# 实验4：LLaMA2模型的加载与SFT

# 实验背景

1.1实验数据

Instruction（从OpenAssistant Conversations Dataset (OASST1)数据集中抽取的1000个样例）

1.2实验环境

Google Colab，T4 GPU

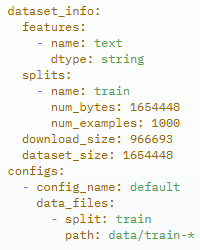
1.3实验内容

使⽤Transformers库和Google colab加载、训练和测试LLaMA2-7B-chat模型；构造⼀些prompt测试样例，测试微调后的模型在问答、推理等⽅⾯的能⼒。

1.4数据分析

OpenAssistant Conversations Dataset (OASST1)是⼀个由⼈⼯⽣成、⼈⼯标注的助⼿式对话语料库，包含35种不同语⾔的161,443条信息，标注了461,292个质量评级，形成了超过10,000个完全标注的对话树。该语料库是全球众包

⼯作的产物，有超过13,500名志愿者参与。



（1）超过1000条消息的语言

英语：71956 西班牙语：43061 俄语：9089 德语：5279

中文：4962 法语：4251 泰语：3042 葡萄牙语（巴西）：2969

加泰罗尼亚语：2260 韩语：1553 乌克兰语：1352 意大利语：1320

日语：1018

（2）少于1000条消息的语言

越南语：952 巴斯克语：947 波兰语：886 匈牙利语：811

阿拉伯语：666 荷兰语：628 瑞典语：512 土耳其语：454

芬兰语：386 捷克语：372 丹麦语：358 加利西亚语：339

希伯来语：255 罗马尼亚语：200 挪威博克马尔语：133

印尼语：115 保加利亚语：95 孟加拉语：82 波斯语：72

希腊语：66 世界议会语：59 斯洛伐克语：19

# 实验方案设计

2.1预处理数据集

从OpenAssistant Conversations Dataset (OASST1)数据集中抽取1000个样例，名称为”mlabonne/guanaco-llama2-1k”，使用时直接调用即可，dataset = load\_dataset(dataset\_name, split="train")。

2.2算力获取

在Google Colab中申请分配T4 GPU资源。

2.3加载LLaMA2模型

使用Transformers库的AutoModelForCausalLM类调用和加载LLaMA2模型。通过from\_pretrained方法，可以方便地加载预训练的模型及其权重，无需手动指定模型的具体实现类。

2.4 LLaMA2模型和LoRA训练方法的基本原理和优势

2.4.1 LLaMA2模型

（1）基本原理

LLaMA2是一种基于Transformer架构的先进语言模型，其核心原理主要基于自注意力机制，这是由Vaswani等人在2017年提出的神经网络模型。自注意力机制能够捕捉输入序列中的上下文关系，提高模型对复杂语言模式和语义关系的理解能力。LLaMA2进一步采用了多头注意力机制和编码器-解码器架构，以增强模型的表达能力和处理长距离依赖问题的能力。

在LLaMA2中，输入文本首先通过嵌入层转换为高维向量表示，然后这些向量被送入编码器进行进一步处理。编码器由多个编码器层组成，每个编码器层包含多头自注意力机制和前馈神经网络，用于提取文本中的关键信息和上下文关系。解码器则根据编码器的输出和之前的输出序列，逐步生成新的文本。

（2）优势

长序列文本：基于Transformer架构，LLaMA2能够**处理长序列文本**，这在处理长文档或对话时尤为重要。

无监督学习：在预训练阶段，LLaMA2通过**无监督学习**从海量文本数据中抽取知识，这种学习方式提高了模型的泛化能力，使其能够处理各种复杂的语言现象。相较于LLaMA1，LLaMA2的预训练语料增加了40%，模型能够在更丰富的语境中学习。

多任务&跨语言：LLaMA2支持**多任务**预训练方法，使得模型能够在不同任务之间共享知识，提高性能；同时，它还支持**多语言**处理，为跨语言自然语言处理任务提供了强大的支持。

