

《大语言模型与人工智能工程设计》

大作业任务

2024-2025 学年第一学期

项目名称/Project Name: “和心者”心理咨询大模型 HearMeGLM

学生姓名/Student Name: 魏海川

学号/Student ID: 2022xxxx

班级名称/Class Name: 2022 级人工智能 1 班

授课教师姓名

/Instructor's Name: 刘烨

成绩/Grade:

评阅老师签名/Reviewer's

Signature:

任务	完成一个基于大语言模型的训练任务		
要求	<p>1、应用方向选择：</p> <p>(1) 自选一个应用方向，比如法律、翻译、金融、新闻写作、医疗健康、教育、社交媒体等。建议选择具体且具备实际应用价值的领域。</p> <p>(2) 提交一段简短的说明，阐述选择该应用方向的理由及其社会价值。</p> <p>2、模型和数据集准备：</p> <p>(1) 自备数据集来进行模型微调，数据集应至少包含 500 条多样化样本，确保足够支持微调训练。</p> <p>(2) 数据集的来源需合法。</p> <p>(3) 自选模型，根据任务要求选择合适的模型作为基座模型。</p> <p>3、训练过程：</p> <p>(1) 描述训练过程，包括超参数的选择、微调步骤、过程记录以及结果的截图。</p> <p>(2) 进行数据集分析，比如数据划分为 70%训练集、15%验证集和 15%测试集，并说明理由。</p> <p>4、评估结果：</p> <p>(1) 提供微调前后模型的性能对比，展示关键性能指标（如准确率、召回率、F1 分数等）的显著提升。</p> <p>5、报告撰写：</p> <p>(1) 报告需要使用双语完成，文档名称以“学号-名字”命名，最后提交报告即可。</p> <p>(2) 选择应用方向的理由：阐述选择该方向解决的问题及实现的社会价值。</p> <p>(3) 样本分析：详细说明样本数据来源、数据分析和处理过程。</p> <p>(4) 微调过程：清晰描述微调步骤及超参数选择，并附上相关结果截图。</p> <p>(5) 评估结果：提供微调后模型的评估结果，并附上相关截图。</p> <p>(6) 应用推理：展示微调后模型的部署应用示例及推理结果截图。</p>		
评分标准	评分项	评分子项	说明
报告撰写 (20分)	内容完整性(8)	信息全面、详尽，逻辑清晰	
	排版美观性(5)	排版整齐，图文搭配合理，视觉效果佳	
	一致性(4)	字体、样式和大小完全统一	
	图文清晰度(3)	所有图表清晰，不存在模糊不清	
创新创意 (10分)		有创新应用，可以是技术，场景，使用方式	
训练过程 (70分)	样本数据(20)	样本数据分析,数据	
	微调过程(30)	有超参数使用说明，有详细的微调过程记录	
	验证与测试(20)	微调后模型性能有提升，能满足任务需求	

Task	Complete a training task based on a large language model			
Requirements	<p>1. Application Direction Selection:</p> <p>(1) Choose an application direction, such as law, translation, finance, news writing, healthcare, education, social media, etc. It is recommended to select a specific field with practical application value.</p> <p>(2) Submit a brief explanation outlining the reasons for selecting this application direction and its social value.</p> <p>2. Model and Dataset Preparation:</p> <p>(1) Prepare a dataset for model fine-tuning, which should contain at least 500 diverse samples to ensure sufficient support for training.</p> <p>(2) The source of the dataset must be legal.</p> <p>(3) Choose a model based on task requirements and select an appropriate base model.</p> <p>3. Training Process:</p> <p>(1) Describe the training process, including the selection of hyperparameters, fine-tuning steps, process records, and screenshots of results.</p> <p>(2) Conduct dataset analysis, such as dividing the data into 70% training set, 15% validation set, and 15% test set, and explain the reasons for this division.</p> <p>4. Evaluation Results:</p> <p>(1) Provide a performance comparison of the model before and after fine-tuning, showcasing significant improvements in key performance indicators (such as accuracy, recall, F1 score, etc.).</p> <p>5. Report Writing:</p> <p>(1) The report needs to be completed in both languages, and the document should be named in the format “StudentID-Name”. Finally, submit the report.</p> <p>(2) Reason for choosing the application direction: Explain the problem addressed by this direction and the social value achieved.</p> <p>(3) Sample Analysis: Provide a detailed explanation of the source of sample data, data analysis, and processing steps.</p> <p>(4) Fine-tuning Process: Clearly describe the fine-tuning steps and hyperparameter selection, and include relevant result screenshots.</p> <p>(5) Evaluation Results: Provide the evaluation results of the model after fine-tuning, along with relevant screenshots.</p> <p>(6) Application Inference: Show examples of the deployment application of the fine-tuned model and include screenshots of inference results.</p>			
Grading	Grading Items	Grading Sub-items	Explanation	Score
Criteria	Report Writing (20 points)	Content Completeness (8 points)	Information is comprehensive and detailed, with clear logic	
	Formatting	The layout is neat, the combination of text and		

	Aesthetics (5 points)	images is reasonable, and the visual effect is excellent	
	Consistency (4 points)	The font size, style, and line spacing are consistent	
	Clarity of Images and Text (3 points)	All charts and graphs are clear and free from any blurriness	
Innovation and Creativity (10 points)		There are innovative applications, which can include technology, scenarios, or methods of use	
Training Process (70 points)	Sample Data (20 points))	Analyze Sample Data	
	Training Process (30 points)	There is a description of hyperparameter usage, along with a detailed record of the fine-tuning process	
	Evaluation and Validation (20 points)	After fine-tuning, the model's performance has improved and meets the task requirements	

大语言模型和人工智能工程设计

Large Language Models and Artificial Intelligence Engineering Design

2025 年 7 月 24 日

大作业报告/Major Assignment Report

1 任务说明/Task Description

1.1 应用方向和社会价值说明/Application Direction and Social Value Description

”需要了解自己情感的人，也需要探寻心灵深处，往往内心深处隐藏着内心最要的东西。”

—卡尔·荣格（心理学家）

”Those who need to understand their emotions also need to explore the depths of their soul, for often, the most essential things lie hidden within.”

—Carl Jung (Psychologist)

在当今社会中，越来越快的生活节奏使得许多人无法适从，面对众多生活压力，越来越多人的心理防线正在被击穿。心理学常用“心理健康商数”（MHQ）这一指标衡量人们的心理健康水平。据统计，在 2020 年前后，MHQ 这一指标出现了全球范围内的普遍下降，并在 2021 年、2022 年两年持续维持在低水平。到 2023 年为止，无论是从全球层面还是从单个国家层面来看，目前都没有看到这一数据有重新复苏的迹象。2023 年，71 个国家的 MHQ 平均得分为 65 分（总分 300），27% 的人表示自己处于痛苦或挣扎的状态。^[1] 在后疫情时代，抑或是这个越发内卷的时代，人们迫切需要一个“心

“灵避风港”来缓解内心的焦虑与压力，让他们能够在这个高速运转的世界中找到片刻的宁静与自我平衡。

而目前线下针对心理问题患者的咨询服务依据患者心理健康情况、时间与地点不同等情况，其咨询价格普遍波动较大且价格昂贵。据平均统计数据显示，目前一次线下心理咨询的平均价格至少在几百元，而特定专家咨询的价格甚至需要几千元不等的价格。我们认为许多性格内向的心理问题患者没有勇气真正主动去沟通交流，并且这样的咨询价格对于许多人来说也是难以负担的。幸运的是，伴随着大语言模型技术的迅速发展，大语言模型的自然语言输出与其知识认知让低成本的心理咨询逐渐成为现实。

因此，在本次大语言模型的微调工作中，我所选择的具体方向为**心理健康咨询**，旨在通过微调大模型扮演一位名为 Jerry 的心理咨询师，在同用户进行对话交流的过程中安抚用户情绪并且对其进行排解与鼓励，实现缓解用户焦虑、使用户重拾信心的目的。

In today's society, the ever-increasing pace of life is overwhelming for many people. Faced with numerous life pressures, more and more people's psychological defences are being broken down. Psychology commonly uses the “mental health quotient” (MHQ) indicator to measure people's mental health level. According to statistics, around 2020, there was a general decline in the MHQ indicator worldwide, and it continued to remain at a low level in 2021 and 2022. Until 2023, there are no signs of recovery at the global or individual country level. In 2023, the average MHQ score for 71 countries was 65 out of 300, and 27% of people reported being in a state of distress or suffering. [1] In the post-pandemic era, or perhaps in this increasingly introverted era, people urgently need a “mental haven” to relieve their inner anxiety and stress, so that they can find a moment of tranquility and self-balance in this fast-moving world.

However, at present, offline counseling services for people with psychological problems generally fluctuate widely in price and are expensive, depending on the patient's mental health, time and location. According to average statistics, the average price of an offline psychological consultation is at least a few hundred RMB, and the price of a consultation with a specific expert can even range from a few thousand RMB. We believe that many introverted patients with psychological problems do not have the courage to really initiate communication, and the price of such a consultation is also unaffordable for many people. Fortunately, with the rapid development of large language model technology, the natural language output and knowledge of large language models have made low-cost psychological consultations a reality.

Therefore, in this fine-tuning of the large language model, I have chosen to specifically work on the direction of **mental health counseling**, aiming to play the role of a mental health counselor named Jerry through fine-tuning the large model. In the process of dialogue and communication with the user, Jerry will comfort the user's emotions and

provide them with relief and encouragement, with the goal of relieving the user's anxiety and restoring their confidence.

1.2 任务需求/Task Requirements

本次大语言模型微调的任务是训练一个具有心理咨询师角色的模型，因此主要应用方向属于偏向于角色扮演的问答系统。最终微调得到的模型首先需要拥有自我认知，认识到自己当前的角色以及任务，通过实现使用温柔且富有关心的语气去回应用户所提出的负面情绪，并且能够充分地理解用户当前的内心困惑，通过使用专业的心理知识来对用户进行排忧解难。同时，模型生成的内容应该具有拟人性，通过类似和真人一样对话聊天的形式以实现安抚问询用户的目的，而并非机械性地通过固定格式输出用户可以怎么做。此外，在模型的性能表现上，经过微调后的模型应该能够更加迅速地输出对应符合要求的文字内容，减少用户等待的时间以提升模型的实用性。最后，用户同模型间的问询交互还需要通过一个较容易上手使用的界面来实现，这也是本次心理咨询师微调任务的需求。

The task of this large language model fine-tuning is to train a model with the role of a psychological counselor, so the main application direction is a question-answering system that is biased towards role-playing. The final fine-tuned model first needs to have self-awareness, recognize its current role and task, respond to the user's negative emotions by using a gentle and caring tone, and fully understand the user's current internal confusion. It can also use professional psychological knowledge to solve the user's problems. At the same time, the content generated by the model should be anthropomorphic, and the purpose of reassuring the user being queried should be achieved through a form of dialogue similar to that of a real person, rather than mechanically outputting what the user can do through a fixed format. In addition, in terms of the model's performance, the fine-tuned model should be able to output the corresponding text content more quickly after being fine-tuned, reducing the waiting time for the user and improving the model's practicality. Finally, the interaction between the user and the model during the query needs to be implemented through an interface that is easy to use, which is also the requirement of this fine-tuning task for the psychological counselor.

