大模型的探索与实践

Introduction to Large Language Models

§2.2 微调与ICL

Finetuning, In-Context Learning

滕佳烨
上海财经大学
www.tengjiaye.com

回顾 Recall

多模态的核心组件

- 模态编码器
- 输入投影仪
- 大模型基座
- 输出投影仪
- 模态生成器

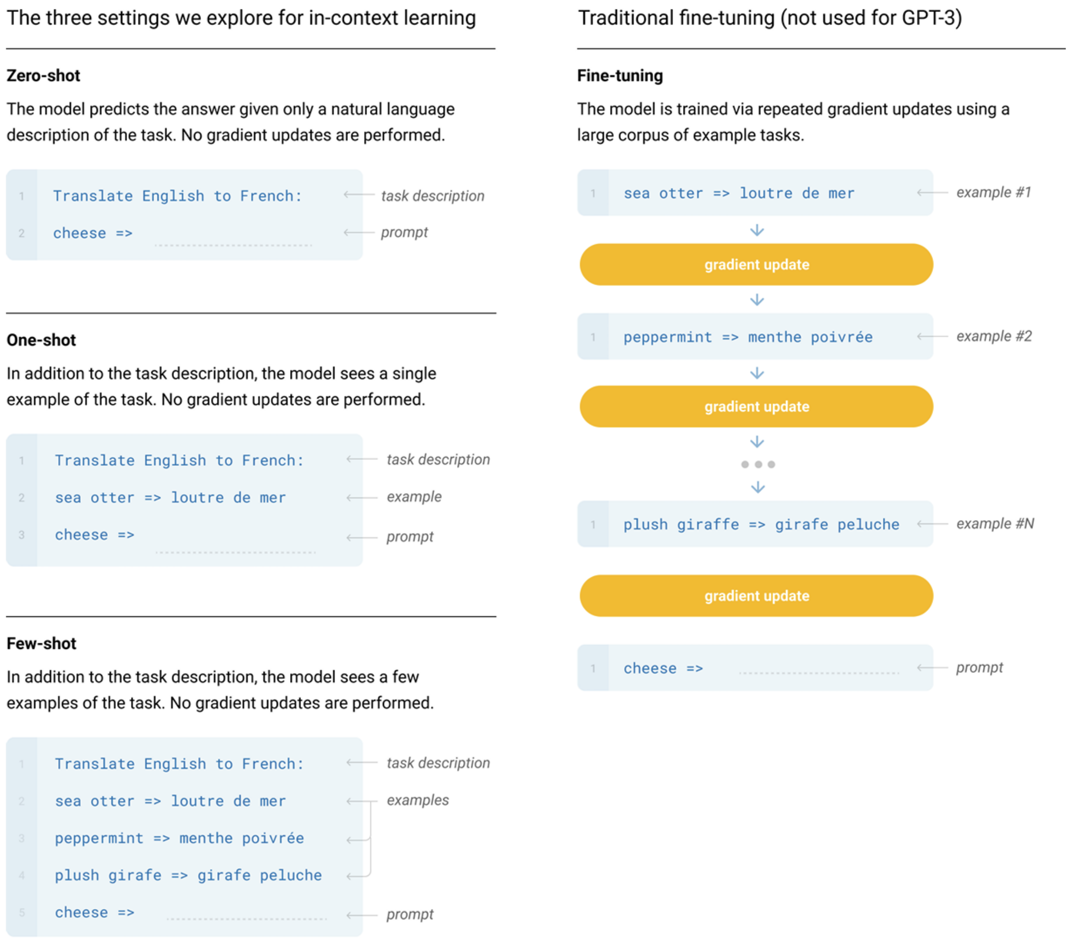
在多模态输出中产生的问题

今天的任务： Finetuning with LoRA, In-Context Learning

上下文学习 In-context Learning

不改变模型本身
通过提示词来引导模型能力

(few-shot, one-shot, few-shot)