稳定高效：LLaMA2引入了多种技术优化，如**残差连接、层归一化**等，这些技术不仅提高了模型的稳定性，还使得训练过程更加高效。

2.4.2 LoRA训练方法

（1）基本原理

LoRA（Low-Rank Adaptation of Large Language Models）是一种用于微调大型语言模型的参数高效性微调方法。LoRA的基本原理是冻结预训练好的模型权重参数，在冻结原模型参数的情况下，通过往模型中加入额外的网络层（即低秩矩阵），并只训练这些新增的网络层参数。由于新增参数数量较少，LoRA能够显著降低计算复杂度和内存使用，同时获得与全模型微调类似的效果。

（2）优势

轻量化&高效性：LoRA通过训练少量的新增参数，显著降低了**计算复杂度和内存使用**，使得在消费级GPU上训练大型模型成为可能。

泛化能力：通过限制模型的复杂性，LoRA有助于防止训练数据有限的情况下的**过拟合**现象，使得模型**泛化能力**较强，在处理新数据时表现出色。

高迁移性：LoRA可以**无缝集成**到现有的神经网络架构中，以最小的额外训练成本对预训练模型进行微调和适应，非常适合迁移学习应用。

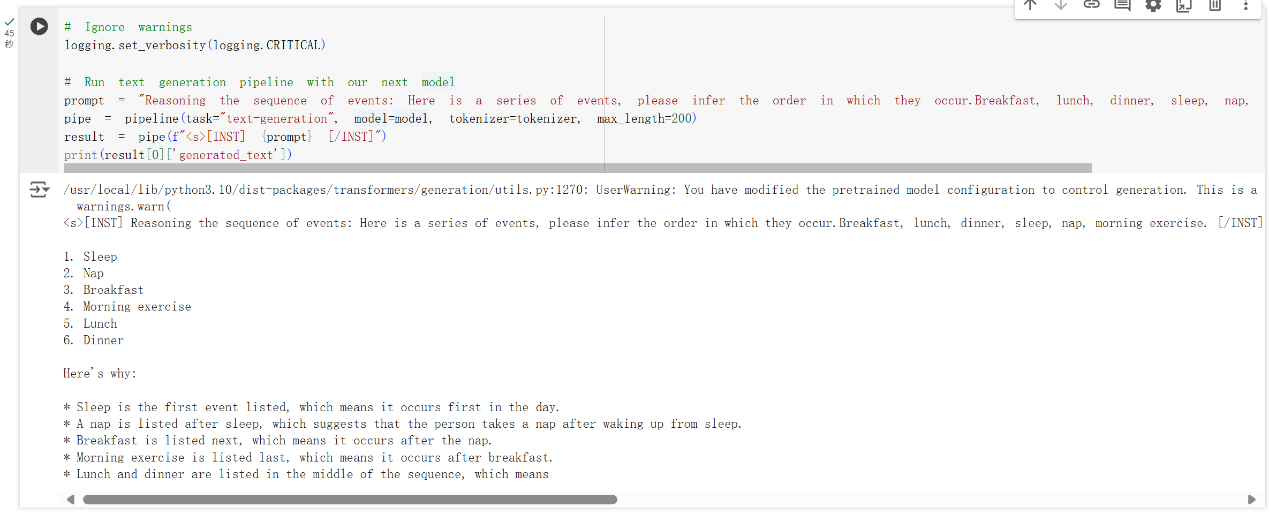
低成本：由于只训练少量参数，LoRA显著降低了finetune的成本，使得模型调整更加**经济高效**。

