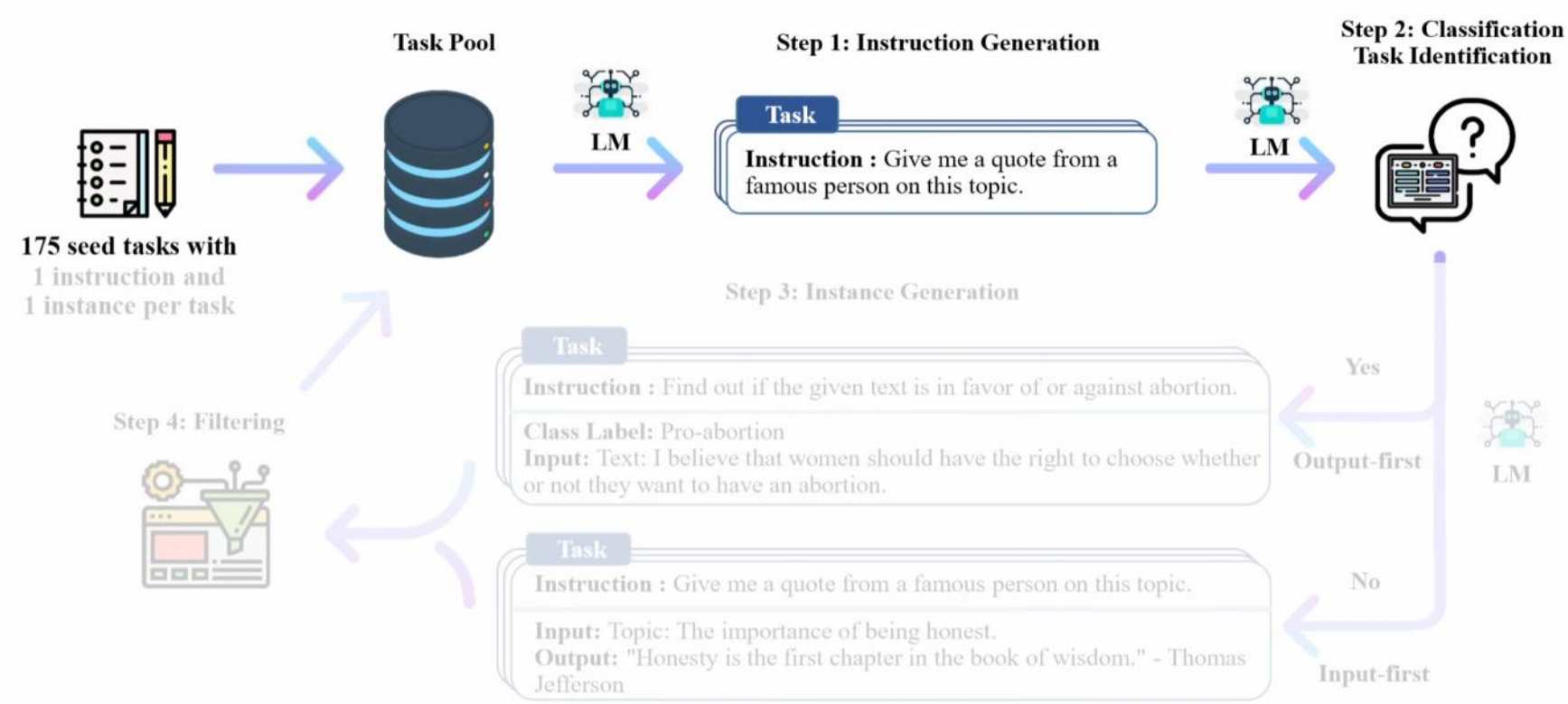
Self-Instruct

- 标准格式啥样呢
  - 指令：将下述文本总结成3点；生成下述主题的文章；提取文本中所有的人名；
  - 输入：今天天气不错挺风和日丽的，我们下午没有课。。。
  - 输出：1.今天天气好；2.今天是晴天；3.我们下午没有上课
  - 为了让下游任务能训练的更好，其实我们希望输入的就是如上的三元组

