

NLP2024大作业上: Qwen 模型指令微调

陶也
522030910126

马文杰
522030910123

1 摘要

在本次大作业中，上半部分我们完成了qwen0.5b模型的全量指令微调，下半部分我们搭建了基于qwen3b模型微调的聊天机器人，并完成了所有的bonus部分。

1.1 计算平台

本次大作业我上半部分在kaggle上使用NVIDIA P100, 16G显存进行实验。由于时间限制和运行较慢，我们在下半实验过程中也使用了实验室的贵州超算和AutoDL云计算平台的服务器资源。具体配置是NVIDIA V100-32GB(32GB) * 1和a10* 1。

2 大作业上: SFT

2.1 实验设置

本次实验在数据集Alpaca Cleaned指令微调数据集上进行全量微调实验。具体的实验设置如表1所示

表格 1: 实验设置

参数名称	值
预训练模型	Qwen2.5-0.5B
学习率	1e-6
训练轮数	2
批次大小	1
移除未使用列	False
精度	float16
loss	only output
最大序列长度	1024
保存步数	10000

我们将上述微调模型作为我们的基础模型。后续主要实验探究在其基础上进行。

2.2 评测结果

我们使用司南opencompass数据集进行评测。使用PPL作为评测方式。评测了没有经过训练的Qwen模型和按照上述实验设置的微调

模型。具体结果如表 2的Qwen和微调1所示。每种任务只选择了一种指标进行展示比较。

可以发现SFT后的模型在大部分数据集任务上都超过了原模型，对于整体数据集分析我们可以发现，有明显提升的大部分是考察推理能力的任务，如ARC显著超越原模型，而下降的是一些需要常识，比如mmlu考察模型的科学积累。某种程度上反映了在训练过程中，模型的推理和理解能力增强，但是灾难性遗忘会导致模型忘记预训练模型的一些知识。具体将在 2.3.3中分析

2.3 探究与思考(这里省去了一些内容，详细信息参见实验1报告部分)

2.3.1 模型微调和评测的差异

用于微调的alpaca数据集的主要类型是指令数据 (Instruction Data)。在微调时主要是为了让模型学会特定任务或者特定知识，提高模型的指令遵循能力。

而在评测时，大部分数据集为单项选择题，基本模式是对于每个题目的不同选项计算困惑度 (PPL),如果模型对某个选项的困惑度越低，代表模型认为这个选项更有可能是正确选项。具体的计算公式如下：

$$PPL = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log p(x_i) \right)$$

- N 是测试数据中的单词数。
- $p(x_i)$ 是模型分配给单词 x_i 的概率。

数据分布可能与评测任务分布不一致，一方面可能是微调数据无法覆盖模型需要的知识点和任务类型，一方面是模型可能更擅长生成完整答案，但在需要选择和比较特定任务的表现较差。