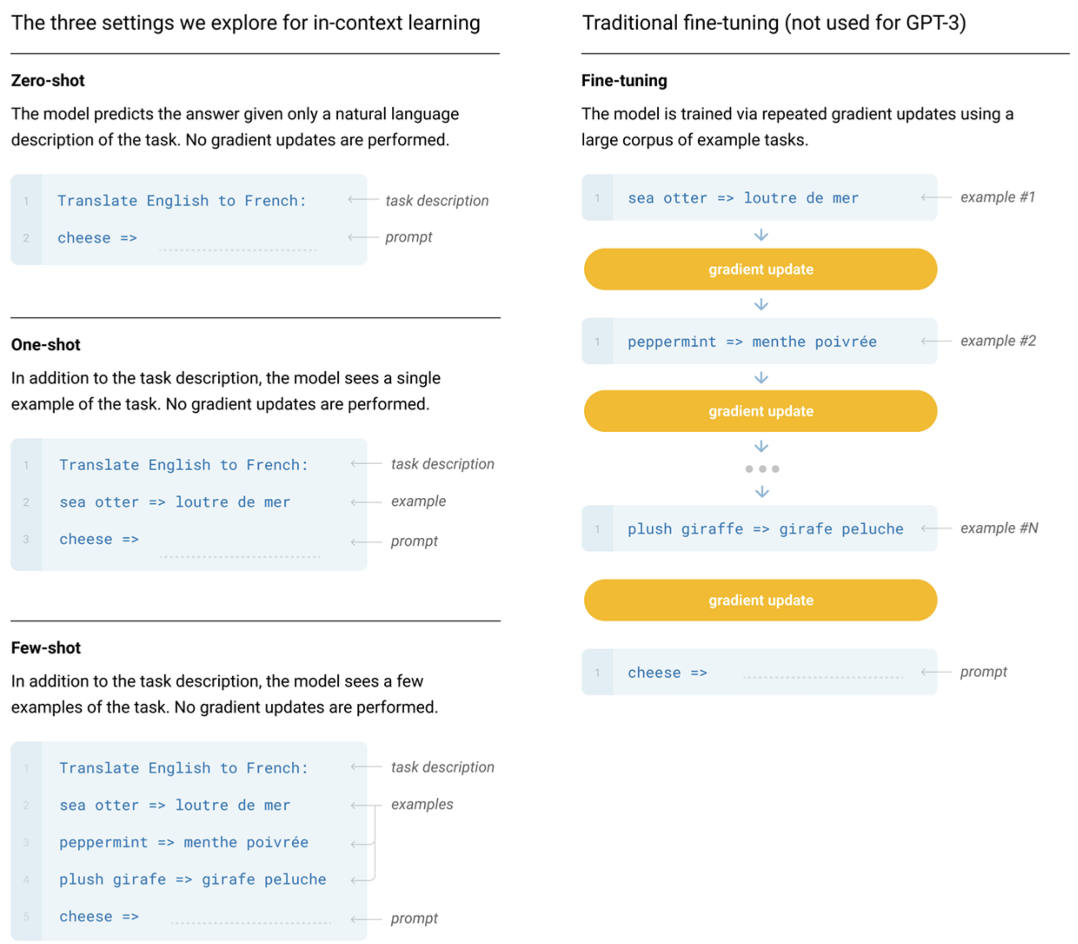


https://arxiv.org/pdf/2005.14165
https://arxiv.org/pdf/2302.13971

上下文学习 In-context Learning

ICL Prompt include:

