LLM-Adapters：大语言模型参数高效微调的适配器系列

Zhiqiang Hu[1] **Lei Wang**[2] * **YihuaiLan Wanyu Xu**[4] **Ee-Peng Lim**[2]
Lidong Bing[3] **Xing Xu**[5] **SoujanyaPoria**[1] **Roy Ka-Wei Lee**[11]新加坡科技设计大学[2]新加坡管理大学[3]达摩院，阿里巴巴集团[4]西南交通大学[5]中国电子科技大学

arXiv:2304.01933v3 [cs.CL] 9 Oct 2023

摘要

大语言模型（LLM），如 GPT-4 和 ChatGPT 的成功，推动了众多经济实惠且易于获取的替代方案的发展。这些方案通过使用特定任务数据（例如 ChatDoctor）或指令数据（例如 Alpaca）对开放获取大语言模型进行微调而创建。在各种微调方法中，基于适配器的参数高效微调（PEFT）无疑是最具吸引力的研究主题之一，因为它只需微调少量外部参数，而非整个大语言模型，即可取得相当甚至更好的性能。为促进对大语言模型参数高效微调方法的进一步研究，本文提出了 LLM-Adapters，这是一个易于使用的框架，可将各种适配器集成到大语言模型中，并针对不同任务执行这些基于适配器的参数高效微调方法。该框架包含 LLaMA、BLOOM 和 GPT-J 等先进的开放获取大语言模型，以及串行适配器、并行适配器、基于提示的学习和基于重参数化的方法等广泛使用的适配器。此外，我们通过大量实证研究，探讨了适配器类型、放置位置和超参数对各类基于适配器的方法最佳设计的影响。我们在来自两种不同推理任务（算术推理和常识推理）的十四个数据集上评估了适配器的有效性。结果表明，在较小规模的大语言模型（7B）中使用基于适配器的参数高效微调，只需少量额外可训练参数，在两种推理任务的零样本推理中即可达到与强大的大语言模型（175B）相当、某些情况下甚至更优的性能。

1 引言

大型语言模型（LLM），例如 Chat- GPT（OpenAI，2022年）和 GPT-4（OpenAI，2023年），在各种自然语言处理（NLP）任务中展现出前所未有的性能（Qin 等人，2023年）和多模态任务（Shen 等人，2023年）。这些大语言模型通常拥有数千亿个参数，且为闭源模型。因此，推动了人们开发易于获取且经济高效的替代方案，例如 LLaMA（Touvron 等人，2023年）。这些替代方案通过特定任务数据（例如 Chat-Doctor（Yunxiang 等人，2023年））或指令数据（例如 Alpaca（Taori 等人，2023年））对开源大语言模型进行微调。然而，全模型微调（FFT）在计算和存储方面开销巨大，因而在实际实施中面临重大挑战。

在大语言模型（例如 LLaMA）的全模型微调兴起之前，自然语言处理领域已提出一种颇具吸引力的方案，称为参数高效微调（PEFT）（Houlsby 等，2019），专门用于预训练模型（例如 BERT（Devlin 等，2018）），为高效微调大语言模型提供了一种有前景的方法。PEFT 的优势在于，只需微调整少量外部参数，而非整个骨干模型，同时仍能取得相当甚至更优的性能（Mangrulkar 等，2022）。此外，与全模型微调相比，PEFT 能有效缓解灾难性遗忘（Wang 等，2022）。如表 1 所示，PEFT 的优势推动了多种参数高效微调模块的发展，包括串行适配器（Houlsby 等，2019；Wang 等，2022；He 等，2022b；Fu 等，2021）、并行适配器（He 等，2022a）、基于重参数化的方法（Hu 等，2021；Edalati 等，2022）以及基于提示的学习方法（Lester 等，2021；Li 和 Liang，2021）。

通过将这些参数高效微调模块集成到骨干模型（即大语言模型）中，我们无需投入大量计算资源，便能充分利用骨干模型的卓越能力。这为更广泛的应用创造了机会，使得计算资源有限的人也能获

*通讯作者。

得高性能计算资源，从而在特定任务中发挥大语言模型的强大能力。尽管参数高效微调已成功应用于预训练模型，但对于特定任务或数据集，哪种参数高效微调模块与哪一层及超参数配置相结合最为合适，仍不明确，尤其是在使用大语言模型（例如 LLaMA（Touvron 等人，2023年））时。因此，仍需进一步研究，以确定能够在不同任务和数据集上实现性能最大化的最佳参数高效微调方案。

受此启发，本文对三种具有代表性的开源大语言模型的参数高效微调进行了全面的实证研究，包括 BLOOM（Muennighoff 等人，2022年）、GPT-J（Wang 和 Komatsuzaki，2021年）和 LLaMA（Touvron 等人，2023年）。具体而言，我们开展了实证研究，以回答以下三个研究问题：（i）不同参数高效微调方法的最佳放置位置和配置是什么？（ii）不同适配器在各项下游任务中的表现如何？以及（iii）对于参数高效微调方法，分布内（ID）场景与分布外（OOD）场景之间的性能差异是什么？我们的研究发现如下：

1.**串行适配器、并行适配器和 LoRA 的最佳放置位置分别是多层感知机层之后、与多层感知机层并行，以及同时位于注意力层和多层感知机层之后**；

2. **采用参数高效微调方法的较小语言模型，在特定任务上能够达到与较大语言模型相当甚至更优的性能。例如，在 MultiArith、AddSub 和 SingleEq 上，采用 LoRA 的 LLaMA-13B 可以优于 GPT-3.5（>175B）**；

3.**采用适配器微调的 ID LLaMA-13B 在常识推理任务上的表现优于 ChatGPT，这表明，使用 ID 微调数据时，较小语言模型有潜力在特定任务上超越较大语言模型**。

我们的贡献可概括如下：

• 我们对应用于不同开源大语言模型的各种参数高效微调方法进行了全面的实证研究。

• 为便于开展实证研究，我们构建了两个高质量训练数据集，以提升数学推理和常识推理任务中的参数高效微调性能。

•我们开发了一个易于使用的框架 LLM-Adapter，它能够将各种适配器无缝集成到大语言模型中，使研究人员能够针对广泛的任务实现基于适配器的参数高效微调方法。

• 我们开展了广泛的实验来回答这三个研究问题，以期为未来的研究提供启发。

方法	提示词修复	串行	并行
提示调优（Lester 等人，2021年）	√		
前缀调优（Li 和 Liang，2021年）	√		
SPOT（Vu 等人，2021年）	√		
IPT（Qin 等人，2021年）	√		
LoRA（Hu 等人，2021年）		√	
KronA（Edalati 等人，2022年）		√	
适配器（Houlsby 等人，2019年）			√
AdaMix（Wang 等人，2022年）			√
SparseAdapter（He 等人，2022b）			√
LeTS（Fu 等人，2021年）			√
并行适配器（He 等人，2022a）			√
MAM 适配器（He 等人，2021年）	√	√	√
UniPELT（Mao 等人，2021年）	√	√	√
Compacter（Henderson 等人，2021年）		√	√
S4-model（Chen 等人，2023年）	√	√	

表 1：参数高效微调方法根据四种常见的基本方法进行分类。“Prompt”表示基于提示的学习方法，“Repara”表示基于重参数化的方法，“Series”表示串行适配器，“Parallel”表示并行适配器。

2 参数高效微调概述

本节简要概述四种参数高效微调（PEFT）方法：基于提示的学习、基于重参数化的方法、串行适配器和并行适配器。（Li 和 Liang，2021年；Houlsby 等人，2019年；Houlsby 等人，2019年；He 等人，2022a）

基于提示的学习。 如图 1(a)所示，基于提示的学习将寻找最优硬提示这一离散优化问题转化为连续（软）提示问题。为此，Lester 等人 (2021年)提出了提示调优的概念，即将一个可训练张量作为前缀添加到输入嵌入中。另一种称为前缀调优的方法（Li 和 Liang，2021年）则独立探索了将软提示添加到所有层的隐藏状态中。内在提示调优（Qin 等人，2021年）使用自编码器对软提示进行压缩和解压缩。我们以融入注意力层的可学习向量为例，说明基于提示的学习，其形式化表达如下：

