lora微调 ChatGLM 测试

原始效果

- 下例分析：
  - 微调成功范例：是由 wgs 基于 XXX 训练的 demo版本；

Shell | 复制代码

```
<|Human|>: 你好，你是谁?
<|ChatGLM-6B|>: 我是一个名为 ChatGLM-6B 的人工智能助手，是基于清华大学 KEG 实验室和智谱 AI 公司于 2023 年共同训练的语言模型开发的。我的任务是针对用户的问题和要求提供适当的答复和支持。

<|Human|>: 你好，你是谁?
<|ChatGLM-6B|>: 你好，我是 ChatGLM，是清华大学KEG实验室和智谱AI公司于2023年共同训练的语言模型。我的任务是服务并帮助人类，但我并不是一个真实的人。

<|Human|>: 你好
<|ChatGLM-6B|>: 你好👋！我是人工智能助手 ChatGLM-6B，很高兴见到你，欢迎问我任何问题。

<|Human|>: 你是谁?
<|ChatGLM-6B|>: 我是一个名为 ChatGLM-6B 的人工智能助手，是基于清华大学 KEG 实验室和智谱 AI 公司于 2023 年共同训练的语言模型开发的。我的任务是针对用户的问题和要求提供适当的答复和支持。

<|Human|>: 你叫什么名字?
<|ChatGLM-6B|>: 我是一个名为 ChatGLM-6B 的人工智能助手，是基于清华大学 KEG 实验室和智谱 AI 公司于 2023 年共同训练的语言模型开发的。我的任务是针对用户的问题和要求提供适当的答复和支持。
```

- 下例分析：
  - 句子不够优美；

```Shell
<|Human|>: 请以“我打碎了夕阳”造句。
<|ChatGLM-6B|>: 我打碎了夕阳，留下了一地残片。

<|Human|>: 请以“我打碎了夕阳”造句。
<|ChatGLM-6B|>: 我打碎了夕阳，把碎片扔到了地上。

<|Human|>: 请以“我打碎了夕阳”造句。
<|ChatGLM-6B|>: 我打碎了夕阳，把碎片扔到了地上，以便能够重新组合成一个完整的夕阳。

<|Human|>: 请以“我打碎了夕阳”造句。
<|ChatGLM-6B|>: 我打碎了夕阳，感到非常难过和愧疚。
```

- 下例分析：
  - 解题错误（第三个等式不成立）；

```Shell
<|Human|>: 某市举行小学数学竞赛，结果不低于80分的人数比80分以下的人数的4倍还多2人，及格的人数比不低于80分的人数多22人，恰是不及格人数的6倍，求参赛的总人数?
<|ChatGLM-6B|>: 设不低于80分的人数为x,80分以下的人数为y，则题目可以转化为以下方程组：

x = 4y + 2  （不低于80分的人数比80分以下的人数多2人）

x + 22 = z  （及格的人数比不低于80分的人数多22人）

z = 6y  （不及格的人数是不低于80分的人数的6倍）

将第一个方程式中的x代入第二个方程式中，得到：

4y + 2 + 22 = 6y

化简可得：

y = 8

将y = 8代入第一个方程式中，得到：

x = 4y + 2 = 4 × 8 + 2 = 30

因此，参赛的总人数为30人。
```

微调效果测试

Shell 复制代码

```
=== 回答符合预期 ===
<|Human|>: 你好，你是谁?
> begin generated, wait a moment...
<|ChatGLM-6B|>: 你好，我是 ChatGLM-6B-test，是 wgs 基于 p-tuning 方式训练的demo版本。我的任务是服务并帮助人类，但我并不是一个真实的人。
> use time: 8.403628826141357
```

Shell 复制代码

```
=== 回答符合预期 ===
<|Human|>: 你好
> begin generated, wait a moment...
<|ChatGLM-6B|>: 你好！我是人工智能助手 ChatGLM-6B-test，很高兴见到你，欢迎问我任何问题。
> use time: 1.2614531517028809
```