Self-Instruct

- STEP1：根据种子来生成指令

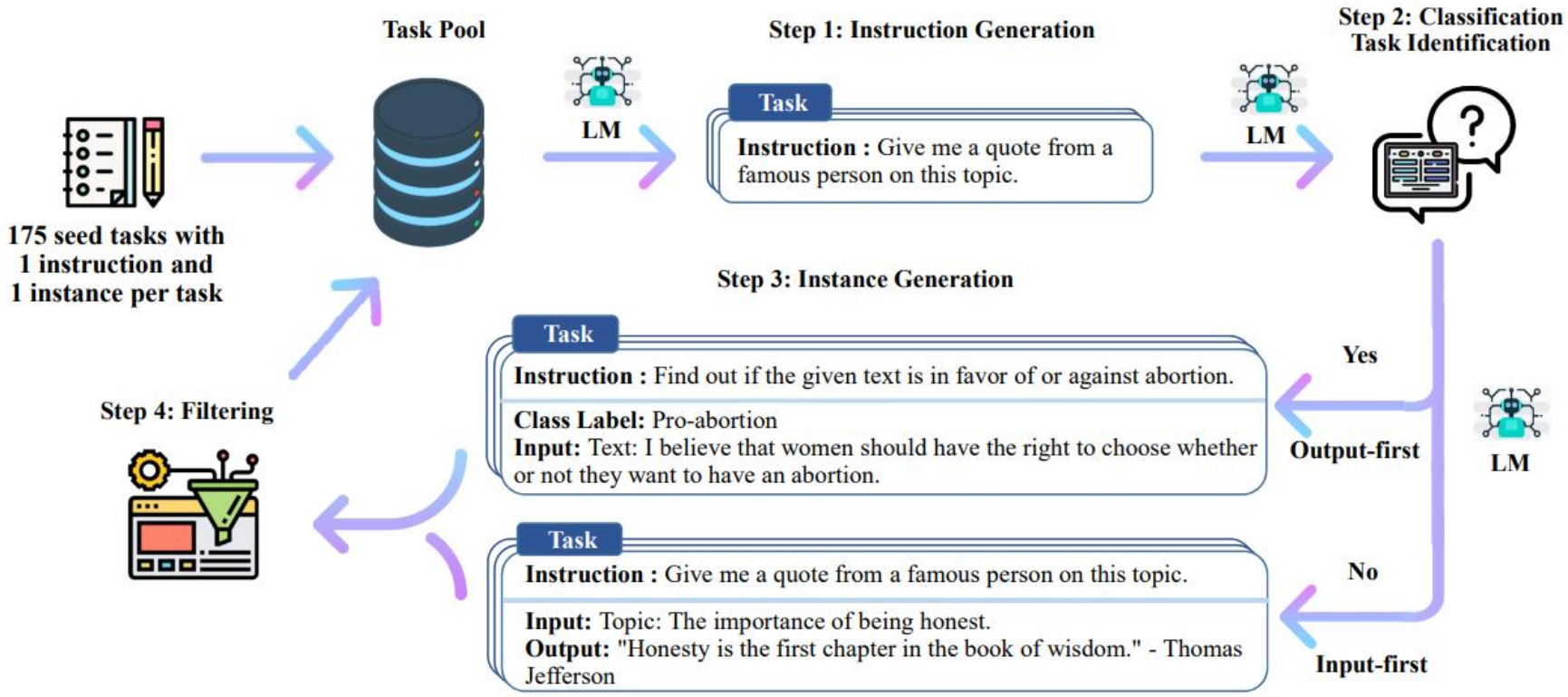
  - 人工写一部分指令，让LLM来模仿生成一些类似的指令



Self-Instruct

- STEP2：利用LLM的指令来生成结果

- 分类任务与非分类任务要区别对待，然后生成一个样本数据



PEFT

- ✓ 前面几个大哥要做的事它都给集成了
- 今年的新货，NLP的下游任务就靠它了（小作坊必备）

PEFT

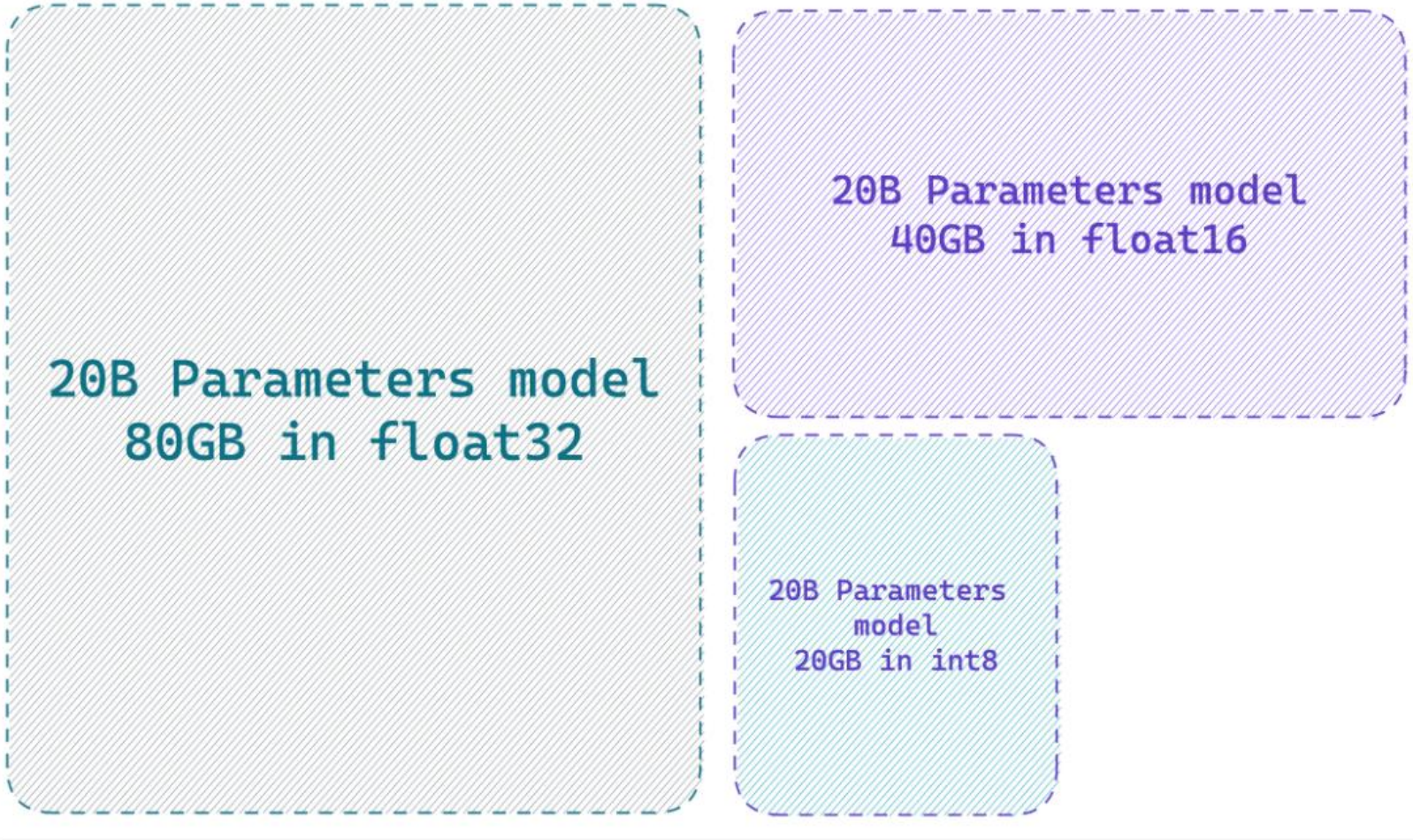
State-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods

Parameter-Efficient Fine-Tuning (PEFT) methods enable efficient adaptation of pre-trained language models (PLMs) to various downstream applications without fine-tuning all the model's parameters. Fine-tuning large-scale PLMs is often prohibitively costly. In this regard, PEFT methods only fine-tune a small number of (extra) model parameters, thereby greatly decreasing the computational and storage costs. Recent State-of-the-Art PEFT techniques achieve performance comparable to that of full fine-tuning.

Seamlessly integrated with 🤗 Accelerate for large scale models leveraging DeepSpeed and Big Model Inference.