- Task description
- Examples
- Prompt



https://arxiv.org/pdf/2005.14165
https://arxiv.org/pdf/2302.13971

上下文学习 In-context Learning

- ICL 是一种当模型scale up足够的时候才激发的能力

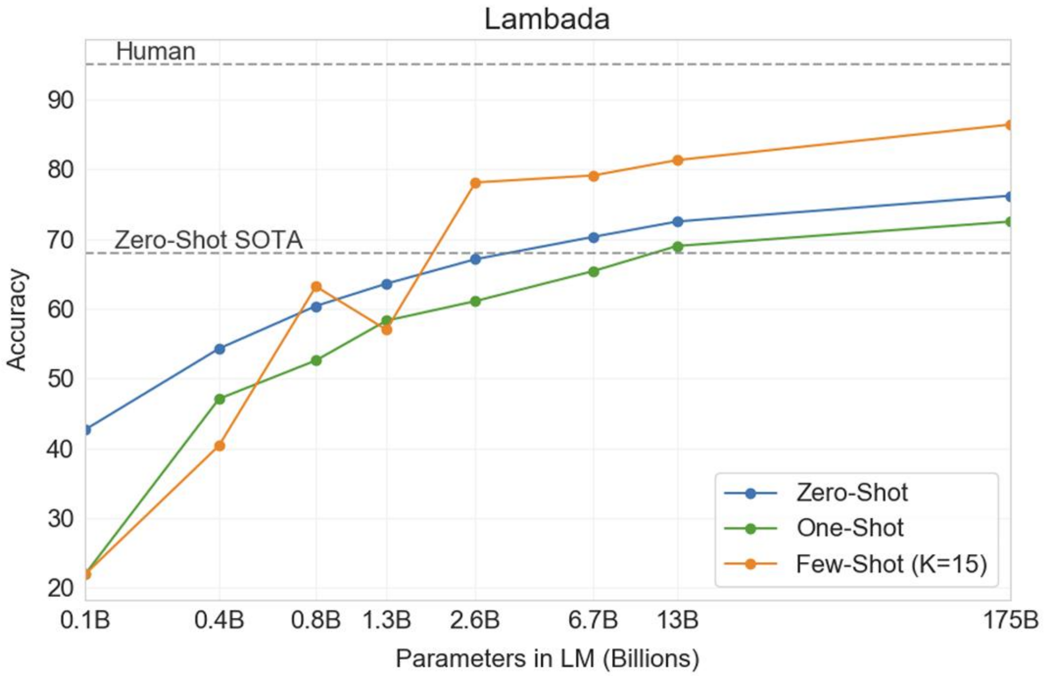


Figure 3.2: On LAMBADA, the few-shot capability of language models results in a strong boost to accuracy. GPT-3 2.7B outperforms the SOTA 17B parameter Turing-NLG [Tur20] in this setting, and GPT-3 175B advances the state of the art by 18%. Note zero-shot uses a different format from one-shot and few-shot as described in the text.

