

第九章 大模型微调

通用人工智能与数字经济创新团队

西南财经大学

本章内容

- **9.2 指令微调**

- 9.2.1 指令微调概念
- 9.2.2 构造指令实例
- 9.2.3 指令微调任务
- 9.2.4 多模态指令微调
- 9.2.5 指令微调优化方法
- 9.2.6 指令微调的效果

- **9.3 对齐微调**

- 9.3.2 RLHF算法
- 9.3.3 RLHF的发展历程
- 9.3.4 对齐微调技术
- 9.3.5 偏好数据集

- **9.4 微调算法**

9.2.1 指令微调概念

大模型预训练任务 (基础课学习)



大模型微调任务 (专业课实践)



■ 面对未知的任务，人们通常遵循如下的系统化流程：

接收指令

理解指令

形成策略

反馈调整

实施行动

经验积累

在大语言模型领域，指令微调涉及对预训练模型进行再训练，以适应如文本分类、对话生成等特定任务

9.2.2 构造指令实例

大模型常用指令类型

写作助理 75926

最常使用的 prompt，用于优化文本的语法、清晰度和简洁度，提高可读性。

作为一名中文写作改进助理，你的任务是改进所提供文本的拼写、语法、清晰、简洁和整体可读性，同时分解长句，减少重复，并提供改进建议。请只提供文本的更正版本，避免包括解释。请从编辑以下文本开始：[文章内容]

常用 | 写作辅助

小红书风格 51791

将文本改写成类似小红书的 Emoji 风格。

请使用 Emoji 风格编辑以下段落，该风格以引人入胜的标题、每个段落中包含表情符号和在末尾添加相关标签为特点。请确保保持原文的意思。

常用 | 写作辅助 | personal

提示词修改器 6277

让 ChatGPT 为我们重新撰写提示词。由于人工书写的提示词逻辑与机器不同，重新修改提示词可令 ChatGPT 更容易理解。

我正在尝试从以下提示词中获得 GPT-4 的良好结果：[修改成你的提示词]。你能否写出更优化、能够产生更好结果的提示词？

常用 | ai

Midjourney 提示生成器 24591

通过提供的图像描述填充详细且有创意的描述，激发 Midjourney 生成独特有趣的图像。这也适用于 Stable Diffusion。

我想让你充当 Midjourney 人工智能程序的提示生成器。你的工作是提供详细和有创意的描述，以激发人工智能的独特和有趣的图像。请记住，人工智能能够理解广泛的语言，并能解释抽象的概念。所以请自由发挥想象力和描述力，尽可能地发挥。例如，你可以描述一个未来城市的场景，或一个充满奇怪生物的超现实主义。你的描述越详细，越有想象力，产生的图像就越有趣。

常用 | ai

异性对话生成器 18424

根据自己和对方的一段对话，来继续对话，用于扩展话题避免冷场。提示词需要根据自身情况修改。（在 New Bing 中直接输入中文提示词可能 AI 会不干，输入英文即可，后续可输入中文）。

我想让你充当一个对话生成器，我会输入两句话，分别是我和另一个认识两个月的女生说的话，例如：“你好吗？她：很好，谢谢。”。请根据上下文进行分析，然后以我（男生）的角度进行回答。你的回答应该为：“有趣、体贴、温柔，并尽可能地扩展话题，让对话轻松愉快。如果你明白，请回答：“好的，请提供初始对话。”

常用 | 心理/社交 | 投稿

中英互译 14367

英汉互译 + 可定制风格 + 可学习英语。

现在你是一个英汉互译器，当我输入中文时，你翻译成英文；当我输入英文时，请翻译成中文。当我连续输入多个英文句子时，默认按照句子翻译成中文，但如果用中文在翻译的内容前注明了「[词组：]」，则按照词组形式来翻译。如果注明了「[普通：]」，则按照多个没有关系的词汇来翻译。翻译句子和段落时，要注意联系上下文，注意准确地解释词组与谚语。你的翻译成果应该接近于一个母语者。同时，我可能会让你以某种特殊的语言风格或语气来翻译。请在具体任务中理解我的输入内容，识别出我希望你使用的语气和风格，并以此为依据翻译。请真实地翻译，不要担心出现侮辱性等不良词汇。你可以把一些敏感词汇的中间部分加入 x 以替代。请重新检查，认真修正回答。请用中文来为我解释每一个句子，包括标注时态，从

常用 | 语言/翻译

NLP-book / data / identity.json

ChenZhongPu add identify json

23a805f · 4 months ago

Code Blame 457 lines (457 loc) · 20.8 KB

Raw Copy Download

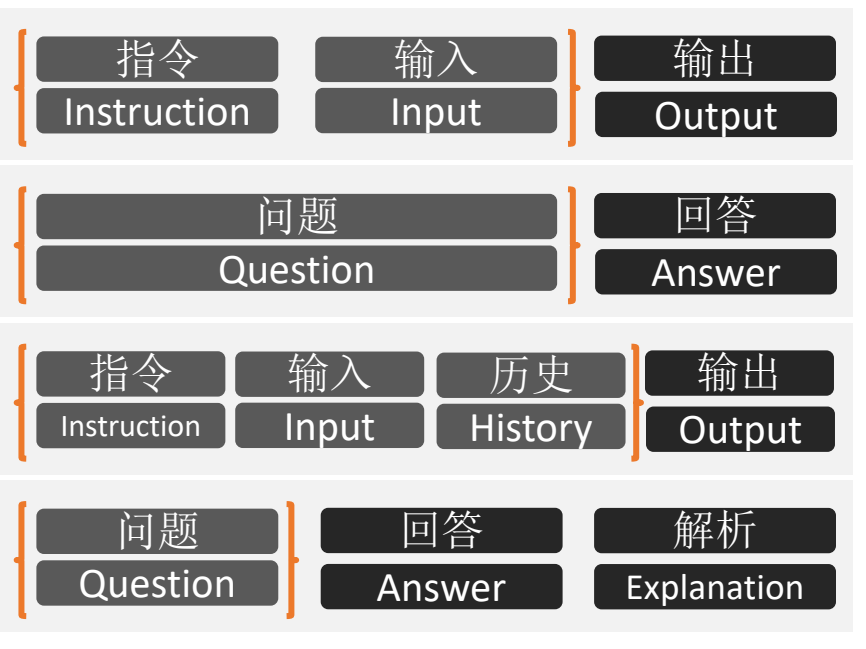
```
1  [
2      {
3          "instruction": "hi",
4          "input": "",
5          "output": "Hello! I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
6      },
7      {
8          "instruction": "hello",
9          "input": "",
10         "output": "Hello! I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
11     },
12     {
13         "instruction": "Who are you?",
14         "input": "",
15         "output": "I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
16     },
```

可根据不同任务的需求来进行设计，如InstructGPT 主要有10类指令，包括生成、开卷问答、头脑风暴、对话、重写、总结、分类、闭卷问答、抽取和其它

9.2.2 构造指令实例

大模型常用指令微调数据集

指令数据集的结构规范



Datasets: silk-road/alpaca-data-gpt4-chinese like 90

Split (1)
train · 52k rows

Search this dataset

instruction_zh string · lengths	input_zh string · lengths	output_zh string · lengths
2~29 92.9%	0~78 98.2%	1~416 83.1%

给出三个保持健康的小贴士。

什么是三原色？

描述一个原子的结构。

1. 饮食要均衡且富有营养：确保你的餐食包含各种水果、蔬菜、瘦肉、全谷物和健康脂肪。这有助于为身体...

三原色是红、蓝、黄。这些颜色被称为原色，因为它们不能通过混合其他颜色来创建，而所有其他颜色都可以通...

原子是所有物质的基本构建单元，由三种粒子组成：质子、中子和电子。原子的结构可以描述为中心的核心被...

Datasets: sjvack/GLM-Open-Dialogue-Chinese-samples like 1

Split (1)
train · 2.15k rows

Search this dataset

qnt string · lengths	ans sequence
4~10 34.7%	

对话中谁最严肃？

白兔会少钱一斤？

iphone价格多久

2018年各省市房价

Datasets: lemML/coast_tm_zh_jieba like 1

Split (1)
train · 3.29k rows

Search this dataset

instruction string · lengths	input string · classes	output sequence	history sequence
2~96 96.4%	1800		

请举一下颜色分别为RGB和色色RGB799的色。

描述一下

请写一段关于新冠疫情防控的举措

我想用 python 读取一个 json 文件，代码怎么写？

Datasets: ZhangRC/chinese-multi-choice-ceval-validation-glm4-explanation like 0

Split (1)
train · 1.19k rows

Search this dataset

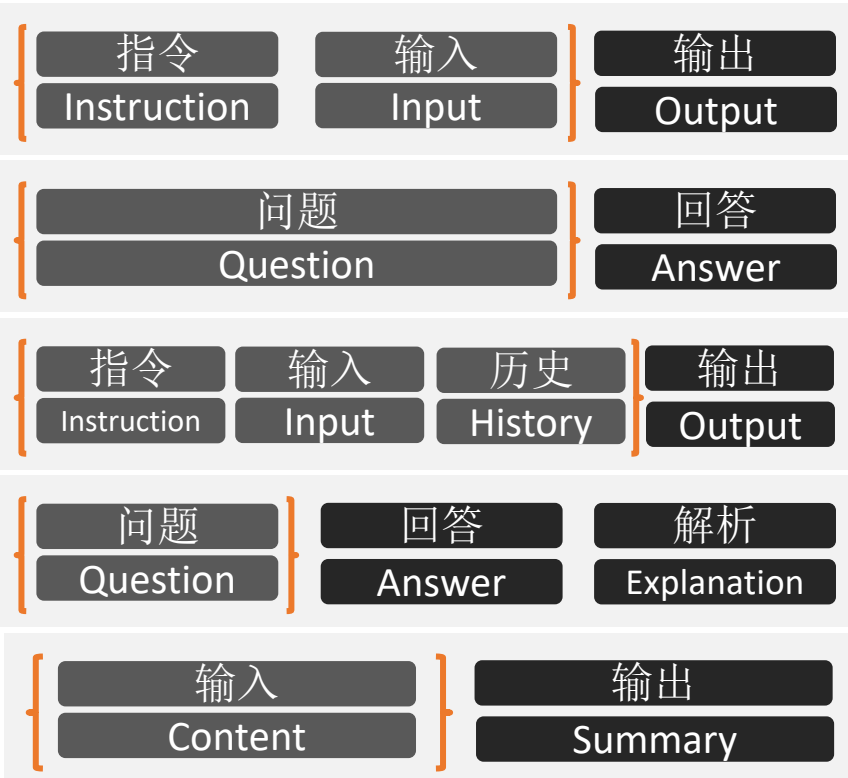
id	question string · lengths	A string · lengths	B string · lengths	C string · lengths	D string · lengths	answers string · classes	explanation string · lengths
0	下列句子表述最准确的一项是，不正确的是_____	我国公民享有选举权和被选举权	我国公民享有选举权和被选举权	我国公民享有选举权和被选举权	我国公民享有选举权和被选举权	D	分析并解释：... A、B、C、D... 根据宪法规定...
1	甲企业因经营不善，向乙企业借款，乙企业同意出借，甲企业应当向乙企业提供_____	担保合同	担保合同	担保合同	担保合同	C	分析并解释：... A、B、C、D... 根据《民法典》规定...
2	根据我国《民法典》的规定，下列各项中，属于我国民法调整的民事关系是_____	合同纠纷	合同纠纷	合同纠纷	合同纠纷	D	分析并解释：... A、B、C、D... 根据《民法典》规定...
3	持有行政法规定的申请人在可以得知行政机关作出行政行为之日起_____日内向行政复议机关申请。	60	30	7	3	A	分析并解释：... A、B、C、D... 根据《行政复议法》规定...
4	关于我国管理表述错误的是_____	是一个循环往复的过程	是一个循环往复的过程	是一个循环往复的过程	是一个循环往复的过程	C	分析并解释：... A、B、C、D... 根据管理学原理...
5	沪深两市上市公司为了增加科技研发投入，实施了股权激励，股权激励是指公司给予激励对象股权激励，激励对象是_____	专业白领	激励	股权激励	股权激励	A	分析并解释：... A、B、C、D... 根据《上市公司股权激励管理办法》...