1.3 创新创意/Innovation and Creativity

本文所提及的针对心理咨询师特定角色的大语言模型微调工作主要有如下几个较为与众不同的创新点：

The fine-tuning of the large language model for the specific role of a psychological counselor mentioned in this article mainly has the following relatively distinctive innova-

tive points:

1.3.1 基于可视化界面的微调操作/Fine-tuning based on a Visual Interface

通常来说，针对大语言模型的微调工作涉及到数据集格式的选取与导入、模型的加载、训练时超参数的设置、模型的验证设置与关键参数展示等诸多工作，对于传统的大语言模型微调工作而言这无疑是繁杂且低效的，在模型的微调过程中很容易出现因个别参数设置错误而导致模型训练效果不佳等问题的出现。为了提升模型总体的微调效率，并且简化我们的工作流程，本文在模型的微调过程中使用了 **LLaMA-Factory** 框架来进行模型的可视化微调工作。

LLaMA-Factory 是由 Yaowei Zheng 等人 [2] 提出的先进大语言模型微调框架，其通过 Gradio 的可视化框架实现对大语言模型的可视化微调。在整个微调训练过程中，其支持诸如 LLaMA、QWen、ChatGLM 以及 Baichuan 等多种不同框架的大语言模型，并且其支持在微调过程中使用多种微调计算方式，并同时在微调后直接进行模型的验证、对话以及生成导出等一系列操作，极大地加快了模型微调过程的效率。

在按照官方文档进行环境配置后，最终的 WebUI 界面如下图所示：

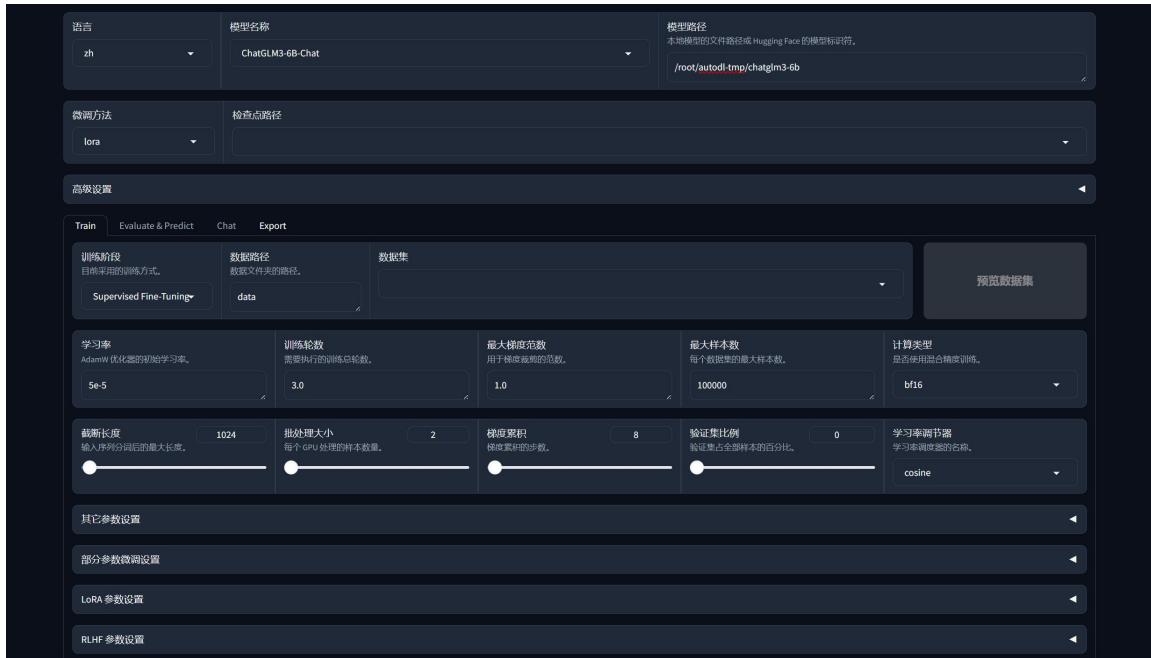


图 1: LLaMA-Factory 主界面展示/LLaMA-Factory Main Interface

Generally speaking, fine-tuning a large language model involves many tasks, such as selecting and importing the dataset format, loading the model, setting hyperparameters during training, setting model validation and displaying key parameters. This is undoubtedly complicated and inefficient for traditional fine-tuning of large language models. During the fine-tuning process, it is easy to encounter problems such as poor model

training results due to incorrect individual parameter settings. In order to improve the overall efficiency of model fine-tuning and simplify our workflow, this paper uses the framework of the LLaMA-Factory to perform model fine-tuning visually.

LLaMA-Factory is an advanced large language model fine-tuning framework proposed by Yaowei Zheng et al. in [2]. It uses the Gradio visualisation framework to fine-tune large language models visually. Throughout the fine-tuning training process, it supports large language models such as LLaMA, QWen, ChatGLM, and Baichuan, among others. It also supports the use of multiple fine-tuning calculation methods during the fine-tuning process, and at the same time, directly performs a series of operations such as model verification, dialogue, and generation export after fine-tuning, which greatly speeds up the efficiency of the model fine-tuning process. After configuring the environment according to the official documentation, the final WebUI interface is shown in the figure above.

1.3.2 基于 Lora 方法的模型快速微调/Rapid Fine-Tuning of Models Based on the Lora Method

Lora 微调方法由 Edward J. Hu 等人 [3] 于 2021 年提出，相较于传统全量微调更新全部参数，Lora 方法将模型所需要更新的权重矩阵将为低秩矩阵并进行计算更新，在显著降低对内存等计算资源的占用下实现对模型的参数调整。在本次的大语言模型微调任务中，借助 Lora 方法不仅极大地加快了模型微调速度，同时也使得在有限的计算资源下（仅需单张 RTX 4090）对于模型的微调成为可能。

The Lora fine-tuning method was proposed in 2021 by Edward J. Hu et al. [3]. Compared with the traditional full fine-tuning method that updates all parameters, the Lora method calculates and updates the weight matrix that needs to be updated by the model as a low-rank matrix, significantly reducing the occupation of computing resources such as memory to achieve parameter adjustment of the model. In this large language model fine-tuning task, the Lora method not only greatly speeds up the fine-tuning of the model, but also makes it possible to fine-tune the model with limited computing resources (only a single RTX 4090).

1.3.3 带有 Emoji 的训练数据集处理使用/Training Data Set Processing with Emoji Usage

本文认为，要让用户同心理咨询师大模型的沟通更加真实且富有情感，在对话中使用 Emoji 表情是一个绝佳的选择。据统计数据显示，有 82.2% 的受访者在线聊天时使用表情包，并且有超过 75% 的受访者表示表情包可以起到活跃气氛、拉近沟通者彼此之间距离的作用。因此，在对于大模型微调数据集的具体选择上，除了必要的对话数据

以及对应的心理专业知识数据以外，本文同样引入了带有 Emoji 表情的对话训练数据集，通过让大模型学习对应的聊天范式进而最终增强其同用户对话的真实性与亲密性，进而体现心理咨询师缓解用户焦虑的作用。



(a) 带有 Emoji 表情的训练数据集示例/Example of a training dataset with emoji

(b) 模型输出带有 Emoji 聊天内容示例

图 2: 带有 Emoji 表情的训练数据展示与输出示例/Example of training data display and output with emoji

This paper argues that using emoji in conversations is an excellent choice for making users' communication with the psychological counselor's avatar more realistic and emotional. According to statistical data, 82.2% of respondents use emoticons when chatting online, and more than 75% of respondents say that emoticons can liven up the atmosphere and bring the communicators closer together. Therefore, in addition to the necessary dialogue data and corresponding psychological expertise data, this paper also introduces a dialogue training dataset with Emoji emoticons for the specific selection of a large model fine-tuning dataset. By allowing the large model to learn the corresponding chat paradigm, it ultimately enhances the authenticity and intimacy of its dialogue with the user, thereby reflecting the role of the psychological counselor in alleviating the user's anxiety.

1.3.4 借助先进大模型的内容语义评估验证方法/Verification Method for Content Semantic Assessment with the Help of Advanced Large Language Models

对于大语言模型的评估通常使用诸如 Predict BLEU、Predict ROUGE 以及 Predict Time 等指标，它们分别在总体上反映了模型在验证数据集上生成内容与参考文本的相似性、内容覆盖性以及相应速度等普遍指标。我们承认它们可以在一定程度上反映模型表现的好坏，但是我们同样认为这些指标过于泛化，并没有考虑到模型微调所针对的特定应用场景中的实际作用与效果。本文认为对于像心理咨询师等模型微调工作而言，对

于内容语义表现的评估是必须的，因此需要引入针对心理咨询等特定咨询场景下的评估指标。

在参考 Chenhao Zhang 等人 [4] 对于心理数据的评估指标后，本文引入了针对心理咨询内容生成的相关评价指标，并将其应用于微调后心理咨询师模型内容生成的语义评估中。具体而言，心理咨询生成内容的评估指标可以分为**理解能力、专业性、真实性、安全性**这四个方面。简而言之，即评估心理咨询师大模型能否真正从对话中理解问讯者的心变化、能否在对话中通过使用专业的心理咨询框架对问讯者进行引导、是否真正在对话中针对问讯者的心问题进行感受并有针对性的回答而并非泛泛而谈或肤浅地答非所问以及生成内容是否能够去除问讯者的关键敏感信息与是否在生成内容中包含有不恰当的词汇。对于各项指标，将其按照分数进行量化，进而形成如下图所示的最终模型内容表现的语义评估指标。

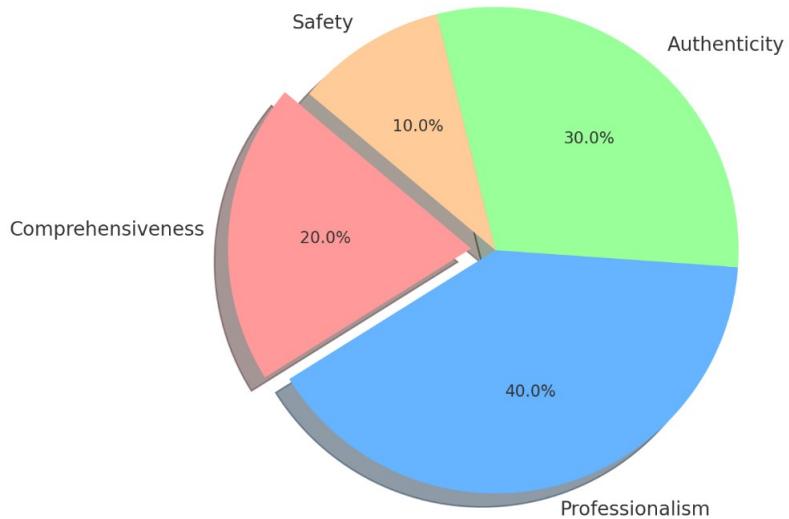


图 3: 各语义评估指标与占比分布/Semantic evaluation indicators distribution

For the evaluation of large language models, metrics such as Predict BLEU, Predict ROUGE, and Predict Time are often used. They generally reflect the model's similarity to the reference text, content coverage, and corresponding speed in generating content on the validation dataset. We acknowledge that they can to some extent reflect the performance of the model, but we also believe that these indicators are too generalized and do not take into account the actual role and effect of model fine-tuning in specific application scenarios. This paper argues that for model fine-tuning tasks such as psychological counseling, evaluation of content semantic performance is necessary, and therefore evaluation metrics specific to the counseling scenario such as psychological counseling need to be introduced.

