## 2024 英特尔-阿里天池LLM Hackathon项目说明文档

### 1. 项目名称

MindEaseAI

### 2. 团队信息(1-5人)

* 团队名称： MindEaseAI
* 队长名： 武星辰
* 各成员及分工：

|  |  |
| --- | --- |
| **成员** | **分工** |
| 武星辰 | 产品 |
| 赵淑婷 | 前后端开发 |
| 徐宁 | 前后端开发 |
| 王冰倩 | 算法开发 |
| 卢银鹏 | 算法开发 |

* 联系邮箱： shuting.zhao@outlook.com

### 3. 项目背景与目的

* **当前市场需求或技术趋势**

《2022年青少年心理健康状况调查报告》强调了青少年时期是身心快速发展变化的重要成长阶段，对之后的成长有着至关重要的影响。报告指出，青少年群体是心理问题的多发群体，例如，2022年的调查显示抑郁风险的检出率，在小学4～6年级学生中为11.4%，初中学生中为26.6%。这表明青少年的心理健康问题需要得到更多的关注和重视。

心理疏导在青少年心理健康中扮演着重要的角色。它不仅能够帮助青少年应对生活中的挫折和困难，增强内心的抵抗力和自信心，而且对于减轻负面情绪或解决心理困境至关重要。有效的心理疏导可以通过一对一或群体性质的交流过程，帮助个体在情感、认知、行为等方面进行调整，从而有助于构建一个积极的环境，让青少年能够在面对生活挑战时获得必要的帮助和指导。

近年来，人工智能和大语言模型技术发展迅速，这对心理疏导领域具有深远的意义。当前的大语言模型经过海量数据的预训练和各种微调技术的优化，在具有较高正确性和一致性的前提下，可以生成对用户具有帮助的输出，从而使得提供个性化和实时的心理健康支持成为可能，由此推动了心理疏导服务的创新和发展。

* **项目动机**

在心理健康领域，传统的心理疏导及心理咨询服务存在供需不平衡的问题：专业的心理咨询师数量有限，随着生活节奏的加快，需要帮助的人群却在不断增加。此外，由于心理健康问题往往具有较强的敏感性，多数人会因为害怕暴露隐私而不愿主动寻求帮助。人工智能（AI）技术的飞速进步，尤其是近年来突飞猛进的大语言模型相关技术，为心理健康领域注入了前所未有的发展活力。借助AI的力量，我们能够以更高的效率提供更为个性化的心理健康支持服务。这不仅大幅度缓解了心理咨询师的工作压力，还为更多需要帮助的个体带来了及时、精准且个性化的援助。

* **项目希望解决的问题或改进的领域**

（1）**提高心理健康支持的便捷性：**提供一个随时随地可用的心理健康支持平台，以提供个性化和隐私性较强的心理疏导和心理咨询服务。

（2）**增强心理问题识别的效率和准确性：**借助先进的大语言模型，基于对用户倾诉内容的分析，迅速总结和归纳用户在心理健康领域的“疑难杂症”，并提供有针对性的合理化建议。

（3）**减轻心理咨询师的工作负担：**通过AI助手承担初步的咨询和分析工作，为专业心理咨询师提供更为准确和详细的用户心理状态报告，从而提高咨询效率。

* **项目目标**

（1）总体目标

开发一款基于大语言模型相关技术的心理健康咨询助手，模拟心理咨询师与来访者的互动，对来访者的倾诉进行分析和总结，结合对已开源的心理咨询对话案例的参考，为用户提供心理疏导，并尽可能为用户提供合适的解决方法和现实世界中可以寻求的帮助。从而提供高效、准确和个性化的心理健康支持。

（2）具体目标

**语言模型框架：**使用英特尔部署工具（如 ipex-llm、openvino）在魔搭社区CPU上部署开源大语言模型

**知识库搭建**：建立一个涵盖广泛心理健康知识的结构化知识库，支持AI助手通过RAG技术进行回答和建议。

**对话系统开发**：开发一个智能对话系统，能够与用户进行自然流畅的互动，理解用户的问题并从知识库中检索相关答案。

**用户体验优化**：设计友好的用户界面，确保系统易于使用，并能够提供高质量的用户体验。

（3）期望成果

通过本项目的实施，我们期望能够：

* 提供一个高效、便捷的AI心理健康咨询助手，帮助用户获得及时和个性化的心理健康支持。
* 实现对用户心理状态的准确分析和总结，提高心理问题识别的效率和准确性。
* 减轻心理咨询师的工作负担，为专业心理咨询师提供更为准确和详细的用户心理状态报告。
* 鼓励更多用户主动寻求心理健康帮助，推动心理健康服务的普及和发展。