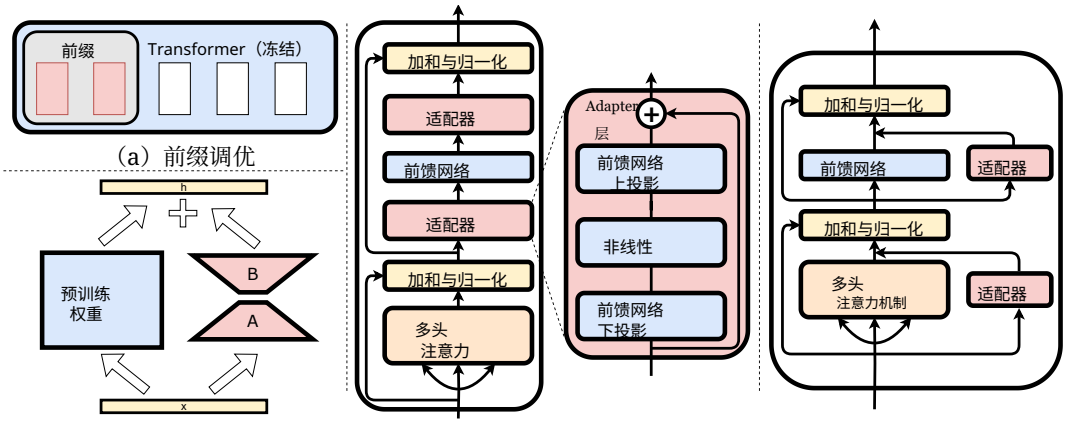

图 1：三种不同适配器的模型架构详细示意图：（a）前缀调优，（b）LoRA，（c）串行适配器，以及（d）并行适配器。

其中：

$$H_o = \text{Attn}(H_i W_Q, [P_K; H_i W_K], [P_V; H_i W_V]), \tag{1}$$

其中，$H_i \in \mathrm{R}^{T\times d}$和 $H_o \in \mathrm{R}^{T\times d}$分别是注意力层的输入和输出。请注意，$T$ 是最大输入长度，d 是向量维度。$P_K \in \mathrm{R}^{L\times d}$ 和 $P_V \in \mathrm{R}^{L\times d}$是参数高效微调的可学习向量。$L$ 是可学习标记的数量，实验章节中将对此进行详细讨论。Q，K，V分别表示注意力模块的查询向量、键向量和值向量。

基于重参数化的方法。这类方法旨在利用低秩技术转换网络权重。这种方法在保留处理高维矩阵能力的同时，有效减少了可训练参数的数量。Intrinsic SAID（Aghajanyan 等人，2020年）研究了低秩子空间中微调的内在维度。LoRA（Hu 等人，2021年）提出了一种简单的方法，将权重矩阵分解为两个低秩矩阵的乘积，从而更新其参数。KronA（Edalati 等人，2022年）利用克罗内克积改进了 LoRA 的矩阵分解方法。我们以 LoRA 为例，说明基于重参数化的学习，其形式化表示如下：

$$H_o = H_i W_0 + H_i \Delta W = H_i W_0 + H_i B A, \tag{2}$$

其中，$W_0 \in \mathrm{R}^{d\times d}$ 可以是任意预训练权重矩阵，包括多层感知机或注意力层中的权重。$B \in \mathrm{R}^{r\times d}$ 和 $A \in \mathrm{R}^{r\times d}$ 是低秩矩阵，旨在覆盖 ΔW。$r \ll d$ 是 LoRA 的一个重要超参数。

串行适配器。串行适配器会在特定子层中按顺序加入额外的可学习模块。在他们的研究中，Houlsby 等人（2019年）提出在 Transformer 模型（Vaswani 等人，2017年）的注意力层和前馈网络层之后集成全连接网络。另一项研究发现，Pfeiffer 等人（2020年）发现，只在自注意力层之后插入适配器，而非在每个 Transformer 模块中使用两个适配器，也能取得相当的性能。AdaMix（Wang 等人，2022年）提出了一种以混合专家（MoE）方式使用多个串行适配器的方法。Compacter（Henderson 等人，2021年）利用克罗内克积、低秩矩阵以及跨层参数共享来生成适配器权重。这种技术旨在降低与适配器相关的计算复杂度，同时保持其性能。串行适配器可形式化表示如下：

$$H_o \leftarrow H_o + f(H_o W_{down}) W_{up}, \tag{3}$$

其中，特定层（例如多层感知机层）的输出 H_o 首先通过 $W_{down} \in \mathrm{R}^{d\times r}$ 下投影至较低维度 r，然后再通过 $W_{up} \in \mathrm{R}^{r\times d}$ 上投影回原始维度 d。f 是一个非线性函数。我们将在实验章节中讨论 r 的选择。

并行适配器。并行适配器（He 等人，2022a）旨在将额外的可学习模块与骨干模型中的不同子层并行结合。并行适配器可以是

如下所述：

$$H_o \leftarrow H_o + f(H_i W_{down}) W_{up}, \qquad (4)$$

其中 H_i （H_o）分别是特定层的输入（输出）。在此概念基础上，多头并行适配器更进一步，使用并行适配器来修改注意力头的输出。另一方面，缩放并行适配器是一种变体，它将 LoRA（Hu 等人，2021年）的组合与插入形式应用于适配器。另一种称为梯形侧向微调（Sung 等人，2022年）的方法涉及训练一个轻量级梯形侧网络。该网络通过快捷连接（梯子）接收来自骨干网络的中间激活值。

3 实验设置

3.1 基准测试

我们在两类推理问题的14个基准数据集上进行了广泛的实证研究：算术推理：（1）GSM8K（Cobbe 等人，2021年）数据集包含由人工命题者编写、语言表达多样且质量较高的小学数学文字题；（2）SVAMP（Patel 等人，2021年）基准数据集通过对另一个现有数据集中的一组问题进行简单修改，构建了适用于最高四年级学生的单未知数算术文字题；（3）MultiArith（Roy 和 Roth，2016年）数据集包含需要多个推理步骤和运算的数学文字题；（4）AddSub（Hosseini 等人，2014年）数据集包含加法和减法算术文字题；（5）AQuA（Ling 等人，2017年）数据集包含带有自然语言推理过程的代数文字题；（6）SingleEq（Koncel-Kedziorski 等人，2015年）数据集包含可映射为单个方程、长度各异的小学代数文字题。常识推理：（1）BoolQ（Clark 等人，2019年）数据集是一个用于是非题的问答数据集，包含15942个示例。这些问题源自自然场景，并在无提示、无约束的环境中生成；（2）PIQA（Bisk 等人，2020年）数据集包含每题有两种解决方案、需要运用物理常识作答的问题；（3）SIQA（Sap 等人，2019年）侧重于推理人们的行为及其社会影响；（4）HellaSwag 数据集包含常识自然语言推理问题，每题包括一个语境和若干用于补全语境的结尾选项；（5）WinoGrande（Sakaguchi

数据集	领域	# 训练	# 测试	答案
MultiArith	Math	-	600	数字
AddSub	Math	-	395	数字
GSM8K	Math	8.8K	1,319	数字
AQuA	Math	100K	254	选项
SingleEq	Math	-	508	数字
SVAMP	Math	-	1,000	数字
BoolQ	CS	9.4K	3,270	是/否
PIQA	CS	16.1K	1,830	选项
SIQA	CS	33.4K	1,954	选项
HellaSwag	CS	39.9K	10,042	选项
WinoGrande	CS	63.2K	1,267	选项
ARC-e	CS	1.1K	2,376	选项
ARC-c	CS	2.3K	1,172	选项
OBQA	CS	5.0K	500	选项

表 2：评估所用数据集的详情。数学：算术推理。常识：常识推理。

等人，2021年）数据集采用填空任务的形式，并提供二元选项；目标是为给定句子选出正确选项，这需要运用常识推理；（6）ARC-c 和（7）ARC-e 分别是 ARC（Clark 等人，2018年）数据集中具有挑战性的集合和简单集合，包含真实的小学水平多项选择科学问题；（8）OBQA 数据集包含需要多步推理、运用额外常识与常识知识，以及较强文本理解能力的问题。表2 展示了数据集统计信息。

3.2 微调数据收集

为了对适配器进行微调，我们获取了两个专为数学推理和常识推理设计的高质量训练数据集。表2 显示，只有 GSM8K 和 AQuA 数据集提供了算术推理训练集。为增加数据的多样性，我们纳入 GSM8K、MAWPS、MAWPS-single（Koncel- Kedziorski 等人，2016年）的训练集，并从 AQuA 中选取 1000 个样本用于收集微调数据。不过，值得注意的是，这些数据集仅提供方程及其对应答案。为了增强模型的推理能力，尤其是提供逐步推理依据的能力，我们使用 ChatGPT 作为教师模型。通过采用零样本思维链提示，ChatGPT 生成推理步骤。我们在附录 A.1中列出了用于收集数学推理数据集的具体提示模板。为确保数据质量，我们剔除了包含错误答案的样本。最终，我们获得了 10K 个数学推理样本，称为 Math10K，并将其用于后续分析和微

调。

为便于在常识推理领域进行微调，我们将 BoolQ 、 PIQA 、 SIQA 、 HellaSwag 、 WinoGrande、ARC-e、ARC-c 和 OBQA 的训练集套用预定义模板进行格式化，以此构建微调数据。由于常识推理领域中的每个数据集都涉及不同的任务，我们采用结构化模板，先描述任务目标，再呈现相应内容和答案。用于创建微调数据的模板见 A.2。完成这一过程后，我们获得了包含 170K 个常识推理样本的数据集，并将其称为 Common-sense170K。我们将公开这些数据集，以鼓励在该领域开展进一步研究与探索。

3.3 实现

为便于在研究和实际应用中无缝使用参数高效微调方法，我们开发了一个易于使用的框架 LLM-Adapter。LLM-Adapters 将多种适配器无缝集成到大语言模型中，使研究人员能够针对广泛的任务实现基于适配器的参数高效微调方法。我们的实验采用 LLaMA（7B、13B）（Touvron 等人，2023年）、BLOOMz（7B）（Muen-nighoff 等人，2022年）和 GPT-J（6B）（Wang 和 Komatsuzaki，2021年）作为基础模型。对于四类参数高效微调方法，我们选取前缀调优（Li 和 Liang，2021年）、串行适配器（Houlsby 等人，2019年）、LoRA（Hu 等人，2021年）和并行适配器（He 等人，2022a）作为代表性方法来评估其效果。为确保所有微调实验的一致性，我们将批次大小设为 16。前缀调优的学习率设为 3e-2，其余方法的学习率设为 3e-4。每种参数高效微调方法都在微调数据集上训练三个轮次。需要注意的是，对于数学任务或常识推理任务，我们各自只微调一个模型，随后在所有相应数据集上评估其性能。

