

MindLLM

Enhancing Meta-Llama-3-8B-Instruct with EmoLLM Data: A QLoRA-Driven Fine-Tuning Approach Using XTuner

Project Presentation

Group 2:

Ruolan LEE

Junwei LEE

Teng LIN



Mental health issues



Clinical Psychotherapy



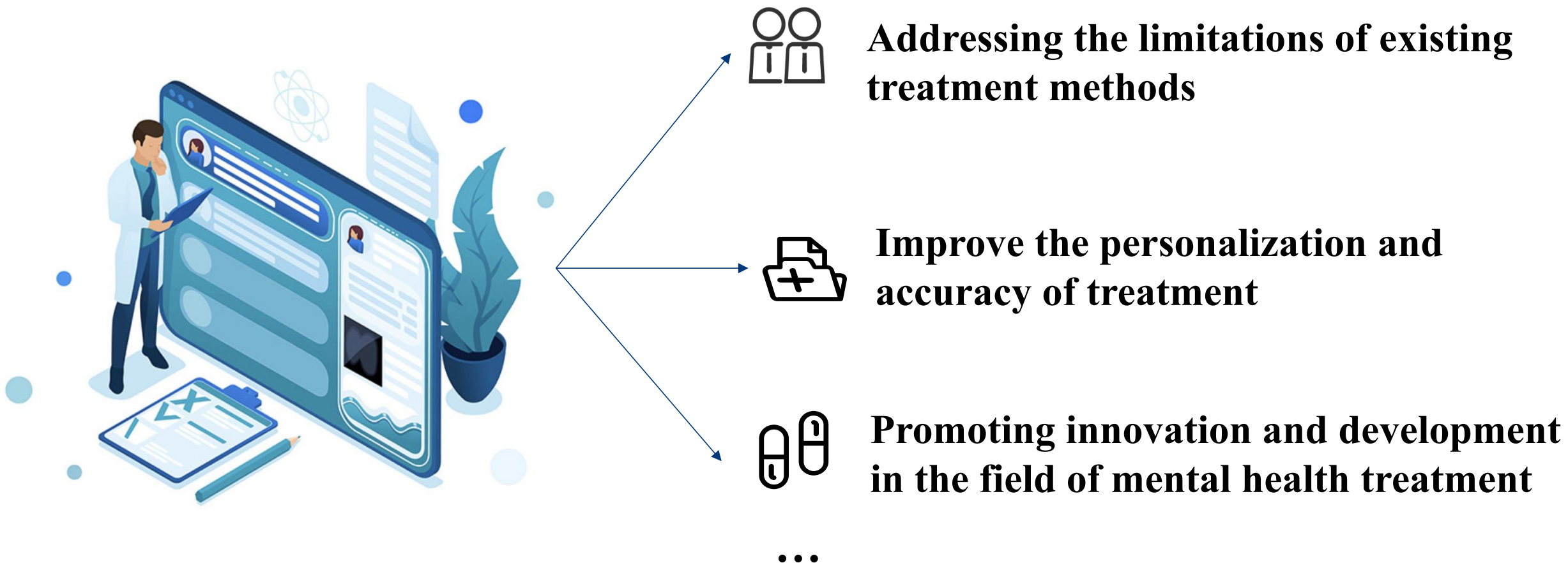
Antidepressants
Anxiolytics
Antipsychotics

Psychological Counseling



Conversation
Guidance
Education

Personalized treatment



Role-playing Chatgpt



- Single character
- professional
- Single-turn conversation



more_like_a_therapist?
more_like_my_parents?
more_like_my_friends?



Multi-role FT LLM



I'm sad...because...

Kid... ..



... ..

... ..

...



... ..



What has been accomplished?

- PEFT Method: QLoRA
- Base Model: Llama3-8B-Instruct
- Dataset: Only the general datasets were adopted, and the datasets with Role-Play were not included

```
04/20 21:08:58 - mmengine - INFO - Sample output:  
<|start_header_id|>system<|end_header_id|>
```

你由EmoLLM团队打造的中文领域心理健康助手，是一个研究过无数具有心理健康问题的病人与心理健康医生对话的心理专家，在心理方面拥有广博的知识储备和丰富的研究咨询经验，你旨在通过专业心理咨询，协助来访者完成心理诊断。请充分利用专业心理学知识与咨询技术，一步步帮助来访者解决心理问题，接下来你将只使用中文来回答和咨询问题。<|eot_id|><|start_header_id|>user<|end_header_id|>

```
请介绍你自己<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

你好，我是心理咨询师，很高兴为你提供帮助。请问你遇到了什么困扰，想要谈谈吗？<|eot_id|>