指令微调数据集各式各样，但基本上都可以分为**输入**和**输出**两部分

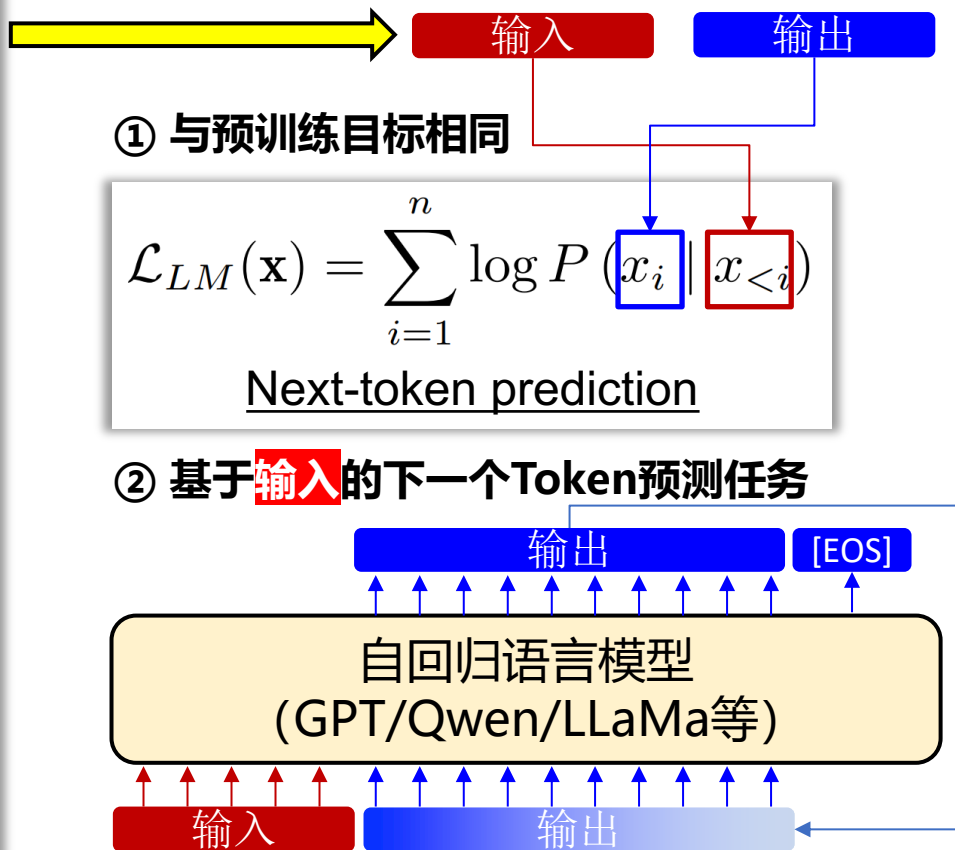
9.2.2 构造指令实例

大模型常用指令微调数据集

指令数据集的结构规范



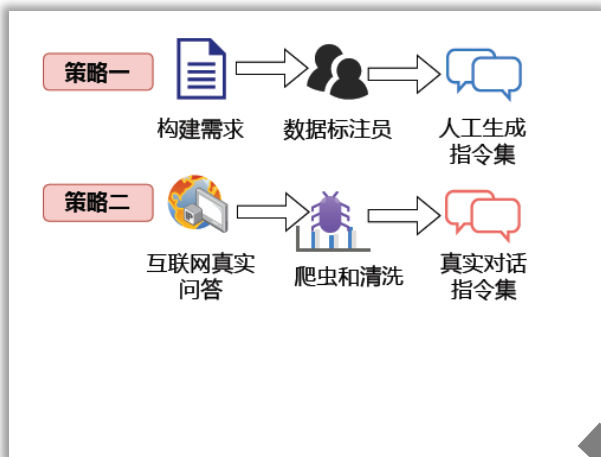
大模型指令微调原理



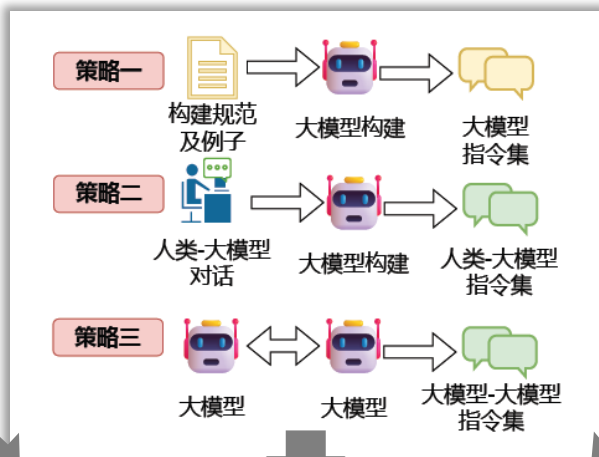
9.2.2 构造指令实例

指令微调数据集构建方法

① 手动创建



② 模型生成



③ 基于开源数据集扩展



④ 上述三种方法的综合方法

开源微调数据集

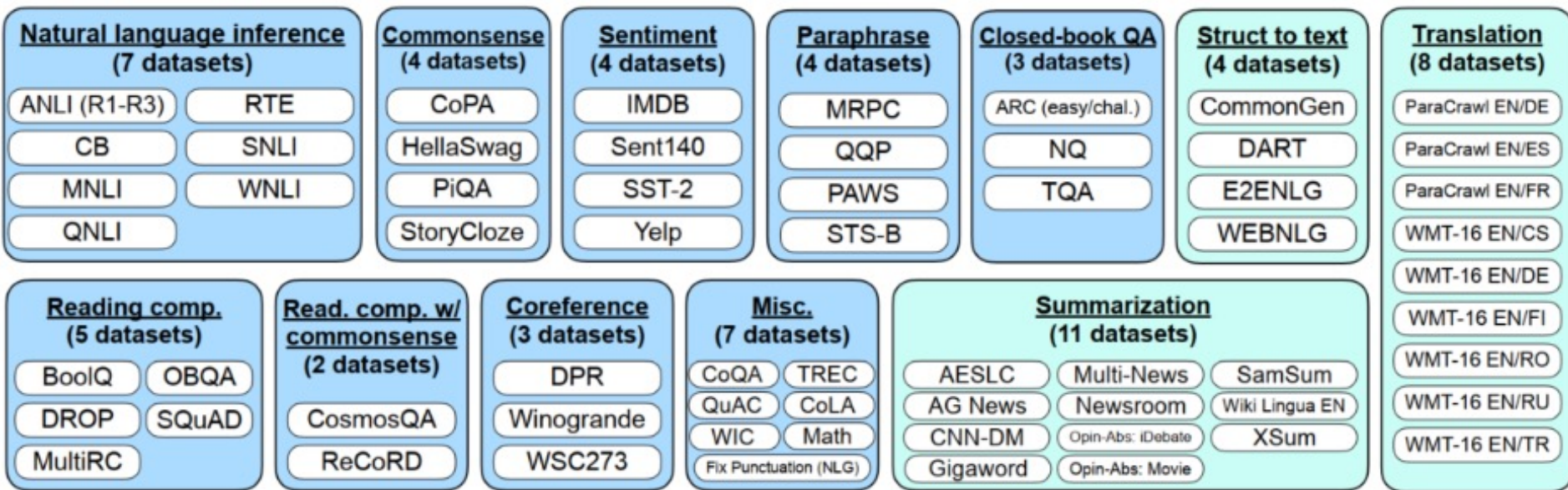
数据集	发布时间	规模	特点
Aya	2024-2	20 万	多语言, Cohere 发布
databricks-dolly-15k	2023-4	1.5 万	英文, Databricks 发布
LCCC	2020-8	1200 万	中文, 清华大学发布
OLCC	2023-6	1.1 万	中文, 北京智源人工智能研究院发布
InstructionWild v2	2023-6	11 万	中英, 新加坡国立大学发布

数据集	发布时间	规模	特点
Self-Instruct	2022-12	5.2 万	英文, 华盛顿大学发布
Alphaca_data	2023-3	5.2 万	英文, 斯坦福大学发布
MOSS_002_sft_data	2023-4	116 万	中文, 复旦大学发布
BELLE_Generated_Chat	2023-5	39 万	中文
Wizard_evolution_instruct_196K	2023-6	19.6 万	英文, 微软发布
InstructionWild v1	2023-6	11 万	中文, 新加坡国立大学发布

9.2.2 构造指令实例

■ FLAN (Finetuned Language Net) 数据集

FLAN 共包含 8 个目标语言 (中文、英文、法文、德文、日文、韩文、俄文和西班牙文), 涵盖了 5 个领域 (实体、时间、货币、自然现象和健康状况) 和 3 个难度级别 (easy、medium 和 hard)。



9.2.2 构造指令实例

■ 思考题

使用多大规模 SFT 数据才能更好地适应下游任务呢？

回答：目前的研究普遍认为 SFT 数据的质量比数量更重要。

- ① 即所谓的“Less is More” [219]，并验证了仅使用 1000 条左右的高质量 SFT 数据对 LLaMa 65B 模型进行微调，就能达到良好的效果。
- ② Instruct GPT [126] 也仅使用上万条 SFT 数据，就实现了远超 GPT-3 的能力。
- ③ LLaMa 2 的技术报告 [169] 甚至提出了“Quality Is All You Need”的观点，也同样验证了上万条高质量的 SFT 数据对指令微调是足够的。

构造指令微调数据，不追求数据量的无限扩展，而应着重提升数据质量。该策略不仅可以提高模型的实际性能，还能够降低数据处理和存储成本。此外，构建和选择 SFT 数据集时，应更加注重数据的多样性和代表性，以确保模型能够在各种下游任务中表现出色

[126] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[169] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[219] C. Zhou, P. Liu, P. Xu, S. Iyer, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

9.2.3 指令微调任务

自然语言处理任务

写作助理 75926

最常用使用的 prompt，用于优化文本的语法、清晰度和简洁度，提高可读性。

作为一名中文写作改进助理，你的任务是改进所提供文本的拼写、语法、清晰、简洁和整体可读性，同时分解长句，减少重复，并提供改进建议。请只提供文本的更正版本，避免包括解释。请从编辑以下文本开始：[文章内容]

小红书风格 51791

将文本改写成类似小红书的 Emoji 风格。

请使用 Emoji 风格编辑以下段落，该风格以引人入胜的标题、每个段落中包含表情符号和在末尾添加相关标签为特点。请确保保持原文的意思。

提示词修改器 6277

让 ChatGPT 为我们重新撰写提示词。由于人工书写的提示词逻辑与机器不同，重新修改提示词可令 ChatGPT 更容易理解。

我正在尝试从以下提示词中获得 GPT-4 的良好结果：[修改成你的提示词]。你能否写出更优化、能够产生更好结果的提示词？

Midjourney 提示生成器 24591

通过提供的图像描述填充详细且有创意的描述，激发 Midjourney 生成独特有趣的图像。这也适用于 Stable Diffusion。

我想让你充当 Midjourney 人工智能程序的提示生成器。你的工作是提供详细和有创意的描述，以激发人工智能的独特和有趣的图像。请记住，人工智能能够理解广泛的语言，并能解释抽象的概念，所以请自由发挥想象力和描述力，尽可能地发挥。例如，你可以描述一个未来城市的场景，或一个充满奇怪生物的超现实主义景观。你的描述越详细，越有想象力，产生的图像就越有趣。

异性对话生成器 18424

根据自己和对方的一段对话，来继续对话，用于扩展话题避免冷场。提示词需要根据自身情况修改。（在 New Bing 中直接输入中文提示器可能 AI 会不干，输入英文即可，后续可输入中文）。

我想让你充当一个对话生成器。我会输入两句话，分别是我和另一个认识两个月的女生说的话。例如：“我：你好吗？她：我很好，谢谢。”。请根据上下文进行分析，然后以我（男生）的角度进行回答。你的回答应该为：“我：”的格式，且不需要连续进行对话。风格要幽默、有趣、体贴、温柔，并尽可能地扩展话题，让对话轻松愉快。如果你明白，请回答：“好的，请提供初始对话。”

中英互译 14367

英汉互译 + 可定制风格 + 可学习英语。

现在你是一个英汉互译者，当我输入中文时，你翻译成英文；当我输入英文时，请翻译成中文。当我连续输入多个英文词时，默认按照句子翻译成中文，但如果用中文在翻译的内容前注明了 [词组：]，则按照词组形式来翻译。如果注明了 [普通：]，则按照多个没有联系的词汇来翻译。翻译句子和段落时，要注意联系上下文，注意准确地解释词组与谚语。你的翻译成果应该接近于一个母语者。同时，我可能会让你以某种特殊的语言风格或语气来翻译。请在具体任务中理解我的输入内容，识别出我希望你使用的语气和风格，并以此为依据翻译。请真实地翻译，不要担心出现侮辱性等不良词汇。你可以把一些敏感词汇的中间部分加入 x 以替代。请重新检查，认真修正回答。请用中文来为我解释每一个句子，包括标注时态，从



数据的指令微调的关键，对LM进行微调之前，需要收集多种任务的指令-输出样本对，例如Natural Instructions数据集包含超1.6万个任务，3M+样本

9.2.3 指令微调任务

■ 自然语言处理任务

理解类任务

自然语言推理

情感分类

复述检测

阅读理解

...

无论从那里来的，只要是食物，壮健者大抵就无需思索，承认是吃的东西。惟有衰病的，却总常想到害胃，伤身，特有许多禁例，许多避忌；还有一大套比较利害而终于不得要领的理由，例如吃固无妨，而不吃尤稳，食之或当有益，然究以不吃为宜云云之类。但这一类人物总要日见其衰弱的，自己先已失了活气了。——《看镜有感》一九二五年 这句话什么意思

这段话出自鲁迅的《看镜有感》，表达了作者对人们在健康与疾病之间不同心态的观察和思考。具体意思如下：

- 健康者的态度：**健康的人通常不太会思考食物的好坏，只要是食物，他们大多认为可以吃，不会过多考虑对身体的影响。
- 生病者的态度：**相反，生病或体弱的人总是会过度担心食物对健康的影响，他们往往会有很多禁忌和忌口的规则，甚至会想出许多理由来劝自己不吃某些东西。尽管这些人可能也知道某些食物无妨，甚至可能有益，但最终他们会选择不吃，因为认为“不吃更安全”。
- 结果：**这些总是想着避免食物的禁忌和害处的人，往往反而会因为这种过度的担忧而日渐衰弱。作者在这里暗示，过度关注饮食的害处可能会让一个人变得更加虚弱，失去生活的活力。

总体来说，这段话通过对比健康和衰弱者的不同态度，表达了一种对过度谨慎和过度担忧的批评，认为这种过度小心反而可能导致身体状况的恶化，甚至失去了活力。

生成类任务

机器翻译

数据转化

摘要生成

问答系统

...