PEFT

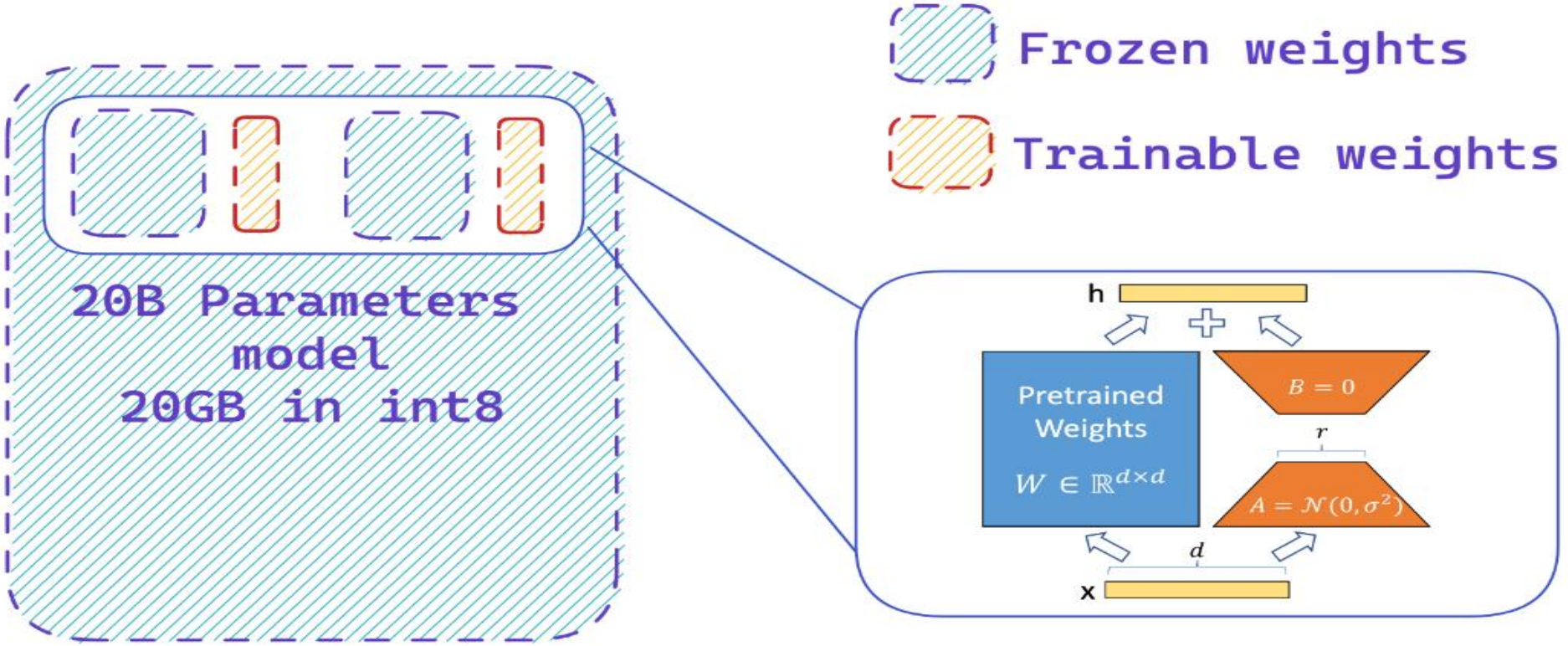
- STEP1：省钱；省电；省心
  - 20B的参数我能跑吗?
  - 80G->40G还不满足
  - 那么还可以再小!
  - PEFT提供好了转换



PEFT

- STEP2：选择合适的微调方法(例如LORA)