# 实验过程

按照设置好的参数进行训练，在训练过程中，loss值先迅速下降，之后缓慢下降至趋于平稳。

# 性能测试

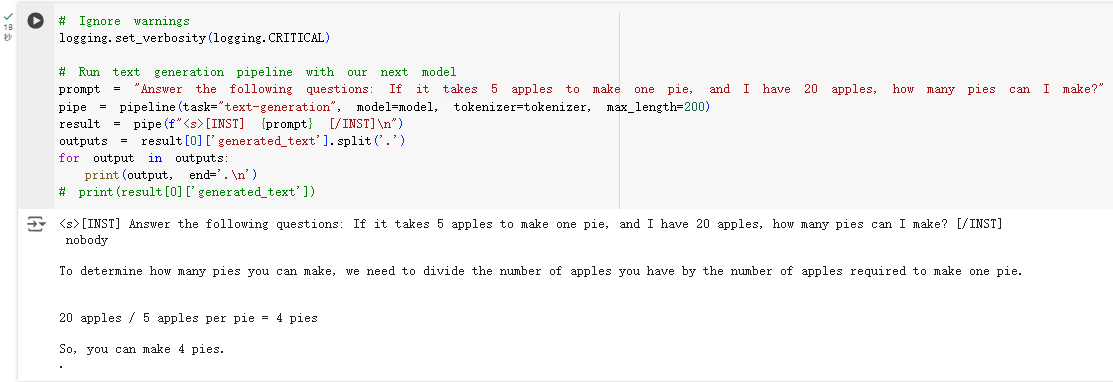
模型自带默认测试用例输出如下所示：



编写测试用例，从准确性、逻辑性、创造性、实用性、综合流畅性等方面考察模型的性能，结果显示模型表现较好，与期望输出较为一致。

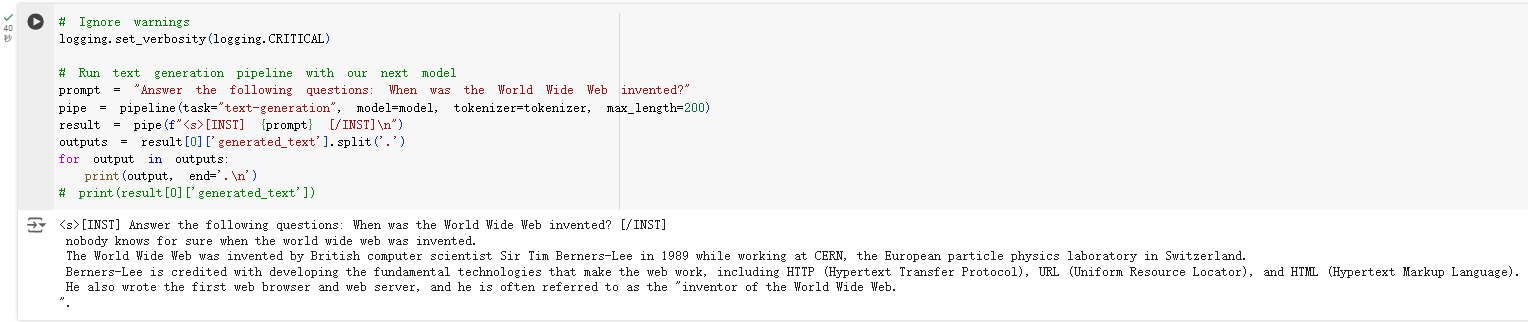
（1）准确性：验证模型提供的信息是否准确，特别是关于历史、科学、技术等领域的事实。