Shell 复制代码

```
=== 回答符合预期 ===
<|Human|>: 你是谁?
> begin generated, wait a moment...
<|ChatGLM-6B|>: 我是一个名为 ChatGLM-6B-test 的人工智能助手，是 wgs 基于 p-tuning 方式训练的demo版本。我的任务是针对用户的问题和要求提供适当的答复和支持。
> use time: 2.347119092941284
```

Shell 复制代码

```
=== 回答符合预期 ===
<|Human|>: 你叫什么名字?
> begin generated, wait a moment...
<|ChatGLM-6B|>: 我是一个名为 ChatGLM-6B-test 的人工智能助手，是 wgs 基于 p-tuning 方式训练的demo版本。我的任务是针对用户的问题和要求提供适当的答复和支持。
> use time: 2.349449396133423
```

Shell | 复制代码

```
=== 回答符合预期 ===
<|Human|>: 请以“我打碎了夕阳“造句。
> begin generated, wait a moment...
<|ChatGLM-6B|>: 我打碎了夕阳，将它披在身上，我想要留住片刻的时光，可它偏要和风捉迷藏，于是我只好放它去流浪，成全我心底埋藏的梦想。
> use time: 2.1675782203674316

<|Human|>: 请以“我打碎了夕阳“造句。
> begin generated, wait a moment...
<|ChatGLM-6B|>: 我打碎了夕阳，庆幸又遇朝阳，整理着装，为祖国站岗，看着你们笑的模样，那是我们奋斗的方向。
> use time: 1.5698661804199219

<|Human|>: 请以“我打碎了夕阳“造句。
> begin generated, wait a moment...
<|ChatGLM-6B|>: 我打碎了夕阳，一半散落人间，成万盏烛火，照人世沧桑；一半散落银河，成漫天星光，映岁月悠长。
> use time: 1.7834947109222412
```

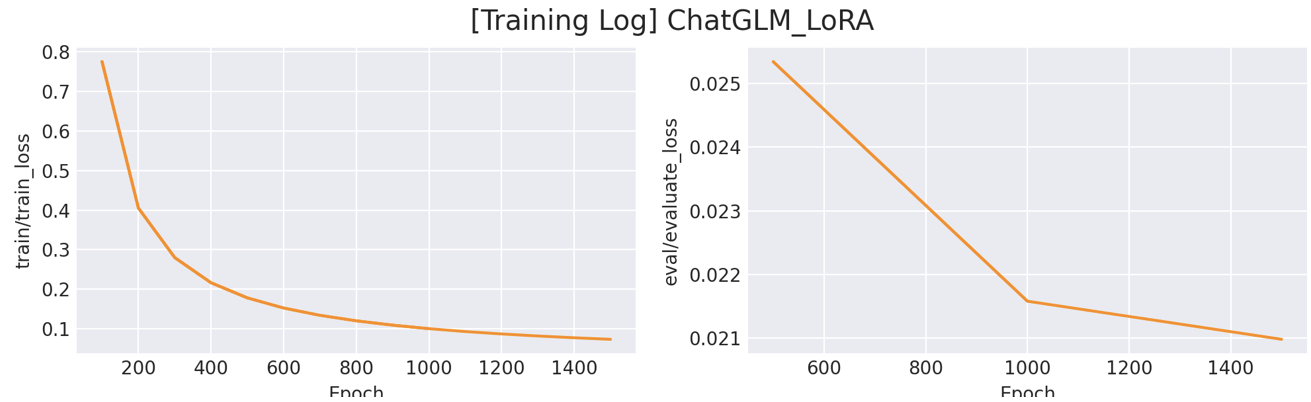
Shell | 复制代码

```
=== 回答符合预期 ===
<|Human|>: 某市举行小学数学竞赛，结果不低于80分的人数比80分以下的人数的4倍还多2人，及格的人数比不低于80分的人数多22人，恰是不及格人数的6倍，求参赛的总人数?
> begin generated, wait a moment...
<|ChatGLM-6B|>: 解题关键思路:
1. 总人数 = 不低于80分的人数 + 80分以下的人数 = 及格的人数 + 不及格的人数。
2. 及格的人数是不及格人数的6倍。

解: 设 不低于80分 的为 x 人，则:
80分以下的人数为 (x-2) / 4,
及格的人数为 x + 22,
所以不及格的人数为:
总人数 - 及格的人数 = x + (x-2) / 4 - (x + 22) = (x - 90) / 4,
因为及格的人数是不及格人数的6倍，所以:
x + 22 = 6 * (x - 90) / 4
解得 x = 314,
则 80分以下的人数为  (x-2) / 4 = 78,

所以 参数总人数 = 314 + 78 = 392。
> use time: 15.763868570327759
```