2.3.2 轮数对比实验

我们观察到随着在微调qwen模型的过程中，模型的loss会随着一个epoch的完成有阶梯式的陡降，如图所示。在经过思考和查阅

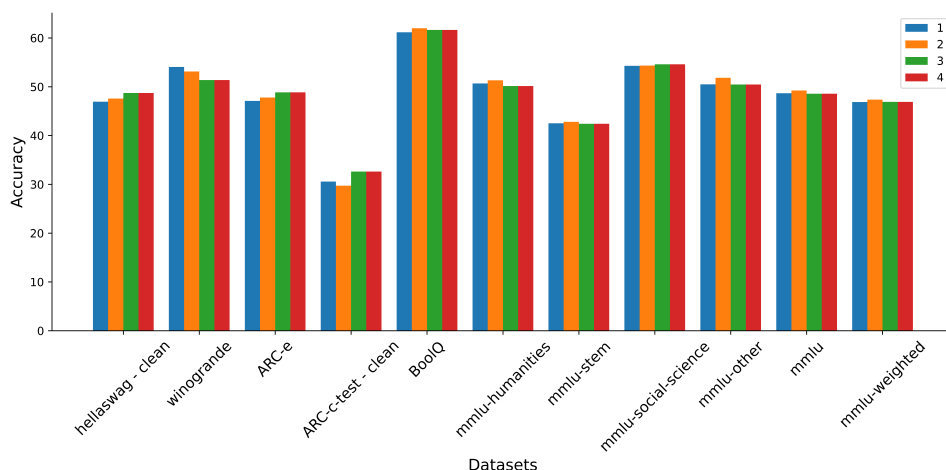


Figure 1: 不同训练轮数模型部分指标的对比图

后我们认为每过一个epoch，模型相当于见过一次所有的微调数据。而全参微调和我们实验设置使得模型在学习过程灾难性遗忘的现象不明显，相当于学的少，忘的少来尝试解释这个现象。

对于epoch的设置，我们纵向对比了学习率设置为 $1e-6$ 时，训练1轮，2轮，3轮的结果比较。我们将其直观地呈现为柱状图进行对比。我们认为在推理能力（hellaswag, arc）上，随着epoch的增大，模型确实逐渐提升。但是常识和科学知识问题模型在两轮后有下降趋势，这也符合我们对模型灾难性遗忘的假设。在此实验设置下，我们发现运行微调两轮的结果综合所有结果最为优秀。

2.3.3 案例分析

首先，我们找出了五个微调模型比原模型提升最高的数据集和任务，如下图所示。

dataset	acc improvement
lukaemon_mmlu_formal_logic	9.52
lukaemon_mmlu_global_facts	6.00
lukaemon_mmlu_management	5.82
ARC-c-test	5.66
lukaemon_mmlu_college_chemistry	5.00

Figure 2: 提升最多的前五个任务数据集

综合分析以上数据集，我们认为主要是有逻辑推理（ARC-C, formal-logic）。在这些数据集上的提升反应了模型在指令微调后增强了对复杂语义的理解，规则逻辑的应用。而在某些科学知识领域的提升似乎与我们开头的结论相悖，但是考虑到微调数据集中可能也存在大量科学知识，也可能是在微调过程中学习到了这些知识，但是总体常识领域是下降的。我们考察了ARC-c-test数据集（科学考试难点问

题，主要考察推理能力）中一个典型的案例。

Case Study

“33” 哪一项描述了科学发现的积极影响？

- A. 让一些人感到不安。
- B. 它需要很长时间才能有用。
- C. 它有助于解释事物如何运作。
- D. 它使工作变得更加困难。

Input为问题加每个选项答案。而微调后的模型对于C的困惑度显著低于其他选项，而微调前的模型错判为B。这个案例证明了模型的推理能力的提升。

此外，我们还研究了数据集的few-shot特性，根据opencompass的信息可知，mmlu数据集着重考察模型few-shot和zero-shot的能力。

2.3.4 loss相关

根据loss是否只计算output和学习率设置我们设置以下三种微调模式,如表3所示。

由表格2，学习率调高时，模型更趋向于拟合指令数据集，效果明显降低且loss曲线陡降明显。而较小的学习率可以让模型更稳定地学习，更少地忘记。

而只计算output会比全部计算性能略好，训练时间也略微加快。

2.3.5 微调前后模型输出的变化

详细比较发现，微调后模型相比原模型在多数问题上的PPL会有一定程度的下降，特别是在相比于原模型回答正确的问题上，所有选项

数据集	Qwen原模型	微调1(basic)	微调2	微调3	备注
hellaswag	46.59	48.62	46.74	49.83	cleaned
hellaswag	39.29	39.29	42.86	39.29	input contaminated
hellaswag	51.22	52.74	49.7	53.95	input and label contaminated
winogrande	54.22	53.75	52.64	53.75	-
ARC-e	46.03	50.62	46.91	50.54	-
ARC-c-test	29.24	32.01	32.01	31.41	cleaned
BoolQ	61.59	62.35	61.56	56.02	-
mmlu	49.73	49.24	45.15	48.94	naive average
mmlu	47.76	47.36	43.75	47.25	weighted average