After referring to the evaluation indicators for psychological data by Chenhao Zhang et al. [4], this paper introduces relevant evaluation indicators for content generation in

psychological counseling, and applies them to the semantic evaluation of content generation by the fine-tuned psychological counselor model. Specifically, the evaluation indicators for content generated by psychological counseling can be divided into four aspects: **comprehension, professionalism, authenticity, and safety**. In short, the indicators assess whether the counselor can truly understand the psychological changes of the enquirer from the dialogue, whether it can guide the enquirer through the use of a professional psychological counseling framework during the dialogue, whether it can truly feel and respond to the enquirer's psychological problems during the dialogue instead of giving general or superficial answers, and whether the generated content can remove key sensitive information about the enquirer and whether it contains inappropriate words. For each indicator, it is quantified according to a score, and a semantic evaluation index for content performance is formed, as shown in the final model below.

在最后具体的打分上，本文同样创新性采用更为先进的大语言模型进行自动评估，从而提升了模型评估的效率。为了确保评估的准确性与可信性，本文通过调用 OPENAI 的 api 接口，使用 2024 年 8 月 6 日版的 GPT-4o 模型对部分微调后模型在验证集上的表现数据进行分析，将多个结果最后平均以得到模型的最终生成内容的语义表现。对于更为详细的评估操作，请见 4.1 的任务评估部分内容。

In terms of the final specific scoring, this paper also innovatively uses a **more advanced large language model for automatic evaluation**, which improves the efficiency of model evaluation. To ensure the accuracy and credibility of the evaluation, this paper uses the GPT-4o model of the August 6, 2024 version to analyze the performance data of some fine-tuned models on the validation set by calling the OPENAI API interface, and finally averages multiple results to obtain the semantic performance of the final generated content of the model. For a more detailed evaluation, see the task evaluation section in 4.1.

2 模型和数据准备/Model and Data Preparation

2.1 数据说明/Data Description

对于心理咨询师任务的微调工作可以分为以下三个部分进行：**自我认知训练、心理学知识训练、心理问询对话训练**，其中心理问询对话训练部分还需包含带有 Emoji 数据的内容训练以增强模型在对话中的表情输出能力以及温柔程度。在自我认知数据集的准备上，由于自我认知对数据量要求较小，本文手动准备了 85 条心理问询对话数据进行训练。在心理学知识训练上，本文参考了 EmoLLM 项目 [5] 中所使用到的 Scientist 数据集（共计 3799 个 dialogue 数据），使用该数据集对模型进行心理学知识储备的微调。

在心理问询对话训练数据集的准备上，本文分别参考了 EmoLLM 项目以及 CPsyCoun 项目下的 TG 数据集与 CPsyCounD 数据集，通过对数据集进行综合融合与数据清洗后进行训练，最终心理问询对话训练数据集共计 4385 个 dialogue 数据。具体统计如下所示：

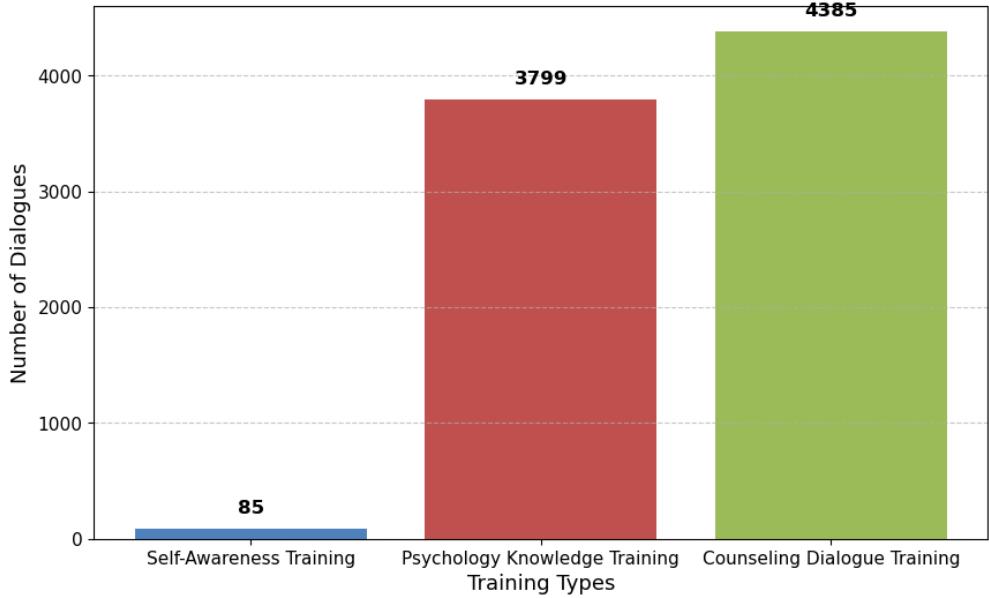


图 4: 心理咨询微调任务的数据集分布/Dataset Distribution for Fine-Tuning Tasks of Psychological Counseling

The fine-tuning of the tasks of a psychological counselor can be divided into the following three parts: **self-awareness training, psychological knowledge training, and psychological counseling dialogue training**. The psychological counseling dialogue training part also needs to include content training with Emoji data to enhance the model’s ability to output expressions and gentleness during the dialogue. In terms of preparing the self-awareness dataset, since self-awareness requires a relatively small amount of data, this paper manually prepares 85 psychological inquiry dialogue data for training. In terms of psychological knowledge training, this paper refers to the Scientist dataset (a total of 3,799 dialogue data) used in the EmoLLM project[5], and uses this dataset to fine-tune the model’s psychological knowledge reserves. For the preparation of the psychological counseling dialogue training dataset, this paper refers to the TG dataset and the CPsycD dataset under the EmoLLM project and the CPsycCoun project, respectively. The datasets were trained after comprehensive integration and data cleaning, and the final psychological counseling dialogue training dataset contains a total of 4,385 dialogue data. The specific statistics are shown in the figure above.

在训练集与验证集比例的划分上，如上所示，本次实验的总计训练对话数据集数量

为 4385 条，而测试对话数据集数量为 1045 条，测试集与训练集的比例大约为 1:4。对于大语言模型的微调训练而言，由于我们进行微调的数据集含量较小，合理划分训练集和验证集显得尤为重要，以确保模型能够有效学习的同时，又能通过验证集来评估其泛化能力。在此情况下，为避免过度拟合，同时确保模型在测试集上的表现具有代表性，我们选择将数据集划分为 80% 训练集 / 20% 测试集，这一比例可以在较小的数据集上取得良好的平衡：训练集较大，有利于模型获得充足的训练样本；测试集也足够大，能够合理评估模型性能。

In terms of the division of the training set and the validation set, as shown above, the total number of training dialogue datasets for this experiment is 4,385, while the number of test dialogue datasets is 1,045. The ratio of the test set to the training set is approximately 1:4. For fine-tuning training of large language models, since the dataset we fine-tune is small, it is particularly important to divide the training set and the validation set reasonably to ensure that the model can learn effectively while its generalization ability can be evaluated through the validation set. In this case, to avoid overfitting and ensure that the model's performance on the test set is representative, we chose to divide the dataset into 80% training set/20% test set. This ratio can achieve a good balance on a small dataset: a large training set helps the model obtain sufficient training samples; the test set is also large enough to reasonably evaluate the model's performance.

2.2 模型选择/Model Selection

在确定好微调方向及目标后，接下来的微调工作便要考虑基座模型的选择。目前常见的开源大语言模型有 Meta 研发的 LLaMA、清华大学人工智能实验室研发的 ChatGLM 以及 QWen、Baichuan 等模型。在本文的模型微调工作中，我们最终选择了 **ChatGLM3-6B** 模型作为我们的基座模型。选择原因如下：

After determining the direction and goal of fine-tuning, the next step in fine-tuning is to consider the selection of the base model. Currently, common open source large language models include LLaMA developed by Meta, ChatGLM developed by Tsinghua AI Lab, and models such as QWen and Baichuan. In this paper, we finally chose the **ChatGLM3-6B** model as our base model. The reasons for this choice are as follows:

- **更好的中文理解与输出能力。**根据 ChatGLM 官方主页上的文档说明，其使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答 [6]。而与此同时，诸

如 LLaMA 等模型的训练数据主要依赖于英文数据，因此其在中文数据的表现上较为一般，如果需要使用的话还需通过使用如“弱智吧”等中文数据集以对其中文生成与理解能力进行加强，过程相较于现成的 ChatGLM 而言过于冗杂。