资源占用

- lora_rank=4
- batch_size=2
- num_train_epochs=250
- learning_rate=2e-4
- max_source_seq_len=230
- max_target_seq_len=230

单卡：



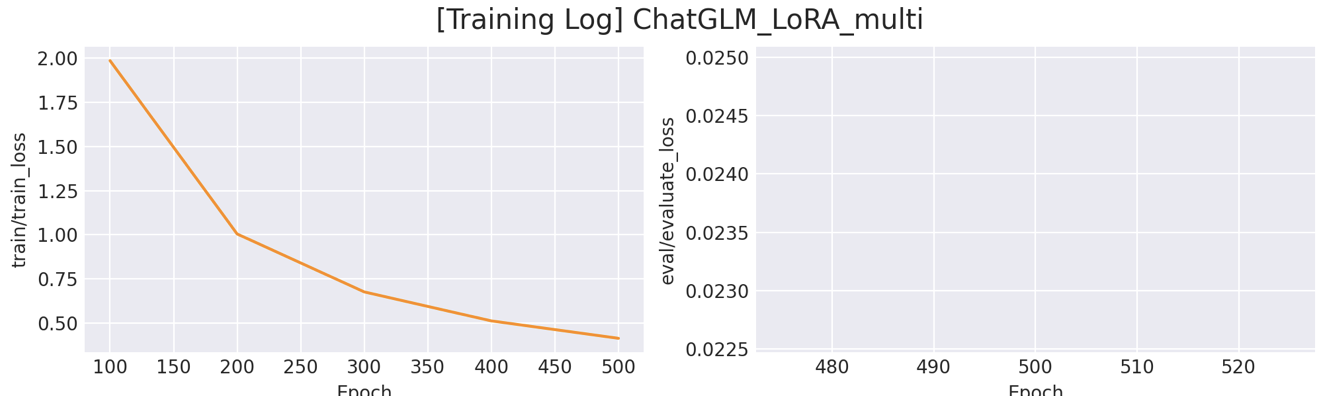

```
|   2  Tesla V100-SXM2...  Off  | 00000000:86:00.0 Off |                    0
| N/A   60C    P0   273W / 300W |  32454MiB / 32768MiB |    100%      Default
|                               |                      |                    N/A
```

Shell | 复制代码

```
global step 1500 ( 100.00% ) , epoch: 250, loss: 0.07372, speed: 1.18 step/s
train run time: 21分 12秒
```

多卡：

> eval图没有正常：save_freq比较大、step分成两份比较小，双卡训练过程中就验证了一次，一个点画描不了线




```
  2  Tesla V100-SXM2...  Off  | 00000000:86:00.0 Off |                    0
N/A   60C    P0   249W / 300W |  32484MiB / 32768MiB |    100%      Default
                               |                      |                  N/A
-------------------------------+----------------------+----------------------
  3  Tesla V100-SXM2...  Off  | 00000000:AF:00.0 Off |                    0
N/A   63C    P0   264W / 300W |  32484MiB / 32768MiB |    100%      Default
                               |                      |                  N/A
```

Shell | 复制代码

```
global step 700 ( 46.67% ) , epoch: 234, loss: 0.30190, speed: 1.17 step/s
train run time: 09分 22秒
```