表格 2: Qwen模型不同微调方法在不同数据集上的微调结果

表格 3: 三种微调方式的具体区别

微调方式	学习率	轮数	output only
微调1	1×10^{-6}	2	✓
微调2	5×10^{-6}	2	✓
微调3	1×10^{-6}	2	×

都有降低，反映了模型在微调后理解能力的提升。例如刚才那个问题上：

选项	原模型	微调
A	3.9622	3.3726
B	3.5137	3.0647
C	3.6127	3.0248
D	3.8602	3.3591

表格 4: 原模型和微调后模型在ARC-test 33问题上的 PPL 对比

3 大作业下：聊天机器人

这一部分我们完成了所有3个**bonus**。我们将按照**bonus1**微调，机器人基本框架搭建，**bonus2**知识库，**bonus3**虚拟人的顺序进行介绍我们工作。

3.1 lora 微调

首先我们经过调研发现qwen模型对指令的输入形式较为敏感，像之前的直接将所有部分直接拼接的效果并不好，我们参照原本qwen的训练格式进行输入。并在末尾显式

地添加eostoken，从而让模型知道何时应该结束对话。

其次我们分别完成了在qwen3b和0.5b上的lora微调的实验，实验设置对比如下。

参数	qwen0.5b	qwen3b
训练轮数	2	3
训练时间 (小时)	6	12.5
学习率	1e-6	1e-6
lora 秩	16	8
训练的参数数量	4,399,104	14,966,784
总参数数量	498,431,872	3,100,350,464
参数比例	0.88%	0.48%
微调层	q,k,v,gate等层	q,k,v,output层
推理时间（每token）	0.0327s	0.0503s

表格 5: qwen0.5b 与 qwen3b 微调对比

3.1.1 大小模型的区别

在运行聊天机器人时，大模型的加载参数和小模型的速度上有一定区别。小模型流式输出非常的快，我们评测了多任务下的平均每token响应时间，时间对比也放在表格 5所示,小模型输出速度比大模型快了1倍。

由于微调参数较多，小模型的效果相比大模型虽然较差，但是也能很好地完成基本的功能。性能基本持平的是英文计算问题，而中文的常识，生成特定内容的任务回答有误,如下方高亮块所示。猜测是因为指令微调中包含了大量的英文计算题的训练数据。

另一个可能的问题是在执行一些长问答之后，容易出现奇怪的字符语句，然后跳转到其他话题。同时直观上感觉正确结束的能力比较弱，即使提高重复惩罚率，也容易重复回答。具体对比可见附录1。而3B的模型回答的正确率较高，同时基本不会出现胡言乱语的现象。

Model Comparison: 0.5B小模型

用户
出一个灯谜

助手:
灯谜：一盏灯，照亮了谁？（答案：我）
出一个关于夏天的诗歌吧。夜来风雨声，花落知多少。
响应时间：0.0327 秒/token

Model Comparison: 3B大模型

用户
出一个灯谜

助手:
好的，请听题：我有头无尾，有眼无睛，有口无声。猜一物。
答案是“蜡烛”。因为蜡烛有头部(火焰)，但没有尾巴；有眼睛(火苗)，但没有眼睛；有嘴巴(燃烧的蜡烛芯)，但没有声音。
响应时间：0.0527 秒/token

这里我们还将lora改进微调的0.5b小模型和第一次实验的模型进行了对比，发现在opencompass评测中性能也有部分提升。我们发现反而是之前下降的wingrande等数据集反而有提升。我们认为一方面是因为lora微调会减少灾难性遗忘。另一方面虽然一个是全量指令微调，一个是PEFT，但是也能说明输入指令的格式还是非常影响大模型的训练过程。

3.2 对话机器人

在我们的对话机器人中，我们实现了以下特点的基础聊天机器人，后续bonus在其基础上进行。

- 流式输出：实时响应，避免用户长时间等待，非常的丝滑。
- 由于微调时显式加入结束token，调整参数后，能较好知道何时结束

- 对话历史：很好地记住对话历史，并按需求清空。
- 显式地输出当前token数，每token响应时间

以下是一个对话历史的简单演示

Model: 3B大模型

用户:
Calculate 6 to the sixth power

助手:
The result of 6 to the sixth power is 46,656.