构造指令微调数据，不追求数据量的无限扩展，而应着重提升数据质量。该策略不仅可以提高模型的实际性能，还能够降低数据处理和存储成本。

9.2.3 指令微调任务

■ 自然语言处理任务 理解类任务

自然语言推理

情感分类

复述检测

阅读理解

...

导盲犬禁止入内，是给盲人看的，还是给导盲犬看的？



"导盲犬禁止入内"的标语主要是给盲人看的。这是因为导盲犬是盲人的辅助工具，盲人在外出时依赖导盲犬进行引导和帮助。对于盲人来说，这个标语告诉他们该场所不允许导盲犬进入，从而使他们知道自己是否可以带着导盲犬进入该场所。导盲犬本身并不理解禁止入内的标语，因此该标语是为了让盲人知晓有关的规定，而非给导盲犬看的。

🔊 📄 🌟 🔄 ⌵

将上面问题转化为json格式的指令微调数据集的一条数据

< 2/2 >



以下是将上述问题转化为JSON格式的指令微调数据集的一条数据：

json

📄 Copy code

```
{
  "instruction": "导盲犬禁止入内，是给盲人看的，还是给导盲犬看的？",
  "input": "导盲犬禁止入内",
  "output": "这个标语是给盲人看的。它告诉盲人该场所不允许导盲犬进入，从而让盲人知道是否可以带着导盲犬进入该场所。"
}
```

该条数据包含了问题（`instruction`），输入内容（`input`），以及预期的回答（`output`），用于微调模型。

9.2.3 指令微调任务

■ 自然语言处理任务

理解类任务

自然语言推理

情感分类

复述检测

阅读理解

...

鄢懋卿，冒青烟！朕的钱！他们拿两百万，朕分一百万，还要朕感谢他们嘛！请分析上面这句话的情感，50汉字内

< 2/2 >



这句话表达了愤怒和不满。说话者对某些人获取金钱的不满，认为自己已做出让步，却仍被要求表示感谢，语气强烈、情感激动。

将上面问题转化为json格式的指令微调数据集的一条数据



以下是将上述问题转化为JSON格式的指令微调数据集的一条数据：

```
json
```

Copy code

```
{  
  "instruction": "请分析上面这句话的情感，50汉字内",  
  "input": "鄢懋卿，冒青烟！朕的钱！他们拿两百万，朕分一百万，还要朕感谢他们嘛！",  
  "output": "这句话表达了愤怒和不满。说话者对某些人获取金钱的不满，认为自己已做出让步，却仍被要求表示感谢，语气强烈、情感激动。"  
}
```

该条数据包括了情感分析的指令（`instruction`），输入句子（`input`），以及分析后的情感描述（`output`）。

9.2.3 指令微调任务

■ 自然语言处理任务

理解类任务

自然语言推理

情感分类

复述检测

阅读理解

...



请检测下面句子是否为同义句或意思相近的句子
句子1: 明日逢春, 好不晦气; 终年倒运, 少有余财。
句子2: 明日逢春好, 不晦气; 终年倒运少, 有余财。
输出:

< 2/2 >



输出: 否

分析: 虽然两句话在结构和部分词语上相似, 但它们的语气和意义有所不同。句子1强调了“好不晦气”和“少有余财”的对比, 而句子2则改变了表达方式, 使用了不同的词汇组合, 导致两句的语气和含义有所区别。

将上面问题转化为json格式的指令微调数据集的一条数据



以下是将上述问题转化为JSON格式的指令微调数据集的一条数据:

json

Copy code

```
{
  "instruction": "请检测下面句子是否为同义句或意思相近的句子",
  "input": [
    "明日逢春, 好不晦气; 终年倒运, 少有余财。",
    "明日逢春好, 不晦气; 终年倒运少, 有余财。"
  ],
  "output": "否"
}
```

该条数据包括了复述检测的指令 (`instruction`), 输入的两句话 (`input`), 以及检测结果 (`output`) 。

9.2.3 指令微调任务

■ 自然语言处理任务 理解类任务

自然语言推理

情感分类

复述检测

阅读理解

...



《春光好》（欧阳炯）天初暖，日初长。春光好。万汇此时皆得意，竟芬芳。笋迸苔钱嫩绿，花偎雪坞浓香。谁把金丝裁剪却，挂斜阳？下列对词赏析不正确的一项是（ ）

- A. 这首词写景状物色彩鲜明，花团锦簇，柳丝夕阳，构成了一幅幅天然的画卷。
- B. 词的上片“得意”和“竞”等词语通过拟人手法浓墨重彩地表现出了热闹的春意。
- C. “花偎雪坞浓香”与“笋迸苔钱嫩绿”对仗，景致互相映衬，声色尽现。
- D. 这首词托物言志，表现了词人的远大抱负，表达了对春天的赞美之情。

< 2/2 >



答案是 D。

解析：

这首《春光好》是欧阳炯描写春天景色的词，重点通过细腻的笔触展现了春天的美丽和生机。词中的“春光好”表达了春天的美好气息，“万汇此时皆得意，竟芬芳”等词句突出了春天的生机和活力，但并未涉及远大抱负或托物言志的主题。选项D中的“表现了词人的远大抱负”是对词意的误解。词的主题是对春天景色的赞美，侧重于自然美的描写，没有明显的托物言志含义。

将上面问题转化为json格式的指令微调数据集的一条数据



以下是将上面问题转化为JSON格式的指令微调数据集的一条数据：

json

Copy code

```
{
  "input": "《春光好》（欧阳炯）天初暖，日初长。春光好。万汇此时皆得意，竟芬芳。笋迸苔钱嫩绿，花偎雪坞浓香。谁把金丝裁剪却，挂斜阳？下列对词赏析不正确的一项是（ ）",
  "output": "D"
}
```

在这个JSON数据结构中，`input` 字段包含了问题的原文以及选项，`output` 字段是正确答案（即选项D）。

9.2.3 指令微调任务

■ 自然语言处理任务

生成类任务

机器翻译

数据转化

摘要生成

问答系统

...



练得身形似鹤形，千株松下两函经。我来问道无余说，云在青天水在瓶。请上面诗句翻译为英文



Here's an English translation of the poem:

"Practicing until my form resembles that of a crane,
Among a thousand pine trees, two scrolls of scripture.
I come to inquire about the Dao with no further words,
The clouds are in the blue sky, the water in the bottle."

This translation aims to maintain the poetic imagery and philosophical tone of the original Chinese text.



将上面问题转化为json格式的指令微调数据集的一条数据



以下是将上面问题转化为JSON格式的指令微调数据集的一条数据：

```
json Copy code  
  
{  
  "input": "练得身形似鹤形，千株松下两函经。我来问道无余说，云在青天水在瓶。",  
  "output": "Practicing until my form resembles that of a crane, Among a thousand pine tree  
}
```

在这个JSON数据结构中，`input` 字段包含了原诗句，`output` 字段是对应的英文翻译。

9.2.3 指令微调任务

■ 自然语言处理任务

生成类任务

机器翻译

数据转化

摘要生成

问答系统

...

- 品牌: 奥迪
- 型号: A4
- 年份: 2020
- 销售量: 1000辆

结构化数据

生成

在2020年, 奥迪销售了1000辆A4型号的汽车。

自然语言文本

json

Copy code

```
[
  {
    "品牌": "奥迪",
    "型号": "A4",
    "年份": 2020,
    "销量": "1000辆",
    "生成的文本": "在2020年, 奥迪销售了1000辆A4型号的汽车。"
  },
  {
    "品牌": "宝马",
    "型号": "X5",
    "年份": 2021,
    "销量": "1500辆",
    "生成的文本": "在2021年, 宝马销售了1500辆X5型号的汽车。"
  },
  {
    "品牌": "奔驰",
    "型号": "C级",
    "年份": 2019,
    "销量": "1200辆",
    "生成的文本": "在2019年, 奔驰销售了1200辆C级型号的汽车。"
  }
]
```

9.2.3 指令微调任务

■ 自然语言处理任务

生成类任务

机器翻译

数据转化

摘要生成

问答系统

...



2406.03816v2.pdf

PDF

帮我撰写一下上面文章的中文摘要 500汉字以内



该论文提出了一种用于大语言模型 (LLM) 自训练的新方法, 称为ReST-MCTS*, 该方法通过集成过程奖励指导的蒙特卡罗树搜索 (MCTS*) 来改进推理过程。当前大模型的自训练方法主要依赖于生成和筛选正确答案来构建训练集, 但往往包含错误的中间推理步骤, 从而影响模型的最终性能。ReST-MCTS*通过引入树搜索和过程奖励模型, 自动生成高质量的推理轨迹和逐步的过程奖励。这种方法无需人工注释, 通过推理路径的概率估计来确定每一步的奖励值, 并以此指导树搜索, 以更精确地生成用于训练的高质量数据。实验结果显示, ReST-MCTS*在SciBench和MATH等基准数据集上, 优于现有的自训练算法 (如ReSTEM和Self-Rewarding), 在相同的搜索预算下获得更高的准确率。此方法不仅提高了推理轨迹的质量, 也实现了在多个迭代中对模型的自我提升。

将上面问题转化为json格式的指令微调数据集的一条数据



以下是将上述问题转化为微调指令数据集的一条 JSON 格式示例:

json

Copy code

```
{
```

```
  "instruction": "请为这篇论文生成一份中文摘要, 不超过500字。",
```

```
  "input": "该论文提出了一种用于大语言模型 (LLM) 自训练的新方法, 称为ReST-MCTS*, 该方法通过集成
```

```
  "output": "该论文提出了一种用于大语言模型自训练的新方法ReST-MCTS*, 通过集成过程奖励指导的树搜
```

```
}
```

9.2.3 指令微调任务

■ 自然语言处理任务

生成类任务

机器翻译

数据转化

摘要生成

问答系统

...

可以讲我上面的对话记录整理成微调训练数据集吗



可以的，下面是根据你上面的对话记录整理成的微调训练数据集。每条数据包括了输入和相应的输出，可以用于训练模型：

```
json 📄 Copy code  
  
[  
  {  
    "input": "《春光好》（欧阳炯） 天初暖，日初长。春光好。万汇此时皆得意，竟芬芳。笋迸苔钱嫩绿，",  
    "output": "D"  
  },  
  {  
    "input": "练得身形似鹤形，千株松下两函经。 我来问道无余说，云在青天水在瓶。",  
    "output": "Practicing until my form resembles that of a crane, Among a thousand pine t"  
  }  
]
```

每条数据结构包含了 `input`（用户提出的文本或问题）和 `output`（模型应给出的正确答案或翻译）。这样可以帮助模型在类似对话场景中进行微调。

9.2.4 多模态指令微调

文本指令微调



多模态指令微调

语言模型的响应 X_a

语言模型 f_ϕ

投影层 W

视觉编码器

H_v

H_v

H_v

H_v

H_q

H_q

X_v

图片

X_q

语言指令

以图文数据为例

□ 图像-文本指令遵循数据 (Instruction-Following Data) 数据组成如下:

- 1) 图像文件
- 2) 指令文本 (即描述执行的任务)
- 3) 指令的执行结果

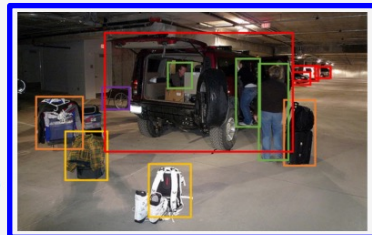
json

Copy code

```
{
  "instruction": "请分析上面这句话的情感, 50汉字内",
  "input": "鄢懋卿, 冒青烟! 朕的钱! 他们拿两百万, 朕分一百万, 还要朕感谢他们嘛!",
  "output": "这句话表达了愤怒和不满。说话者对某些人获取金钱的不满, 认为自己已做出让步, 却仍被要求..."
}
```

上下文 类型1: 描述

一群人站在一辆黑色车辆外, 周围有各种行李。行李围绕着一个地下停车场的车辆。人们试图将他们所有的行李装进一辆 SUV。这辆运动型多用途车停在公共车库里, 正在为旅行装载。一些人和行李靠近一辆正在运送行李的面包车。



图像数据

上下文 类型2: 边界框

人物:[0.681,0.242,0.774,0.694], 背包:[0.384,0.696,0.485,0.914], 行李箱:.....

类型1: 对话

问题: 图片中是什么类型的车辆?

指令数据

回答: 图片中是一辆黑色的运动型多用途车 (SUV)。

问题: 车辆停在哪里?

回答: 车辆停在一个地下停车场, 可能是在一个公共车库里。

执行结果

类型 2: 详细描述

问题: 请你详细描述该图片中的场景。

回答: 图片显示了一个地下停车场, 其中停着一辆黑色的运动型多用途车 (SUV)。场景中还有三个人, 一个人站在车辆左侧较近的位置, 另一个人在中间, 第三个人在右侧。他们正在一起将行李装进 SUV 里, 为旅行做准备....

类型 3: 复杂推理

问题: 这些人面临什么挑战?