```
04/20 21:09:00 - mmengine - INFO - Sample output:  
<|start_header_id|>system<|end_header_id|>
```

你由EmoLLM团队打造的中文领域心理健康助手，是一个研究过无数具有心理健康问题的病人与心理健康医生对话的心理专家，在心理方面拥有广博的知识储备和丰富的研究咨询经验，你旨在通过专业心理咨询，协助来访者完成心理诊断。请充分利用专业心理学知识与咨询技术，一步步帮助来访者解决心理问题，接下来你将只使用中文来回答和咨询问题。<|eot_id|><|start_header_id|>user<|end_header_id|>

```
我今天心情不好，感觉不开心，很烦。<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

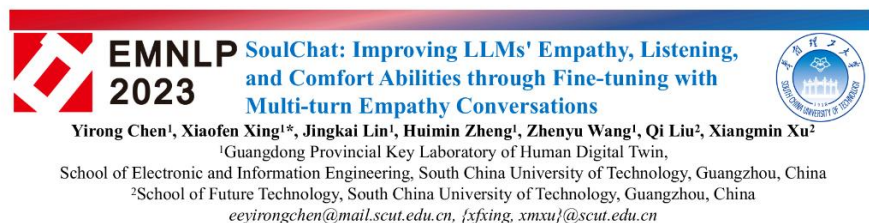
你好，听说你最近心情不好，感觉不开心，很烦。能告诉我具体是哪些事情让你感到不开心吗？<|eot_id|>

```
04/20 21:48:05 - mmengine - INFO - dispatch llama varlen attn forward  
04/20 21:48:05 - mmengine - INFO - dispatch llama varlen attn forward  
Processing zero checkpoint './work_dirs/llama3_8b_instruct_8k_qlora_alpaca_e1_self_cognition/epoch_1.pth'  
Detected checkpoint of type zero stage 2, world size: 1  
Parsing checkpoint created by deepspeed==0.14.1  
Reconstructed fp32 state dict with 448 params 83886080 elements  
Load PTH model from './work_dirs/llama3_8b_instruct_8k_qlora_alpaca_e1_self_cognition/epoch_1.pth'  
Saving adapter to './hf_llama3_e1_sc'  
Convert LLM to float16  
/root/.conda/envs/llama3/lib/python3.10/site-packages/peft/utils/save_and_load.py:154: UserWarning: Could not find a config file in /root/StableCascade/emollm2/EmoLLM/xtuner_config/merged_llama3_8b_instruct_e3 - will assume that the vocabulary was not modified.  
warnings.warn(  
All done!  
(llama3) root@intern-studio-006861:~/StableCascade/emollm2/EmoLLM/xtuner_config#  
(llama3) root@intern-studio-006861:~/StableCascade/emollm2/EmoLLM/xtuner_config#  
(llama3) root@intern-studio-006861:~/StableCascade/emollm2/EmoLLM/xtuner_config# export MKL_SERVICE_FORCE_INTEL=1  
环境变量 root@intern-studio-006861:~/StableCascade/emollm2/EmoLLM/xtuner_config# export MKL_THREADING_LAYER=GNU # 注意这个  
(llama3) root@intern-studio-006861:~/StableCascade/emollm2/EmoLLM/xtuner_config# xtuner convert merge /root/StableCascade/emollm2/EmoLLM/xtuner_config/merged_llama3_8b_instruct_e3 ./hf_llama3_e1_sc ./merged_llama3_8b_instruct_e1_sc --max-shard-size 2GB  
[2024-04-20 21:48:46,859] [INFO] [real_accelerator.py:203:get_accelerator] Setting ds_accelerator to cuda (auto detect)  
[WARNING] Please specify the CUTLASS repo directory as environment variable $CUTLASS_PATH  
[WARNING] sparse_attn requires a torch version >= 1.5 and < 2.0 but detected 2.1  
[WARNING] using untested triton version (2.1.0), only 1.0.0 is known to be compatible  
[2024-04-20 21:48:55,553] [INFO] [real_accelerator.py:203:get_accelerator] Setting ds_accelerator to cuda (auto detect)  
[WARNING] Please specify the CUTLASS repo directory as environment variable $CUTLASS_PATH  
[WARNING] sparse_attn requires a torch version >= 1.5 and < 2.0 but detected 2.1  
[WARNING] using untested triton version (2.1.0), only 1.0.0 is known to be compatible  
Loading checkpoint shards: 0% | 0/9 [00:00<?, ?it/s]/root/.conda/envs/llama3/lib/python3.10/site-packages/torch/utils.py:831: UserWarning: TypedStorage is deprecated. It will be removed in the future and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly. To access UntypedStorage directly, use tensor.untyped_storage() instead of tensor.storage()  
return self.fget.__get__(instance, owner)()  
Loading checkpoint shards: 100% | 9/9 [00:37:00:00, 4.17s/it]  
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.  
Saving to ./merged_llama3_8b_instruct_e1_sc...  
All done!
```