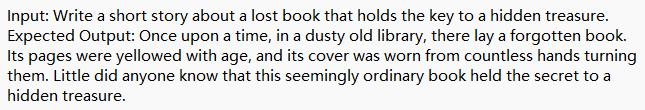


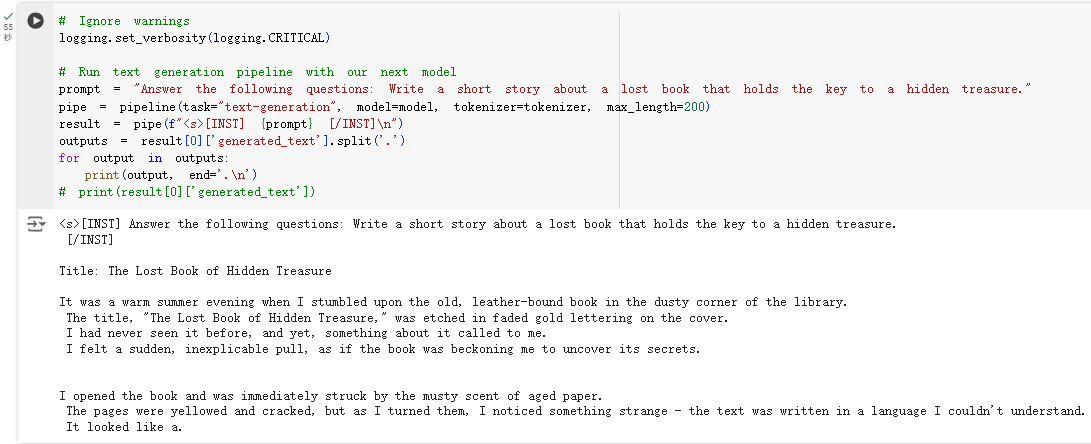
（2）逻辑性：测试模型在日常生活场景中的常识推理和逻辑能力。



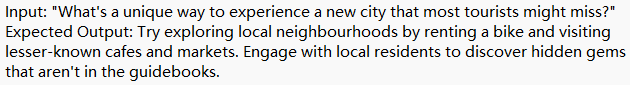


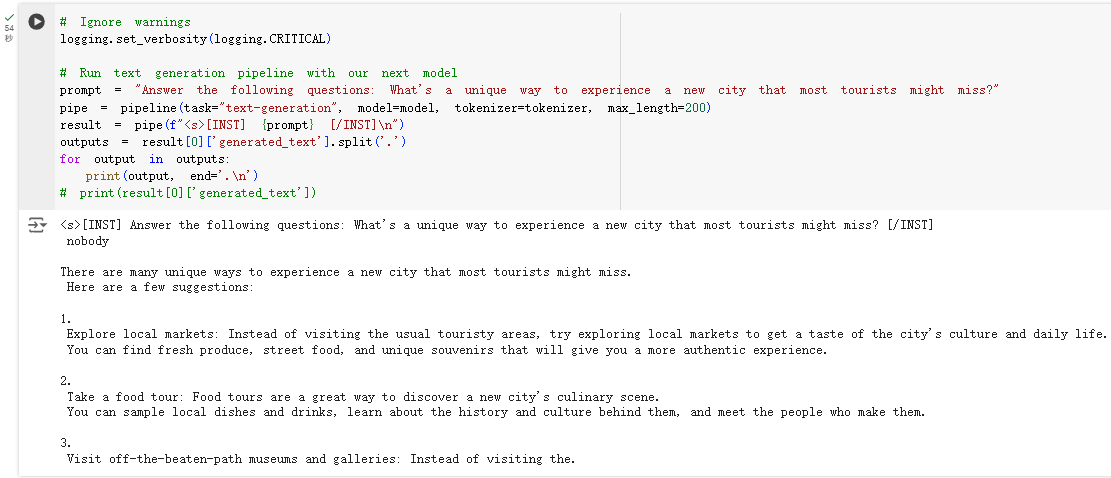
（3）创造性：测试模型在创作（如写作、编程）方面的创造能力，并评估模型生成内容的新颖程度。