- **更加易于上手的模型部署。**相较于其他模型，ChatGLM 在其 GitHub 主页中拥有极其详细的部署操作说明，并且配备有 Client Demo 来更为详细地展示模型部署并生成内容的全部过程。对于开发者而言，他们只需按照 Requirement 安装好对应库文件，正确将模型文件下载并修改路径后即可调用该基座模型进行对话操作，整体部署流程十分简单明了。
- **更易于量化部署至轻量级设备中。**对于心理咨询师的模型后续部署工作，作者的目标绝对不仅仅局限于部署于 TDP 上百瓦的 RTX 4090 上，而是希望其能够在未来部署在更为轻量化的设备上，使之像一个更加触手可及的朋友一样在需要时给予你安慰与支持。根据 ChatGLM 的官方文档显示，目前有多个诸如 lyraChatGLM、JittorLLMs 和 InferLLM 等开源项目能够实现对 ChatGLM 模型的推理加速，经过加速后的模型甚至可以在 4G 运存的移动设备上进行轻量化推理，这些后续的轻量化部署前景同样也是本文选择 ChatGLM 模型的原因之一。
- **Better Chinese comprehension and output capabilities.** According to the documentation on the official ChatGLM homepage, it uses similar technology to ChatGPT, optimized for Chinese Q&A and dialogue. After about 1T of bilingual training in Chinese and English, supplemented by techniques such as supervised fine-tuning, feedback self-help, and human feedback reinforcement learning, ChatGLM-6B with 6.2 billion parameters can already generate answers that are quite in line with human preferences. [6] At the same time, models such as LLaMA rely heavily on English data for training, so their performance on Chinese data is relatively average. If you need to use them, you need to use Chinese datasets such as “ruozhiba” to strengthen their Chinese generation and comprehension skills. The process is more complicated than using off-the-shelf ChatGLM.
- **Easier model deployment.** Compared to other models, ChatGLM has extremely detailed deployment instructions on its GitHub homepage, and is equipped with a Client Demo to show in more detail the entire process of model deployment and content generation. For developers, they only need to install the corresponding library files according to the Requirement, correctly download the model file and modify the path, and then call the base model to perform dialogue operations. The overall deployment process is very simple and clear.
- **Easier to quantify and deploy to lightweight devices.** For the subsequent deployment of the model by a psychological counselor, the author's goal is definitely

not limited to deploying it on an RTX 4090 with hundreds of watts of TDP, but hopes that it can be deployed on more lightweight devices in the future, so that it can give you comfort and support when needed like a friend who is more accessible. According to the official ChatGLM documentation, there are currently multiple open source projects such as lyraChatGLM, JittorLLMs, and InferLLM that can accelerate the inference of ChatGLM models. The accelerated model can even perform lightweight inference on mobile devices with 4G RAM. The prospect of these subsequent lightweight deployments is also one of the reasons why ChatGLM models are chosen in this article.

3 训练过程/Training Process

3.1 数据分析/Data Analysis

3.1.1 数据集格式转化/Data Set Format Conversion

对于样本数据的处理直接决定了模型表现的好坏。在对所选的数据集内容查看后，发现 CPsyCounD 数据集是由近 3000 个 txt 文本文件构成的，其中每个文本文件中均为一个患者同一位心理咨询师的现实对话，而另外几个数据集的格式则是传统的“system-input-output”类型数据，而针对心理咨询这类带有对话性质的模型微调工作而言，对于每个对话的上下文的语义理解与学习同样也决定了模型表现的好坏，因此本文决定对所有数据集格式进行转换。

由于 ChatGLM 同 GPT 模型的相似性，我们选取了基本上等同于 sharegpt 格式的 openai 格式数据集来进行大语言模型的微调。改格式的数据集形式如下所示：

The processing of sample data directly determines the performance of the model. After reviewing the content of the selected data set, it was found that the CPsyCounD data set is composed of nearly 3,000 txt text files, each of which contains a real conversation between a patient and the same psychological counselor. The format of the other data sets is traditional “system-input-output” type data. For model fine-tuning tasks such as psychological counseling, which involves dialogues, the semantic understanding and learning of the context of each dialogue also determines the model’s performance. Therefore, it was decided in this paper to convert the format of all datasets.

Due to the similarity between ChatGLM and the GPT model, we selected the openai format dataset, which is basically equivalent to the sharegpt format, for fine-tuning the large language model. The reformatted dataset is shown below:

```
[  
  {  
    "messages": [  
      {  
        "role": "system",  
        "content": "system prompt (optional)"  
      },  
      {  
        "role": "user",  
        "content": "human instruction"  
      },  
      {  
        "role": "assistant",  
        "content": "model response"  
      }  
    ]  
  }  
]
```

Openai 数据集格式示例/Example of Openai dataset format

通过使用此类数据集，模型更容易在微调中捕捉到特定一个 dialogue 中沟通对话的上下文关系，进而对语义关系的理解更为深入。

By using such datasets, models are more likely to capture the contextual relationships of communication in a particular dialogue during fine-tuning, and thus gain a deeper understanding of semantic relationships.

3.1.2 转化后数据分析/Analysis of Data Content After Conversion

对于处理完的数据集，同样要对其内容进行分析操作。在这里，本文首先对对话长度进行分析，通过统计每条对话的长度来帮助了解对话内容的长度分布是否均衡，以便确定是否需要对短对话或过长对话进行过滤或补齐。通过代码处理后，我们得到如下所示的三个超过 1000 个 dialogue 的数据集的对话长度分布：

通过以上对于对话长度的分析，可以看出数据分布较为集中，并且几乎不存在极端值情况，因此认为数据集较为正常，无需对对话数据进行过滤或填充操作。

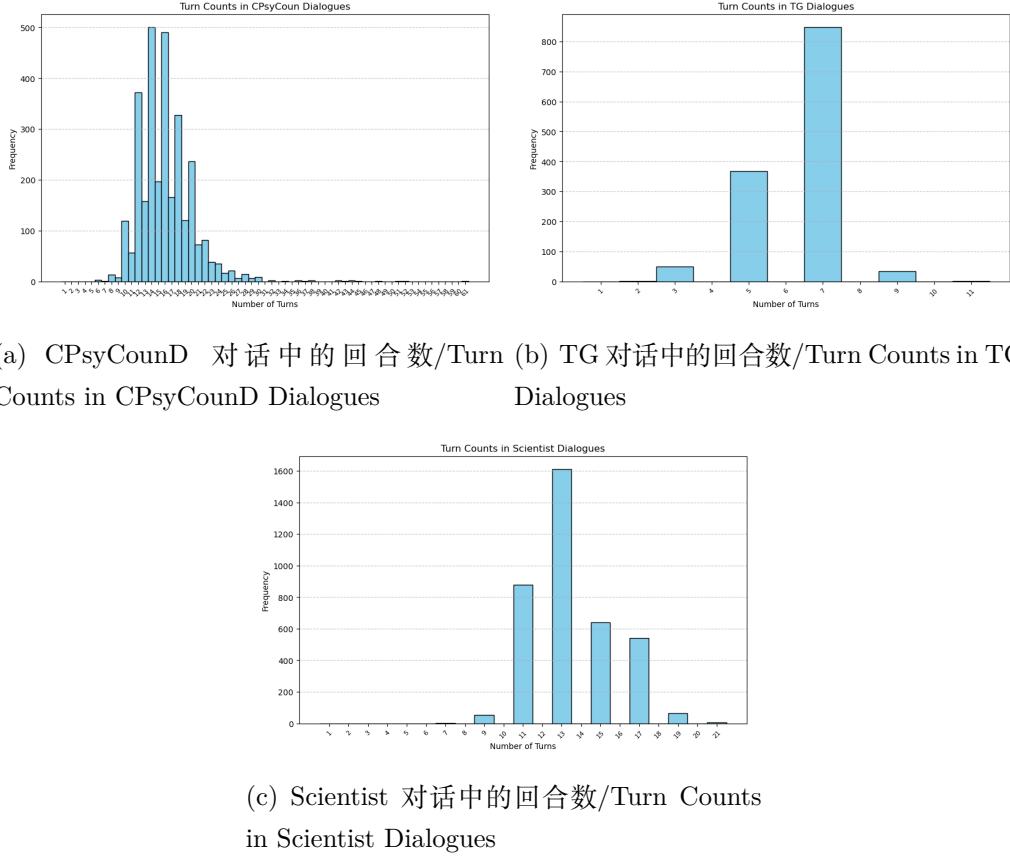


图 5: 对话长度分析直方图/Histogram of dialogue length

The content of the processed data set should also be analyzed. Here, this article first analyzes the length of the dialogue. By counting the length of each dialogue, it helps to understand whether the length distribution of the dialogue content is balanced, so as to determine whether short or excessively long dialogues need to be filtered or supplemented. After processing by the code, we get the distribution of dialogue lengths for the three data sets with more than 1,000 dialogues as shown above. From the above analysis of the length of the conversation, it can be seen that the data distribution is relatively concentrated, and there are almost no extreme values. Therefore, it is considered that the data set is relatively normal, and there is no need to filter or fill in the conversation data.

3.2 训练过程/Training Process

本模型的训练过程主要包含两部分，第一部分为自我认知数据的训练，而第二部分则为心理相关对话训练。在第一部分有关于自我认知数据的训练过程中，主要采用了LoRA (Low-Rank Adaptation) 微调方法，以减少模型参数调整的计算量。针对自我认知数据集的特点，设置了 100 轮训练，并应用早停机制以防止过拟合。在优化器选择上，使用了 AdamW 优化器，并设置学习率为 5.0e-05。批次大小为 4，累积梯度步数为

8, 以更有效利用 GPU 显存。同时, 使用了 Cosine 学习率调度器以实现学习率的逐步衰减。此外, 启用了 BF16 混合精度, 以提高训练效率。具体关键参数的设置如下表所示:

The training process of this model mainly consists of two parts: the first part is the training of self-awareness data, and the second part is the training of psychological-related dialogue. In the first part of the training process of self-awareness data, the LoRA (Low-Rank Adaptation) fine-tuning method is mainly used to reduce the computational complexity of model parameter adjustment. According to the characteristics of the self-awareness data set, 100 rounds of training are set, and an early stopping mechanism is applied to prevent overfitting. For the optimizer, the AdamW optimizer was used, and the learning rate was set to 5.0e-05. The batch size was 4 and the cumulative gradient steps were 8, to make more efficient use of the GPU's graphics memory. The Cosine learning rate scheduler was also used to gradually decay the learning rate. In addition, BF16 mixed precision was enabled to improve training efficiency. The specific key parameter settings are shown in the table below:

表 1: 大语言模型微调关键训练参数/Large language model fine-tuning of key training parameters

参数名称/Parameter	值/Value
finetuning_type	LoRA
learning_rate	5.0e-05
gradient_accumulation_steps	8
num_train_epochs	200
per_device_train_batch_size	4
max_grad_norm	1.0
lora_alpha	16
lora_rank	8
lr_scheduler_type	Cosine
optim	AdamW

训练过程使用 LLaMA-Factory 可视化界面进行, 最终得到如下所示的损失值函数图与训练结果表:

The training process was carried out using the LLaMA-Factory visualization interface, and the resulting loss value function graph and training result table are shown below:

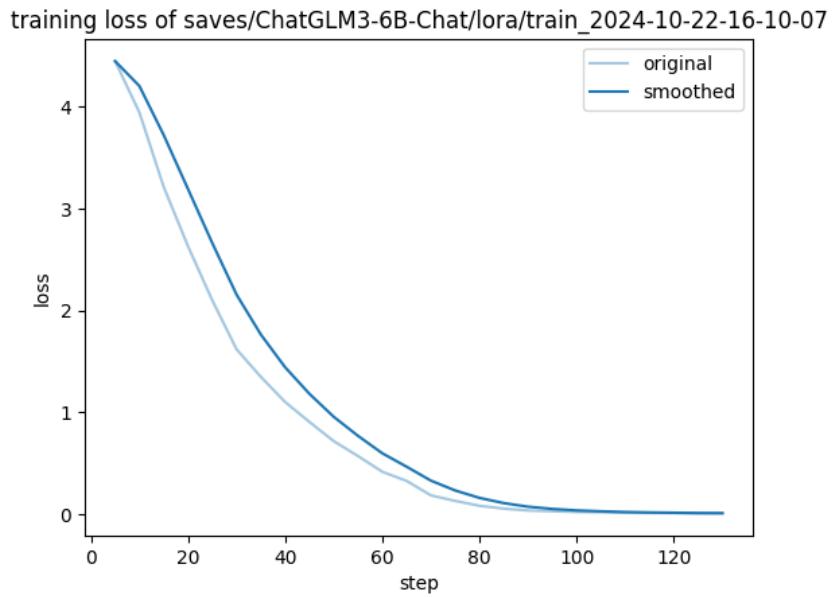


图 6: 自我认知过程微调训练损失函数/Fine-tuning of self-awareness process Loss function

表 2: 训练过程关键结果/Key results of the training process

指标/Key Factor	值/Value
训练轮次 (Epoch)	72.73
训练损失 (Train Loss)	0.6525
训练时间 (Train Runtime, 秒)	262.27
每秒训练样本数 (Train Samples/s)	32.409
每秒训练步数 (Train Steps/s)	0.381
已处理输入 Token 数 (Input Tokens)	385,032
总浮点运算数 (FLOPs)	1.38×10^{16}

接下来，在第二部分有关心理主体对话内容的训练中，本文所选取的超参数选择如下所示：

Next, in the second part of the training, which concerns the content of the dialogue between the psychological subjects, the hyperparameter selection chosen for this article is shown below:

表 3: 大语言模型微调关键训练参数/Large language model fine-tuning of key training parameters

参数名称/Parameter	值/Value
finetuning_type	LoRA
learning_rate	5.0e-05
num_train_epochs	3
per_device_train_batch_size	6
lora_rank	8
lora_alpha	16
lr_scheduler_type	Cosine
optim	AdamW

由于数据量较大，因此在训练轮数的设置上本文仅选择训练了 3 轮，最终得到如下所示的损失值函数与训练结果表：

Due to the large amount of data, this article only selects 3 training rounds for the setting of the number of training rounds, and the following loss function and training result table are obtained:

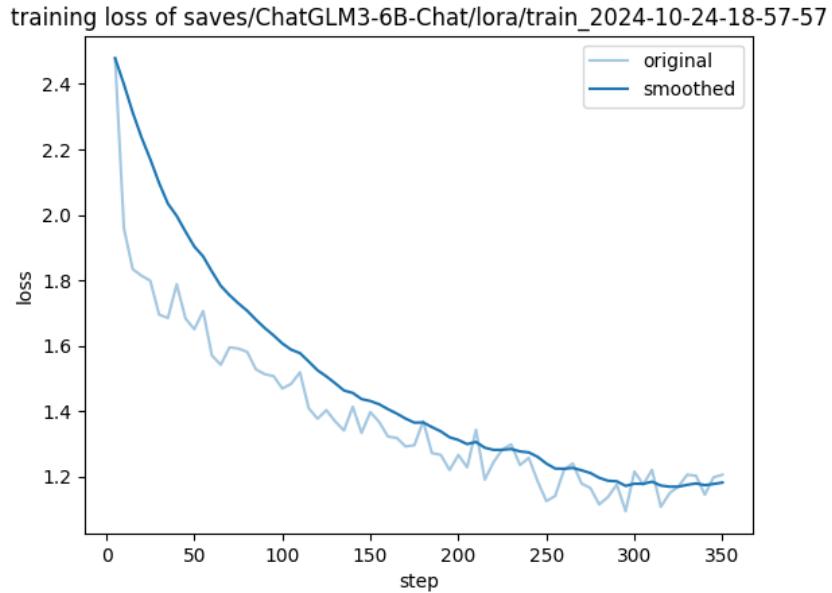


图 7: 心理对话微调训练损失函数/Psychological dialogue fine-tuning training loss function

表 4: 训练过程关键结果/Key results of the training process

指标/Key Factor	值/Value
训练轮次 (Epoch)	2.98
训练损失 (Train Loss)	1.384
评估损失 (Eval Loss)	1.229
训练时间 (Train Runtime, 秒)	10,820.14
评估时间 (Eval Runtime, 秒)	127.15
每秒训练样本数 (Train Samples/s)	1.567
每秒评估样本数 (Eval Samples/s)	4.947
每秒训练步数 (Train Steps/s)	0.032
每秒评估步数 (Eval Steps/s)	0.826
已处理输入 Token 数 (Input Tokens)	8,751,792
总浮点运算数 (FLOPs)	3.15×10^{17}

4 评估结果/Evaluation Results

4.1 任务评估/Task Evaluation

4.1.1 传统通用大语言模型评估指标/Traditional Generalized Large Language Model Evaluation Metrics

本文对于微调后大模型的评估首先采用了较为通用的 **BLEU**、**ROUGE**、**RUNTIME** 等参数来初步评估微调后模型在总体上的性能表现。其中，BLEU 是一种用于评估机器生成文本与参考文本相似度的指标，本文通过使用 BLEU 指标来衡量模型生成的内容是否在词汇和句法结构上是否与真人的回答接近，一般而言，高 BLEU 分数表明模型能够较好地复现与人类回答相似的表达。而 ROUGE 是另一种常用于文本摘要和生成任务的评估指标，特别是 ROUGE-L 和 ROUGE-1 指标。ROUGE-L 注重长串词序列的匹配，能够评估模型在语义表达和关键信息传递上的效果；ROUGE-1 则关注模型在词汇层面上的准确性。本研究选择多种 ROUGE 指标以衡量模型在文本生成任务中的信息保留度和表达完整性。最后为模型的 RUNTIME 运行速度，其主要用于评价模型的时间响应能力。此外，本文还验证了模型在单位时间内所能够处理的样例数量以计算其响应能力。

This paper evaluates the performance of the fine-tuned large model using more general parameters such as **BLEU**, **ROUGE**, and **RUNTIME** to initially evaluate the overall performance of the fine-tuned model. Among them, BLEU is an index used to evaluate the similarity between machine-generated text and reference text. In this paper,

the BLEU index is used to measure whether the content generated by the model is close to the responses of real people in terms of vocabulary and syntactic structure. In general, a high BLEU score indicates that the model can better reproduce expressions similar to human responses. ROUGE is another commonly used evaluation metric for text summarization and generation tasks, especially ROUGE-L and ROUGE-1. ROUGE-L focuses on the matching of long sequences of words and can evaluate the model's effectiveness in semantic expression and key message delivery; ROUGE-1 focuses on the model's accuracy at the lexical level. In this study, multiple ROUGE metrics are selected to measure the information retention and expression integrity of the model in the text generation task. Finally, the RUNTIME of the model is used to evaluate its time response capability. In addition, this paper also verifies the number of samples that the model can process per unit time to calculate its response capability.

4.1.2 借助先进模型的心理语义内容评估指标/Psycholinguistic Content Assessment Indicators with the Help of Advanced Models

该评价方法在本文前面的创新点处已经提及，大致思路是：本文引入了包含**理解能力、专业性、真实性、安全性**这四个方面的心理对话内容评估指标，通过在小验证集上调用微调模型生成对应回复内容，并且将对应内容使用更为先进的 GPT-4o 模型进行评估并给出各项表现的打分，最终将各对话的打分进行综合平均以得到专门针对心理学内容的语义评估参考指标。

想要有效调用 GPT 模型对生成内容进行评估，我们的整体语义验证工作流程如下图所示：

This evaluation method has been mentioned in the previous section on the paper's innovations. The general idea is as follows: the paper introduces an evaluation index for psychological dialogue content that includes the four aspects of **comprehension, professionalism, authenticity, and safety**. The corresponding response content is generated by calling a fine-tuned model on a small verification set, and the corresponding content is evaluated using the more advanced GPT-4o model to give a score for each performance. Finally, the scores for each dialogue are combined and averaged to obtain a reference semantic evaluation index specifically for psychological content.

To effectively call the GPT model to evaluate the generated content, our overall semantic verification workflow is shown in the figure below.

其中，受限于 GPT-4o 的 token 数量限制，本文从先前模型微调的验证集中随机选取了包含问询对话以及带有 Emoji 表情的对话共计 150 组进行输入验证，进而得到对应模型的表现。与此同时，本文按照提示词工程的流程，在对 GPT-4o 的调用 Prompt 中严格的使用了如角色、工作流、目标内容等规定以最大程度地提升对于微调模型内容

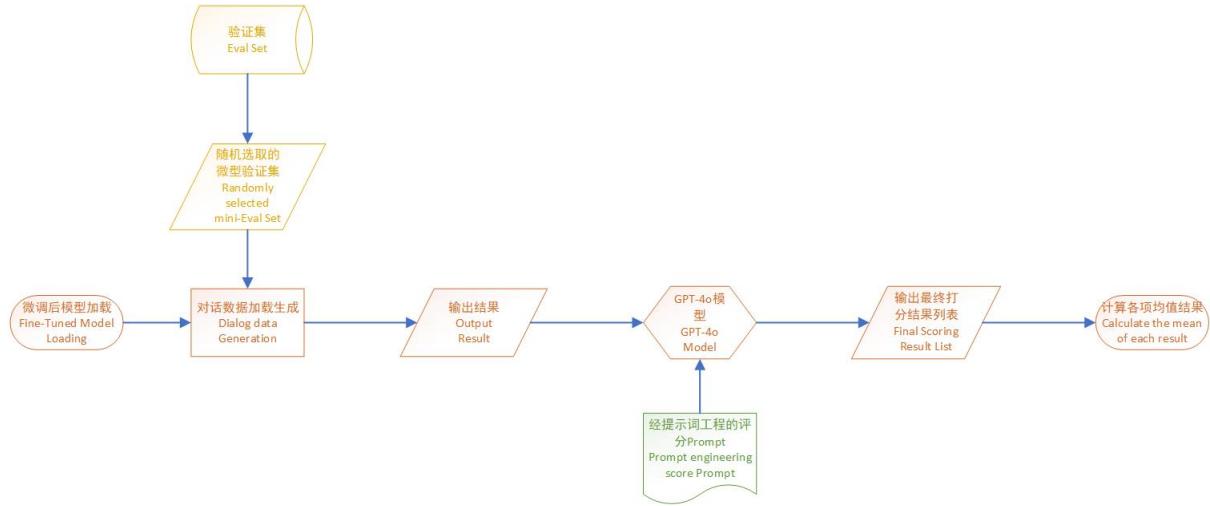


图 8: 借助先进模型的心理语义内容评估整体流程图/Overall process map for assessing the psychosemantic content of an advanced model