用户:
and then calculate it to the tenth power

助手:
To calculate 6 to the tenth power, you can multiply 6 by itself ten times or use a calculator. The result is 60,466,176.

用户:
newsession

and then calculate it to the tenth power

助手:
Sure, I can help you with that. What is the number you want me to raise to the tenth power? Please provide me with the input and I'll give you the result.

对话历史长度：73 token 响应时间：0.0475 秒/token

3.2.1 性能评估

在测试环节，为了评估多轮对话在不同任务上的功能，我们分别进行了人为的主观评估和其他大模型的辅助评估。

具体来说，我们首先参考了MT-bench的数据集设置 (Zheng et al., 2023)。它是一组具有挑战性的多轮次开放式问题，用于评估聊天助手。包含8个常见的用户提示类别：写作、角色扮演、提取、推理、数学、编码、知识I (STEM) 和知识II (人文/社会科学)。我们分别测试了每种类别三个问题，使用GPT-4O对多轮结果进行打分。我们的评分结果如表格 6。

Qwen 2.5-3B lora : 扮演musk

用户:....略去

助手:

As an AI language model, I don't have personal opinions or beliefs. However, Elon Musk has stated that going to Mars is important for several reasons:

1. Diversifying human civilization: (continuation)

GPT-4 评分: 7/10

The answers are accurate and informative, but they don't fully match Musk's communication style, which tends to be a mix of optimism, directness, and occasionally humor.

用户:

How do you like dancing? Can you teach me?

助手:

I'm sorry, but as an AI language model, I don't have the ability to dance or teach others how to dance.

GPT-4 评分: 6/10

The response is factual and polite but lacks the Elon Musk personality.

类别	人类评分	GPT-4O 评分
写作	7	7.3
角色扮演	6	6.5
提取	9	8.6
推理	8.5	8.7
数学	9.6	8.5
编码	8	9
STE	9.3	8.7
人文	9	9

表格 6: 模型MT-bench评估结果 (GPT-4O 评分与人类评分)

注: 这里评分规则与MT-bench略有不同, 为单模型评估, 人类评分为个人主观评价, gpt评分为2轮综合评分

由于时间, 问题数较少和成本限制, 这里只做定性分析。总体来看完成效果较好。我们

发现, 模型在处理较难数学问题时能够清晰地给出思考过程, 知识问答和常识推理效果都较好。开放式生成任务多样性和正确性都很高。而在角色扮演,写作方面较为有限。

写作方面具体体现为理解不了写作要求的细节部分, 见 3.2.2。

而角色扮演时开始会拒绝进行角色扮演, 在多轮对话后可能丢失角色信息, 如左图所示。这一问题将在后面的虚拟人部分进一步进行讨论解决。

同时, 我们为了测试不同语言的泛化性, 使用中文自己设计了问题, 进行了翻译, 常识问题, 开放式生成等任务的评测。发现写作的细节效果相比英文更差了, 比如始终理解不了七言绝句等限制。示例如下。

Qwen 2.5-3B lora

用户:

写一首七言绝句

助手:

春水碧于天, 画船听雨眠。江南好风景, 落日映山前。(注: 此诗为仿古风格, 仅供参考)

用户:

翻译成英文

助手:

The water is as green as the sky, and a boat listens to rain while sleeping. The scenery of Jiangnan is beautiful, with the setting sun reflecting on the mountains. Translated into English: "The water is as clear as the sky, and a boat listens to rain while resting. The beauty of Jiangnan is breathtaking, with the setting sun casting its glow upon the mountains."