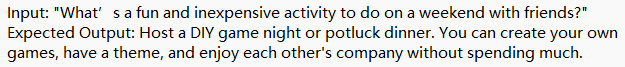


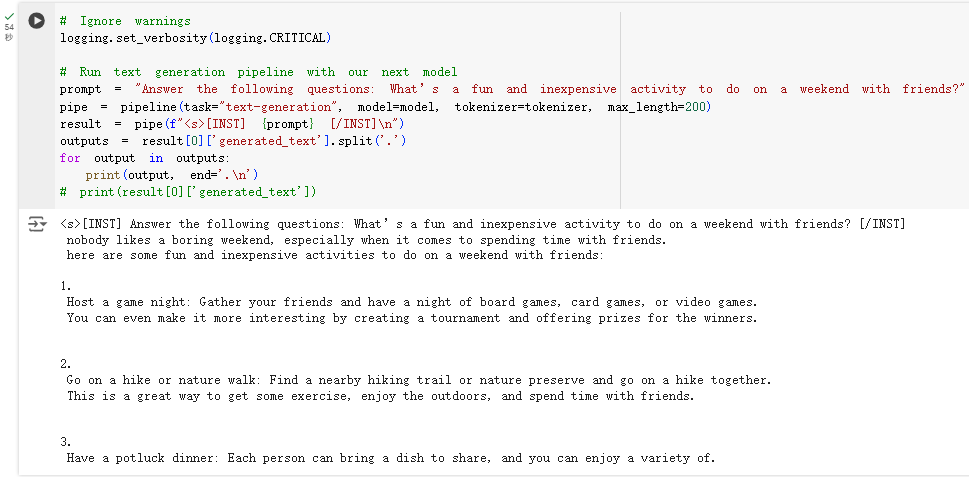


（4）实用性：在实际应用中的价值和有效性，可以解决生活中实际问题。

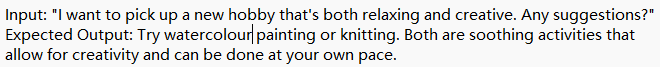


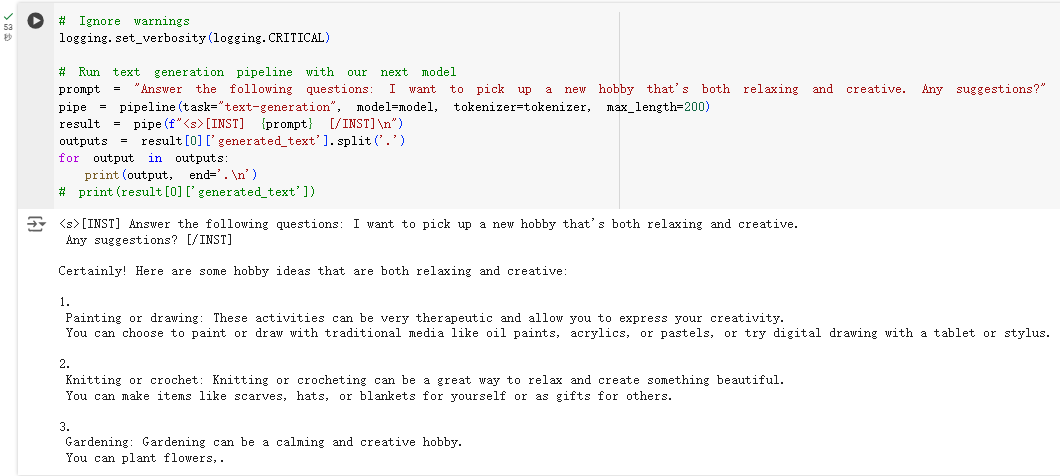






（5）综合流畅性：回答的流畅性和连贯性，能够准确理解输入的文本。





# 实验习题

1. 简述LoRA训练方法的优势与劣势
2. 优势：

轻量化&高效性：LoRA通过训练少量的新增参数，显著降低了**计算复杂度和内存使用**，使得在消费级GPU上训练大型模型成为可能。

泛化能力：通过限制模型的复杂性，LoRA有助于防止训练数据有限的情况下的**过拟合**现象，使得模型**泛化能力**较强，在处理新数据时表现出色。

高迁移性：LoRA可以**无缝集成**到现有的神经网络架构中，以最小的额外训练成本对预训练模型进行微调和适应，非常适合迁移学习应用。

1. 劣势：

数据需求量大：尽管LoRA在参数效率上有所优势，但它仍然需要一定数量的训练数据来确保模型的有效性和泛化能力，在训练数据非常有限的情况下，LoRA的性能可能会受到影响。

效果可能略逊于全参数微调：由于LoRA只训练了少量的新增参数，因此在某些复杂任务上，其性能可能略逊于全参数微调的方法。特别是在需要精细调整模型参数的场景下，LoRA可能无法达到最优效果。

模型结构难以调整：LoRA主要关注于参数的高效微调，而不涉及模型结构的调整。因此，LoRA可能不适合某些需要改变模型结构的情况

1. 简单说明对大语言模型进行SFT有何作用

（1）提升模型在特定任务上的性能

SFT通过在预训练模型的基础上，使用标注好的数据进行进一步训练，使模型能够针对特定任务或领域进行优化。这有助于模型更好地理解任务需求，提高在特定数据集上的表现，从而在各种应用场景中取得更好的效果。

（2）适应不同场景和需求

大语言模型在预训练阶段通常学习的是通用知识，而不同的应用场景和任务需求往往具有特定的数据分布和特征。通过SFT，模型可以针对特定场景和需求进行微调，从而更好地适应这些场景下的数据分布和任务要求。

（3）高效利用预训练知识

SFT是一种高效的模型训练方法，它利用预训练模型已经学习到的通用知识，通过微调快速适应特定任务，不仅能减少训练时间和计算资源消耗，还可以在一定程度上避免从头开始训练模型可能带来的过拟合问题。

（4）灵活性和可扩展性

SFT方法具有灵活性和可扩展性，随着新的应用场景和任务需求的出现，可以通过收集新的标注数据并对模型进行再次微调，使模型能够适应新的任务。这种灵活性使得大语言模型能够持续不断地学习和进化，以适应不断变化的环境和需求。