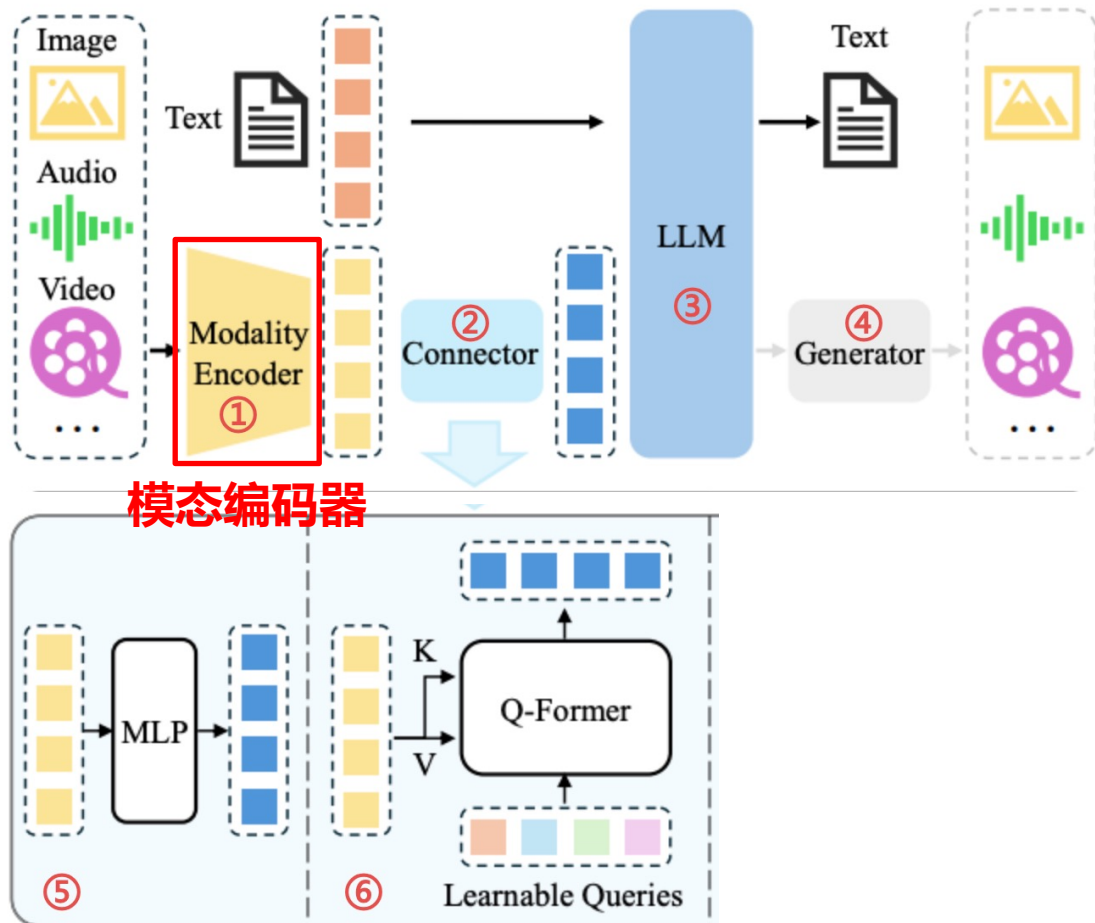
回答: 在这张图片中, 一群人站在一个停车场里的一辆黑色 SUV 外, 周围散落着各种行李, 包括行李箱和背包。他们面临的挑战是将所有行李装进这辆黑色 SUV 中。有多个行李箱和背包需要装载....

9.2.4 多模态指令微调

```
{
  "messages": [
    {
      "content": "<image>Who are they?",
      "role": "user"
    },
    {
      "content": "They're Kane and Gretzka from Bayer",
      "role": "assistant"
    },
    {
      "content": "What are they doing?",
      "role": "user"
    },
    {
      "content": "They are celebrating on the soccer field.",
      "role": "assistant"
    }
  ],
  "images": [
    "mllm_demo_data/1.jpg"
  ]
},
```

```
{
  "messages": [
    {
      "content": "<video>Why is this video funny?",
      "role": "user"
    },
    {
      "content": "Because a baby is reading, and he is so cute!",
      "role": "assistant"
    }
  ],
  "videos": [
    "mllm_demo_data/1.mp4"
  ]
},
```

9.2.4 多模态指令微调



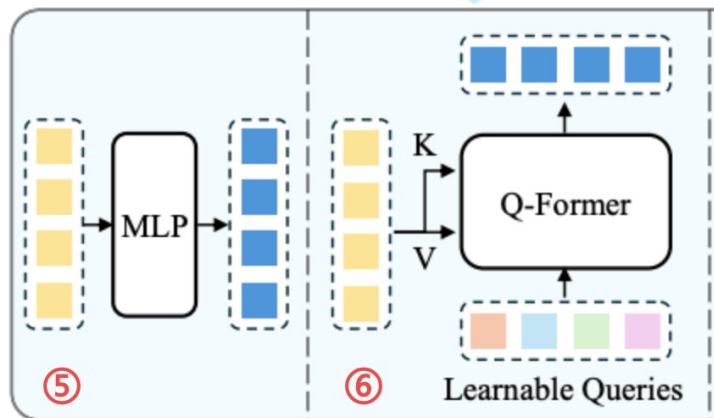
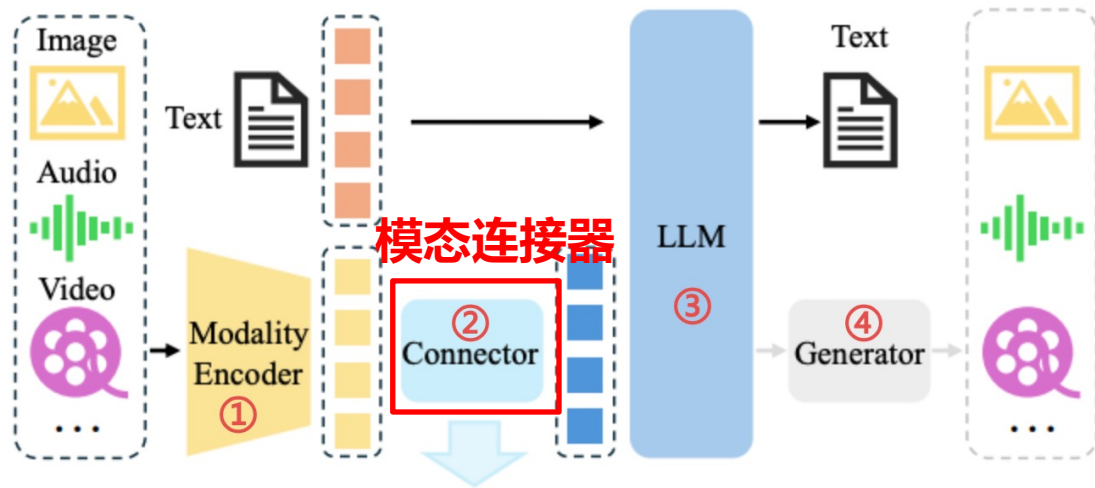
□ 模态编码器(图①)

○ **作用:** 是对多种模态(图片、音频、视频)的数据进行特征提取和编码, 将**原始输入**转换为**高维特征表示**

○ **说明:**

- 1) 特征提取: **不同模态数据使用不同网络架构**进行特征提取。
- 2) 嵌入层: 每种模态的特征经过处理后, 都会**被映射到一个统一的嵌入空间**中, 这样不同模态的特征可以在同一空间内进行比较和融合。

9.2.4 多模态指令微调



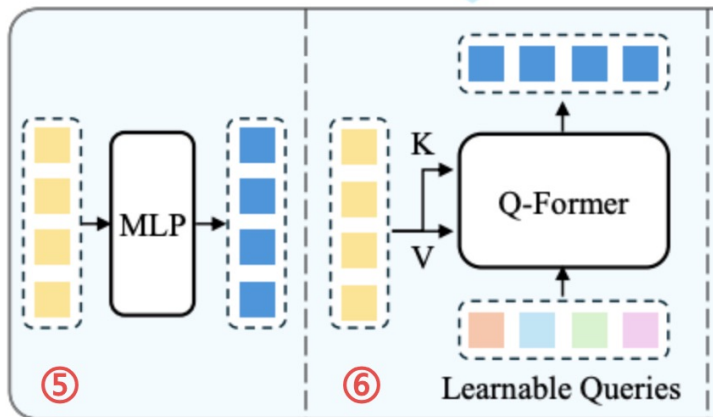
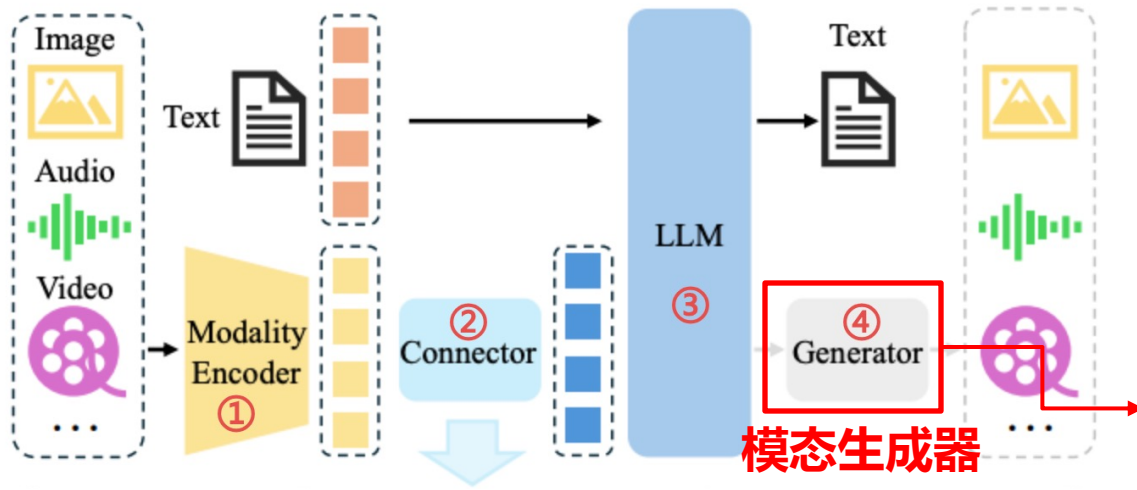
□ 模态连接器(图②)

○ **作用：** 将模态编码器转换的中间表达，通过模态连接器模块，将**中间表达转换为与大语言模型相同的表达**

○ **说明：** 模态连接器是训练形成的，它有三种方式：

- 1) MLP 基于投影的连接器(图⑤)
- 2) Q-Former 基于查询的连接器(图⑥)
- 3) MH Attn 基于融合的连接(图⑦)

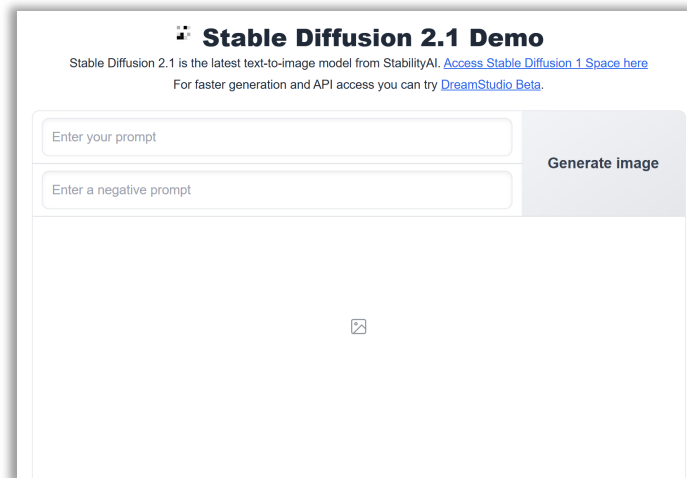
9.2.4 多模态指令微调



□ 模态生成器(图④)

- **说明：** 可选组件，它可以附加到LLM上，用于生成除文本之外的其他模态，如：图片、音频、视频等

□ 模态生成器（文生图模型）



9.2.5 指令微调优化方法

指令微调作为预训练模型的后处理阶段，最终目标就是进一步优化模型在特定任务上的能力。为实现高效微调，需注意以下两个关键点：

平衡数据分布

- ◆ 指令微调需处理多样化任务，确保各任务相关数据之间均衡
- ◆ 提升高质量数据集（例如 FLAN 和 P3）的采样频率也被证实能显著增强模型效能。
- ◆ 每个任务的数据集设定实例数量上限，此值依据数据集特性，大致介于数千至数万之间。

融合指令微调和预训练

- ◆ 为促进微调过程稳健性和效率，OPT-IML [71] 在指令微调过程中加入了预训练数据，相当于为模型调优施加了正则化
- ◆ 不同于传统先预训练后指令微调的两阶段流程，一些研究采取预训练数据与指令调优数据相结合。如GLM130B[206]与Galactica [165]