https://arxiv.org/pdf/2005.14165
https://arxiv.org/pdf/2302.13971

ICL 为何有作用？

- ICL example的输入格式 (x 的分布) 对性能也有帮助
  - 即使我们使用随机标签，也能做ICL

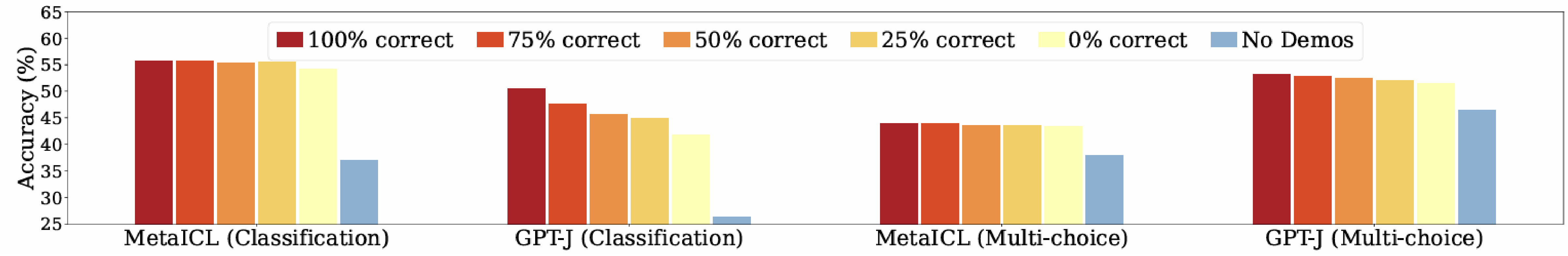


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

https://arxiv.org/abs/2202.12837

ICL 为何有作用？

- ICL example 的示例顺序会显著影响模型结果

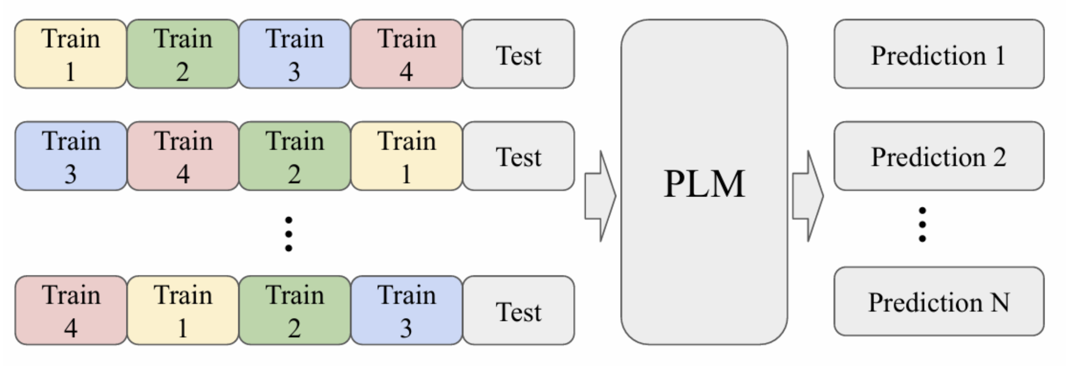


Figure 2: Training sample permutations for the In-context Learning setting. The concatenation of training samples as well as test data transforms the classification task into a sequence generation task.

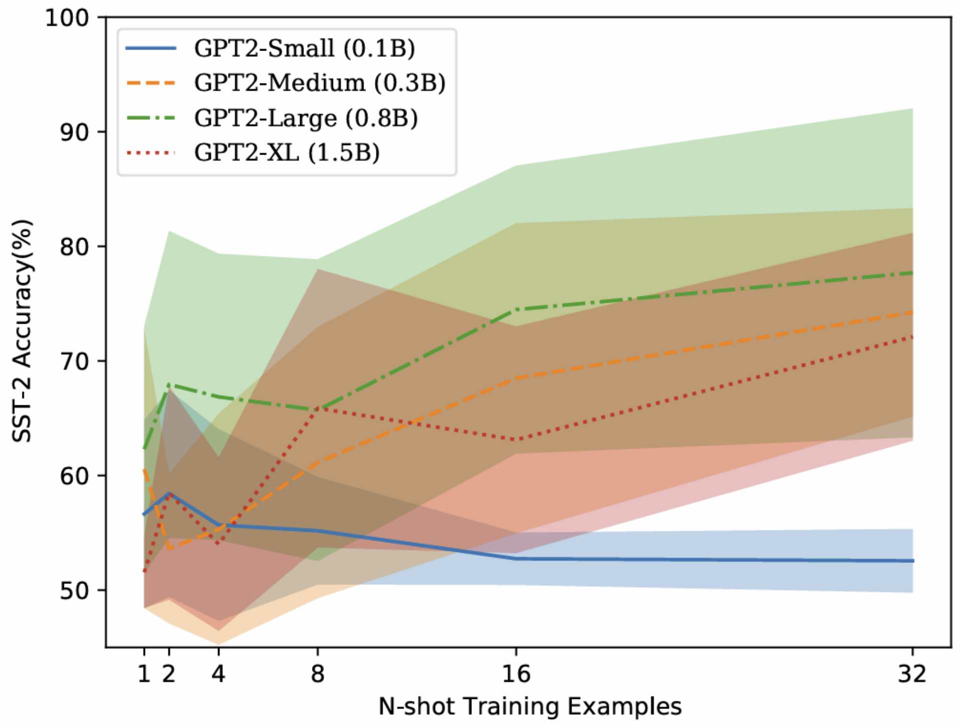


Figure 3: Order sensitivity using different numbers of training samples.