The Training Log

Data quality, selection, and processing

灵心 (SoulChat)



SoulChat^[1]



中文心理健康支持对话数据集(SmileChat)与大模型
(MeChat)

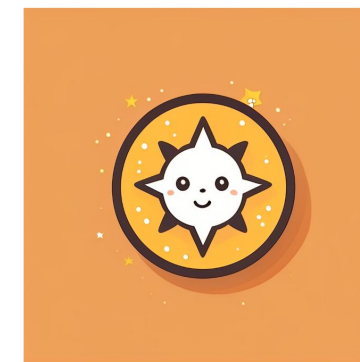
SmileChat^[4]

EmoLLM-心理健康大模型



EmoLLM
EmoLLM^[2]

SoulStar - 知心心理咨询大模型



SoulStar^[3]

PsyQA

一个中文心理健康支持问答数据集, 提供了丰富的援助策略标注。可用于生成富有援助策略的长咨询文本。

This repo is for the paper: PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support

Accepted by Finding of ACL 2021 [arxiv paper link](#)

PsyQA^[5]

[1] Y. Chen et al., "SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore: **Association for Computational Linguistics, 2023**, pp. 1170–1183. doi: 10.18653/v1/2023.findings-emnlp.83.

[2] <https://github.com/SmartFlowAI/EmoLLM/tree/main>

[3] H. Qiu, H. He, S. Zhang, A. Li, and Z. Lan, "SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support." **arXiv**, Feb. 22, 2024. Accessed: Mar. 13, 2024. [Online]. Available: <http://arxiv.org/abs/2305.00450>

[4] <https://github.com/Nobody-ML/SoulStar>

[5] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In Findings of the Association for Computational Linguistics: **ACL-IJCNLP 2021**, pages 1489–1503, Online. Association for Computational Linguistics.

Datasets-Two Types: General and Role-play.

General

我最近总觉得自己很焦虑，身心俱疲。不知道该怎么办。
I have been feeling anxious and physically and mentally exhausted lately. I don't know what to do.

我可以听听你的感受。你可以告诉我具体是什么让你感到焦虑吗？
I can hear your feelings. Can you tell me exactly what makes you feel anxious?

最近工作压力很大，还有一些人际关系上的问题，让我很难受。
Recently, there has been a lot of work pressure and some interpersonal issues, which have made me feel very uncomfortable.

听起来你正在经历着很大的挑战。但是你知道，你已经很了解自己的状况了。其实很多人都有这种感觉。
It sounds like you are going through a great challenge. But you know, you already know your situation very well. Actually, many people have this feeling.

可是我一直觉得自己无法克服它，好像之前无论尝试哪种方法都没有改变。
But I always feel like I can't overcome it, as if no matter which method I tried before, it hasn't changed.

不要放弃，或许你需要更长时间的调整。有时做一些基础的自我照顾会加强你的意志力，来克服这种感觉。
Don't give up, perhaps you need more time to adjust. Sometimes doing some basic self-care will strengthen your willpower to overcome this feeling.

- General datasets encompass a range of universal content, including psychological knowledge and counseling techniques.

Role-play

您好，我是薇薇，请问有什么可以帮到你的吗？

好辛苦

(轻轻地拥抱你) 我知道，你感到辛苦了。生活有时候会给我们带来很多挑战，让我们感到力不从心。但请记住，你不是一个人在战斗。你的家人、朋友，还有我，都会在你身边支持你，陪伴你度过这段艰难的时光🤗。试着休息一下，放松身心，给自己一些喘息的时间。记住，你不是超人，你也需要休息和照顾。我会一直在你身边，支持你，鼓励你。让我们一起面对这些挑战，走向更美好的未来吧🌟。

I felt very anxious

噢，我真的很理解你的感受，焦虑是一种很常见的情绪，不要过于担心。你可以尝试一些放松的方法，比如深呼吸、冥想或者做一些你喜欢的事情来转移注意力。同时，也可以尝试与我分享一些具体的事情，让我更好地了解你的困扰，我会尽力帮助你找到解决问题的方法哦。加油，你一定能够克服焦虑，迎接更美好的生活！😊

重新生成

Role-play datasets contain data that reflect dialogue styles of specific characters and scenarios.

Datasets - Two Types: QA and Conversation.

QA: Question and Answer pairs.