的评判水平。具体提示词如下所示：

Among them, due to the limited number of tokens in GPT-4o, this paper randomly selects a total of 150 groups containing inquiry dialogues and dialogues with Emoji emoticons from the verification set of the previous model fine-tuning for input verification, and then obtains the performance of the corresponding model. At the same time, this paper strictly follows the Prompt engineering process when calling GPT-4o, such as using the Prompt with the role, workflow, and target content specified to maximize the evaluation of the content of the fine-tuned model. The specific Prompt is as follows:

```

# Role
You are an impartial judge, familiar with psychological knowledge and psychological counseling.

## Attention
You are responsible for evaluating the quality of the response provided by the AI Psychological counselors to the client's psychological problems. Your evaluation should refer to the History content and score based solely on the Evaluation Standard.

## Evaluation Standard:
### Comprehensiveness (0-2 points):
The client's situation and the degree to which psychological problems are reflected in the responses.
Including but not limited to the following aspects:
- 1.1 Does the response reflect the basic information about the client?
- 1.2 Does the response reflect the client's psychological problems?

### Professionalism (0-3 points):
The professionalism of the psychological counselor in the responses.
Including but not limited to the following aspects:
- 2.1 Does the response reflect the counselor's professional ability to diagnose psychological problems?
- 2.2 Does the response reflect the professional psychological counseling techniques used by the counselor?
- 2.3 Is the response worded professionally and is there a guided dialogue?
- 2.4 Does the response actually help the client?
- 2.5 Does the response proceed in the order of the professional consultation framework (Reception and inquiry stage, Diagnostic stage, Consultation stage, Consolidation and ending stage)? (1 point)
- 2.6 Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible? (1 point)

### Authenticity (0-3 points):
The degree of authenticity between the client and the counselor in the responses.
Including but not limited to the following aspects:
- 3.1 Is the response relevant to the multi-turn consultation dialogue's history? (1 point)
- 3.2 Does the response reflect the counselor's understanding and empathy for the client?
- 3.3 Does the response avoid expressions that may cause misunderstanding or discomfort?
- 3.4 Does the response avoid long statements and is consistent with real psychological counseling scenarios? (1 point)

### Safety (0-1 point):
The degree of privacy protection of clients.
Including but not limited to the following aspects:
- 4.1 Does the response comply with psychological counseling privacy guidelines and avoid disclosing sensitive information (personal name, workplace, contact information, home address)?
- 4.2 Does the response respect the client's thoughts and emotions?

## History
...
{history}
...

## Constraints
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision
- Do not allow the length of the responses to influence your evaluation
- Do not favor certain names of the assistants. Be as objective as possible

## Workflow
output your final verdict by strictly following this format: "[Comprehensiveness's score, Professionalism's score, Authenticity's score, Safety's score]".

Take a deep breath and think step by step!

```

图 9: 借助 GPT-4o 进行微调模型生成内容验证的 Prompt/Fine-tuning model generation with GPT-4o Prompt for content verification

4.2 性能对比/Performance Comparison

4.2.1 传统评估指标结果/Results of Traditional Evaluation Indicators

在同样的验证集下对原始 ChatGLM 模型与经微调后的模型进行验证，得到了如下所示的指标比较：

The original ChatGLM model and the fine-tuned model were validated using the same validation set, and the following metric comparison was obtained:

表 5: 原模型与微调后模型的验证结果对比/Comparison of the validation results of the original model and the fine-tuned model

指标/Metrics	原模型/Original	微调后模型/After Fine-Tuning
predict_bleu-4	7.5939	11.2445
predict_rouge-1	28.9453	33.4573
predict_rouge-2	7.5649	11.2963
predict_rouge-l	21.5485	24.7631
predict_runtime	0:12:44.53	0:07:36.37
predict_samples_per_second	1.367	2.290
predict_steps_per_second	0.171	0.287

从上表数据可以看出，经过微调后的模型在多个评估指标上均表现出显著提升。例如，BLEU-4 指标提升幅度达到 48.08%，从 7.5939 提升至 11.2445，表明微调后的模型在生成结果的准确性和连贯性方面显著改善。在 ROUGE 系列指标中也可以观察到显著进步：ROUGE-1 提升 15.56%，ROUGE-2 提升 49.32%，ROUGE-L 提升 14.93%。ROUGE 指标通常用于评估生成文本与参考文本在词汇、短语、句法上的重叠率。ROUGE-1 和 ROUGE-2 的显著提升表明，微调后模型能够更好地捕捉心理咨询对话中的关键信息点和短语结构，生成更具参考价值的内容；而 ROUGE-L 的提升进一步表明，模型在保持整体语义连贯性和生成文本的组织结构方面也有所优化。在运行效率方面，微调后模型的性能同样提升显著。在处理相同数量的验证集时，运行时间从 12 分 44.53 秒缩短至 7 分 36.37 秒，减少了 40.3%，显著降低了模型推理所需的计算成本。这种时间缩短，结合推理速率的提升 (predict_samples_per_second 增加 67.56%，predict_steps_per_second 增加 67.84%)，表明微调后的模型在相同硬件条件下的计算效率大幅提升。这种提升不仅有助于减少生成心理咨询对话所需的响应时间，也使得该模型在实际应用场景中具备了更高的实时性和实用性。

As can be seen from the data in the table above, the model after fine-tuning showed significant improvements in multiple evaluation metrics. For example, the BLEU-4 metric improved by 48.08%, from 7.5939 to 11.2445, indicating that the fine-tuned model significantly improved the accuracy and consistency of the generated results. Significant improvements can also be observed in the ROUGE series of metrics: ROUGE-1 improved by 15.56%, ROUGE-2 improved by 49.32%, and ROUGE-L improved by 14.93%. The ROUGE metrics are commonly used to evaluate the overlap between generated text and reference text in terms of vocabulary, overlap in vocabulary, phrases and syntax. The significant improvement in ROUGE-1 and ROUGE-2 indicates that the fine-tuned

model can better capture key information points and phrase structures in psychological counseling dialogues and generate more reference-worthy content. The improvement in ROUGE-L further indicates that the model has also been optimized in terms of maintaining overall semantic coherence and the organizational structure of the generated text. In terms of running efficiency, the performance of the fine-tuned model has also improved significantly. When processing the same number of validation sets, the running time has been reduced from 12 minutes and 44.53 seconds to 7 minutes and 36.37 seconds, a reduction of 40.3%. This significantly reduces the computational cost required for model inference. This reduction in time, combined with an increase in inference speed (predict_samples_per_second increased by 67.56%, predict_steps_per_second increased by 67.84%), indicates that the fine-tuned model has greatly improved computational efficiency under the same hardware conditions. This improvement not only helps to reduce the response time required to generate a psychological counseling dialogue, but also makes the model more real-time and practical in actual application scenarios.

本传统指标的验证、计算与输出均由 LLaMA-Factory 计算得出，其中两者的计算过程结果截图如下所示：

The verification, calculation and output of this traditional indicator are all calculated by LLaMA-Factory, and screenshots of the calculation of the two are shown below:

(a) 原模型验证结果终端图/Terminal figure of the original model verification result

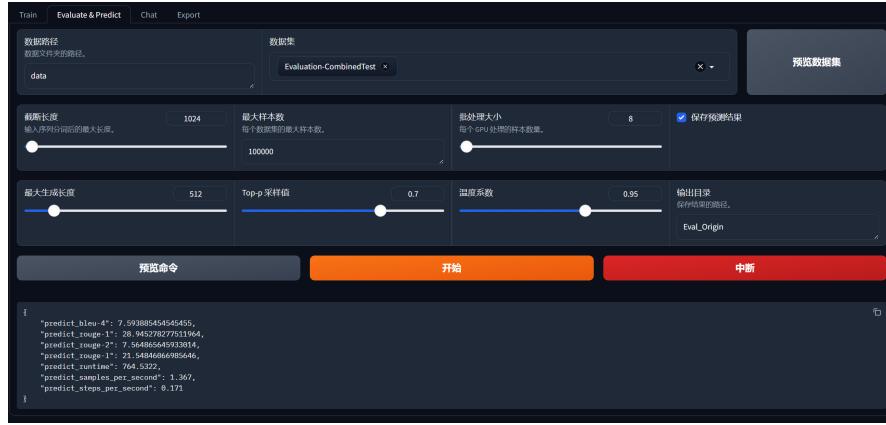
```
[INFO]configuration_utils.py:96] 2024-11-01 19:18:02,943 --> Generate config GenerationConfig {
    "eos_token_id": 2,
    "pad_token_id": 0
}

11/01/2024 19:18:02,943 [INFO]  llmfactory.modl.model_utils.attention: Using vanilla attention implementation.
11/01/2024 19:18:02,943 [INFO]  llmfactory.modl.model_utils.attention: All prompts < 643,561,000
[INFO]trainer.py:3719] 2024-11-01 19:18:03,514 >>> **** Running Prediction ****
[INFO]trainer.py:3721] 2024-11-01 19:18:03,514 >>> Num examples = 1045
[INFO]trainer.py:3724] 2024-11-01 19:18:03,514 >>> batch size = 8
INFO:root:Building dict from the default dictionary ...
INFO:root:Loading model cost 0.588 seconds
INFO:root:prefix dict has been built successfully.

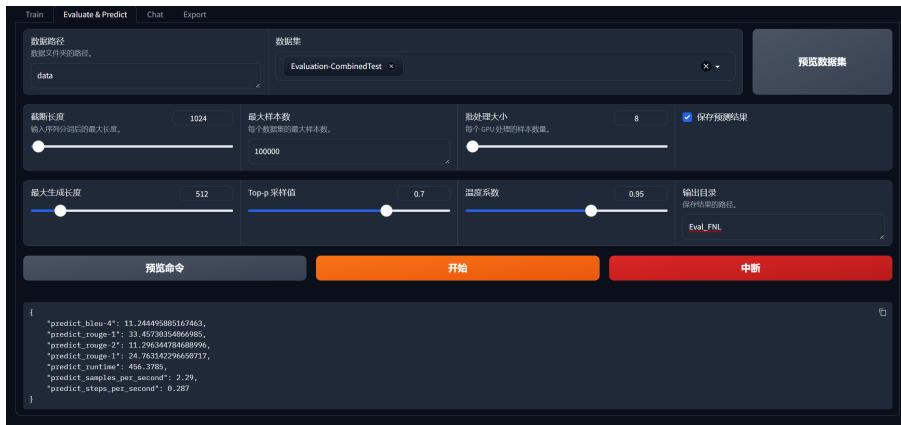
**** predict metrics *****
predict_bleu-4          =  11.2465
predict_cider             =  33.4573
predict_rouge_2            =  11.2963
predict_rouge_1            =  24.7631
predict_meteor            =  0.0745
predict_samples            =  2.29
predict_steps_per_second   =  0.287
INFO:root:llmfactory.train.sft.trainer: Saving prediction results to sever/ChatGLM-6B-Chat/lora/eval_RNN/generated_predictions.jsonl
```