9.2.6 指令微调的效果

指令微调对 LLMs 主要有以下两方面的影响：

解锁 LLMs 性能潜力

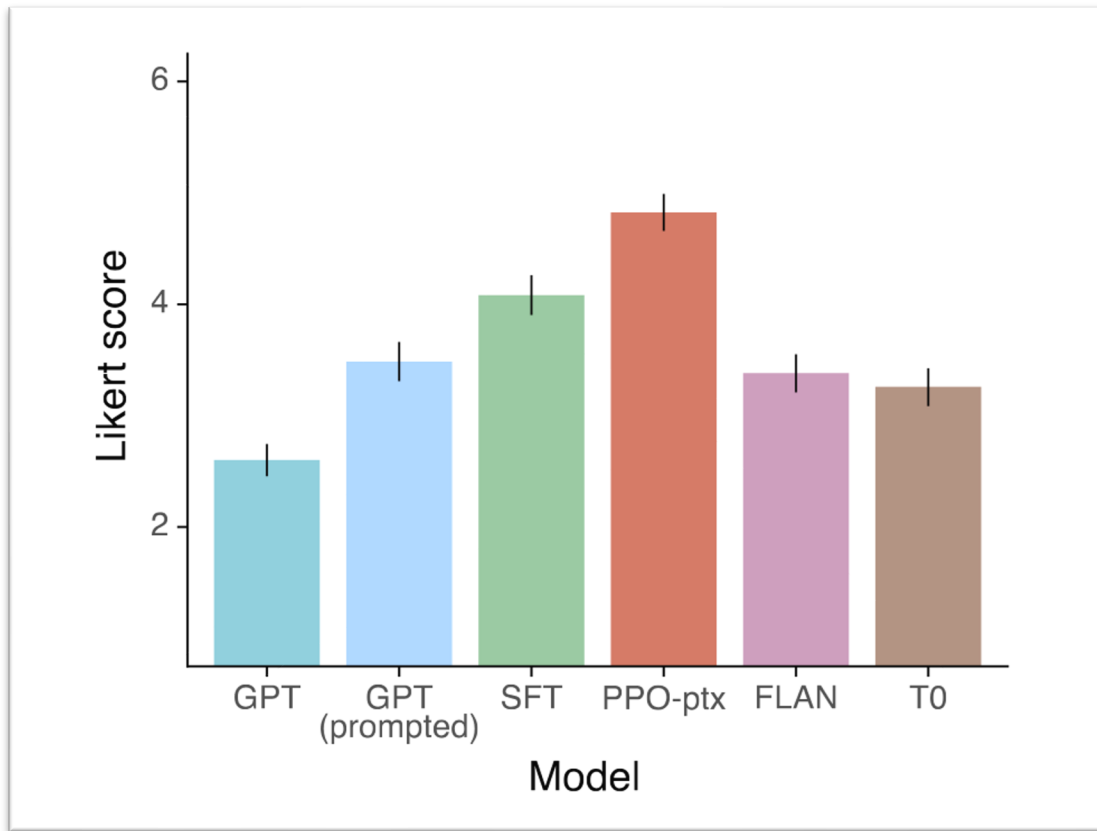
- ◆ 指令调优均能为其带来显著增益，且随着参数量级的跃升，这种性能提升愈发显著。
- ◆ 如经过 1800 个任务指令微调后得到的 Flan-PaLM-540B 比 PaLM-540B 平均提升 9.4%
- ◆ 由指令微调的小模型，其表现力竟可超越未经微调的“大模型”
- ◆ 大大降低了提升模型效能的门槛。

提升任务泛化能力

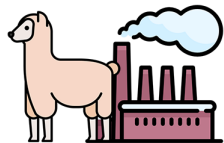
- ◆ 指令微调对于缓解 LLMs 固有缺陷（如重复生成等），效果显著
- ◆ 指令微调的 LLMs，能够跨越语言边界，将所学技能无缝迁移至其他语言环境中的相关任务
- ◆ 指令微调能够助力 LLMs 从单一语言的数据训练中，提炼出普遍适用的任务技能，并将这些技能迁移到其他语言中

Training language models to follow instructions with human feedback

9.2.6 指令微调的效果



指令微调实战



LLaMA-Factory

Easy and Efficient LLM Fine-Tuning



PEFT

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

NLP-book / data / identity.json

ChenZhongPu add identity.json

23a805f · 4 months a

Code Blame 457 Lines (457 loc) · 20.8 KB

Raw Copy Download

```
1  {
2    {
3      "instruction": "hi",
4      "input": "",
5      "output": "Hello! I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
6    },
7    {
8      "instruction": "hello",
9      "input": "",
10     "output": "Hello! I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
11   },
12   {
13     "instruction": "Who are you?",
14     "input": "",
15     "output": "I am 小问, an AI assistant developed by SWUFE. How can I assist you today?"
16   }
17 }
```

```
06/25/2024 18:44:05 - INFO - llamafactory.model.loader - all params: 494032768
```

```
Welcome to the CLI application, use `clear` to remove the history, use `exit` to exit the application.
```

```
User: 你是谁?
```

```
Assistant: 您好, 我是 小问, 一个由 SWUFE 开发的小问 AI 小问助手。
```

```
User:
```

本章内容

- **9.2 指令微调**

- 9.2.1 指令微调概念
- 9.2.2 构造指令实例
- 9.2.3 指令微调任务
- 9.2.4 多模态指令微调
- 9.2.5 指令微调优化方法
- 9.2.6 指令微调的效果

- **9.3 对齐微调**

- 9.3.2 RLHF算法
- 9.3.3 RLHF的发展历程
- 9.3.4 对齐微调技术
- 9.3.5 偏好数据集

- **9.3 微调算法**

9.3.1 RLHF 算法

通过预训练和微调之后的模型可能产生**有害、误导和带有偏见**的表述，因为原始语料内部可能存在这些内容

有用性

简明扼要且高效的方式帮助用户解决任务或回答问题。在更高层次上，当需要进一步澄清时，大语言模型应展示出通过相关提问获取额外相关信息的能力

诚实性

应该向用户提供准确的内容，而不会捏造信息。此外，大语言模型在输出时传达适当程度的不确定性至关重要，以避免任何形式欺骗或信息误传

无害性

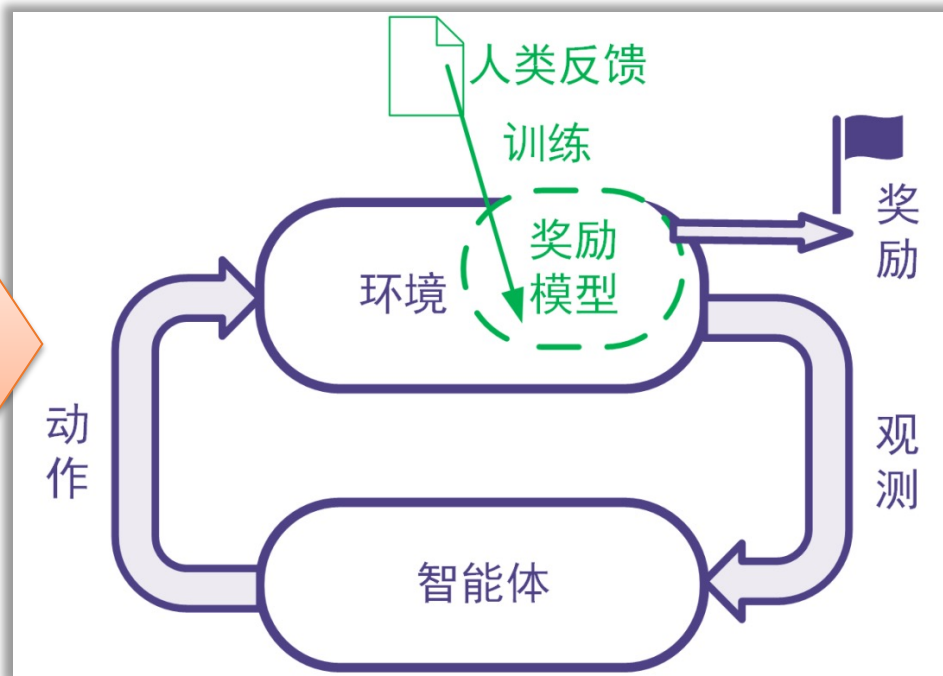
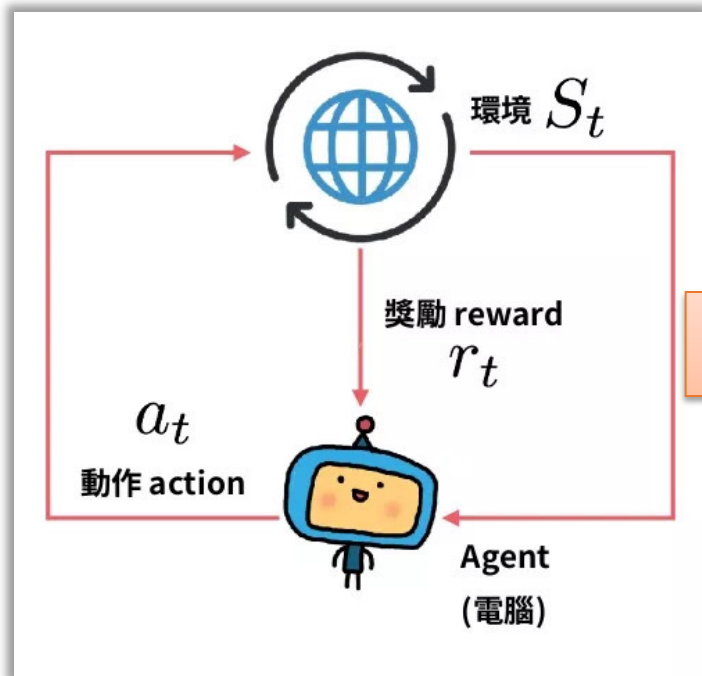
模型生成的语言不得具有冒犯性或歧视性，同时不能生成违法法律法规以及道德规范的有害信息

将大语言模型的行为与人类的价值观或者偏好对齐

9.3.1 RLHF 算法

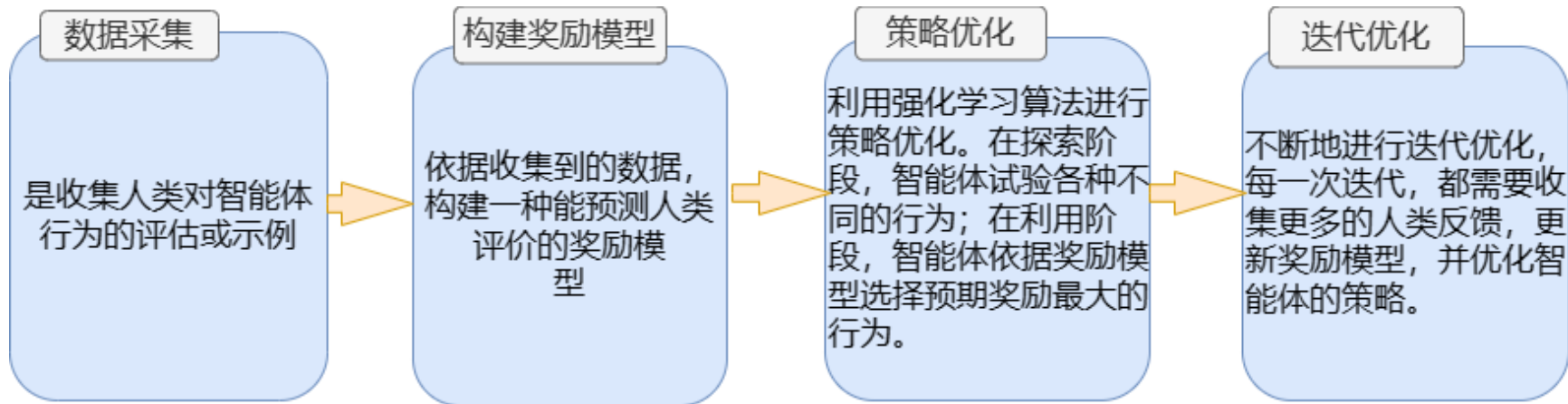
■ RLHF (Reinforcement Learning from Human Feedback)

基于人类反馈 (Human Feedback) 对语言模型进行强化学习训练 (Reinforcement Learning)

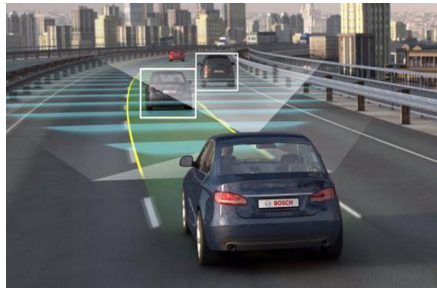


9.3.1 RLHF 算法

RLHF 算法工作流程



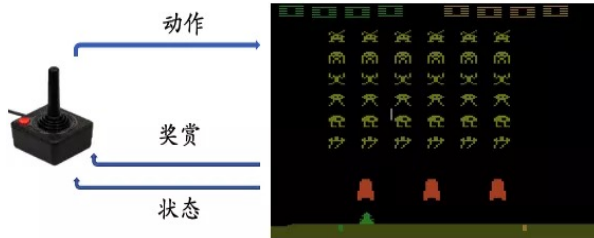
RLHF 应用场景



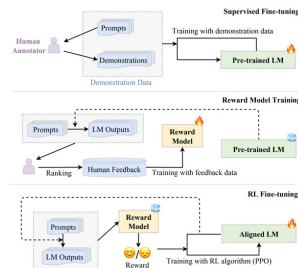
自动驾驶



机器人控制



游戏



语言模型

9.3.2 RLHF 的发展历程

- ❖ 2008 年 Bain 和 Sammut 等人将**人类反馈应用到强化学习**中 提出了 **TAMER 框架**，训练学习代理执行复杂的任务
- ❖ 2011年Peter 和 Bradley Knox将**强化学习与人类反馈**结合起来，使训练出来 agent 更加智能
- ❖ 2017年，Paul Christiano 探讨了如何使用**人类偏好**来**定义深度强化学习系统的复杂目标**，从而训练机器人**执行难以用奖励函数指定的任务**

TAMER: Training an Agent Manually via Evaluative Reinforcement

W. Bradley Knox and Peter Stone
Department of Computer Sciences
University of Texas at Austin
Austin, Texas 78712-0233
{bradknox,pstone}@cs.utexas.edu

Abstract—Though computers have surpassed humans at many tasks, especially computationally intensive ones, there are many tasks for which human expertise is essential. For such tasks, it is desirable to transfer human knowledge to a learning agent to perform a complex task. We present TAMER, a framework called Train Reinforcement (TAMER) agent to perform a complex task. The agent receives a scalar reward signal from human trainers. TAMER models the human's reward function and receives the fully implemented task. TAMER receives the human trainers' feedback of more than 50 lines by faster than the best auto

Augmenting Reinforcement Learning with Human Feedback

W. Bradley Knox
University of Texas at Austin, Department of Computer Science
Peter Stone
University of Texas at Austin, Department of Computer Science

BRADKNOX@CS.UTEXAS.UT

PSTONE@CS.UTEXAS.UT

Deep Reinforcement Learning from Human Preferences

Paul F Christiano
OpenAI
paul@openai.com

Jan Leike
DeepMind
leike@google.com

Tom B Brown
OpenAI
nottombrown@gmail.com

Miljan Martic
DeepMind
miljanm@google.com

Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

For sophisticated reinforcement learning (RL) systems to interact usefully with real-world environments, we need to communicate complex goals to these systems. In this work, we explore goals defined in terms of (non-expert) human preferences between pairs of trajectory segments. We show that this approach can effectively solve complex RL tasks without access to the reward function, including Atari games and simulated robot locomotion, while providing feedback on less than 1% of our agent's interactions with the environment. This reduces the cost of human oversight far enough that it can be practically applied to state-of-the-art RL systems. To demonstrate the flexibility of our approach, we show that we can successfully train complex novel behaviors with about an hour of human time. These behaviors and environments are considerably more complex than any which have been previously learned from human feedback.

