数据格式

1.SFT（有监督微调）的数据集格式?

对于大语言模型的训练中，SFT（Supervised Fine-Tuning）的数据集格式可以采用以下方式：

1. 输入数据：输入数据是一个文本序列，通常是一个句子或者一个段落。每个样本可以是一个字符串或者是一个tokenized的文本序列。
2. 标签数据：标签数据是与输入数据对应的标签或类别。标签可以是单个类别，也可以是多个类别的集合。对于多分类任务，通常使用one-hot编码或整数编码来表示标签。
3. 数据集划分：数据集通常需要划分为训练集、验证集和测试集。训练集用于模型的训练，验证集用于调整模型的超参数和监控模型的性能，测试集用于评估模型的最终性能。
4. 数据集格式：数据集可以以文本文件（如CSV、JSON等）或数据库的形式存储。每个样本包含输入数据和对应的标签。可以使用表格形式存储数据，每一列代表一个特征或标签。

下面是一个示例数据集的格式：

```
Input,Label
"This is a sentence.",1
"Another sentence.",0
...
```

在这个示例中，**输入数据是一个句子，标签是一个二分类的标签**（1代表正例，0代表负例）。每一行代表一个样本，第一列是输入数据，第二列是对应的标签。

需要注意的是，具体的数据集格式可能会因任务类型、数据来源和使用的深度学习框架而有所不同。因此，在进行SFT训练时，建议根据具体任务和框架的要求来定义和处理数据集格式。

2.RM（奖励模型）的数据格式?

在大语言模型训练中，RM（Reward Model，奖励模型）的数据格式可以采用以下方式：

1. 输入数据：输入数据是一个文本序列，通常是一个句子或者一个段落。每个样本可以是一个字符串或者是一个tokenized的文本序列。
2. 奖励数据：奖励数据是与输入数据对应的奖励或评分。奖励可以是一个实数值，表示对输入数据的评价。也可以是一个离散的标签，表示对输入数据的分类。奖励数据可以是人工标注的，也可以是通过其他方式（如人工评估、强化学习等）得到的。
3. 数据集格式：数据集可以以文本文件（如CSV、JSON等）或数据库的形式存储。每个样本包含输入数据和对应的奖励数据。可以使用表格形式存储数据，每一列代表一个特征或标签。

下面是一个示例数据集的格式：

```
Input,Reward
"This is a sentence.",0.8
"Another sentence.",0.2
...
```

在这个示例中，输入数据是一个句子，**奖励数据是一个实数值，表示对输入数据的评价。**每一行代表一个样本，第一列是输入数据，第二列是对应的奖励数据。

需要注意的是，具体的数据集格式可能会因任务类型、数据来源和使用的深度学习框架而有所不同。因此，在使用RM进行大语言模型训练时，建议根据具体任务和框架的要求来定义和处理数据集格式。

3.PPO（强化学习）的数据格式?

在大语言模型训练中，PPO（Proximal Policy Optimization，近端策略优化）是一种常用的强化学习算法。PPO的数据格式可以采用以下方式：

1. 输入数据：输入数据是一个文本序列，通常是一个句子或者一个段落。每个样本可以是一个字符串或者是一个tokenized的文本序列。
2. 奖励数据：奖励数据是与输入数据对应的奖励或评分。奖励可以是一个实数值，表示对输入数据的评价。也可以是一个离散的标签，表示对输入数据的分类。奖励数据可以是人工标注的，也可以是通过其他方式（如人工评估、模型评估等）得到的。
3. 动作数据：动作数据是模型在给定输入数据下的输出动作。对于语言模型，动作通常是生成的文本序列。动作数据可以是一个字符串或者是一个tokenized的文本序列。
4. 状态数据：状态数据是模型在给定输入数据和动作数据下的状态信息。对于语言模型，状态数据可以是模型的隐藏状态或其他中间表示。状态数据的具体形式可以根据具体任务和模型结构进行定义。
5. 数据集格式：数据集可以以文本文件（如CSV、JSON等）或数据库的形式存储。每个样本包含输入数据、奖励数据、动作数据和状态数据。可以使用表格形式存储数据，每一列代表一个特征或标签。