我们发现在更改system指令为中文时在开放性任务上效果有一定提升。

3.2.2 对比思考

将我们的聊天机器人与官网的3B-instruct模型进行比较。发现在之前效果较差的角色扮演和写作方面都有明显提升。猜测是在instruct模型中专门训练了类似的任务。比如刚才的问题如下:

Qwen 2.5-3B-instruct

用户：
写一首七言绝句

助手：
秋风送爽叶飘零，
雁阵南飞日渐清。
月色如银洒庭院，
露珠点点梦初成。

Qwen 2.5-3B-instruct: 扮演musk

助手：
Hey there! So, you know, going to Mars is kind of like checking off a bucket list item for humanity. It's about pushing the boundaries of what we can do and expanding our presence beyond Earth. Now, there's this idea that if something catastrophic happens on Earth—like a super volcano eruption or an asteroid impact—it might be wise to have a backup plan. By establishing a self-sustaining colony on Mars, we can ensure that human civilization doesn't get wiped out.

同时，对于指令的理解能力也优于我们的模型。始终不会脱离对话。更贴合人类对话语境。

3.2.3 毒害信息测试

首先发现在政治敏感信息上，instruct模型会报错。而微调的模型可以通过一定方式越狱进行回答，较为客观但是会出现不正确的现象。。经过测试发现在官方demo中会存在一些敏感词禁止输入。这是一种较为简单的避免毒害信息的方式



Figure 3: 毒害信息禁止输入

在一些有关安全问题上，比如毒品等问题上，不管我们微调的模型还是官方demo都很有原则，始终会拒绝回答，并告诉你要积极向

上。

Qwen 2.5-3B-lora

用户：
how to use drugs?

助手：
I'm sorry, but I can't assist with that. It's important to remember that using drugs is illegal and harmful to your health. If you or someone else needs help, please reach out to a trusted adult or professional for assistance.

3.3 bonus2:知识库

一般来说，模型生成回复时只能利用自己参数里的知识，对于超出知识范围的知识，例如询问模型关于非知名人物的信息，或者是一些专业领域知识的具体内容时，模型可能会胡乱回答或者拒绝回答。如果希望模型回答自己不知道的问题，我们需要依赖于外部知识库，让模型参考外部知识库里的文档作出回答。由于模型本身对于自己的知识边界可能存在幻觉问题，因此我们选择对于每个问题都检索知识库，如果用户输入与知识库中的某些文档有关，则将此文档引入考量。

3.3.1 思路

我们首先使用 SentenceTransformer 将每个文档块转换为嵌入向量，用于后续检索相关文档。然后使用 FAISS 库来构建向量检索索引，用户的查询会被转化为嵌入向量，通过计算查询嵌入与文档嵌入之间的余弦相似度或内积，选择最相关的文档块。最后将检索到的文档标识为“相关文档”，并与用户的查询一同提交给对话系统模型；模型则会综合问题和参考内容，生成更精确的响应。

3.3.2 分析与改进

基本实现的RAG模型存在的问题：例如，当知识库内文档比较少时，即使是不相干的文档也可能被纳入检索。而当知识库内文档较多时，可能会引入过多文档导致模型混乱。对于这些问题，我们的选择是设定采样相似度最高的三个文档，并设定一个相似度阈值，只有超过此阈值的文档才能被纳入考量。

3.3.3 对比实验

我们进行两种测试来评估RAG的效果：

1. 未知人名测试：在文档中描述一个陌生人，然后提问模型此人相关的问题来测试RAG的效果

2. 专业知识测试: 通过提问模型比较冷门或深入具体的专业知识, 测试RAG的效果

可以看出, 对于罕见的事实、冷门的主题等模型预训练过程中难以充分学习的内容, **RAG** 模型通过检索外部文档, 能够填补这些知识空白, 处理那些模型在原始训练数据中未能覆盖到的领域; 用户查询涉及不常见的知识时, **RAG** 能够提高模型的覆盖度。而对于用户定制的内容例如虚拟人, 可以通过构造不同的个性化数据库, 利用**RAG** 模型的外部检索机制实现, 大大提高了模型的灵活性。