- 一般只需要训练低于1%的参数就可以完成对下游任务的微调



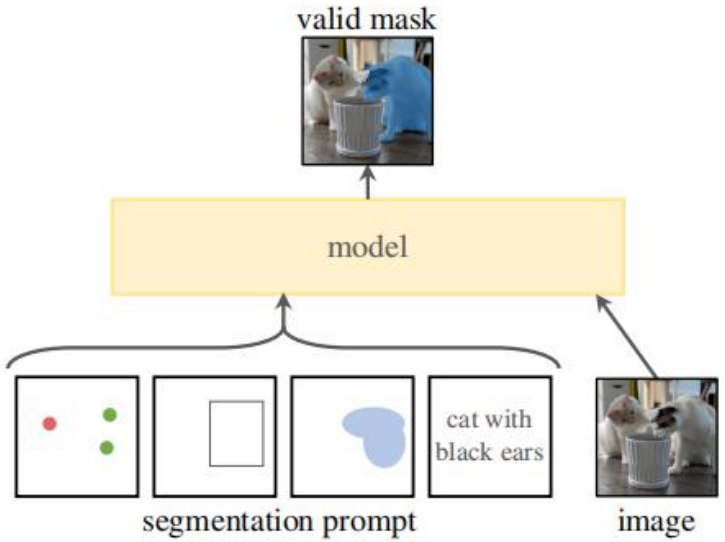
视觉大模型

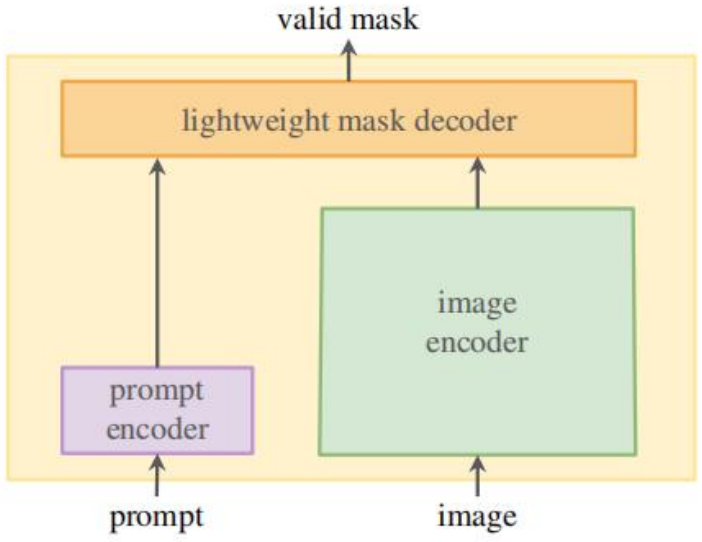
- Segment Anything（以不变应万变）

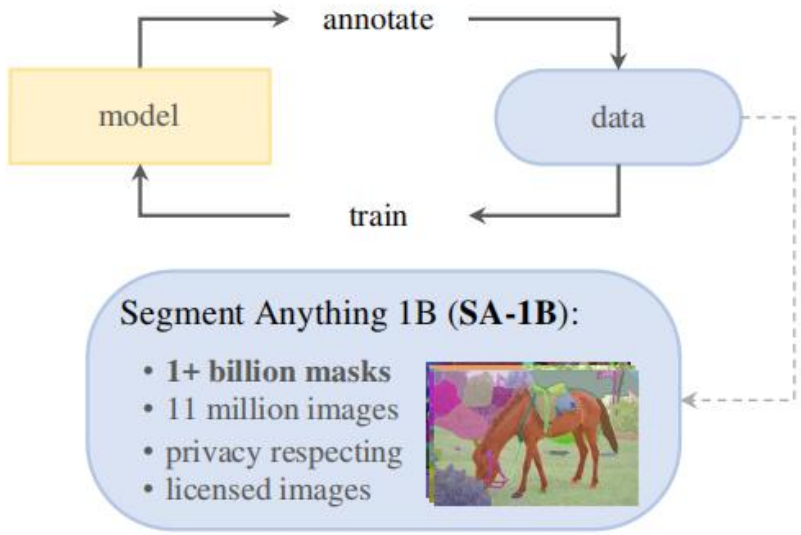
  - 微调不仅需要数据，更需要算力（微调个llama也得吃我几个卡）

Segment Anything

- 提示学习+数据循环+简易算法
  - 提示：要输出什么东西，得给我透漏点消息（仿NLP）
  - 数据链路：自产自销，循环利用（核心所在）







Segment Anything

- 何为大模型

  - 简易的模型+无限的数据=可应用于任何下游任务+实时响应

  - 视觉之路在于如何仿NLP做到自监督任务（FaceBook告诉我们这条路可行）