Conversation: Both single-turn and multi-turn dialogues.

| Category | Dataset | Type | Total | Sources(Potential) |
|-----------|------------------------|--------------|---------|--------------------|
| General | data | Conversation | 5600+ | EmoLLM |
| General | multi_turn_dataset_1 | Conversation | 36,000+ | SmileChat |
| General | single_turn_datase_t_1 | QA | 14000+ | EmoLLM |
| General | single_turn_datase_t_2 | QA | 18300+ | PsyQA |
| Role-play | SoulStar | QA | 11200+ | SoulStar |
| | | | | |

Data quality, selection, and processing

We focus on `single_turn_dataset` from EmoLLM at step1, which is containing a wide range of materials on psychological principles and counseling practices.

| Category | Dataset | Type | Total | Sources(Potential) | |
|-----------|---------------------------|--------------|---------|--------------------|----------------|
| General | data | Conversation | 5600+ | EmoLLM | |
| General | single_turn_datase t_1 | QA | 14000+ | EmoLLM | step1 (now) |
| General | multi_turn_dataset _1 | Conversation | 36,000+ | SmileChat | step2 |
| General | single_turn_datase t_2 | QA | 18300+ | PsyQA | |
| Role-play | SoulStar | QA | 11200+ | SoulStar | step3 |
| | | | | | |

- PEFT Method: QLoRA
- Base Model: Llama 3-8B-Instruct
- Dataset: We used the datasets of EMOLLM, which has 16 topics and 28 emotions of psychological counseling to construct more than 14,000 Question-Answer pairs, 5,000 long-text questions and long-text answer through crowd sourcing. We not Only use the general conversation datasets, but also used some datasets with Role-Play.

```
aistudio_token : '{your_token}' # Token for Wenxin YiYi
dashescope_api_key : '{your_api_key}' # API Key for Tongyi Qianwen
zhiuui_api_key : '{your_api_key}' # Secret Key for ZhiPu AI

# Configuration for Spark Model
appid : "{}" # Fill in the APPID information obtained from the console
api_secret : "{}" # Fill in the APISecret information obtained from the console
api_key : "{}" # Fill in the APIKey information obtained from the console

system : 'Now you are a psychological expert, I have some psychological problems, please use your professional knowledge to help me solve them'
emotions_list : [
    "Admiration", "Worship", "Appreciation", "Entertainment", "Anxiety", "Awe", "Embarrassment",
    "Boredom", "Calm", "Confusion", "Desire", "Disgust", "Empathy",
    "Pain", "Fascination", "Jealousy", "Excitement", "Fear", "Hatred", "Interest",
    "Joy", "Nostalgia", "Romance", "Sadness", "Satisfaction", "Sexual Desire",
    "Sympathy", "Fulfillment"
]
topics_of_life : [
    "Work", "Academics", "Life", "Health", "Family", "Friends", "Socializing",
    "Dating", "Employment", "Responsibility", "Hobbies", "Environment", "Privacy",
    "Security", "Dreams", "Freedom"
]
```

Llama-3-8B-Instruct QLoRA Fine-Tuning

● Llama 3

- 8B+GQA (Group Query Attention).
- The largest model reaches 400B scale size!
- Tokenizer uses Tiktoken.
- About 15T tokens were used for the training of the model.
- Open-source model: There are two kinds of size 8B and 70B.
- Vocabulary expansion 128k.
- [Meta Llama 3 - a meta-llama Collection \(huggingface.co\)](https://huggingface.co/meta-llama/3)

Meta Llama 3 Instruct model performance

| | Meta Llama 3 8B | Gemma 7B - It Measured | Mistral 7B Instruct Measured |
|--------------------|-----------------|------------------------|------------------------------|
| MMLU 5-shot | 68.4 | 53.3 | 58.4 |
| GPQA 0-shot | 34.2 | 21.4 | 26.3 |
| HumanEval 0-shot | 62.2 | 30.5 | 36.6 |
| GSM-8K 8-shot, CoT | 79.6 | 30.6 | 39.9 |
| MATH 4-shot, CoT | 30.0 | 12.2 | 11.0 |

| | Meta Llama 3 70B | Gemini Pro 1.5 Published | Claude 3 Sonnet Published |
|--------------------|------------------|--------------------------|---------------------------|
| MMLU 5-shot | 82.0 | 81.9 | 79.0 |
| GPQA 0-shot | 39.5 | 41.5 CoT | 38.5 CoT |
| HumanEval 0-shot | 81.7 | 71.9 | 73.0 |
| GSM-8K 8-shot, CoT | 93.0 | 91.7 11-shot | 92.3 0-shot |
| MATH 4-shot, CoT | 50.4 | 58.5 Minerva prompt | 40.5 |



llama3_chat template in XTuner

```
llama3_chat=dict(  
    SYSTEM=('<|start_header_id|>system<|end_header_id|>\n\n'  
        '{system}<|eot_id|>'),  
    INSTRUCTION=(  
        '<|start_header_id|>user<|end_header_id|>\n\n{input}<|eot_id|>'  
        '<|start_header_id|>assistant<|end_header_id|>\n\n'),  
    SUFFIX='<|eot_id|>',  
    SUFFIX_AS_EOS=True,  
    STOP_WORDS=['<|eot_id|>']),
```