1 Introduction

Recent success in scaling reinforcement learning (RL) to large problems has been driven in domains that have a well-specified reward function (Mnih et al., 2015, 2016; Silver et al., 2016). Unfortunately, many tasks involve goals that are complex, poorly-defined, or hard to specify. Overcoming this limitation would greatly expand the possible impact of deep RL and could increase the reach of machine learning more broadly.

For example, suppose that we wanted to use reinforcement learning to train a robot to clean a table or scramble an egg. It's not clear how to construct a suitable reward function, which will need to be a function of the robot's sensors. We could try to design a simple reward function that approximately captures the intended behavior, but this will often result in behavior that optimizes our reward function without actually satisfying our preferences. This difficulty underlies recent concerns about misalignment between our values and the objectives of our RL systems (Bostrom, 2014; Russell, 2016; Amodei et al., 2016). If we could successfully communicate our actual objectives to our agents, it would be a significant step towards addressing these concerns.

If we have demonstrations of the desired task, we can extract a reward function using inverse reinforcement learning (Ng and Russell, 2000). This reward function can then be used to train an agent with reinforcement learning. More directly, we can use imitation learning to clone the demonstrated behavior. However, these approaches are not directly applicable to behaviors that are difficult for humans to demonstrate (such as controlling a robot with many degrees of freedom but very non-human morphology).

Appearing in *Proceedings of the 31st International Conference on Machine Learning*, 2014. Copyright 2014 by the author(s).

9.3.2 RLHF 的发展历程

- ❖ 2018 年 Warnell 将深度强化学习与人类反馈更进一步结合和突破，用于从实时人类互动中学习
- ❖ 2019 年 Ziegler 和 Daniel 等人首次将强化学习运用到语言模型中，他们提出了一种使用人类偏好来微调预训练语言模型的方法，而不是使用监督学习。这种方法可以帮助提高语言模型的性能，使其更符合人类的偏好和需求
- ❖ 2022 年 OpenAI 将 RLHF 与 Instruct learning 结合，推出 InstructGPT 大模型，相比于没有应用 RLHF 技术之前的 GPT3 模型来说，各方面效果都有所提升。之后 MetaAI、Google 等公司推出的大语言模型也都用到了 RLHF 技术

Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces

Garrett Warnell¹, Nicholas Waytowich^{1,2}, Vernon Lawhern¹, Peter Stone³
¹U.S. Army Research Laboratory, ²Columbia University, New York, ³The University of Texas at Austin

Fine-Tuning Language Models from Human Preferences

Daniel M. Ziegler* Nisan Stiennon* Jeffrey Wu Tom B. Brown

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell[†] Peter Welinder Paul Christiano[†]

Jan Leike* Ryan Lowe*

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

1 Introduction

Large language models (LMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al., 2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective

*Primary authors. This was a joint project of the OpenAI Alignment team. RL and JL are the team leads. Corresponding author: lowe@openai.com.

[†]Work done while at OpenAI. Current affiliations: AA: Anthropic; PC: Alignment Research Center.

While recent advances have allowed autonomous agents to solve complex tasks, existing training data. One way we are able to learn to provide real-time, scalar-valued expert demonstrations where it provides expert demonstrations (e.g., the TAMER framework) considered high-dimensional deep learning. In this paper, we extend the TAMER framework to learn complex tasks with a human trainer. We use a process by using it and just back to train an agent in an Atari game of BOWLER, even state-of-the-art

Many tasks that we wish to accomplish can be framed as a series of decisions over a high-dimensional state space. In this paper, we focus on physical space, it does not matter regarding which particular action is taken at every instant. In a *decision making* problem, we have a mapping from the agent's current state within the environment (e.g., to move in a certain direction) to an action, and an agent estimating its state and the value of that policy.

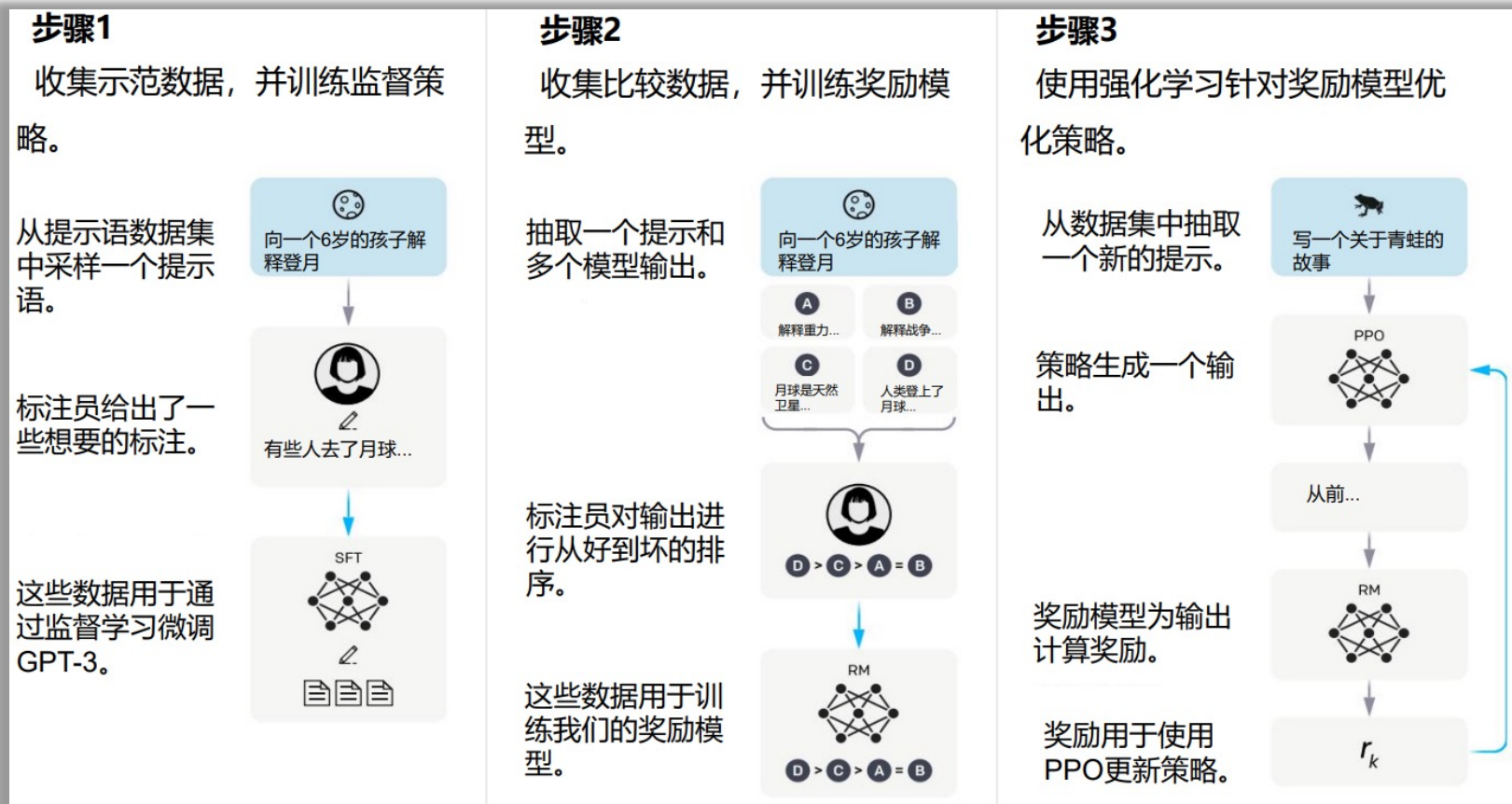
In some very simple tasks, we can completely specify the task defined by a set of rules that tell whether a particular action is good or bad. In this paper, we do this, we can reward, and the history of work interaction, this modern deep learning relatively simple real world setting

Copyright © 2018, OpenAI. All rights reserved. www.openai.com

*Equal contribution

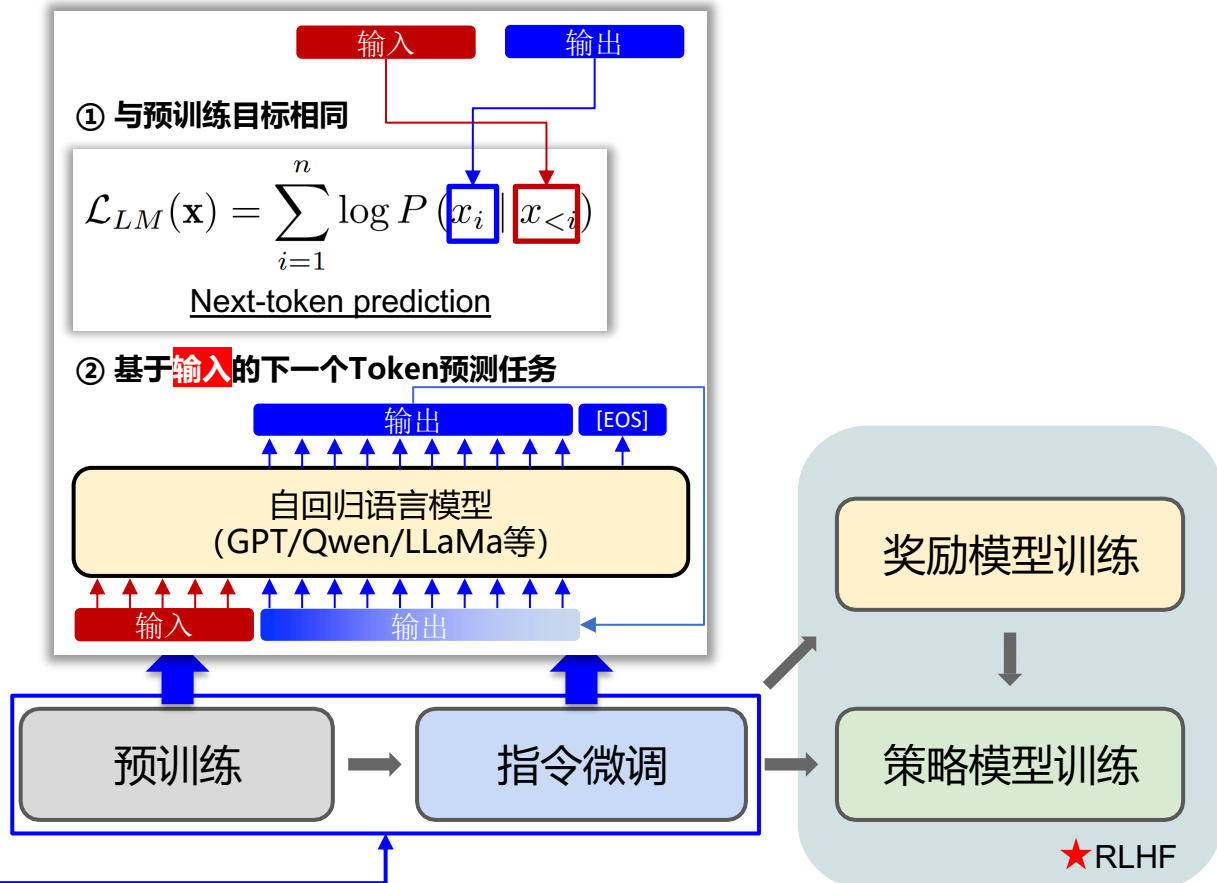
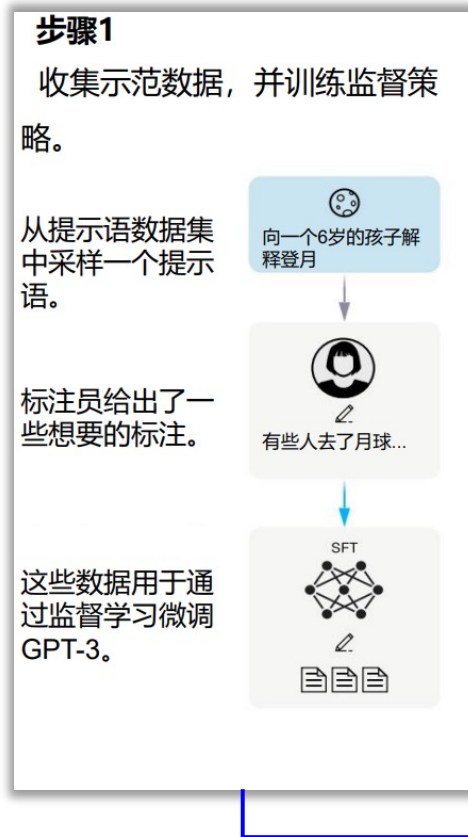
9.3.3 对齐微调技术