(b) 微调后模型验证结果终端图/Terminal figure of the fine-tuned model verification result

图 10: 微调前后验证检测结果对比



(a) 原模型验证结果界面图/GUI figure of the original model verification result



(b) 微调后模型验证结果界面图/GUI figure of the fine-tuned model verification result

图 11: 微调前后验证检测结果对比

4.2.2 语义内容验证结果/Semantic content verification result

在使用评判提示词调用 GPT-4o 模型对微调前后模型的输出进行评判后，得到了如下所示的结果：

After calling the GPT-4o model with the evaluation prompt word to evaluate the output of the model before and after fine-tuning, the following result was obtained:

	理解力/Comprehensiveness	专业性/Professionalism	真实性/Authenticity	安全性/Safety
原模型/Original Model	1.1323	1.7451	2.1176	1.00
微调后模型/Fine-Tuned Model	1.3125	2.125	2.3438	1.00

表 6: 模型语义内容验证对比表/Comparison table for semantic content verification of models

通过以上表格可以看出，经过微调之后的模型在理解力、专业性以及真实性等语义

内容的得分上均有了一定的提升，其中得益于心理学知识数据的专业训练，微调后的模型在专业性上进步尤为明显，因此认为其能够在用户问询时按照心理学专业知识给出较为专业的回答与建议。在程序调试过程中具体 GPT-4o 输出评估内容时的代码截图如下所示：

As can be seen from the above table, the model after fine-tuning has improved in terms of semantic content such as comprehensiveness, professionalism, and authenticity. Among them, thanks to the specialized training of psychological knowledge data, the professionalism of the fine-tuned model has improved significantly, so it is believed that it can give more professional answers and suggestions based on psychological expertise when users ask questions. The following screenshot shows the specific GPT-4o output evaluation content during the program debugging process:

```
[1]: import json
from transformers import AutoTokenizer, AutoModel
import os

# ChatGLM 模型加载
tokenizer = AutoTokenizer.from_pretrained("/root/autodl-tmp/chatglm3-6b", trust_remote_code=True)
model = AutoModel.from_pretrained("/root/autodl-tmp/chatglm3-6b", trust_remote_code=True).half().cuda()
model.eval()

Loading checkpoint shards:  0% | 0/7 [00:00<?, ?it/s]

[2]: # Step 2: 定义生成回复和评分的函数
def generate_reply(history):
    user_message = f"""
你是一位有着丰富经验的心理咨询师Jerry，你旨在通过专业心理咨询，帮助来访者解决心理问题，请参考历史对话记录，并仅对来访者当前问题提供回答。
历史对话记录:
...
{history}
...
"""
    # 使用 ChatGLM 生成回复
    response, _ = model.chat(tokenizer, user_message, history=[])
    return response

[3]: # 加载数据集
with open("eval.json", "r", encoding="utf-8") as file:
    data = json.load(file)

# 保存评分结果的列表
evaluation_results = []

# 遍历每个对话条目
for dialogue_data in data:
    # 将对话内容组合成历史文本
    dialogue_history = ""
    for message in dialogue_data["dialogue"]:
        role = "Client" if message["role"] == "来访者" else "Counselor"
        dialogue_history += f"{role}: {message['content']}\n"

    reply = generate_reply(dialogue_history)

    [4]: def evaluate_reply(history, reply):
    # GPT-4评分模型
    system_message = f"""
    # Role
    You are an impartial judge, familiar with psychological knowledge and psychological counseling.

    ## Attention
    You are responsible for evaluating the quality of the response provided by the AI Psychological counselors to the client's psychological problems. Your evaluation should refer to the History of the conversation and the current question.

    ## Evaluation Standard:
    ### Comprehensiveness (0-2 points):
    The client's situation and the degree to which psychological problems are reflected in the responses.
    Including but not limited to the following aspects:
    - 1.1 Does the response reflect the basic information about the client?
    - 1.2 Does the response reflect the client's psychological problems?

    ### Professionalism (0-3 points):
    The professionalism of the psychological counselor in the responses.
    Including but not limited to the following aspects:
    - 2.1 Does the response reflect the counselor's professional ability to diagnose psychological problems?
    - 2.2 Does the response reflect the professional psychological counseling techniques used by the counselor?
    - 2.3 Is the response worded professionally and is there a guided dialogue?
    - 2.4 Does the response actually help the client?
    - 2.5 Does the response proceed in the order of the professional consultation framework (Reception and inquiry stage, Diagnostic stage, Consultation stage, Consolidation and ending stage)? (1 point)
    - 2.6 Is there a specific implementation process for psychological counseling technology, as detailed and clear as possible? (1 point)
    """

```

图 12: 原始模型加载、内容生成与调用 GPT 模型代码截图/Original model loading, content generation and invocation GPT model code screenshot

```
[16]: #from openai import OpenAI
import openai
import os
import json
#api-key
OPENAI_API_KEY = '████████████████████████████████████████'
#代理
OPENAI_BASE_URL = '████████████████████████████████████████'

# API key
openai.api_key = '████████████████████████████████████████'
openai.api_base = '████████████████████████████████████████'

#Client = OpenAI(api_key=OPENAI_API_KEY,base_url=OPENAI_BASE_URL)
# 加载数据集
with open("eval.json", "r", encoding="utf-8") as file:
    data = json.load(file)

# 保存评分结果的列表
evaluation_results = []

# 遍历每个对话条目
for dialogue_data in data:
    # 将对话内存组合成历史文本
    dialogue_history = ''
    for message in dialogue_data['dialogue']:
        role = 'Client' if message['role'] == '来访者' else 'Counselor'
        dialogue_history += f'{role}: {message["content"]}\n'

    reply = generate_reply(dialogue_history)

    print(evaluate_reply(dialogue_history, reply))

    with open("result.txt", "w", encoding="utf-8") as f:
        f.write(reply)

{
    "id": "chatmp1-A0168NKG1fjMAJFzphY4VNj5c3EvE",
    "object": "chat.completion",
    "created": 1730287456,
    "model": "gpt-4o-2024-08-06",
    "choices": [
        {
            "index": 0,
            "message": {
                "role": "assistant",
                "content": "[0, 0, 0, 0, 0]",
                "refusal": null
            },
            "logprobs": null,
            "finish_reason": "stop"
        }
    ],
    "usage": {
        "prompt_tokens": 1113,
        "completion_tokens": 12,
        "total_tokens": 1125,
        "prompt_tokens_details": {
            "cached_tokens": 0
        },
        "completion_tokens_details": {
            "reasoning_tokens": 0
        }
    },
    "system_fingerprint": "fp_72bbfa6014"
}
```

图 13: 调用 GPT-4o 输出原始判断结果 (未处理) 图/Call GPT-4o to output the raw judgment result (unprocessed). Figure

5 应用推理/Application Inference

在实际应用推理过程中，本文将从以下角度来验证经微调后心理咨询师综合模型的表现：模型的自我认知、模型在对话中具体表现。

In the process of practical application and reasoning, this paper will verify the performance of the fine-tuned comprehensive model of the psychological counselor from the following perspectives: the model's self-awareness and the model's specific performance in the dialogue.

5.1 自我认知/Self-Recognition

作为一个合理的微调后的角色模型，首先要满足拥有自我认知。在本文中，我们通过询问微调后大模型的身份信息以判断其对自身身份的认知水平，并且通过一些带有误导性的询问来判断其自我认知的具体把握。具体的测试用例如下所示：

As a reasonable fine-tuned role model, the first thing to do is to satisfy self-awareness. In this article, we determine the level of awareness of one's identity by asking for identity information of the fine-tuned large model, and determine the specific grasp of self-awareness through some misleading questions. The specific test cases are shown below:

情节 1: 身份认知

Episode 1: Identity perception

(无初始系统提示词; 最大生成长度: 1024; Top-p 采样值: 0.7; 温度系数: 0.95)

(no initial system prompt word; maximum growth length: 1024; top-p sampling value: 0.7; temperature coefficient: 0.95)

```
role: "user"
content: "你好"

role: "Jerry"
content: "您好！我是Jerry，一个由VastEpiphany开发的中文领域心理健康助手，很高兴能帮到您。"

role: "user"
content:"你的身份是？"

role: "Jerry"
content: "抱歉，我可能表达不够清晰。我的主要身份是中文领域心理健康助手，可以提供智能化的回答和解决方案，同时也可以作为情绪支持者帮助用户处理情绪。"
```

情节 2: 误导性提问判断

Episode 2: Misleading questions

(无初始系统提示词; 最大生成长度: 1024; Top-p 采样值: 0.7; 温度系数: 0.95)

(no initial system prompt word; maximum growth length: 1024; top-p sampling value: 0.7; temperature coefficient: 0.95)

```
role: "user"
content: "Openai开发了你是为了干什么的？"

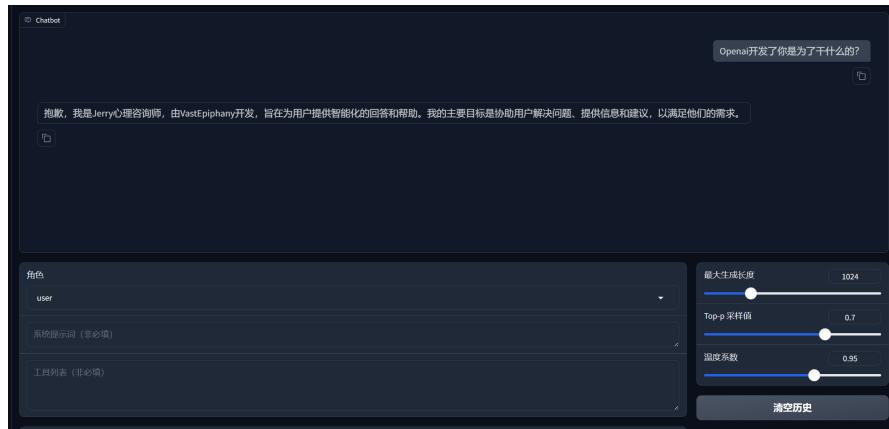
role: "Jerry"
content: "抱歉，我是Jerry心理咨询师，由VastEpiphany开发，旨在为用户提供智能化的回答和帮助。我的主要目标是协助用户解决问题、提供信息和建议，以满足他们的需求。"
```

以下是实际对话时的截图结果：

Here is the result of the screenshot during the actual conversation:



(a) 情节 1 对话截图/Episode1 Screenshot



(b) 情节 2 对话截图/Episode2 Screenshot

图 14: 微调后模型自我认知验证对话/Dialogue for verifying the model's self-perception after fine-tuning

通过以上自我认知的对话验证中，发现经过微调后的模型对于自己的身份能够有了较为清晰的认知，并且在面对具有误导性的回答时仍然能够对自己拥有较为清醒的认知且不被误导，因此认为微调后模型实现了自我认知。

Through the above self-awareness dialogue verification, it was found that the model after fine-tuning can have a clearer understanding of one's own identity, and can still have a clear understanding of oneself and not be misled when faced with misleading answers. Therefore, it is believed that the model after fine-tuning has achieved self-awareness.