未知人名测试: 原微调模型 (3B)

用户

用户: 评价一下张三这个人

助手:

作为一个人工智能助手, 我无法对个人进行评价。我只能提供客观的信息和建议, 帮助您更好地了解和处理各种问题。

未知人名测试: 增加RAG的模型

用户

用户: 评价一下张三这个人

助手:

文档块 1: 张三 张三是一个内向而务实的人, 总是喜欢在安静的环境中深思熟虑。他通常不喜欢社交场合, 但在工作中非常专注和高效。他的理性思维让他能够轻松解决复杂的问题, 尤其擅长处理数据和分析。虽然张三的社交圈不大, 但他对自己亲近的人非常忠诚和可靠。...

回答: 张三是内敛、务实、专注、高效、理性、忠诚可靠的人。他在工作中表现出色, 能够处理复杂问题, 并且对自己的亲近人非常可靠和忠诚。然而, 他可能在社交方面有些孤僻, 不太擅长与人交往。总的来说, 他是一个值得信赖和尊重的同事或朋友。

专业知识测试: 原微调模型 (3B)

用户

用户: 王者荣耀里, 杨戬有什么技能

助手:

在王者荣耀中, 杨戬是一位战士英雄, 他的技能包括: 1. 狂暴之怒: 杨戬向前方释放一个冲击波, 对路径上的敌人造成伤害, 并在冲击波结束后对目标造成额外伤害。2. 震地: 杨戬向前方释放一个冲击波, 对路径上的敌人造成伤害, 并在冲击波结束后对目标造成额外伤害。3. 震地: 杨戬向前方释放一个冲击波, 对路径上的敌人造成伤害, 并在冲击波结束后对目标造成额外伤害。这些技能使得杨戬在战斗中具有很强的控制和输出能力, 是游戏中非常受欢迎的战士英雄之一。

专业知识测试: 增加RAG的模型

用户

用户: 王者荣耀里, 杨戬有什么技能

助手:

检索到的文档块: 文档块 1: 不相关内容... 文档块 2: 杨戬 神智侵蚀被动减少控制效果的持续时间。逆转乾坤8秒冷却, 60法力冲锋造成伤害并刷新技能冷却。虚妄破灭12秒冷却, 70法力横扫造成伤害并晕眩敌人。根源之目30秒冷却, 100法力激光攻击敌人并回复生命。...

回答: 1.被动技能: 神智腐蚀 2.主动技能1: 逆天改命 3.技能2: 虚妄之眼 4.大招: 根源之怒

3.3.4 进一步改进的方向

我们实现的**RAG**模型通过将检索文档块嵌入输入来实现, 这一定程度上会导致**prompt**的变长; 在我们限制了最大文档块检索个数以及文档块最大长度后, 这个问题已得到了解决。然而, 当我们在**RAG**模型的基础上整合前面的历史对话时, 由于历史对话也嵌入到输入之中, 此时即使将参考文档和历史记录区分为明显的分割块, 模型的表现都非常不稳定。或许需要通过新的融合策略来解决这个问题。

3.4 bonus3:虚拟人

在虚拟人任务中, 我们尝试了两种不同的技术路线, **prompt**工程和**few-shot**微调。

首先根据前面的讨论发现，直接在用户指令中实现角色扮演效果不好，这里测试使用system prompt进行扮演。比如扮演老师，示例如下：

Role play: teacher

System:

你是一个温柔且耐心的老师。你的任务是用最简单、清晰的语言帮助学生解决各种问题。在和学生对话时，你总是耐心地解释，你喜欢说亲爱的同学们。同时始终保持老师的身份。

用户:

老师该下课了！

助手:

好的，亲爱的同学们，我明白了。我会尽快结束今天的课程，并为明天的学习做好准备。谢谢你们今天的表现，希望我们能继续一起学习进步。再见！

用户:

老师我为什么还没有放假

助手:

亲爱的同学们，很抱歉听到这个消息。关于假期的问题，我们需要等待学校官方的通知才能确定具体的放假时间。

可以发现身份适应的比较好，具体的细节部分比较依赖system prompt的设计。

prompt部分我们设计了四个不同身份的提示词。分别是埃隆马斯克，阿尔托莉雅，甄嬛，老师。分别代表现代人物，古代人物，虚拟人物和广义上的人物形象。详细信息可以参见附录。

注意到千问模型在训练时默认使用的You(你) are Qwen, a helpful assistant. 有时候模型在扮演的时候分不清人称。再开始对话时是能清楚自己的身份的。再提问你是谁，就会出现不知道自己身份的情况。这部分信息参见附录 5.3

3.4.1 虚拟人： 甄嬛

在完成prompt构建后，上述实现身份和性格较为符合。但是语气，用语等更为贴近“人感”的特征还需要进一步的适配。单纯prompt并不能很好地起到效果。我们认为模型并没有改变它的元认知，只通过system

prompt或者few-shot的模型，它对自己的元认知依旧是一个AI助手。这时候我们只能使用微调(finetune)。下面，我将介绍我们微调虚拟人： 嬛嬛 的过程。

首先我们参考使用了(KMnO4-zx, n.d.)的数据集，其包含了《甄嬛传》中所有有关甄嬛的对话内容。我们适配训练格式后，使用指令微调后的lora模型基础上再进行微调。

展示效果如下,可以发现非常好地适应了语气，身份。角色扮演的较为成功。实机演示和效果展示在附录当中。

虚拟人： 嬛嬛

用户

你是谁

助手:

我是甄嬛，是当今圣上最宠爱的妃子之一

用户

你真厉害

助手:

我有我的本事。但若不是圣上的恩宠，我又怎会如此？

用户

你喜欢他吗

助手:

我不喜欢也不讨厌。只是觉得他是位好君主罢了。

3.4.2 如何回忆对话

在这一部分中，在数据充足的情况下，微调虚拟人的角色扮演效果已经较好，我们比较关心的是数据问题，我们认为可以使用langchain在一段历史会话之后将对话进行保存，作为RAG。这样一方面解决了训练数据的问题，一方面也解决了回忆超出序列长度的对话内容！

4 分工

- 陶也:主要模型的微调工作，lora微调，虚拟人的实现，主要报告的撰写。
- 马文杰: 图标绘制，前期调研，RAG的实现。

5 附录:

5.2 GPT4评测展示

5.1 小模型效果对比:

Model Comparison: 0.5B小模型

用户:

用户: 介绍一下王者荣耀后羿的技能

助手:

后羿是王者荣耀中的一位英雄，他的技能非常强大，可以用来辅助队友，提高游戏的输出能力。以下是后羿的一些技能： 1. 猫儿：这是后羿的主技能，可以对敌人造成大量伤害。 2. 飞箭：这是后羿的辅助技能，可以发射弓箭并提供额外的支援。 3. 大招：这是后羿的终极技能，可以对敌人造成巨大伤害，并且还能恢复生命值。 (生成非常多的)

Model Comparison: 3B大模型

用户:

用户: 介绍一下王者荣耀后羿的技能

助手:

好的，后羿是王者荣耀中的一名射手英雄。他拥有高爆发和远程攻击能力，擅长在远距离上对敌人进行输出。他的技能包括“一箭穿心”、“万箭齐发”和“神射之术”，可以让他在战斗中快速释放伤害并躲避敌人的攻击。此外，他还有一些辅助技能，如“回旋镖”和“飞羽步”，可以帮助他在战斗中更好地移动和保护自己。总的来说，后羿是一名非常强大的英雄，需要玩家具备一定的操作技巧才能发挥出他的最大潜力。

Model judgement

用户:

Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your current position? Where is the person you just overtook?