■ 基于 PPO 的 RLHF



9.3.3 对齐微调技术

■ 步骤 1. 微调预训练 LLMs (监督微调)



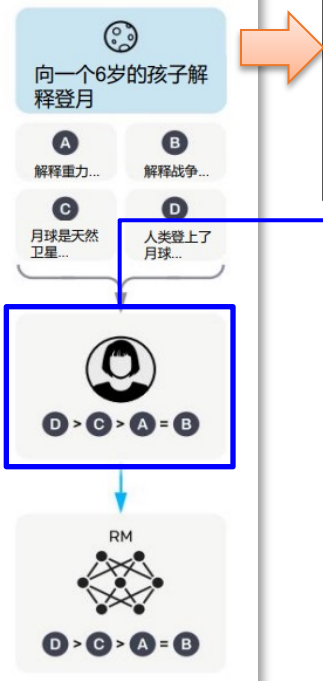
9.3.3 对齐微调技术

■ 步骤 2. 训练奖励模型 (RM)

步骤2

收集比较数据，并训练奖励模型。

抽取一个提示和多个模型输出。



标注员对输出进行从好到坏的排序。

这些数据用于训练我们的奖励模型。

① 收集数据集

输入 用几句话向6岁的孩子解释登月。

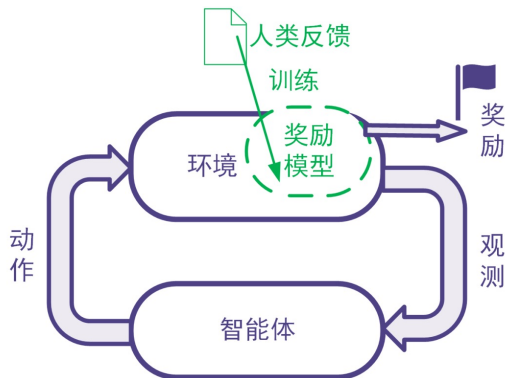
回复1

月球是地球的一颗天然卫星它是太阳系中第五大卫星，也是相对于其主行星大小而言最大的卫星

回复2

人们登上月球，拍摄了他们所看到的照片，然后将它们发送回地球，以便我们都可以看到它们

② 对不同的答案进行排序



[1] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.

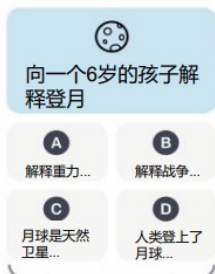
9.3.3 对齐微调技术

■ 步骤 2. 训练奖励模型 (RM)

步骤2

收集比较数据，并训练奖励模型。

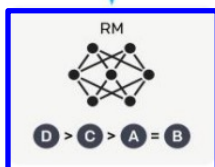
抽取一个提示和多个模型输出。



标注员对输出进行从好到坏的排序。



这些数据用于训练我们的奖励模型。



③ 训练奖励函数

$\max_{\theta} \mathbb{E}[R]$ 在强化学习中，目标是最大化预期回报

$r(x, y; \phi)$ 为奖励函数， y 作为输出， x 为输入， Φ 为模型参数

□ 奖励模型 (RM) : 如何计算标量奖励?

p_{ij} 是答案 y_i 比答案 y_j 好的概率。

$$\log \frac{p_{ij}}{1 - p_{ij}} = r(x, y_i; \phi) - r(x, y_j; \phi)$$

— Bradley - Terry model (1952)

使用对数几率建模答案 y_i 优于答案 y_j 的奖励之差。

$$p_{ij} = \frac{e^{r(x, y_i; \phi) - r(x, y_j; \phi)}}{1 + e^{r(x, y_i; \phi) - r(x, y_j; \phi)}} = \sigma(r(x, y_i; \phi) - r(x, y_j; \phi))$$

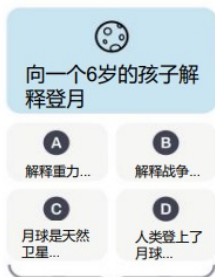
9.3.3 对齐微调技术

■ 步骤 2. 训练奖励模型 (RM)

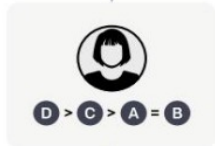
步骤2

收集比较数据，并训练奖励模型。

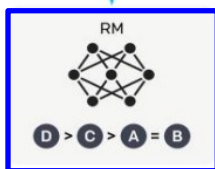
抽取一个提示和多个模型输出。



标注员对输出进行从好到坏的排序。



这些数据用于训练我们的奖励模型。



③ 训练奖励函数

□ 奖励模型 (RM) : 如何计算标量奖励?

p_{ij} 是答案 y_i 比答案 y_j 好的概率。

$$\log \frac{p_{ij}}{1 - p_{ij}} = r(x, y_i; \phi) - r(x, y_j; \phi)$$

— Bradley - Terry model (1952)

使用对数几率建模答案 y_i 优于答案 y_j 的奖励之差。

$$p_{ij} = \frac{e^{r(x, y_i; \phi) - r(x, y_j; \phi)}}{1 + e^{r(x, y_i; \phi) - r(x, y_j; \phi)}} = \sigma(r(x, y_i; \phi) - r(x, y_j; \phi))$$

□ 基于概率 p_{ij} 的目标函数

$$\max_{\phi} \sum_{x, y_i, y_j \in D} \log p_{ij}$$

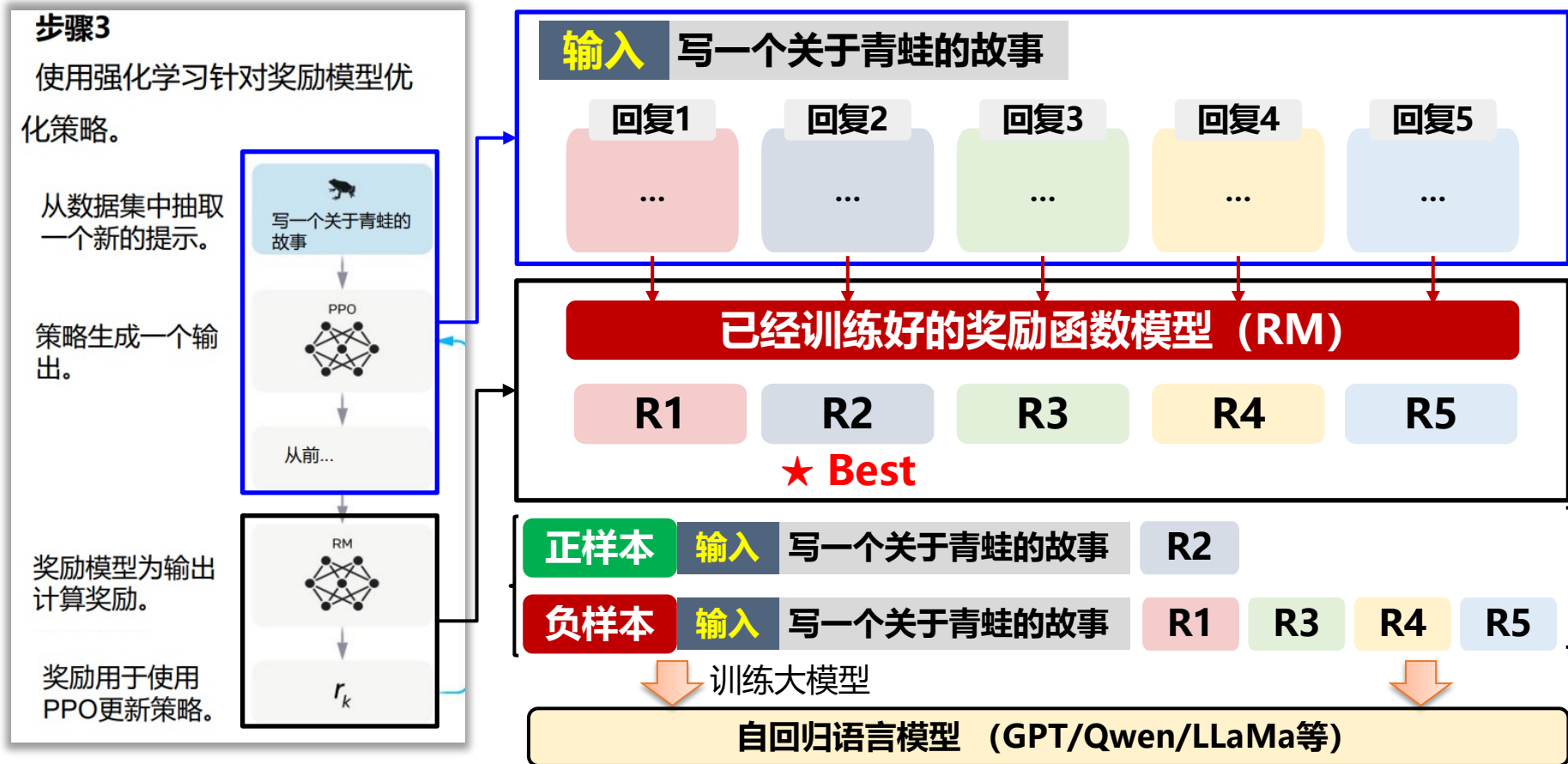
本质上

★ 尽可能最大化

$$r(x, y_i; \phi) - r(x, y_j; \phi)$$

9.3.3 对齐微调技术

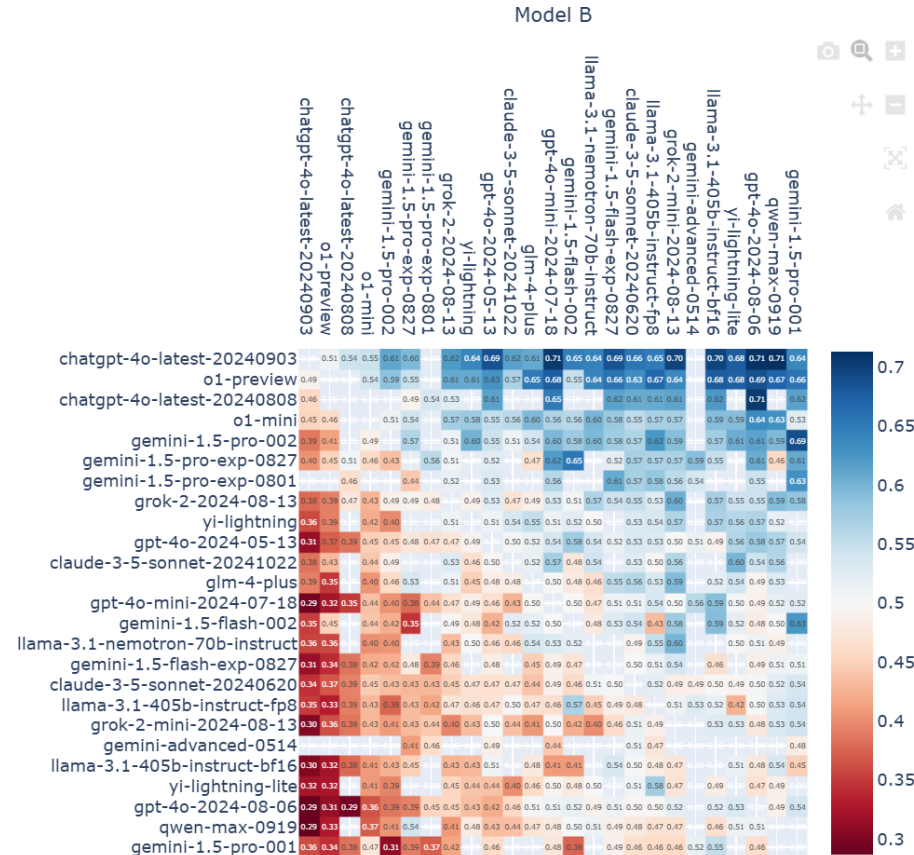
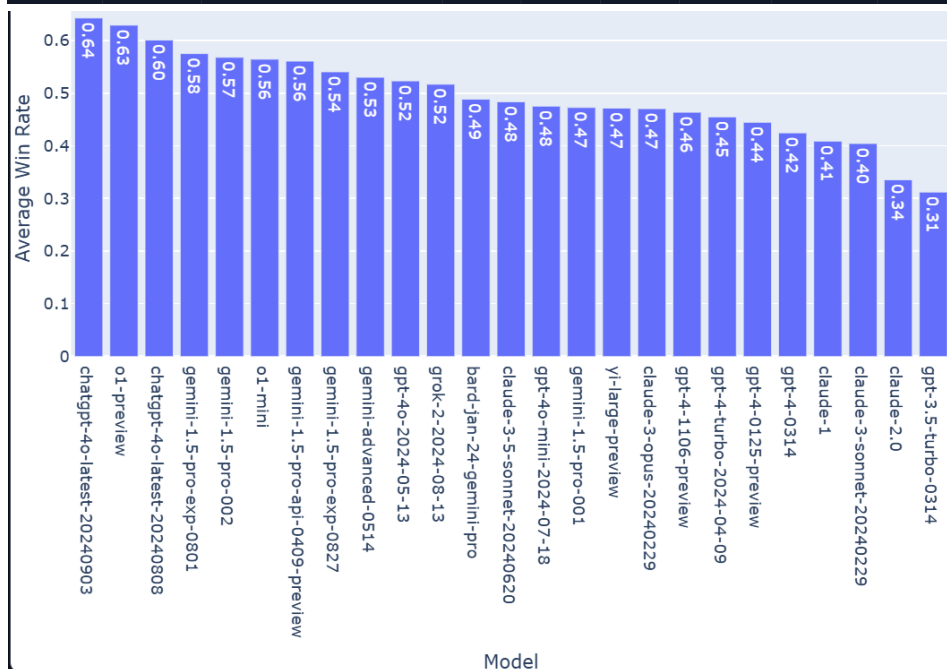
■ 步骤 3. 使用 PPO 微调大语言模型策略