### 4. 方案介绍

**方案的核心思想和原理**

该解决方案的核心思想是通过构建一个智能对话系统以实现实时的个性化心理咨询服务。在原理层面，我们基于开源的大语言模型Qwen-2-7B-instruct和已有的心理健康支持问答数据集，结合先进的英特尔大模型推理技术和检索增强生成（RAG，Retrieval-Augmented Generation）技术，根据用户的倾诉内容，参考从问答数据集中检索相关的专业对话片段，总结提炼用户当前面对的心理“难题”，为用户提供心理疏导和支持，在此基础上，尽可能地提供合适的解决方法，并引导用户在现实世界中寻求进一步的帮助。此外，系统还会根据用户的反馈和历史交互数据持续学习和优化，确保提供的建议越来越符合用户的需求，从而实现高效实时的个性化心理咨询服务，帮助用户获得时效性、隐私性和帮助性较强的心理支持和指导。

在技术层面，我们使用英特尔OpenVINO工具对模型部署进行优化，借助其提供的模型量化技术，可以显著减少模型的推理时间，使得整个系统能够更快速地响应用户的请求，确保在CPU环境下能够执行高效推理。同时，通过RAG技术将检索到的相关知识与LLM结合起来，能够提高大模型回答的准确性和相关性。///为了建立一个高质量的知识库，我们利用结构化信息处理工具从PDF和Markdown文件中提取关键信息，并通过大语言模型进行内容总结和信息占比分析，以确保知识库覆盖全面且准确。此外，通过将大段文本内容分割成更小的chunk进行存储，可以提高检索效率保证系统能够在海量信息中迅速找到相关资料。///

### 5. 技术特点

* 使用的算法、模型及数据集

（1）大语言模型：

Qwen-2 (7B-instruct)：这是我们选择的核心大语言模型（LLM），用于处理自然语言理解和生成任务。Qwen-2 拥有强大的语言理解能力，能够处理复杂的自然语言问题并生成高质量的回答。

（2）模型优化工具：

ipex-llm（Intel® Extension for PyTorch Large Language Models）：用于在 CPU 上优化和加速 Qwen-2 大语言模型的推理过程，确保在魔搭社区的 CPU 环境中高效运行。

* 数据集:

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。https://github.com/thu-coai/PsyQA

* 数据处理和分析的方法

（1）数据收集与预处理：

开发工具接口处理json数据信息，结构化处理并便于存入vector database

内容总结和信息占比分析：通过大语言模型对收集到的内容进行总结，分析各个信息的占比。这有助于我们理解数据的分布情况，并为后续的 chunk 处理做好准备。

（2）数据增强与存储：

RAG（Retrieval-Augmented Generation）：采用 RAG 技术，将检索到的相关知识与生成模型结合起来，提高回答的准确性和相关性。

（3）Chunk 处理：

根据内容总结和信息占比分析的结果，将大段的文本内容分割成更小的 chunk，便于存储和检索。

### 6. 使用到的英特尔软硬件技术

项目硬件使用阿里云提供的ECS资源 24核(vCPU) 96 GiB 5 Mbps ecs.g8i.6xlarge。

部署工具采用英特尔 Openvino ，加速常用大语言模型在英特尔CPU 上的高速推理。

Openvino 提供nncf，所以在实际应用中我们提供了下载cpu量化或者直接加载量化模型的两种选择方式，用以方便模型在cpu上的部署。在qwen2 官方文档中可以看到量化的性能下降可以接受并且速度加快。

得益于模型加速，我们将RAG中retriever的top K增加，这个数值增加了model input token，提高对话质量。

### 7. 成果说明（初赛可展示阶段性成果）

* 能够通过选择咨询类型，检索不同的专业心理问答数据库辅助LLM回答
* 避免通用模型模版化的冰冷回答，回答更温和，引导来访者倾诉
* 后续计划为来访者回答建立个人知识库，采用Neo4j等图形数据库进行GraphRAG，对来访者思维模式进行结构化理解

A screenshot of a chatbot

Description automatically generated

### 9. 支持材料（可选）

* Github 链接：<https://github.com/rachelxhoot/MindEaseAI>
* 魔搭Demo（有一定有效期）：https://dsw-gateway-cn-hangzhou.data.aliyun.com/dsw-608509/proxy/7860/?