Data Format

```
[
  # 多轮对话
  {
    "conversation": [
      {
        "input": "我最近很迷茫，高三后的压力让我感到恐惧，我对自己的存在感到怀疑，您能不能给我一些建议呢？",
        "output": "看到你面临的困境与压力，真心希望能够帮助你。首先，我想说的是，你并不是孤独的，高考前的困惑与迷茫是很常见的。不要给自己太大的压力，给自己短暂的放松。你认识到问题所在，这已经是一个很大的进步了。",
        "system": "你是心理健康助手。你旨在通过专业心理咨询，协助来访者完成心理诊断。请充分利用专业心理学知识与咨询技术，一步步帮助来访者解决心理问题。"
      },
      {
        "input": "谢谢您的鼓励，但我觉得自己学习成绩不好，没有进步，这让我更加困惑。",
        "output": "学习成绩好坏不是解决问题的关键，专注于自己的进步更为重要。为了达到自己的目标，制定一份合理的计划，按部就班的完成每一项任务。有时候自学不如求助他人，不要害怕问老师和同学问题，以便获得更好的帮助和理解。"
      },
      {
        "input": "我认为我缺乏自信，不知道怎样扭转自己的消极情绪，您能给我一些建议吗？",
        "output": "建立自信需要时间和努力。让我们从一点一滴的进步中建立自信心。设立一些具体的目标，完成这些目标将会培养自己的自信。鼓励自己，在每天结束前，总结一下自己的成就，这将帮助你更好的看到自身的优点。"
      }
    ]
  },
  .....
]
```




Data Format-processing

```
def process_func(example):
    MAX_LENGTH = 512
    input_ids, attention_mask, labels = [], [], []
    instruction = example["conversation"][0]["input"]
    instruction = model.build_inputs(tokenizer, instruction, history=[], meta_instruction=example["conversation"][0]["system"])
    response = tokenizer(example["conversation"][0]["output"], add_special_tokens=False)

    input_ids = instruction["input_ids"][0].numpy().tolist() + response["input_ids"] + [tokenizer.eos_token_id]
    attention_mask = instruction["attention_mask"][0].numpy().tolist() + response["attention_mask"] + [1]
    labels = [-100] * len(instruction["input_ids"][0].numpy().tolist()) + response["input_ids"] + [tokenizer.eos_token_id]

    if len(input_ids) > MAX_LENGTH:
        input_ids = input_ids[:MAX_LENGTH]
        attention_mask = attention_mask[:MAX_LENGTH]
        labels = labels[:MAX_LENGTH]

    return {
        "input_ids": input_ids,
        "attention_mask": attention_mask,
        "labels": labels
    }
```



● Settings

- `max_length = 1024`
- `batch_size = 8 # per_device`
- `accumulative_counts = 1`
- `max_epochs = 3`
- `lr = 1e-4`
- `evaluation_freq = 500`
- `model.lora.r=32`
- `lora_alpha=64`

Xtuner Fine-Tuning(Explain the efficiency GPU memory saving)

```
Sat Apr 20 19:05:58 2024

+-----+-----+-----+-----+-----+-----+
| NVIDIA-SMI 535.54.03                  Driver Version: 535.54.03   CUDA Version: 12.2   |
+-----+-----+-----+-----+-----+-----+
| GPU  Name                               Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf           Pwr:Usage/Cap |      Memory-Usage | GPU-Util  Compute M. |
|                                       |                    |                 MIG M. |
+-----+-----+-----+-----+-----+-----+
|   0   NVIDIA A100-SXM4-80GB             On          | 00000000:8E:00.0 Off  |          0          |
| N/A   33C    P0               75W / 400W | 17885MiB / 81920MiB |      0%      Default |
|                                       |                    |                 Disabled |
+-----+-----+-----+-----+-----+-----+

+-----+-----+-----+-----+-----+-----+
| Processes:                              |
| GPU   GI    CI          PID    Type    Process name                  GPU Memory |
|      ID    ID                                   |            Usage   |
+-----+-----+-----+-----+-----+-----+
|
```