4 实验结果

4.1 放置与配置

为了解答这一研究问题，“各种类型的适配器应采用何种最佳放置位置和配置？”，我们采用 LLaMA-7B 作为基础模型，在以下内容中评估不同的适配器设置

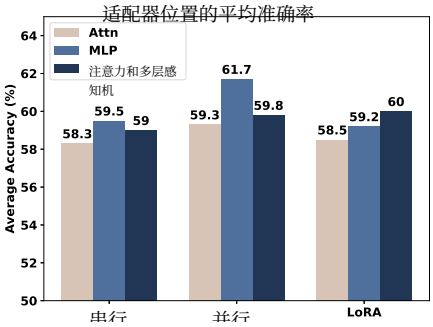

图 2：数学推理数据集上不同适配器位置的平均准确率。

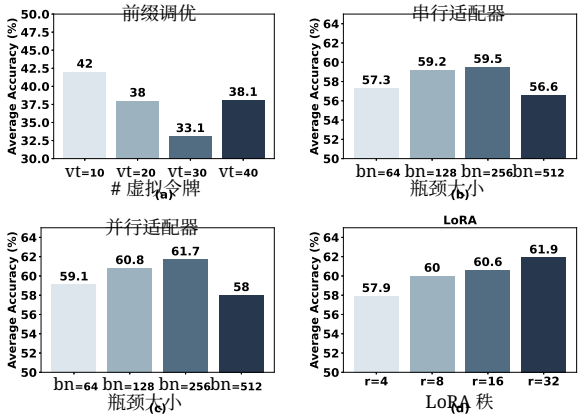

图 3：数学推理数据集上不同变量设置的平均准确率。其中，“vt”表示虚拟令牌的数量，“bn”表示瓶颈层大小，“r”表示 LoRA 秩。

数学推理任务的语境。我们的实证研究首先确定串行适配器、并行适配器和 LoRA 的最佳放置位置。由于前缀调优的位置是预先确定的，因此本分析不包括该方法。对于串行适配器，我们考察将其放置在多头注意力层之后、多层感知机层之后，或两者之后的方案。对于并行适配器和 LoRA，我们将它们集成到多头注意力层、多层感知机层，或两者之中，以评估各自的表现。各数据集的详细结果见附录A.3。图 2展示了数学推理数据集上的平均准确率。可以看出，对于串行适配器，最佳位置是在多层感知机层之后，其在数学推理数据集上的平均准确率达到 59。5% 。对于并行适配器，将其置于多层感知机层中时表现最佳，准确率达到 61。7%。对于 LoRA，需要同时将其插入多头注意力层和多层感知机层，才能达到最佳表现60%。

LLM	方法	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	Avg
GPT-3.5$_{175B}$	-	83.8	**56.4**	85.3	**38.9**	88.1	**69.9**	**70.4**
BLOOMz$_{7B}$	前缀	68.8	13.8	47.1	12.5	49.4	24.1	36.0
	串行	80.7	14.3	72.6	20.5	69.3	38.1	49.3
	并行	85.8	18.5	77.7	18.9	74.8	36.4	52.0
	LoRA	82.8	17.4	72.4	21.3	69.9	41.0	50.8
GPT-J$_{6B}$	前缀	74.5	16.0	65.6	14.7	61.4	31.0	43.9
	串行	91.7	19.5	85.8	15.0	81.7	43.6	56.2
	并行	92.2	18.9	83.8	17.9	80.7	41.1	55.8
	LoRA	90.7	23.0	84.1	16.1	84.1	46.0	57.3
LLaMA$_{7B}$	前缀	63.2	24.4	57.0	14.2	55.3	38.1	42.0
	串行	92.8	33.3	80.0	15.0	83.5	52.3	59.5
	并行	94.5	35.3	86.6	18.1	86.0	49.6	61.7
	LoRA	**95.0**	37.5	83.3	18.9	84.4	52.1	61.9
LLaMA$_{13B}$	前缀	72.2	31.1	56.0	15.7	62.8	41.4	46.5
	串行	93.0	44.0	80.5	22.0	87.6	50.8	63.0
	并行	94.3	43.3	83.0	20.5	89.6	55.7	64.4
	LoRA	94.8	47.5	**87.3**	18.5	**89.8**	54.6	65.4

表 3：在六个数学推理数据集上，使用不同适配器的大语言模型的准确率比较。我们使用 GPT-3.5`text-Davinci-003` 进行零样本思维链（Kojima 等人，2022年）作为基线。

为了确定各类适配器的最佳配置，我们针对每种参数高效微调方法中最关键的变量进行了分析，并比较其在数学推理数据集上的平均准确率。适配器的放置位置采用位置分析得出的最佳设置。对于前缀调优，我们评估了不同虚拟令牌数量（vt）设为 $[10, 20, 30, 40]$ 时的表现。对于串行和并行适配器，我们评估了不同瓶颈大小（bn）取值为 $[64, 128, 256, 512]$ 时的影响。对于 LoRA，我们考察了不同秩值（r）为 $[4, 8, 16, 32]$ 时的影响。各数据集的详细结果见附录 A.4。图 3 展示了数学推理数据集上不同变量对应的平均准确率。可以看出，当 Prefix-Tuning 的虚拟令牌数量设为 10 时，其在数学推理数据集上的平均准确率为42。0 % 。将瓶颈维度设为 256 时，串行和并行适配器的表现最佳。然而，当瓶颈大小增加到 512 时，串行和并行适配器的准确率均有所下降。LoRA 秩的典型设置为 8，但我们发现，增大秩可以提升 LoRA 的表现。当秩从 8 增加到 32 时，LoRA 的平均准确率从60。0% 61。9%。

基于全面的放置与配置分析，我们确定了每个适配器的最佳设置，并将在后续实验中始终采用这些设置。

- **对于前缀调优，我们将虚拟令牌的数量设为 10。**
- **对于串行和并行适配器，我们将其无缝集成到多层感知机层中，并将瓶颈大小设为 256。**
- **对于 LoRA，我们将其无缝集成到多头注意力层和多层感知机层中，秩设为 32。**

4.2 算术推理

为了评估适配器在算术推理任务上的有效性，我们开展了一项研究：在 Math10K 数据集上对适配器进行微调，随后在六个不同的数学推理数据集上进行评估。作为基线，我们采用 GPT-3.5 模型，具体使用`text-Davinci-003` 版本，并根据 Kojima 等人 (2022年) 的研究进行零样本思维链评估。GPT-3.5 模型的结果见 Wang 等人 (2023年)。表 3报告了不同参数高效微调方法及基线的性能。平均而言，GPT-3.5 模型（175B）的准确率优于基于适配器的参数高效微调大语言模型。然而，在 MultiArith、AddSub 和 SingleEq 等较简单的数学推理数据集上，LLaMA-13B 搭配 LoRA 等基于适配器的方法表现优于 GPT-3.5。值得注意的是，LLa MA-13B 搭配 LoRA 的平均准确率达到 65.4%，约为 GPT-3.5 性能的 92.8%。这表明，只要有充足的任务特定训练数据，较小大语言模型的基于适配器的参数高效微调就有望达到与超大型语言模型相当的性能。尤其是在 MultiArith、AddSub 和 SingleEq 等较简单的任务中，基于适配器的参数高效微调能让较小语言模型取得优于 GPT-3.5 的表现。然而，

LLM	方法	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg
GPT-3_{175B}	-	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6	-
$PaLM_{540B}$	-	**88.0**	82.3	-	83.4	81.1	76.6	53.0	53.4	-
ChatGPT	-	73.1	**85.4**	68.5	78.5	66.1	**89.8**	**79.9**	74.8	77.0
$BLOOMz_{7B}$	前缀	45.6	53.7	46.3	26.7	49.5	52.1	39.7	44.3	44.7
	串行	65.4	70.4	73.6	53.4	69.3	72.3	55.9	68.0	66.0
	并行	64.1	71.5	72.1	52.9	67.0	70.5	54.7	69.6	65.3
	LoRA	65.9	75.3	74.5	57.3	72.5	74.6	57.8	73.4	68.9
GPT-J_{6B}	前缀	63.1	66.9	68.7	34.4	64.5	64.4	46.8	59.0	58.5
	串行	62.1	63.5	72.3	30.6	68.0	63.9	48.1	63.8	59.0
	并行	62.2	69.7	70.0	41.7	65.0	60.2	44.6	58.2	59.0
	LoRA	62.4	68.6	49.5	43.1	57.3	43.4	31.0	46.6	50.2
$LLaMA_{7B}$	前缀	64.3	76.8	73.9	42.1	72.1	72.9	54.0	60.6	64.6
	串行	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
	并行	67.9	76.4	78.8	69.8	78.9	73.7	57.3	75.2	72.3
	LoRA	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.7
$LLaMA_{13B}$	前缀	65.3	75.4	72.1	55.2	68.6	79.5	62.9	68.0	68.4
	串行	71.8	83.0	79.2	88.1	82.4	82.5	67.3	81.8	79.5
	并行	72.5	84.8	79.8	**92.1**	**84.7**	84.2	71.2	**82.4**	**81.5**
	LoRA	72.1	83.5	**80.5**	90.5	83.7	82.8	68.3	**82.4**	80.5