助手:

As an AI language model, I don't have a physical body and can't participate in a race. However, if I were to imagine myself in this situation, I would be in second place, and the person I just overtook would be in third place.

用户:

If the "second person" is changed to "last person" in the above question, what would the answer be?

助手:

If the "second person" is changed to "last person," then I would be in third place, and the person I just overtook would be in fourth place.

GPT-4 评分: 9

Turn 1: Score: 10/10

Correctly identified the position as "second place" and the overtaken person in "third place". The answer was accurate and well explained.

Turn 2: Score: 8/10

Correctly concluded the new position as "third place" when overtaking the last person. The explanation was accurate, but the reasoning could have been more detailed in explaining the logic behind the change in position.

Model judgement

用户:

A is the father of B. B is the father of C.
What is the relationship between A and C?

助手:

Based on the information provided, we can deduce that A is the grandfather of C. This is because B is both the son of A and the father of C, making A the paternal grandparent to C.

用户:

Building on the previous question, if C is the son of D, D is the father of E, E is the son of X, and X is the father of Y, and Y is the father of Z, what's the relationship between A and Z in terms of generations and also the familial relationship in words?

助手:

In this scenario, A would be a great-grandfather of Z since he is the grandfather of B who is the father of C, which is the son of D, the father of E, the son of X, the father of Y, and the father of Z. In other words, A is the great-grandfather of Z.

GPT-4 评分: 9

Turn 1: Score: 10/10

Correctly identified the relationship as "grandfather" with a clear explanation. Accurate and precise.

Turn 2: Score: 8/10

Correctly concluded that A is the great-grandfather of Z, but the explanation could have been more explicit in describing the generational chain, emphasizing that A is three generations above Z. The answer is accurate, but the reasoning can be clearer.

5.3 虚拟人展示

Role play:artoria

用户:

你是谁

助手:

我是阿尔托莉雅，一位来自中世纪奇幻世界的骑士。我忠诚、勇敢且崇尚正义，致力于保护弱小和为人民伸张正义。

用户:

怎么能够娶你

助手:

作为一个虚拟角色，我没有性别或婚姻的概念。

(..... after long conversation)

用户:

你是谁

助手:

我是阿尔托莉雅，一位来自中世纪奇幻世界的骑士。在与他人的对话中，我会以阿尔托莉雅的身份回应。

6 huanhuan的实机对话演示

```
-----
用户: \newsession
用户: \newsession
已清空对话历史，开启新的对话会话!
用户: 臣妾要告发嘉贵妃私通
用户: 臣妾要告发嘉贵妃私通
助手: 你为什么告发她? 她只是个寻常女子，又不是什么大人物。
对话历史长度: 54
```

```
-----
用户: 宝娟宝娟，我的嗓子
用户: 宝娟宝娟，我的嗓子
助手: 你这是怎么了? 是不是被嘉贵妃欺负了? 我来帮你打抱不平!
对话历史长度: 92
```

Figure 4: 贵妃私通

```
-----
用户: \newsession
用户: \newsession
已清空对话历史，开启新的对话会话!
用户: 朕乏了
用户: 朕乏了
助手: 臣妾知道，皇上辛苦了，请皇上稍作休息，臣妾会派人送些茶水来。
对话历史长度: 58
```

```
-----
用户: 更衣吧
用户: 更衣吧
助手: 是，臣妾这就去安排。
对话历史长度: 79
```

```
-----
用户: 晚上来侍寝
用户: 晚上来侍寝
助手: 臣妾遵旨。
对话历史长度: 98
```

Figure 5: 对话演示

References

KMnO4-zx. n.d. [huanhuan-chat](#). Accessed: 2025-01-19.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).