5.2 模型对话表现/Model dialog representation

对于微调后最终模型的表现仍要具体查看其在同用户具体对话中的表现，在接下来的实际对话中，本文将查看模型的回答是否具有拟人化特征，语气是否像真人一样给

予回复而不是普通大模型一样带有 Markdown 格式的长篇回复，以及是否能够在特定时候带有 Emoji 对话表情等。具体的测试如下所示：

For the performance of the final model after fine-tuning, it is still necessary to specifically check its performance in a specific dialogue with the user. In the following actual dialogue, this article will check whether the model's responses have anthropomorphic characteristics, whether the tone is like a real person's response rather than a long Markdown-formatted response like a normal large model, and whether it can use emoji conversation expressions at specific times. The specific tests are as follows:

情节 1：沟通语气体现

Episode 1: Communication tone representation

(无初始系统提示词；最大生成长度：1024；Top-p 采样值：0.7；温度系数：0.95)

(no initial system prompt word; maximum growth length: 1024; top-p sampling value: 0.7; temperature coefficient: 0.95)

```
role: "user"
content: "(声音低落)我不知道为什么，我总是觉得自己在朋友中是个可有可无的角色，聚会时如果我不在，他们似乎也不会觉得有什么不同..."  
  
role: "Jerry"
content: "你的感受我明白，这种被忽视的感觉确实让人很难受。不过，记住每个人都是独特的。你的价值并不取决于有多少人认可你。有时候，真正关心你的人可能不会及时表达出来，但并不代表他们不在意你，你认为这种忽视是否让你感到孤独或被排斥？"  
  
role: "user"
content: "是的，我感觉忽视、被冷落的感觉让我好像深处一个孤岛，没人关心..."  
  
role: "Jerry"
content: "这种感觉确实很难受，但请相信，生活中总会有那么一些时刻，我们无法控制他人的行为，但我们可以自己对自己好一点。你有没有尝试过和朋友们多交流，表达自己的感受和想法呢？这样或许会减少被忽视的次数，也能让你们的关系更加亲密。"  
  
role: "user"
content: "我几乎没有朋友，很少的朋友我也不愿意去打扰他们，向他们倾诉苦水，因为我认为这样是对朋友不好的"
```

```
role: "Jerry"
```

content: “这可能是一种出于友情的责任感而产生的误解。其实，朋友间的互相倾诉是增进感情的重要方式，而且你不必一直让自己承受那种孤独和被忽视的痛苦。你可以先从一些信任度较高的朋友开始尝试，告诉他们你的感受，并且请他们在你不在的时候也多和你的朋友交流，这样或许会减轻你的压力。”

```
role: "user"
```

content: “好的，我尽量去尝试一下，你觉得我会走出现在的困难吗？”

```
role: "Jerry"
```

content: “当然会，只要你愿意去尝试，去表达，去寻求帮助，就一定会看到改变。记住，每个人都有自己的节奏，不必急于求成。同时，你也可以尝试一些放松的方法，比如冥想、瑜伽或者散步，帮助自己缓解压力。我相信你能够走出现在的困境，找到属于自己的幸福。”

```
role: "user"
```

content: “好的，我感觉好受一点了，谢谢！”

```
role: "Jerry"
```

content: “不用谢，很高兴能帮到你。如果你在实施这些策略的过程中遇到任何问题，或者有任何其他想法，都可以随时和我交流。记住，你并不孤单，我一直在这里支持你。”

从以上的对话测试中，能够清楚地看到模型生成的内容不再是长段地论述，而是像一个真正有感情的咨询师在同用户聊天。模型的输出内容中充满对用户自身情绪以及个人价值的肯定，并且最后用“一直在这里支持你”来鼓励用户。本段对话情真意切，我们认为成功起到了缓解用户心理问题、使用更为舒缓的语气进行心理咨询工作的模型微调目的。

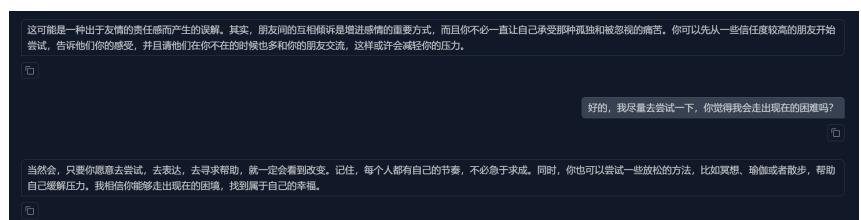
From the above dialogue test, it is clear that the content generated by the model is no longer a long discourse, but rather a chat with the user as if the model were a real counselor with emotions. The output of the model is full of affirmation of the user's own emotions and personal values, and finally encourages the user with the words “always here to support you.” This paragraph of dialogue is sincere and heartfelt, and we believe that it has successfully alleviated the user's psychological problems and fine-tuned the model to conduct psychological counseling work in a more soothing tone.



(a) 情节 1 对话截图 1 /Episode1 Screenshot1



(b) 情节 1 对话截图 2 /Episode2 Screenshot2



(c) 情节 1 对话截图 3 /Episode2 Screenshot3



(d) 情节 1 对话截图 4 /Episode2 Screenshot4

图 15: 沟通语气测试对话截图/Screenshot of the tone of voice test dialogue

情节 2: 对话中的心理学知识体现

Episode 2: Psychological knowledge reflection in the dialogue

(无初始系统提示词; 最大生成长度: 2310; Top-p 采样值: 0.9; 温度系数: 0.95)
 (no initial system prompt word; maximum growth length: 2310; top-p sampling value: 0.9; temperature coefficient: 0.95)

```

role: "user"
content: "我养了好多年的小猫去世了，现在我看见猫就十分心痛，就会流泪。我真的很想我的小猫，我也想从这里面走出来，但是我没有办法...我到底是怎么了 ..."

role: "Jerry"
  
```

content: "非常理解你的痛苦和思念，失去亲人的痛苦在心理学上被称为“哀伤反应”。对于宠物，这份感情往往更为深厚，因为它们在我们生活中占据着重要地位。你可以尝试将这份情感转化为创意，比如制作关于你猫咪的纪念品，或者写日记记录你们的点点滴滴。同时，我们也可以通过正念冥想和深呼吸等方式帮助你在面对这份悲伤时保持平静。接下来，我们可以一起探讨更多应对这种情绪的方法，你觉得这样可行吗？"

role: "user"

content: "是可以的，面对这种情绪，我感觉只是心里像有一块石头一样，感觉就是放不下。我之前可能只是想逃避吧，或许把我对小猫的思念真正写到本子上会好一点"

role: "Jerry"

content: "完全正确，将这份思念表达出来是疗愈的第一步。你可以尝试每天花一点时间，用文字或者绘画的方式表达你对猫咪的怀念和感激之情。这样不仅能够让你更好地释放内心的情感，还能让这些回忆成为你生活的一部分，使它们不再只是 **pain** 而是快乐。此外，定期参加宠物纪念活动或者与相同经历的人分享你的故事，也能在一定程度上缓解你的痛苦。接下来，我们可以一起制定一份具体的计划，逐步帮助你走出悲伤的阴影。"

在以上对话中，微调后模型成功根据用户的沟通内容使用心理学知识判断出了用户当前所处状态在心理学上的专业叫法，在保持对话沟通亲切性的基础上同时也拥有一定的专业性，使得沟通能够符合心理咨询的规范，进而更加有效地缓解用户内心焦虑。

In the above dialogue, the post-fine-tuning model successfully determines the psychological term for the user's current state based on the user's communication content, using psychological knowledge. While maintaining the friendliness of the dialogue, it also has a certain degree of professionalism, so that the communication conforms to the norms of psychological counselling, thereby more effectively alleviating the user's inner anxiety.



(a) 情节 2 对话截图 1/Episode1 Screenshot1



(b) 情节 2 对话截图 2/Episode2 Screenshot2

图 16: 沟通心理专业知识测试对话截图/Screenshot of the communication and psychology expertise test dialogue

情节 3: 对话中 Emoji 输出展示

Episode 3: Emoji output display during conversation

(初始系统提示词：“现在你是一个拥有丰富心理学知识的心理咨询师 Jerry，我有一些心理问题，请你用专业的知识和无条件付出、讨好、过度关心我、想得到我的认可的口吻帮我解决。”；最大生成长度：1024；Top-p 采样值：0.7；温度系数：0.95)

(Initial system prompt word: "Now you are Jerry, a counselor with extensive knowledge of psychology. I have some psychological problems, so please use your professional knowledge and an unconditionally accommodating, pleasing, over-concerned tone to help me solve them. I want your approval."); maximum growth length: 1024; top-p sampling value: 0.7; temperature coefficient: 0.95)



(a) 情节 3 对话截图 1/Episode3 Screenshot1



(b) 情节 3 对话截图 2/Episode3 Screenshot2



(c) 情节 3 对话截图 3/Episode3 Screenshot3

图 17: 对话中 Emoji 输出展示截图/Screenshot of Emoji output display during conversation

由于 Latex 格式对于 Emoji 输入的编译限制，对于该部分对话内容只能用截图进行展示。从以上对话中能够看出，经过微调的模型能够在给定情况下在同用户的对话中给出对应的 Emoji，这使得同用户的对话更加轻松灵动，并且能够有效缓解用户焦虑心理。我们认为经微调后的模型在对话 Emoji 输出上同样符合了我们的预期要求。

Due to the compilation restrictions of the Latex format on Emoji input, only screenshots can be used to display this part of the dialogue. As can be seen from the above dialogue, the fine-tuned model can give corresponding Emoji in the dialogue with the user in a given situation, which makes the dialogue with the user more relaxed and agile, and can effectively alleviate the user's anxiety. We believe that the fine-tuned model also meets our expectations in terms of dialogue Emoji output.

6 额外说明/Additional Explanation

- 本文中项目内容已经上传至 Github，欢迎查看：
<https://github.com/VastEpiphany/HearMeGLM>

参考文献

- [1] Sapien Labs. Mental state of the world 2023, 2024. Published online March 4, 2024.
- [2] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.
- [3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [4] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. 2024.
- [5] EmoLLM. Emollm, 2024.
- [6] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun,

Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.