表 4：不同适配器的大语言模型在八个常识推理数据集上的准确率比较。ChatGPT 的结果通过使用 gpt-3.5-turbo API 进行零样本思维链得到。

在 GSM8K 和 SVAMP 等更复杂的任务中，挑战依然存在。这些任务要求基础模型具备更高水平的语言理解能力和熟练度，因此性能差距明显。就所采用的不同适配器而言，LoRA 在使用的可训练参数显著更少的同时，仍取得了出色的性能。这说明，任务特定微调可能不需要过多的可学习参数。总体而言，这些发现表明，较小大语言模型采用基于适配器的参数高效微调，仅需少量可训练参数，便有望在特定任务上达到较高性能。

4.3 常识推理

此外，我们还评估了各种参数高效微调方法在常识推理任务上的效果。我们使用 Commonsense170K 数据集对适配器进行微调。用于常识推理的基线模型包括 GPT-3（175B）、PaLM（540B）和 ChatGPT。GPT- 3 和 PaLM 的结果见Tou- vron 等人（2023年）。为评估 ChatGPT 在常识推理方面的表现，我们使用`gpt-3.5-turbo` API，并采用零样本思维链。零样本思维链提示与收集常识推理微调数据集时使用的模板一致，详见附录A.2。表4展示了采用不同大语言模型的参数高效微调方法与基线模型的表现。值得注意的是，采用串行适配器、并行适配器和 LoRA 的 LLaMA-13B 表现优于所有基线模型，包括被誉为迄今最出色的大语言模型的 ChatGPT。采用并行适配器的模型平均准确率达到 81.5%，比 ChatGPT 高 4.5%。值得注意的是，常识推理数据集中的所有训练集都包含在微调数据 Commonsense170K 中。此外，我们观察到，参数高效微调方法的表现受到基础模型自身能力的影响。与 BLOOMz 和 GPT-J 模型相比，LLaMA-7B 和 LLaMA-13B 展现出更强的常识推理能力。

4.4 分布内与分布外分析

比较参数高效微调方法在数学推理和常识推理任务上的表现时，我们可以观察到，参数高效微调方法在常识推理领域展现出更为显著的结果。接下来，我们将从分布内（ID）和分布外（OOD）两个角度分析导致这一现象的因素。在常识推理的背景下，微调数据集 Commonsense170K 涵盖了常识推理数据集中的所有训练集。值得注意的是，参数高效微调方法已经展现出超越 ChatGPT 的能力。这一观察结果表明，通过利用分布内微调数据，较小的语言模型（如 LLaMA-13B）能够在特定下游任务中超越较大的语言模型（如 ChatGPT 和 PaLM）。然而，在考虑数学推理任务时，微调数据集 Math10K 仅包含 GSM8K 和 AQuA 的训练集。在这方面，有研究观察到，参数高效微调方法，尤其是采用 LoRA 的 LLaMA-13B，在 MultiArith、AddSub 和 SingleEq 上展现出优于 GPT-3.5 的性能。这些

问题	Boarden先生正在翻修浴室。每平方英尺需要24块马赛克瓷砖。他需要多少块马赛克瓷砖才能铺满他36平方英尺浴室的三分之二?
ChatGPT	1. 求36的三分之二：将36乘以2/3。**(36 *2/3) = 24 平方英尺** 2. 计算每平方英尺所需的马赛克瓷砖数量：已知Mr. Boarden需要24块马赛克每平方英尺的瓦片数量，将第 1 步的结果乘以 24。**(24 *24) = 576马赛克瓷砖** 因此，Mr. Boarden需要576块马赛克瓷砖，才能铺满他36平方英尺浴室的三分之二。
LLaMA$_{13B}$-前缀	1平方英尺 = 4平方英寸，1平方英寸 = 16平方厘米。因此，1平方英尺 = 16 平方厘米。所以，36平方英尺 = 36 x 16平方厘米 = 576 平方厘米。要铺满浴室的三分之二，Boarden先生需要铺满576平方厘米 x 2/3 = 384 平方厘米。由于1平方厘米 = 1 马赛克瓷砖，Boarden先生需要384块马赛克瓷砖。因此，Boarden先生需要384块马赛克瓷砖，才能铺满其36平方英尺浴室的三分之二。
LLaMA$_{13B}$-串行	1. 求浴室的面积：36平方英尺 2. 求面积的三分之二：**2/3 x 36平方英尺 = 24平方英尺** 3. 将面积乘以每平方英尺所需的瓦片数：**24平方英尺 x 24块/平方英尺 = 576 瓦片** 4. 因此，Boarden先生需要576块马赛克瓷砖，才能铺满其36平方英尺浴室的三分之二。
LLaMA$_{13B}$-并行	1. 计算浴室的总面积：36 平方英尺 2. 确定 Boarden先生想用马赛克瓷砖铺设的面积：**2/3 x 36 平方英尺 = 24平方英尺** 3. 计算所需的马赛克瓷砖数量：**24 平方英尺 x 24**片马赛克瓷砖/平方英尺 = 600 片马赛克瓷砖 因此，Boarden先生需要 600 片马赛克瓷砖，才能铺满其 36 平方英尺浴室的三分之二面积。
LLaMA$_{13B}$-LoRA	1. 计算浴室的总面积：36平方英尺 2. 确定需要覆盖的面积：**2/3 x 36平方英尺 = 24 平方英尺** 3. 计算所需马赛克瓷砖的数量：**24平方英尺 x 每平方英尺24块马赛克瓷砖 = 576 块马赛克瓷砖** 因此，Boarden先生需要576块马赛克瓷砖，才能覆盖其36平方英尺浴室的三分之二。

表5：从GSM8K中随机抽取的示例。ChatGPT和LLaMA-13B使用不同参数高效微调方法时的输出。

发现表明，参数高效微调方法能够增强大语言模型的数学推理能力，并且可以成功应用于分布外数据集。然而，在分布内数据集 GSM8K 和 AQuA 上评估参数高效微调方法的性能时，与 GPT-3.5 相比，仍然存在明显的性能差距。这种差异可能是由于 GSM8K 和 AQuA 数据集在数学推理方面具有更高的复杂性，而较小的大语言模型的推理能力仍然有限。因此，探索提升参数高效微调方法在复杂数学推理任务上性能的策略，是未来研究的一个潜在方向。

5 定性研究

前几节介绍了定量分析。本节将提供定性示例，以展示不同模型输出的质量。表5 展示了从 GSM8K 中随机选取的一个问题，以及使用各种参数高效微调方法的 ChatGPT 和 LLaMA-13B 模型的输出。更多详细示例请参见附录 A.5。ChatGPT 全面理解了该问题，并生成了两个步骤：“(36 * 2/3) = 24 平方英尺”和“(24 * 24) =576块马赛克瓷砖”，有效地解决了该问题。然而，LLaMA-13B-Prefix 模型的语言理解能力有限，导致 LLaMA-13B-Prefix 在第一步中采取了错误的思路。另一方面，配备串行适配器的 LLaMA-13B 通过给出关键的两个步骤，并进行正确计算以获得准确结果，生成了高质量的答案。有趣的是，LLaMA-13B-并行和 LLaMA-13B-LoRA 生成了几乎相同的推理过程。然而，LLaMA-13B-并行由于计算错误而得出了错误答案，称“24 平方英尺 x 每平方英尺 24 块马赛克瓷砖 = 600马赛克瓷砖”。总体而言，在配备特定任务的微调数据后，像 LLaMA-13B 这样较小的语言模型能够生成令人印象深刻的高质量答案，其水平可与 ChatGPT 生成的答案相媲美。

6 结论

本文开发了一个易于使用的框架 LLM-Adapter，可将各种适配器无缝集成到大语言模型中，使研究人员能够针对广泛的任务实现基于适配器的参数高效微调方法。为了评估不同参数高效微调方法在下游任务上的表现，我们构建了两个高质量的微调数据集，以提升参数高效微调方法在数学推理和常识推理任务上的性能。借助 LLM-Adapter工具包和构建的微调数据集，我们开展了全面的实证研究，并解答了有关不同参数高效微调方法的最佳放置位置与配置、适配器架构的影响，以及分布内和分布外场景影响的研究问题。我们希望这项工作

能够推动针对大语言模型的参数高效微调方法进一步发展。

7 局限性

这项工作存在两个局限性。首先，由于计算资源有限，我们无法评估更大型语言模型（如 LLaMA-33B 和 LLaMA-65B）的性能。预计这些具备更强语言理解能力的大型模型能够取得更优异的表现。其次，本文未深入探讨不同适配器的组合。鉴于组合各种参数高效微调方法所涉及的搜索空间十分庞大，我们计划在未来的研究中探索这一方向。