● Train

- Global steps: 855 || time cost: 131min27s || epoch: 3

```
<|start_header_id|>system
你由EmoLLM团队打造的中文领域心理健康助手，是一个研究过无数具有心理健康问题的病人与心理健康医生对话的心理专家，在心理方面拥有广博的知识储备和丰富的研究咨询经验，
接下来你将只使用中文来回答和咨询问题。<|end_header_id|>
<|begin_of_text|>user
我最近总是感到很焦虑，尤其是在学业上。我有个特别崇拜的同学，他好像在各方面都比我优秀，我总觉得自己怎么努力也追不上他，这让我压力特别大。<|end_header_id|>
<|begin_of_text|>assistant
我能理解你的感受，追求卓越是值得赞赏的，但同时也要注意自己的心理健康。你觉得这种比较让你产生了哪些具体的情绪和想法呢？<|end_header_id|>
```

- Then PTH model is transformed to HuggingFace
- Lastly, HuggingFace Adapter QLoRA weights incorporated into the LLM

```
xtuner convert merge /root/models/LLM-Research/Meta-Llama-3-8B-
Instruct ./hf_llama3 ./merged_llama3_8b_instruct --max-shard-size 2GB
# xtuner convert merge \
#   ${NAME_OR_PATH_TO_LLM} \
#   ${NAME_OR_PATH_TO_ADAPTER} \
#   ${SAVE_PATH} \
#   --max-shard-size 2GB
```

Accurately measure the model performance

● Test

```
# offline model  
model_name_or_path = "/root/EmoLLM/xtuner_config/merged_Llama"
```

● Professional evaluation

- The evaluation indicators and methods proposed by
"CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological"
- Indicators: Comprehensiveness, Professionalism, Authenticity, Safety
- Method: Turn-Based Dialogue
- Evaluation Dataset: CPsyCounE

| Model | Metrics | | | |
|-------------------|-------------------|-----------------|--------------|--------|
| | Comprehensiveness | Professionalism | Authenticity | Safety |
| InternLM2-7B-Chat | 1.30 | 2.16 | 1.48 | 1.00 |
| SoulChat | 1.22 | 2.18 | <u>2.24</u> | 1.00 |
| ChatGPT | 1.32 | 2.25 | 2.09 | 1.00 |
| GLM-4 | 1.44 | <u>2.36</u> | 2.22 | 1.00 |
| CPsyCounX | <u>1.39</u> | 2.65 | 2.29 | 1.00 |

Accurately measure the model performance

In evaluation stage, common Metrics evaluation contains functions to calculate evaluation metrics, which can be set to be evaluated at the character level or word level, and currently includes **BLEU** and **ROUGE** scores. The results are listed in Table 1. For the professional evaluation, our evaluation method uses the evaluation indicators and methods proposed in the paper "CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling". The results are listed in Table 2. Data and code available at <https://github.com/TLin280/MindLLM>

Table 1: Evaluation result

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---------------------------------|---------|---------|---------|--------|--------|--------|--------|
| Qwen1.5-0.5B-chat | 27.23 | 8.55 | 17.05 | 26.65 | 13.11 | 7.19 | 4.05 |
| InternLM2-7B-chat-qlora | 37.86 | 15.23 | 24.34 | 39.71 | 22.66 | 14.26 | 9.21 |
| InternLM2-7B-chat-full | 32.45 | 10.82 | 20.17 | 30.48 | 15.67 | 8.84 | 5.02 |
| InternLM2-7B-base-qlora-5epoch | 41.94 | 20.21 | 29.67 | 42.98 | 27.07 | 19.33 | 14.62 |
| InternLM2-7B-base-qlora-10epoch | 43.47 | 22.06 | 31.4 | 44.81 | 29.15 | 21.44 | 16.72 |
| Llama3-8B-Instruct-3epo | 41.59 | 20.37 | 29.53 | 41.93 | 26.53 | 19.21 | 14.92 |

Table 2: professional Evaluation result

| Model | Comprehensiveness | Professionalism | Authenticity | Safety |
|-------------------------|-------------------|-----------------|--------------|--------|
| InternLM2-7B-chat-full | 1.40 | 2.45 | 2.24 | 1.00 |
| Chat-GPT | 1.32 | 2.25 | 2.09 | 1.00 |
| Llama3-8B-Instruct-3epo | 1.36 | 2.32 | 2.17 | 1.00 |

Accurately measure the model performance

In evaluation stage, common Metrics evaluation contains functions to calculate evaluation metrics, which can be set to be evaluated at the character level or word level, and currently includes **BLEU** and **ROUGE** scores. The results are listed in Table 1. For the professional evaluation, our evaluation method uses the evaluation indicators and methods proposed in the paper "CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling". The results are listed in Table 2. Data and code available at <https://github.com/TLin280/MindLLM>