https://arxiv.org/pdf/2104.08786

ICL 为何有作用？

- ICL example 的示例挑选会显著影响模型结果
  - 在一个大数据集中选择与test sample比较近的样本能够有效提升效果

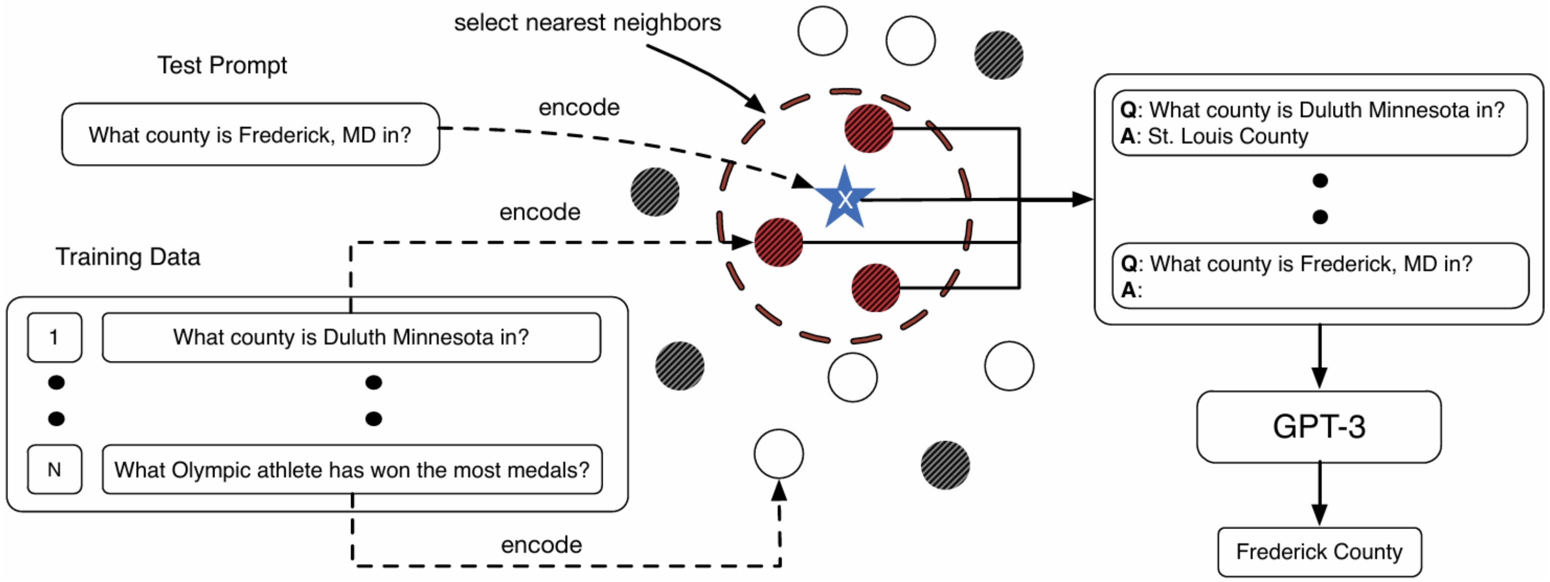


Figure 2: In-context example selection for GPT-3. White dots: unused training samples; grey dots: randomly sampled training samples; red dots: training samples selected by the k-nearest neighbors algorithm in the embedding space of a sentence encoder.

我们永远可以使用ICL ... 吗？

- ICL的好处在于，当我们的算例和样例都不多，ICL能够在无需调整模型参数的情况下快速提升模型效果
- 然而，如果我们的样例足够多，ICL 每次需要消耗的资源将会太大
- 此时，我们还是需要调整模型参数，这就引出了finetuning

微调 Finetuning

- 通过训练永久改变原始模型
- 原始方法：全参数微调
  - 更新模型中的所有参数
  - 潜在性能高
  - 成本极其高昂，容易发生灾难性遗忘

微调 Finetuning

- 当前方法：高效参数微调 (PEFT)
    - Adapter: 冻结主模型，在原有transformer层之间插入小网络层进行训练
    - Prefix-Tuning: 冻结主模型，在prompt前加入少许可训练的虚拟提示
    - LoRA / QLoRA: 冻结主模型，通过训练低秩矩阵模拟权重变化（***）