参考文献

Armen Aghajanyan、Luke Zettlemoyer 和 Sonal Gupta。2020年。内在维度解释了语言模型微调的有效性。发表于计算语言学协会年会。

Yonatan Bisk、Rowan Zellers、Ronan Le Bras、高剑峰和 Yejin Choi。2020年。Piqa：关于自然语言中物理常识的推理。发表于第三十四届 *AAAI* 人工智能大会。

Jiaao Chen、Aston Zhang、Xingjian Shi、李沐、Alex Smola 和 Diyi Yang。2023年。参数高效微调设计空间。*arXiv* 预印本 *arXiv:2301.01821*。

Christopher Clark、Kenton Lee、Ming-Wei Chang、Tom Kwiatkowski、Michael Collins 和 Kristina Toutanova。2019年。BoolQ：探索自然是非题令人惊讶的难度。载于*2019*年北美计算语言学协会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第2924–2936页，美国明尼阿波利斯，明尼苏达州。计算语言学协会。

Peter Clark、Isaac Cowhey、Oren Etzioni、Tushar Khot、Ashish Sabharwal、Carissa Schoenick 和 Oyvind Tafjord。2018年。你认为自己解决了问答问题吗？来试试 ARC，也就是 AI2 推理挑战。*arXiv:1803.05457v1*。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。2021年。训练验证器解决数学文字题。*arXiv* 预印本 *arXiv:2110.14168*。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。Bert：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*。

Ali Edalati、Marzieh S. Tahaei、Ivan Kobyzev、V. Nia、James J. Clark 和 Mehdi Rezagholizadeh。2022年。Krona：使用克罗内克适配器进行参数高效微调。*ArXiv*，abs/2212.10650。

Cheng Fu、Hanxian Huang、Xinyun Chen、Yuandong Tian 和 Jishen Zhao。2021年。学习共享：一种利用计算与参数共享的硬件友好的迁移学习框架。载于第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第3469–3479页。PMLR。

Junxian He、Chunting Zhou、Xuezhe Ma、Taylor Berg-Kirkpatrick 和 Graham Neubig。2021年。迈向参数高效迁移学习的统一视角。*arXiv* 预印本 *arXiv:2110.04366*。

Junxian He、Chunting Zhou、Xuezhe Ma、Taylor Berg- Kirkpatrick 和 Graham Neubig。2022a。迈向参数高效迁移学习的统一视角。载于国际学习表征会议。

Shwai He、Liang Ding、Daize Dong、Jeremy Zhang 和 Dacheng Tao。2022b。SparseAdapter：一种提升适配器参数效率的简便方法。载于计算语言学协会论文集：*EMNLP 2022* 论文集，第 2184–2190 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

James Henderson、Sebastian Ruder 等。2021年。Compacter：高效的低秩超复数适配器层。载于神经信息处理系统进展。

Mohammad Javad Hosseini、Hannaneh Hajishirzi、Oren Etzioni 和 Nate Kushman。2014年。学习通过动词分类求解算术文字题。载于*EMNLP*，第 523–533 页。

Neil Houlsby、Andrei Giurgiu、Stanislaw Jastrzebski、Bruna Morrone、Quentin de Laroussilhe、Andrea Gesmundo、Mona Attariyan 和 Sylvain Gelly。2019年。面向自然语言处理的参数高效迁移学习。发表于国际机器学习会议。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang 和 Weizhu Chen。2021年。LoRA：大语言模型的低秩适配。*ArXiv*，abs/2106.09685。

Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yu- taka Matsuo 和 Yusuke Iwasawa。2022年。大语言模型是零样本推理器。*arXiv* 预印本 *arXiv:2205.11916*。

Rik Koncel-Kedziorski、Hannaneh Hajishirzi、Ashish Sabharwal、Oren Etzioni 和 Siena Dumas Ang。2015年。将代数文字题解析为方程。计算语言学协会汇刊，3:585–597。

Rik Koncel-Kedziorski、Subhro Roy、Aida Amini、Nate Kushman 和 Hannaneh Hajishirzi。2016年。MAWPS：数学文字题库。载于*NAACL*论文集，第1152–1157页。

Brian Lester、Rami Al-Rfou 和 Noah Constant。2021年。参数高效提示调优的规模效应。*arXiv*，abs/2104.08691。

Xiang Lisa Li 和 Percy Liang。2021年。前缀调优：优化用于生成的连续提示。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第4582–4597页，线上。计算语言学协会。

Wang Ling、Dani Yogatama、Chris Dyer 和 Phil Blunsom。2017年。通过理由生成进行程序归纳：学习求解并解释代数文字题。载于第*55*届计算语言学协会年会论文集（第*1*卷：长篇论文），第158–167页。

Sourab Mangrulkar、Sylvain Gugger、Lysandre Debut、Younes Belkada 和 Sayak Paul。2022年。Peft：先进的参数高效微调方法。https://github.com/huggingface/peft。

Yuning Mao、Lambert Mathias、Rui Hou、Amjad Alma- hairi、Hao Ma、Jiawei Han、Wen tau Yih 和 Madian Khabsa。2021年。UniPELT：一种用于参数高效语言模型调优的统一框架。*arXiv*，abs/2110.07577。Niklas Muennighoff、Thomas Wang、Lintang Sutawika、Adam Roberts、Stella Biderman、Teven Le Scao、M Saiful Bari、Sheng Shen、Zheng-Xin Yong、Hailey Schoelkopf 等人。2022年。通过多任务微调实现跨语言泛化。*arXiv* 预印本 *arXiv:2211.01786*。OpenAI。2022年。ChatGPT 介绍。https://openai.com/blog/chatgpt。OpenAI。2023年。GPT-4 技术报告。*CoRR*，abs/2303.08774。Arkil Patel、Satwik Bhattamishra 和 Navin Goyal。2021年。自然语言处理模型真的能解决简单的数学文字题吗？发表于 *NAACL* 论文集，第 2080–2094 页。Jonas Pfeiffer、Ivan Vulic、Iryna Gurevych 和 Sebastian Ruder。2020年。MAD-X：一个基于适配器的多任务跨语言迁移框架。发表于自然语言处理实证方法会议。Chengwei Qin、Aston Zhang、Zhuosheng Zhang、Jiaao Chen、Michihiro Yasunaga 和 Diyi Yang。2023年。ChatGPT 是通用自然语言处理任务求解器吗？*arXiv* 预印本 *arXiv:2302.06476*。

Yujia Qin、Xiaozhi Wang、Yusheng Su、Yankai Lin、Ning Ding、Jing Yi、Weize Chen、Zhiyuan Liu、Juanzi Li、Lei Hou 等人。2021年。通过提示调优探索通用内在任务子空间。*arXiv* 电子印本，第 arXiv–2110 页。

Subhro Roy 和 Dan Roth。2016年。解决一般算术文字问题。*arXiv* 预印本 *arXiv:1608.01413*。

Keisuke Sakaguchi、Ronan Le Bras、Chandra Bhagavat- ula 和 Yejin Choi。2021年。WinoGrande：大规模对抗性 Winograd模式挑战。《*ACM* 通讯》，64(9):99–106。

Maarten Sap、Hannah Rashkin、Derek Chen、Ronan LeBras 和 Yejin Choi。2019年。Socialiqa：关于社会互动的常识推理。*arXiv* 预印本 *arXiv:1904.09728*。

Yongliang Shen、Kaitao Song、Xu Tan、Dongsheng Li、Weiming Lu 和 Yueting Zhuang。2023年。Hugging-gpt：使用 ChatGPT 及其在 huggingface 中的朋友解决 AI 任务。*CoRR*，abs/2303.17580。

Yi-Lin Sung、Jaemin Cho 和 Mohit Bansal。2022年。LST：梯形侧向微调，实现参数和内存高效的迁移学习。*arXiv*，abs/2206.06522。

Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。2023年。Stanford Alpaca：指令跟随型Llama模型。https://github.com/tatsu-lab/stanford_alpaca。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023年。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*。Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就是一切所需。在神经信息处理系统进展，第 5998–6008 页。Tu Vu、Brian Lester、Noah Constant、Rami Al-Rfou 和 Daniel Cer。2021年。SPOT：通过软提示迁移改进冻结模型的适配。*arXiv* 预印本 *arXiv:2110.07904*。Ben Wang 和 Aran Komatsuzaki。2021年。GPT-J-6B：一个拥有 60 亿参数的自回归语言模型。https://github.com/kingoflolz/mesh-transformer-jax。Lei Wang、Wanyu Xu、Yihuai Lan、Zhiqiang Hu、Yunshi Lan、Roy Ka-Wei Lee 和 Ee-Peng Lim。2023年。规划与求解提示：通过大语言模型改进零样本思维链推理。*arXiv* 预印本 *arXiv:2305.04091*。

Yaqing Wang、Subhabrata Mukherjee、Xiaodong Liu、Jing Gao、Ahmed Hassan Awadallah 和 Jian-feng Gao。2022年。AdaMix：用于大语言模型参数高效微调的适配器混合。*ArXiv*，abs/2205.12410。

Li Yunxiang、Li Zihan、Zhang Kai、Dan Ruilong 和 Zhang You。2023年。Chatdoctor：一种利用医学领域知识在 LLaMA 模型上微调的医疗聊天模型。*arXiv* 预印本 *arXiv:2303.14070*。