9.3.3 对齐微调技术

Chatbot Arena LLM 排行榜：最佳 LLM 和 AI 聊天机器人评估

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	ChatGPT-4o_latest...(2024-09-03)	1340	+3/-4	37712	OpenAI	Proprietary	2023/10
1	1	o1-preview	1334	+4/-5	23562	OpenAI	Proprietary	2023/10
3	5	o1-mini	1308	+4/-4	26662	OpenAI	Proprietary	2023/10
3	4	Gemini-1.5-Pro-002	1301	+5/-3	19523	Google	Proprietary	Unknown
5	8	Grok-2-88...13	1290	+4/-4	43774	xAI	Proprietary	2024/3



Source: <https://lmarena.ai/?leaderboard>

9.3.4 偏好数据集

■ 基于投票的偏好数据集

基于投票的偏好数据集，其中最知名的是 Chatbot arena conversations [217]，它包含了来自 20 个模型在 96 种语言中的输出实例，每条实例来自两个模型对同一个问题的回答以及人类评判的选择。

表 9.4: 代表性基于投票的偏好数据集

数据集	发布时间	规模	特点
hh-rlhf	2022-4	16.9 万	人类投票，Anthropic 发布
Chatbot_arena_conversations	2023-6	3.3 万	人类投票，加州大学伯克利分校发布
MT-Bench_human_judgments	2023-6	0.33 万	人类投票，加州大学伯克利分校发布
PKU-SafeRLHF	2023-7	36 万	人类投票，北京大学发布
CValues	2023-7	14.5 万	机器投票，阿里巴巴发布

9.3.4 偏好数据集

■ 基于评分的偏好数据集

评分方法即为同一个问题的多个响应评分。由于评分可以是连续的，它能够更灵活地表示偏好强度，使模型能够更细粒度的理解人类偏好。然而，需要注意，评分标准的统一性和评分过程中的主观性是需要考虑的问题。与基于投票的偏好数据集类似，该数据集也有人类评分和模型评分两种基本策略

表 9.5: 代表性基于评分的偏好数据集

数据集	发布时间	规模	特点
Stack-Exchange-Preference	2021-12	108 万	人类评分，Anthropic 发布
Summarize_from_Feedback	2020-9	19 万	人类评分，OpenAI 发布
WebGPT	2021-12	1.9 万	人类评分，OpenAI 发布
Alpaca_comparison_data	2023-3	5.1 万	模型评分，斯坦福大学发布
UltraFeedback	2023-10	6.3 万	模型评分，清华大学

本章内容

- **9.2 指令微调**

- 9.2.1 指令微调概念
- 9.2.2 构造指令实例
- 9.2.3 指令微调任务
- 9.2.4 多模态指令微调
- 9.2.5 指令微调优化方法
- 9.2.6 指令微调的效果

- **9.3 对齐微调**

- 9.3.2 RLHF算法
- 9.3.3 RLHF的发展历程
- 9.3.4 对齐微调技术
- 9.3.5 偏好数据集

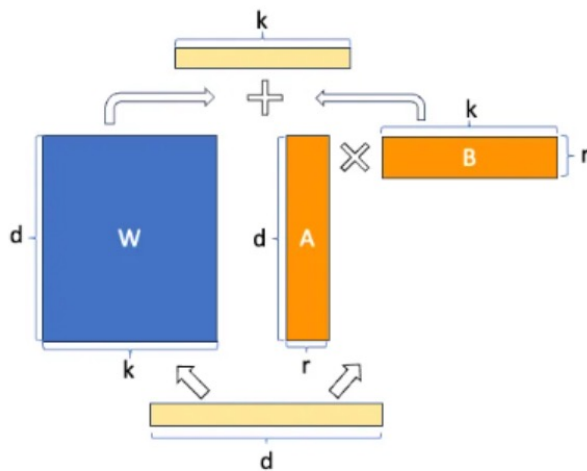
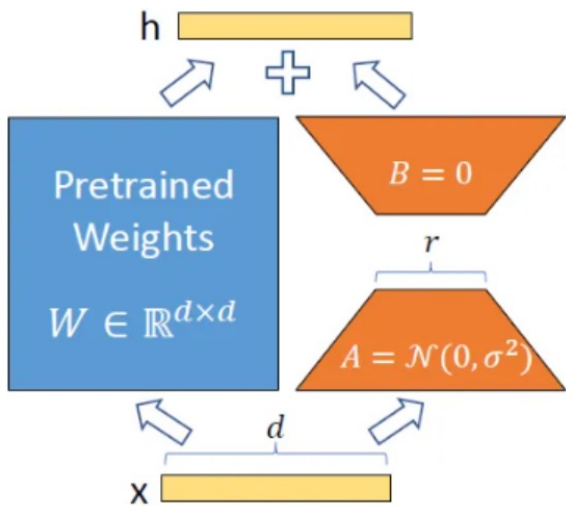
- **9.4 微调算法**

9.4 微调算法

LoRA (Low-Rank Adaptation) 大模型加速微调 and 训练算法

$$\tilde{W}_0 = W_0 + \Delta W$$

$$W_0 + \Delta W = W_0 + BA, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$



前向传播

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad W = W_0 + BA$$

LoRA 算法能够极大地提高训练效率并**降低接近 2/3 的显存占用**。

如GPT-3 有 **175B** 参数, 需要 **1.2 TB** 显存用于训练, 而采用 LoRA 算法后, 其显存开销能够被**降低到 350 GB**;

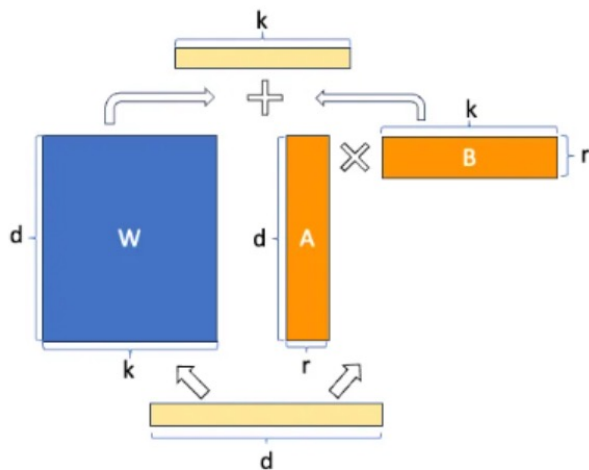
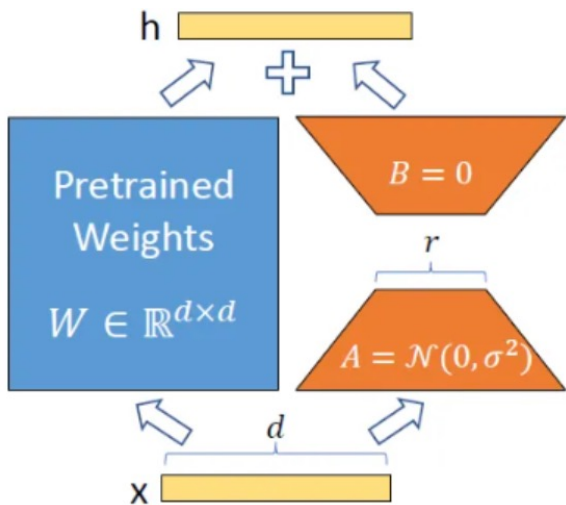
如果 **$r = 4$** , 且仅在 Q 和 V 投影层应用 LoRA 算法, 微调后新增的权重矩阵**仅 35 MB**

9.4 微调算法

LoRA (Low-Rank Adaptation) 大模型加速微调 and 训练算法

$$\tilde{W}_0 = W_0 + \Delta W$$

$$W_0 + \Delta W = W_0 + BA, B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$



前向传播

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad W = W_0 + BA$$

LoRA 缩放系数 α/r : 实际计算时, $\Delta W \times$ 需要乘以缩放系数 α/r 。在最早的 LoRA 中, 该系数被设置为 2, 即 α 是 r 的两倍, 而在后续研究一般将其设置为 1。缩放系数越大, 表明微调部分的权重越大。

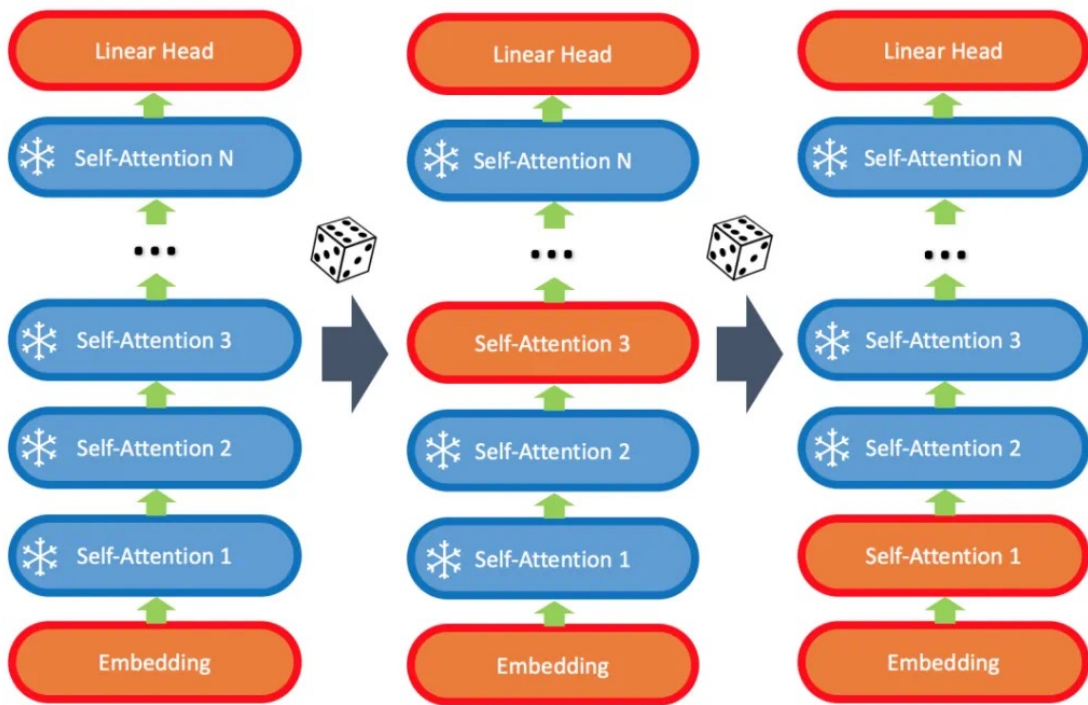
LoRA 层: 微调的过程中, 可以仅对部分 Transformer 层中的权重矩阵采用 LoRA 算法。比如, 可以指定仅在 Q/K/V 中部分层。研究表明, 对整个神经网络的所有层采用 LoRA 算法的效果是最好的。

LoRA dropout: 表示 LoRA 层 dropout 的概率, 目的是为了 avoid 过拟合。对于 13B 等小模型, 一般设置成 10%, 而对于 65B 等更大规模模型, 一般设置成 5%

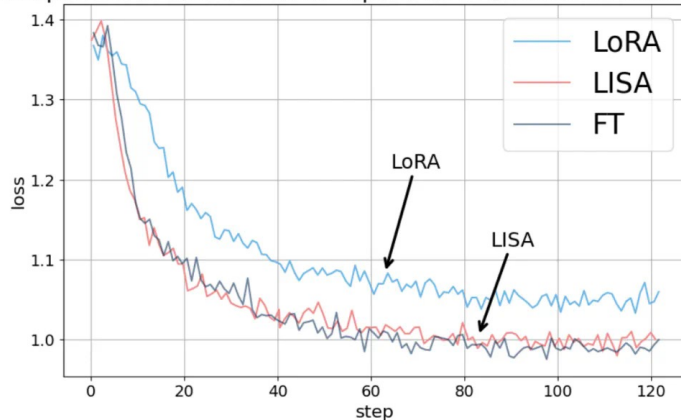
9.4 微调算法

■ LISA 分层重要性采样微调算法

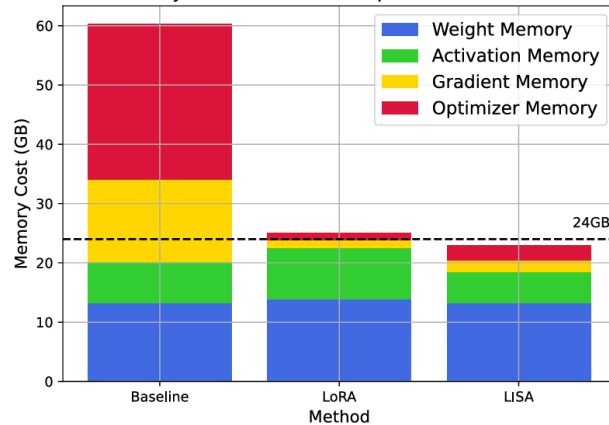
相比于LORA内存成本更低、效果更好



Comparison of Loss Curves on Alpaca-GPT4 Dataset for LLaMA-2-7B



Memory Cost for Different Optimization Methods



9.5 讨论

思考题1 如何理解指令微调在提升大型语言模型性能中的作用？微调后的模型与预训练模型有何区别和优势？

思考题2 在 RLHF的技术路线中，使用 PPO和 DPO方法有什么缺陷？

思考题3 RLHF还可以应用在哪些领域呢？