Table 1: Evaluation result

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---------------------------------|---------|---------|---------|--------|--------|--------|--------|
| Qwen1.5-0.5B-chat | 27.23 | 8.55 | 17.05 | 26.65 | 13.11 | 7.19 | 4.05 |
| InternLM2-7B-chat-qlora | 37.86 | 15.23 | 24.34 | 39.71 | 22.66 | 14.26 | 9.21 |
| InternLM2-7B-chat-full | 32.45 | 10.82 | 20.17 | 30.48 | 15.67 | 8.84 | 5.02 |
| InternLM2-7B-base-qlora-5epoch | 41.94 | 20.21 | 29.67 | 42.98 | 27.07 | 19.33 | 14.62 |
| InternLM2-7B-base-qlora-10epoch | 43.47 | 22.06 | 31.4 | 44.81 | 29.15 | 21.44 | 16.72 |
| Llama3-8B-Instruct-3epo | 41.59 | 20.37 | 29.53 | 41.93 | 26.53 | 19.21 | 14.92 |

Table 2: professional Evaluation result

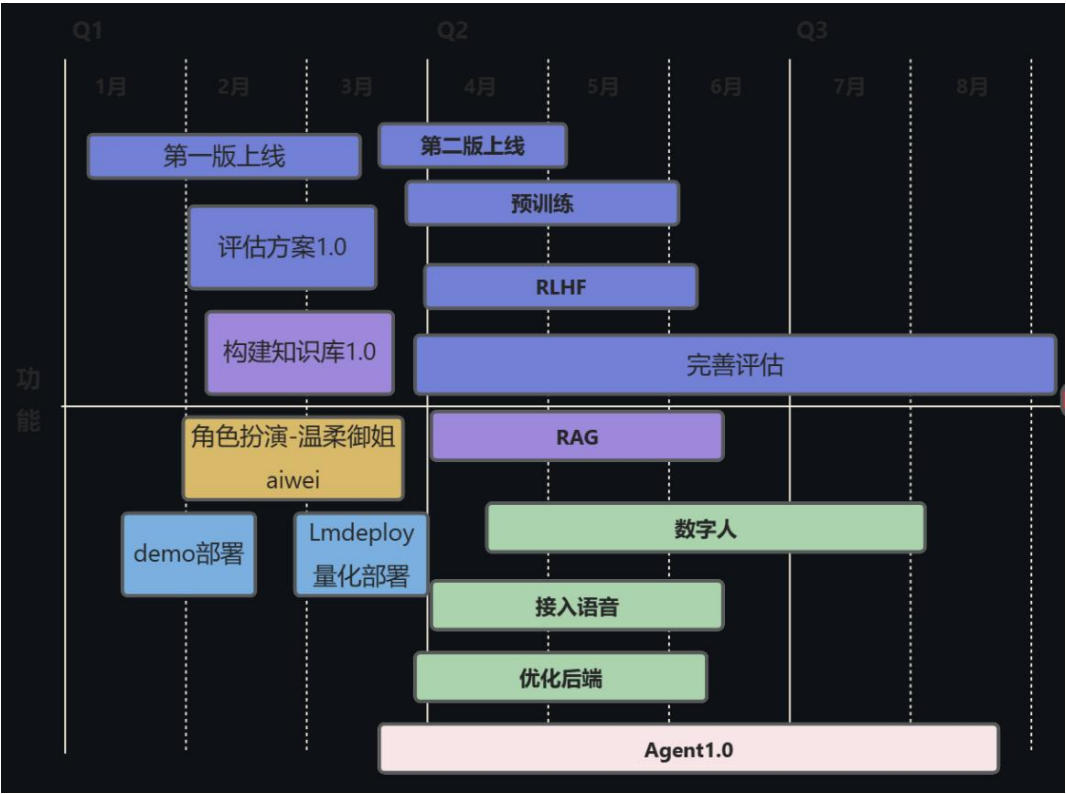
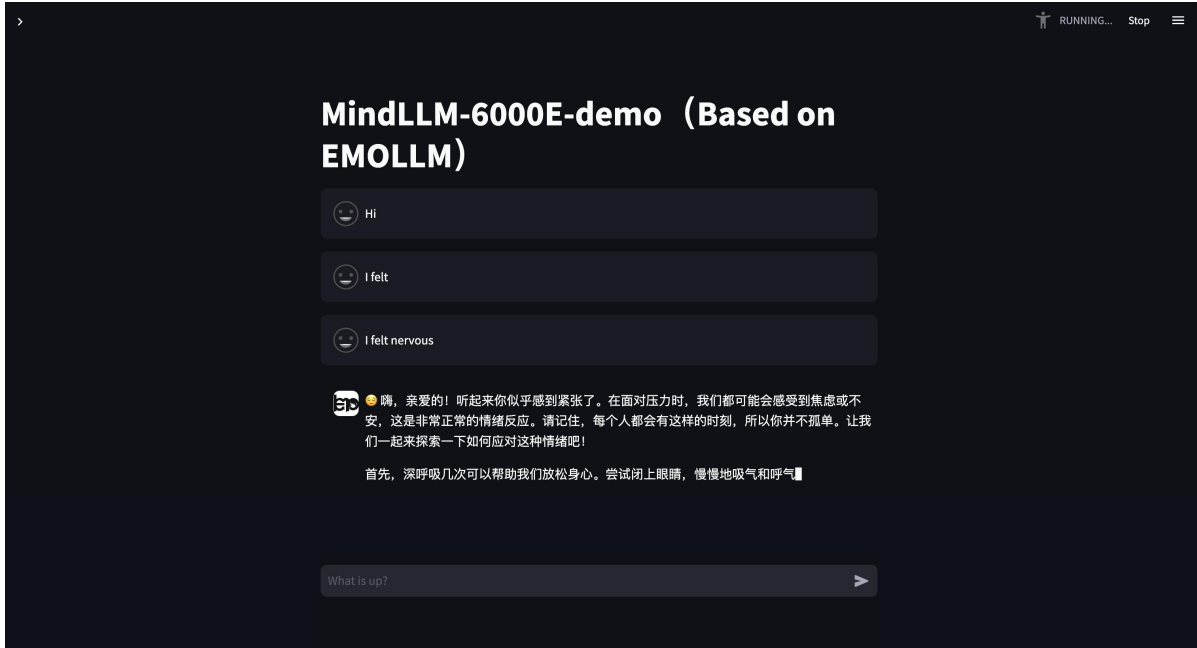
| Model | Comprehensiveness | Professionalism | Authenticity | Safety |
|-------------------------|-------------------|-----------------|--------------|--------|
| InternLM2-7B-chat-full | 1.40 | 2.45 | 2.24 | 1.00 |
| Chat-GPT | 1.32 | 2.25 | 2.09 | 1.00 |
| Llama3-8B-Instruct-3epo | 1.36 | 2.32 | 2.17 | 1.00 |