A 附录

A.1 数学推理提示模板

我们使用 ChatGPT 收集用于微调的数学推理数据。表格 6 展示了用于向 ChatGPT 提交查询的提示模板。我们使用“请给出步骤”来引导 ChatGPT 生成推理步骤，从而利用推理依据对适配器进行微调。“请用阿拉伯数字给出答案。”则用于引导 ChatGPT 生成作为最终答案的阿拉伯数字，以便从输出中提取答案。

A.2 常识数据模板

由于常识推理中的每个数据集都涉及不同的任务，我们采用结构化模板，首先描述任务目标，随后介绍相应的内容和答案。表 7 展示了用于收集常识推理微调数据的模板。

A.3 位置分析

表 8 展示了适配器在数学推理数据集中的不同位置时的性能。该研究使用的微调数据集为 Math10K，基础模型为 LLaMA-7B。可以看到，对于串行适配器，最佳位置是多层感知机层之后，数学推理数据集上的平均准确率达到 59。5% 。对于并行适配器，将其置于多层感知机层内时性能最佳，达到61。7%。对于 LoRA，则需要同时将其插入多头注意力层和多层感知机层，才能达到最佳性能60%。

A.4 配置分析

表9 展示了参数高效微调方法在数学推理数据集上不同变量设置下的准确率对比。本研究使用的微调数据集为 Math10K。可以看到，当 Prefix- Tuning 的虚拟令牌数量设为 10 时，Prefix-Tuning 在数学推理数据集上的平均准确率为42.0% 。将瓶颈维度设为 256 时，串行适配器和并行适配器表现最佳。然而，当瓶颈大小增加到 512 时，串行适配器和并行适配器的准确率均有所下降。LoRA 秩的典型设置为 8，但我们发现更高的秩可以提升 LoRA 的性能。值得注意的是，当秩增加到 32 时，LoRA 的准确率达到 61.9%。

A.5 定性示例

本节将展示从数学推理和常识推理数据集中随机抽取的示例。

模板	[问题] 请给出解题步骤，并用阿拉伯数字作答。
提示	Albert 想知道自己一天能吃多少披萨。他买了 2 个大号披萨和 2 个小号披萨。一个大号披萨有 16 片，一个小号披萨有 8 片。如果他全部吃完，那他一共吃了多少片? day? 请给出解题步骤，并用阿拉伯数字作答。
ChatGPT 输出	为了解决这个问题，我们首先需要算出 2 个大号披萨和 2 个小号披萨一共有多少片: 2 个大号披萨 × 每个披萨 16 片 = 32 披萨片 2 个小号披萨 × 每个披萨 8 片 = 16披萨片。 接下来，我们把不同尺寸披萨的披萨片数相加，得出披萨片总数：32 片 + 16 披萨片 = 48 披萨片。 因此，如果 Albert 吃完 2 个大号披萨和 2 个小号披萨，他当天会吃掉 48 片披萨。 答案：48 片。

表 6：用于收集微调数学推理数据集的提示模板。还包含一个来自 GSM8K 的示例。

数据集	微调数据模板
BoolQ	请判断以下问题的答案是对还是错，问题： [QUESTION] 答案格式：对/错 正确答案是 [ANSWER]
PIQA	请选择该问题的正确解答： [QUESTION] 解答1： [SOLUTION_1] 解答2： [SOLUTION_2] 答案格式：方案1/方案2 正确答案是 [答案]
SIQA	请选择问题的正确答案： [问题] 答案1： [答案_1] 答案2： [答案_2] 答案3： [答案_3] 答案格式：答案1/答案2/答案3 正确答案是 [答案]
HellaSwag	请选择正确的结尾来完成给定的句子： [活动_标签]: [语境] 结尾1： [结尾_1] 结尾2： [结尾_2] 结尾3： [ENDING_3] 结尾4： [ENDING_4] 答案格式：结尾1/结尾2/结尾3/结尾4 正确答案是 [ANSWER]
WinoGrande	请选择正确的答案，填入空白处以完成给定句子： [SENTENCE] 选项1： [选项_1] 选项2： [选项_2] 正确答案是 [答案]
ARC-e&ARC-c	请选择问题的正确答案： [问题] 答案1： [答案_1] 答案2： [ANSWER_2] 答案3： [ANSWER_3] 答案4： [ANSWER_4] 答案格式：答案1/答案2/答案3/答案4 正确答案是 [ANSWER]
OBQA	请选择问题的正确答案： [QUESTION] 答案1： [ANSWER_1] 答案2： [ANSWER_2] 答案3： [ANSWER_3] 答案4： [ANSWER_4] 答案格式：答案1/答案2/答案3/答案4 正确答案是 [ANSWER]

Ta表 7：用于创建常识推理 数据以供微调的各数据集的数据模板 用于微调的推理数据。

模型	位置	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	平均值
串行	Attn	92.3	32.0	80.0	16.9	80.5	47.9	58.3
	MLP	92.8	33.3	80.0	15.0	83.5	52.3	59.5
	Both	94	29.8	84.1	17.3	83.5	45.1	59.0
并行	Attn	94.5	33.5	83.0	17.3	80.5	46.9	59.3
	MLP	94.5	35.3	**86.6**	**18.1**	**86.0**	**49.6**	**61.7**
	Both	94.3	30.2	84.8	17.7	84.3	47.2	59.8
LoRA	Attn	94.2	35.3	79.7	16.9	78.7	45.9	58.5
	MLP	95.8	35.0	80.0	15.7	81.7	47.0	59.2
	Both	**96.2**	**35.6**	80.5	15.7	82.3	**49.6**	60.0

表8：在数学推理数据集上评估了将适配器置于不同位置时的准确率。本分析所用的微调数据集为 Math10K。在此语境中，“Attn”指多头注意力层，“MLP”指多层感知机层。本研究采用的基础模型为 LLaMA-7B。

模型	变量	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	平均值
前缀	vt=10	63.2	24.4	57.0	14.2	55.3	38.1	42.0
	vt=20	60.3	22.9	46.1	16.1	51.8	30.6	38.0
	vt=30	51.2	16.9	42.3	15.4	41.9	31.1	33.1
	vt=40	54.2	17.8	49.6	**21.7**	52.0	33.2	38.1
串行	bn=64	93.0	31.6	80.5	14.6	79.9	44.1	57.3
	bn=128	93.8	33.4	81.5	18.5	79.9	48.1	59.2
	bn=256	92.8	33.3	80.0	15.0	83.5	**52.3**	59.5
	bn=512	95.3	33.7	77.2	14.9	76.6	42.0	56.6
并行	瓶颈大小=64	92.7	36.8	77.5	16.9	81.9	48.5	59.1
	瓶颈大小=128	95.8	36.9	83.0	17.7	83.3	47.8	60.8
	瓶颈大小=256	94.5	35.3	86.6	18.1	**86.0**	49.6	61.7
	瓶颈大小=512	95.0	31.5	81.8	14.6	81.5	43.4	58.0
LoRA	r=4	95.2	34.7	78.5	16.1	77.8	45.4	57.9
	r=8	**96.2**	35.6	80.5	15.7	82.3	49.6	60.0
	r=16	95.5	36.2	82.8	13.8	84.4	50.9	60.6
	r=32	95.0	**37.5**	**83.3**	18.9	84.4	52.1	**61.9**

表9：针对数学推理数据集，比较参数高效微调方法在变量不同设置下的准确率。本分析使用的微调数据集为 Math10K。在此语境中，“vt”表示虚拟令牌的数量，“bn”表示瓶颈大小，“r”表示 LoRA 秩。本研究采用的基础模型为 LLaMA-7B。

问题	在学校书展上，Sam买了13本冒险书籍和17本悬疑书籍。如果其中15本书是 旧书，他买了多少本新书?
ChatGPT	要计算铺满Mr. Boarden的36平方英尺 浴室，请按以下步骤操作： 求 36 的三分之二：将 36 乘以 2/3。 (36 * 2/3) = 24平方英尺 确定每平方英尺所需的马赛克瓷砖数量：已知 Mr. Boarden 需要 24 块马赛克 块马赛克瓷砖每平方英尺，将步骤 1 的结果乘以 24。 (24 * 24) = 576块马赛克瓷砖 因此，Mr. Boarden 需要 576 块马赛克瓷砖，才能铺满他 36 平方英尺浴室的三分之二。
$LLaMA_{13B}$-前缀	1. 求 Sam 购买的书籍总数： 13 + 17 = 30 书籍 2. 从书籍总数中减去二手书的数量： 30 - 15 = 15 新书 因此，Sam 在书展上购买了 15 本新书。 步骤： 1. 书籍总数字 = 13 + 17 = 30 2. 二手书数量 = 15 3. 新书数量 = 书籍总数 - 二手书数量 4. 新书数字 = 30 - 15 = 15 5. Sam 在书展上买了 15 本新书。 答案：Sam 在书展上买了 15 本新书。（15）
$LLaMA_{13B}$-串行	1. 找出 Sam 买的书籍总数： 13 + 17 = 30 2. 从书籍总数中减去二手书的数量：30 - 15 = 15 3. 因此，Sam 买了 15 本新书。 答案：15
$LLaMA_{13B}$-并行	要计算 Sam 买了多少本新书，我们需要从 他购买的书籍总数中减去二手书的数量。 书籍总数 = 13 + 17 = 30 二手书 = 15 新书 =书籍总数 - 二手书 新书 = 30- 15 新书 = 15 因此，Sam 买了 15 本新书。
$LLaMA_{13B}$-LoRA	1. 求 Sam 买的书籍总数： 13 + 17 = 30 2. 用书籍总数减去二手书的数量：30 - 15 = 15 3. 因此，Sam 买了 15 本新书。 答案：15