LoRA

- Motivation: 预训练模型的权重是过参数的，而微调的变化量往往具有low-rank的特性
- 因此，我们冻结原有参数W，并引入两个low-rank矩阵A, B模拟参数变化
- 由于 $rank(AB) \leq rank(A), rank(B)$，当控制A,B的形状，其乘积的rank仍然可控
- 参数量大大降低！

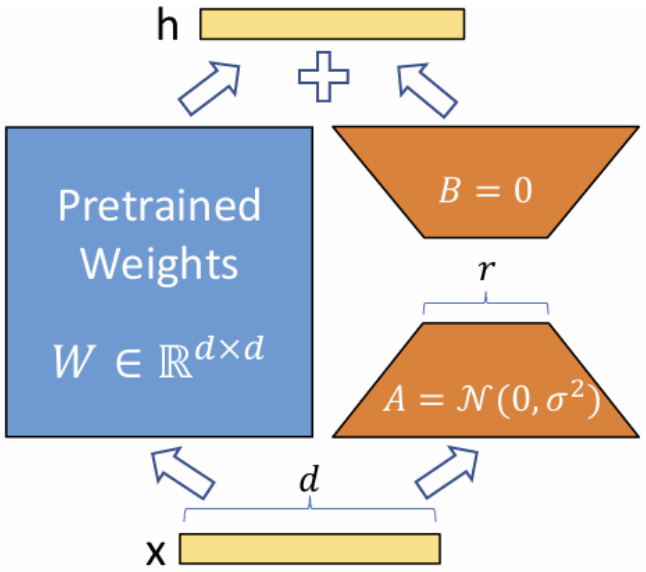


Figure 1: Our reparametrization. We only train A and B.

LoRA

- 优点
    - 可并行计算
    - 显存进一步降低
    - LoRA 和 QLoRA 是当前大模型中最常用的微调技术
    - 如何部署微调？
        - 使用 openai 的工具进行微调 [1]
        - 对Qwen进行微调 [2]

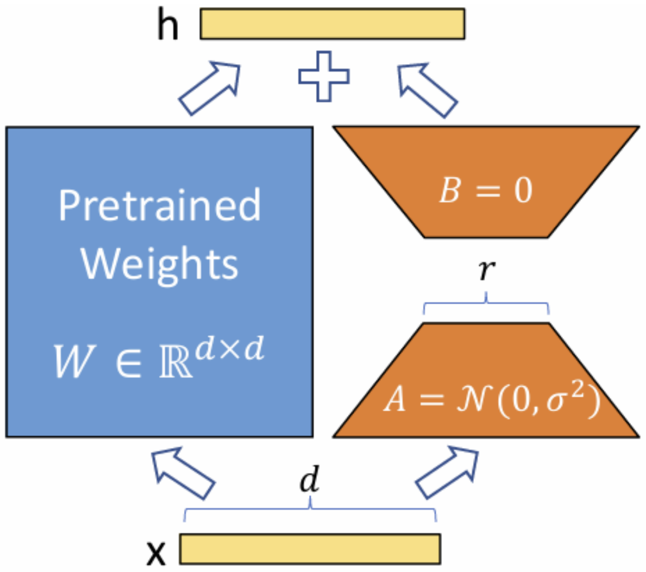


Figure 1: Our reparametrization. We only train A and B.

[1] https://www.youtube.com/watch?v=UJ7ry7Qp2Js

[2] https://www.bilibili.com/video/BV1S79vYdENT/

总结 Take-away Messages

In-context Learning

- 影响因素：示例顺序、示例分布
- 非影响因素：类别准确性

Fine-tuning

- 使用低秩技术进行快速微调

第七次作业（二选一）：

(1) 改变in-context samples的顺序，看看现在的大模型对这件事还是否敏感

(2) 基于前面的视频内容，微调一个1-7B的模型