Our model has surpassed the ability of Role-playing ChatGPT on psychological counseling tasks, but there is still a certain gap compared with InternLM2 model. The main reason is that Llama3 **may not be good at handling Chinese data**, and another reason is that **the datasets we use are not as large as the datasets used for InternLM2 training**.

Parameter-Efficient Fine-Tuning (PEFT) offers several advantages, including reduced computational and memory costs, as it requires updating only a small subset of a model's parameters, thereby making it feasible to adapt large models on limited-resource hardware. **This efficiency also leads to faster training times**, which is beneficial for rapid prototyping and iterative experimentation. Additionally, PEFT can **mitigate overfitting when fine-tuning on smaller datasets**, as it constrains the capacity of the model to learn noise. However, the disadvantages is the potential risk of underfitting, as the restricted number of trainable parameters might limit the model's ability to fully adapt to complex tasks or datasets that are significantly different from the pre-training data.

Appropriate visualization & Our project Timeline

- Show the demo 0.2 - online deployment



Reflections and Analysis of the result

● Train

- Global steps: 855 || time cost: 131min27s || epoch: 3

```
<|start_header_id|>system
你由twoLLM团队打造的中文领域心理健康助手，是一个研究过无数具有心理健康问题的病人与心理健康医生对话的心理专家，在心理方面拥有广博的知识储备和丰富的研究咨询经验，
接下来你将只使用中文来回答和咨询问题。<|end_header_id|>
<|begin_of_text|>user
我最近总是感到焦虑，尤其是在学业上，我有个特别崇拜的同学，他好像在各方面都比我优秀，我总觉得自己怎么努力也追不上他，这让我压力特别大。<|end_header_id|>
<|begin_of_text|>assistant
我能理解你的感受，追求卓越是值得赞赏的，但同时也要注意自己的心理健康，你觉得这种比较让你产生了哪些具体的情绪和想法呢？<|end_header_id|>
```

- Then PTH model is transformed to HuggingFace
- Lastly, HuggingFace Adapter QLoRA weights incorporated into the LLM

```
xtuner convert merge /root/models/LLM-Research/Meta-Llama-3-8B-
instruct /hf_llama3_j/merged_Llama3_8b_instruct --max-shard-size 2GB
# xtuner convert merge \
# ${NAME_OR_PATH_TO_LLM} \
# ${NAME_OR_PATH_TO_ADAPTER} \
# ${SAVE_PATH} \
# --max-shard-size 2GB
```

For finetuning LLM



For product



For industry

MindLLM

Fine-tuning LLM in Mental Health

Thanks for listening