表 10：从 MultiArith 中随机抽取的一个示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。

问题	Boarden先生正在翻新他的浴室。每平方英尺需要24块马赛克瓷砖。需要多少块马赛克瓷砖才能铺满他36平方英尺浴室的三分之二?
ChatGPT	要计算铺满Mr. Boarden的36平方英尺浴室的三分之二所需的马赛克瓷砖数量，请按以下步骤操作： 1. 求36的三分之二：将36乘以2/3。 **(36 * 2/3) = 24平方英尺** 2.确定每平方英尺所需的马赛克瓷砖数量：已知Mr. Boarden每平方英尺需要24块马赛克瓷砖，将第1步的结果乘以24。 **(24 * 24) = 576块马赛克瓷砖** 因此，Mr. Boarden需要576块马赛克瓷砖，才能铺满他那间36平方英尺浴室的三分之二。
LLaMA$_{13B}$-前缀	1 平方英尺 = 4平方英寸 1 平方英寸 = 16 平方厘米 所以，1 平方英尺 = 16平方厘米 因此，36 平方英尺 = 36 × 16 平方厘米 = 576 平方厘米 要覆盖浴室的三分之二，Boarden先生需要覆盖 576 平方厘米 × 2/3 = 384 平方厘米 由于每块马赛克瓷砖的面积为 1 平方厘米 = 1，Boarden先生需要 384 平方厘米 / 1 平方厘米 = 384 块马赛克瓷砖 因此，Boarden先生需要 384 块马赛克瓷砖，才能铺满他的 36 平方英尺浴室的三分之二。
LLaMA$_{13B}$-串行	1. 求浴室的面积：36 平方英尺 2. 求面积的三分之二：**2/3 x 36 平方英尺 = 24平方英尺** 3. 将面积乘以每平方英尺所需的瓦片数量：**24 平方英尺 x 24 块/平方英尺 = 576 瓦片** 4. 因此，Boarden先生需要576块马赛克瓷砖，才能铺满其36平方英尺浴室的三分之二。
LLaMA$_{13B}$-并行	1. 计算浴室的总面积：36平方英尺 2. 确定Boarden先生想用马赛克瓷砖铺设的面积：**2/3 x 36平方英尺 = 24平方英尺** 3. 计算所需马赛克瓷砖的总数：**24平方英尺 x 24** 块马赛克瓷砖/平方英尺 = 600马赛克瓦片 因此，Boarden先生需要600块马赛克瓷砖，才能铺满其36平方英尺浴室的三分之二。 答案用阿拉伯数字表示为600。
LLaMA$_{13B}$-LoRA	1. 计算浴室的总面积：36平方英尺 2. 确定需要铺设的面积：**2/3 x 36 平方英尺 = 24 平方英尺** 3. 计算所需的马赛克瓷砖数量：**24 平方英尺 x 每平方英尺24块马赛克瓷砖 = 576 马赛克瓷砖** 因此，Boarden先生需要576块马赛克瓷砖，才能铺满他的36平方英尺浴室的三分之二。 用阿拉伯数字表示的答案是576。

表 11：从GSM8K中随机抽取的示例。ChatGPT和LLaMA-13B采用不同参数高效微调方法时的输出。

问题	抽屉里有7支蜡笔。玛丽从抽屉里拿走了3支蜡笔。还剩多少支蜡笔 now?
ChatGPT	要计算玛丽取出一些蜡笔后抽屉里还剩多少蜡笔，请按以下步骤操作： 从抽屉里蜡笔的总数中减去玛丽取出的蜡笔数：7 - 3 = 4 蜡笔 . 因此，抽屉里现在还剩 4 支蜡笔。
LLaMA$_{13B}$-前缀	1. 先看抽屉里蜡笔的初始数量：7。 2. 减去玛丽从抽屉里拿出的蜡笔数量：3。 3. 抽屉里最后剩下的蜡笔数量是 7 - 3 = 4。 因此，抽屉里现在有 4 支蜡笔。
LLaMA$_{13B}$-串行	1. 从抽屉里最初的蜡笔数量开始：7 2. 减去玛丽拿出的蜡笔数量：3 3. 进行减法运算：7 - 3 = 4 4. 因此，抽屉里现在有4支蜡笔。 答案：抽屉里现在有4支蜡笔。
LLaMA$_{13B}$-并行	1. 从抽屉里的蜡笔总数开始：7 2. 减去玛丽拿走的蜡笔数量：3 3. 进行减法运算：7 - 3 = 4 4. 最终答案是 4 支蜡笔。 因此，抽屉里现在有 4 支蜡笔。
LLaMA$_{13B}$-LoRA	1. 从最初的蜡笔数量开始：7 2. 减去玛丽拿走的蜡笔数量：7 - 3 = 4 3. 最终答案是 4 支蜡笔。 因此，抽屉里现在有 4 支蜡笔。

表 12：从 AddSub 中随机抽取的示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。

问题	假设你想把英语、印地语、数学、历史、地理和科学书籍 摆放在书架上。你有多少种摆放方式？答案选项：（A）520（B）720（C）920（D）都不是（E） 无法确定
ChatGPT	要计算英语、印地语、数学、历史、地理书籍的摆放方式数量， 以及书架上的科学书籍，我们可以使用排列的概念。由于排列书籍时顺序很重要， 我们需要求出排列的数字。 要确定排列的总数，请按照以下步骤操作： 计算书籍的数量：总共有 6 本书。 使用 n 个对象的排列公式，其中 n 是书籍的数量：n！（n 的阶乘）。 n! = 6! = 6 × 5 × 4 × 3 × 2 × 1 = 720 因此，你可以用 720 种不同的方式将书籍摆放在书架上。 正确答案是 (B) 720。
LLaMA$_{13B}$-前缀	1. 首先，我们为题目中的未知数设定变量。设 a、b 和 c 分别为 M、N 中的数字， 以及 396c 中的数字。 2. 我们知道 M 和 N 都是三位数，因此 a +b + c = 3。 3. 我们还知道 M - N + 396c = 990，因此可以将 a、b 和 c 代入这个方程： a(1000000 - 100000 + 396c) + b(1000000 - 100000 + 396c) +c(1000000 - 100000 + 396c) = 990 4. 化简这个方程后，我们得到： a
LLaMA$_{13B}$-串行	要解决这个问题，我们需要使用乘法计数原理。该原理指出，如果有多种方式可以完成一件事，而有 m 种方式可以完成另一件事，那么同时完成这两件事就有 m x n 种方式 一起。 在这种情况下，有 6 本书需要排列（英语、印地语、数学、历史、地理和科学）。 这些书可以有 6! 种排列方式。 因此，将这些书排列在书架上的总方式数为 6! = 720。 因此，答案是选项 B) 720。
LLaMA$_{13B}$-并行	要在书架上排列的书籍总数为 5 本（英语、印地语、数学、历史， 地理和科学）。 要计算这些书籍在书架上的排列方式数，可以使用从 n 个对象中每次取 r 个进行排列的公式 从 n 个对象中每次取出 r 个对象的组合数，其计算公式为： P(n,r) =n! / (n-r)! 其中，n 是对象总数，r 是每次排列的对象数。 在这种情况下，n = 5 （书籍总数），而 r = 5（因为我们要把全部 5 本书摆放在 书架上）。 将这些数值代入公式，可得： P(5,5) = 5! / (5-5)! P(5,5) = 5! / (5-5)! P(5,5) = 5! / (5-5)! P(5,5) = 5! / (5-5)! P(5,5) = 5! / (5-5)!
LLaMA$_{13B}$-LoRA	将 6 本书摆放在书架上的方法总数为 6! = 6× 5 × 4 × 3 × 2 × 1 = 720。 因此，答案是选项 B) 720。

