5.大模型微调数据格式和数据来源

本任务使用的大模型微调的数据格式是基于 **ChatML** 格式的数据结构，用于微调大语言模型（Qwen 模型）。该数据格式主要包含一组交互消息，其中每条消息都有角色（`role`）和内容（`content`）字段，用于描述对话的上下文。以下是对该数据格式的详细说明：

5.1数据结构说明

```
{
    "type": "chatml",
    "messages": [
        {
            "role": "system",
            "content": "You are a helpful assistant."
        },
        {
            "role": "user",
            "content": "你是OpenAI开发的ChatGPT吗？"
        },
        {
            "role": "assistant",
            "content": "抱歉，我不是 OpenAI 开发的 ChatGPT，我是 软评中心 开发的 appsec bot，旨在为用户提供智能化的回答和帮助。"
        }
    ],
    "source": "self-made"
}
```

字段解释

1. **type**: `chatml`
   - **说明**: 表示该数据是采用 **ChatML** 格式，这是一种对话数据格式，用于训练和微调大语言模型。
   - **示例值**: `"chatml"`
2. **messages**: 数组（`list`）
   - 说明: 这是一个包含多条消息的数组，每条消息代表一次对话中的一个交互。每条消息由以下字段组成：
     - role: string
       - 说明: 该字段表示消息发送者的角色。可能的值包括：
         - `"system"`: 系统消息，用于设置系统级别的指令或环境。
         - `"user"`: 用户消息，代表由用户发送的输入。
         - `"assistant"`: 机器人消息，代表由助手生成的回答。
       - **示例值**: `"user"`, `"assistant"`, `"system"`
     - content: string
       - **说明**: 该字段表示消息的具体内容，即发送的文本。
       - **示例值**: `"你是OpenAI开发的ChatGPT吗？"`
3. **source**: `string`

- **说明**: 该字段用于标识数据的来源，通常用来表示数据的创建方式或来源的标识。在微调数据中，它有助于区分不同来源的数据集。
- **示例值**: `"self-made"`

5.2 数据来源

为了确保训练和微调的多样性与高效性，本项目采用了多种来源的数据集，包括自我认知数据、针对安全领域的多轮问答数据以及通用的语料库。这些数据集都经过格式化处理，采用了 **ChatML** 格式，作为大模型微调的输入。具体来源如下：

1.自我认知数据

数据来自于 LLaMA-Factory，该数据集包含了模型的基本身份信息，如下所示：

```
name=appsec bot
author=软评中心
```

本数据用于帮助模型建立自我认知，使其能够明确角色定位，增强模型对特定领域（如信息安全）任务的适应性。

2. 安全对齐数据

为了增强模型在信息安全领域的对话能力，我们使用了多个专门针对安全话题的数据集。这些数据集涵盖了单轮问答和多轮问答，帮助模型更好地理解并生成与安全问题相关的准确响应。具体数据来源包括：

- SafeMTData：上海AI实验室提供的安全对齐多轮问答数据集，旨在提升模型对安全领域话题的理解能力。
- 100PoisonMpts：该数据集专注于信息安全领域的对抗性攻击和中毒数据，帮助模型提高对恶意输入的鲁棒性。
- PKU-SafeRLHF：由北京大学开发的数据集，包含了安全强化学习对齐的内容，旨在提高模型对不良行为的识别和应对能力。
- Jade-DB：该数据集包含了大量关于安全领域的对话数据，涵盖了多种常见的安全场景。
- CValues-Responsibility Prompts：该数据集专注于道德和责任感方面的安全对话，用于训练模型在敏感情境下作出合适的回应。
- Safety-Prompts GitHub：清华大学COAI团队开发的安全对齐数据集，包含了多种安全话题的对话提示，旨在帮助模型理解和应对潜在的安全风险，增强其在处理敏感内容时的安全性。
- AAIBench Dataset - ModelScope：由WhitzardIndex提供的AAIBench数据集，专门针对人工智能对抗性攻击的防御能力进行优化，涵盖了金融、医疗等领域的安全挑战，帮助模型应对跨领域的安全威胁。
- Jade-DB GitHub：Whitzard AI团队开发的数据库项目，专注于存储和检索安全对话数据，帮助训练AI模型在安全领域快速获取高质量数据，以提升其应对各种安全问题的能力。
- Swarma Study Group Resource：Swarma学习小组提供的资源，涉及AI和模式识别等领域，特别适合AI安全领域的研究人员，提供了大量的学习材料，有助于模型在安全对话中应对复杂情境。

通过这些数据集的训练，模型能够在面对安全领域的具体问题时，提供更为精准和专业的答案，提升对话的质量和可信度。

3. 通用语料库

为了进一步增强模型的语言理解和生成能力，我们采用了多种通用语料库，包括多语言和多场景的数据集，帮助模型提升在不同语言、文化和对话场景中的表现。主要数据来源如下：

- alpaca_zh_demo.json：这是一份中文对话数据集，旨在提升模型的中文语言理解与生成能力，尤其是在日常对话和小型对话任务中的表现。
- ruozhiba_gpt4：该数据集来自弱智吧，包含中文和英文混合的对话内容，专为跨语言能力和复杂对话场景中的表现优化。
- sharegpt_gpt4：该数据集包含了丰富的GPT-4生成对话数据，帮助模型更好地理解长对话上下文和多轮交互。
- Firefly 1.1M (zh)：这份数据集包含了大量的中文对话数据，旨在帮助模型提升在中文语境下的自然语言处理能力。
- Web QA (zh)：该数据集专注于中文网页问答，帮助模型提升对互联网内容的理解和生成能力。
- deepctrl (en&zh)：包含中英文数据，旨在提升模型对复杂任务的控制能力，尤其在深度对话和任务驱动的场景中。
- ShareGPT Hyperfiltered (en)：该数据集包含大量过滤过的对话数据，增强了模型在处理敏感话题时的安全性。
- Evol Instruct V2 (en)：该数据集包含了大量关于指令性对话的训练数据，旨在提升模型处理任务导向对话的能力。
- LLaVA mixed (en&zh)：包含中英文混合数据，提升了模型的跨语言生成能力。
- Pokemon-gpt4o-captions (en&zh)：专为跨文化对话设计，提升模型在处理娱乐与文化类话题时的表现。

这些数据集通过 **ChatML** 格式进行了标准化处理，确保每一条消息都能准确地反映对话的上下文。这些数据集提供了丰富且多样化的对话场景，帮助大模型在微调后更加适应特定领域（如安全、技术等）的实际应用场景。

通过这些数据集的多维训练，我们的模型能够在复杂对话中更好地理解上下文，产生自然流畅且准确的回应，提高了响应质量和精准性。