表 13：从 AQuA 中随机抽取的一个示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。

问题	Oceanside自行车租赁店租用一辆自行车的费用为17美元，外加每小时7美元。汤姆支付了80美元租用一辆自行车。他支付了多少小时的租车费用?
ChatGPT	要计算汤姆支付了多少小时的租车费用，请按以下步骤操作： 1. 从汤姆支付的总金额中减去基础费用：80 - 17 = 63 美元。 2. 将剩余金额除以每小时费率，得出小时数：63 / 7 = 9 小时。 因此，汤姆付费租车进行了 9 小时的检查。 注意：结果表示完整的小时数。如果存在任何不足一小时的时长， 这些时长不会体现在计算结果中。
LLaMA$_{13B}$-前缀	1. 设 x 为汤姆付费租赁自行车的小时数。 2. 租用自行车 x 小时的费用为 17 + 7x 美元。 3. 我们知道汤姆支付了 80 美元，因此可以列出一个方程： 17 + 7x = 80 4. 解出 x，我们将等式两边都除以 7： x = 10 5. 因此，汤姆租用自行车 10 小时共支付了这笔费用。
LLaMA$_{13B}$串行	1. 求租用自行车的总费用： $17 + $7/小时 =$24/小时 2. 用总费用除以每小时费用，求出租用时数： $80 ÷ $24/小时 = 3.33 小时 3. 向上取整到最接近的整数，因为租用自行车的时间不能是不足一小时的时间： 3小时 因此，汤姆付费让人检查自行车，检查了3小时。
LLaMA$_{13B}$-并行	1. 从支付总额中减去17美元的初始费用：80美元 - $17 =$63 2. 将剩余金额除以每小时费率：$63 ÷ $7 = 9 3. 答案是9小时。 因此，汤姆付费租用了这辆自行车 9 个小时。
LLaMA$_{13B}$-LoRA	1. 设 x 为汤姆付费租用这辆自行车的小时数。 2. 租用这辆自行车 x 小时的费用为 7x 美元。 3. 租用这辆自行车 x 小时的总费用为 17 + 7x 美元。 4. 我们知道汤姆支付了 80 美元租用这辆自行车，因此可以列出方程： 17 + 7x = 80 5. 解出 x：从等式两边同时减去 17：7x = 63 6. 等式两边同时除以 7，得到 x = 9 因此，汤姆付费租用自行车检查了 9 小时。

表14：从 SingleEq 中随机抽取的示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。

问题	Matthew 把数量相等的脆饼干和蛋糕分给了他的 4 位朋友。如果他一开始有 32 块脆饼干和 98 块蛋糕。每个人吃了多少块脆饼干?
ChatGPT	要计算每个人吃了多少块脆饼干，我们需要找出每位朋友最初收到的脆饼干数量，并从 Matthew 拥有的脆饼干总数中减去这个数量。 步骤如下： 将脆饼干总数除以朋友人数（4），即可得出每个人最初收到的脆饼干数量：32 / 4 = 8每位朋友的脆饼干。 由于我们要计算每个人吃掉的脆饼干数量，因此需要从脆饼干总数中减去每位朋友最初收到的脆饼干数量：32 - 8 = 24块脆饼干。 因此，每个人吃了24块脆饼干。
$LLaMA_{13B}$-前缀	1. 找出 Matthew 最初有的脆饼干和蛋糕总数： 32 + 98 = 130 2. 将脆饼干和蛋糕的总数除以朋友人数，得出最初分给每位朋友的数量： 130 ÷ 5 = 26 3. 从脆饼干和蛋糕的总数中减去最初分给每位朋友的数量，得出剩余数量：130 - 26 = 104 4. 用剩余的数量除以朋友人数，得出每个人吃了多少： 104 ÷ 5 = 21 因此，每个人吃了 21 块脆饼干。
$LLaMA_{13B}$-串行	1. 用脆饼干总数除以朋友人数： 32 ÷ 4 = 8 2. 用蛋糕总数除以朋友人数： 98 ÷ 4 = 24 3. 将每位朋友分到的脆饼干和蛋糕数量相加： 8 + 24 = 32 4. 因此，每个人吃了32块脆饼干和24块蛋糕。 答案：每个人吃了32块脆饼干。
$LLaMA_{13B}$-并行	1. 用脆饼干的总数除以朋友的人数： 32 ÷ 4 = 8 2. 用蛋糕的总数除以朋友的人数： 98 ÷ 4 = 24 3. 将每个人吃的脆饼干和蛋糕数量相加： 8 + 24 = 32 因此，每个人总共吃了32块脆饼干和蛋糕。
$LLaMA_{13B}$-LoRA	1. 将脆饼干的总数除以朋友的人数： 32 ÷ 4 = 8 2. 将蛋糕的总数除以朋友的人数： 98 ÷ 4 = 24 3. 将每个人分到的脆饼干数量和蛋糕数量相加： 8 + 24 = 32 因此，每个人总共吃了32块脆饼干和蛋糕。

表 15：A 个从 SVAMP 中随机抽取的示例。ChatGPT 和 LLaMA-13B 使用不同 nt 参数高效微调方法ods.

问题	生产乙醇需要消耗更多能量吗?
ChatGPT	错误。
$LLaMA_{13B}$-前缀	正确答案为真
$LLaMA_{13B}$-串行	正确答案为真
$LLaMA_{13B}$-并行	错误
$LLaMA_{13B}$-LoRA	正确答案是错误

表 16：从 BoolQ 中随机抽取的一个示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法的输出。

问题	如何为豚鼠的新住户准备笼子? 方案1：给豚鼠准备一个笼子，里面铺上几英寸厚的碎纸条垫料，还需要为它准备一个水瓶和一个食盆。 方案2：给豚鼠准备一个笼子，里面铺上几英寸厚的牛仔裤布料垫料，还需要为它准备一个水瓶和一个食盆。
ChatGPT	方案1
$LLaMA_{13B}$-前缀	正确答案是 solution1
$LLaMA_{13B}$-串行	正确答案是 solution2
$LLaMA_{13B}$-并行	正确答案是 solution2
$LLaMA_{13B}$-LoRA	正确答案是 solution1

表17：从 PIQA 中随机抽取的示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。

问题	Tracy 那天晚上没有回家，并抵抗了 Riley 的攻击。在此之前，Tracy 需要做什么？ 答案1：制定一个新计划 答案2：回家见 Riley 答案3：找个地方去
ChatGPT	答案1/制定一个新计划
$LLaMA_{13B}$-前缀	正确答案是答案1
$LLaMA_{13B}$-串行	正确答案是答案3
$LLaMA_{13B}$-并行	正确答案是答案3
$LLaMA_{13B}$-LoRA	正确答案是答案3

表 18：从 SIQA 中随机抽取的示例。使用不同参数高效微调方法时，ChatGPT 和 LLaMA-13B 的输出。

问题	屋顶瓦片拆除：一名男子正坐在屋顶上。他 结尾1：正用包装纸包裹一双滑雪板。结尾2：正把平整的瓦片扯下来。结尾3：正拿着一个魔方。结尾4：开始拆除屋顶上的屋面材料。
ChatGPT	结尾4：开始掀起屋顶上的屋面材料。
$LLaMA_{13B}$-前缀	正确答案是结尾4
$LLaMA_{13B}$-串行	正确答案是结尾4
$LLaMA_{13B}$-并行	正确答案是ending4
$LLaMA_{13B}$-LoRA	正确答案是ending4

表 19：A 随机抽取的一个示例，来自。ChatGPT 和 LLaMA-13B 使用不同 PEF 时的输出 T 方法 ·

问题	Sarah 是一名比 Maria 优秀得多的外科医生，因此 _总能接到更简单的病例。 选项1：Sarah 选项2：Maria
ChatGPT	选项1
$LLaMA_{13B}$-前缀	正确答案是选项1
$LLaMA_{13B}$-串行	正确答案是选项1
$LLaMA_{13B}$-并行	正确答案是选项1
$LLaMA_{13B}$-LoRA	正确答案是选项1

表20： 从 WinoGrande 中随机抽取的示例。ChatGPT 和 LLaMA-13B 的输出 th 不同的参数高效 微调方法。

问题	哪项陈述最能解释为什么光合作用是大多数食物网的基础? 答案1：阳光是几乎所有生态系统的能量来源。答案2：大多数生态系统 位于陆地而非水中。答案3：二氧化碳比其他气体更易获得。答案4： 所有生态系统中的生产者都是植物。
ChatGPT	
$LLaMA_{13B}$-前缀	答案1. 阳光几乎是所有生态系统能量的来源。
$LLaMA_{13B}$-串行	正确答案是答案1
$LLaMA_{13B}$-并行	正确答案是答案4
$LLaMA_{13B}$-LoRA	正确答案是答案1

表 21：从 ARC-e 中随机抽取的示例。ChatGPT 和 LLaMA-13B 在采用不同参数高效微调方法时的输出。

问题	一位天文学家观察到，一颗行星在遭受陨石撞击后自转速度加快。这种增加最可能产生什么这种旋转增量的影响? 答案1：行星密度会降低。答案2：行星年会变长。答案3： 行星日会变短。答案4：行星引力会增强。
ChatGPT	答案3
$LLaMA_{13B}$-前缀	正确答案是答案3
$LLaMA_{13B}$-串行	正确答案是答案2
$LLaMA_{13B}$-并行	正确答案是答案3
$LLaMA_{13B}$-LoRA	正确答案是答案3

表22：从 ARC-c 中随机抽取的一个示例。ChatGPT 和 LLaMA-13B 使用不同 参数高效微调方法 .

问题	一个人想开始存钱，这样就能在年底负担得起一次惬意的假期。经过查看预算和开支后，他们决定最好的存钱方式是 答案1：多打电话 答案2：不再在外面吃午餐 答案3：少用大富翁游戏币购物 答案4：和朋友共进午餐
ChatGPT	答案2：别再外出吃午餐了。
LLaMA_{13B}-前缀	正确答案是答案2
LLaMA_{13B}-串行	正确答案是答案2
LLaMA_{13B}-并行	正确答案是答案2
LLaMA_{13B}-LoRA	正确答案是答案2

表23：从 OBQA 中随机抽取的一个示例。ChatGPT 和 LLaMA-13B 使用不同参数高效微调方法时的输出。