基本元素：

概念	含义
State（状态）	表示当前环境的状态，对语言模型来说通常是输入文本或上下文
Action（动作）	模型根据当前状态做出的决策，在语言模型中就是生成的文本
Reward（奖励）	对动作好坏的评估，可以是人工打分、自动评分模型（如RM）输出
Policy（策略）	模型本身，即从状态到动作的概率分布

下面是一个示例数据集的格式：

```
Input,Reward,Action,State
 "This is a sentence.",0.8,"This is a generated sentence.",[0.1, 0.2, 0.3, ...]
"Another sentence.",0.2,"Another generated sentence.",[0.4, 0.5, 0.6, ...]
...
```

在这个示例中，输入数据是一个句子，奖励数据是一个实数值，动作数据是生成的句子，状态数据是模型的隐藏状态。每一行代表一个样本，第一列是输入数据，第二列是对应的奖励数据，第三列是生成的动作数据，第四列是状态数据。

需要注意的是，具体的数据集格式可能会因任务类型、数据来源和使用的深度学习框架而有所不同。因此，在使用PPO进行大语言模型训练时，建议根据具体任务和框架的要求来定义和处理数据集格式。

4.找数据集哪里找?

在训练自己的大语言模型时，可以从以下几个途径找到合适的数据集：

1. **公开数据集**：有许多公开可用的数据集可供使用，涵盖了各种领域和任务。例如，Common Crawl、Wikipedia、OpenWebText、BookCorpus等都是常用的大规模文本数据集，可以用于语言模型的训练。
2. **开放数据平台**：许多组织和机构提供了开放的数据平台，可以获取各种类型的数据。例如，Kaggle、UCI Machine Learning Repository、Google Dataset Search等平台都提供了丰富的数据集资源。
3. **学术界研究**：许多学术研究项目会公开其使用的数据集，可以通过相关论文或项目页面找到这些数据集。例如，NLP领域的一些会议和竞赛（如ACL、EMNLP、CoNLL、GLUE等）提供了公开的数据集供研究使用。
4. **数据收集和爬取**：如果没有合适的公开数据集，您可以自己进行数据收集和爬取。这可以通过爬虫技术从互联网上收集相关的文本数据。需要注意的是，在进行数据收集和爬取时，需要遵守法律法规和网站的使用条款，并确保获得数据的合法使用权。
5. **数据增强**：如果您已经有了一些初始的数据集，但觉得数量不够，可以考虑使用数据增强技术来扩充数据。数据增强可以通过对原始数据进行一些变换、替换、合成等操作来生成新的样本。
    - EDA(Easy Data Augmentation): 同义词替换、同义词随机插入、随机选择两个单词交换位置、随机删除一个单词
    - AEDA(An Easier Data Augmentation): 在$[1, \frac{1}{3} \times \text{len}]$中随机选择一个数作为插入的位置的数目，在每一个插入位置从{'.', ';', '?', ':', '!', ','}中随机选择一个插入
    - 回译(Back Translation): 将文本翻译成另一种语言，然后再翻译回来。可以翻译成多种语言，从而得到多条回译样本
    - Masked Language Model: 利用预训练好的BERT, Roberta等模型，对原句子进行部分掩码，然后让模型预测掩码部分，从而得到新的句子。但是，这种方法存在的一个问题是，决定要屏蔽文本的哪一部分并不简单。可以考虑使用启发式方法来确定掩码部分，否则，生成的文本可能无法保留原始句子的含义。（启发式方法：基于词性或词频等方法。基于词性选择对句子语义影响不大的介词、冠词、连词等，基于词频选择频率较高的功能词）
    - More: 文本数据增强方法总结

无论从哪个途径获取数据集，都需要注意数据的质量、版权和隐私等问题。确保您有合法的使用权，并遵守相关的法律和伦理规范。

5.微调需要多少条数据?

根据 Scaling Laws，随着模型大小、数据集大小和用于训练的计算浮点数的增加，模型的性能会提高。并且为了获得最佳性能，所有三个因素**必须同时放大**。一般来说对于**给定模型的理想训练数据集 token 数量大约是模型中参数数量的20倍**。

6.有哪些大模型的训练集?

以下是一些常用的大语言模型训练集的示例：

1. Common Crawl：这是一个由互联网上抓取的大规模文本数据集，包含了来自各种网站的文本内容。它是一个常用的数据集，可用于语言模型的训练。
2. Wikipedia：维基百科是一个包含大量结构化文本的在线百科全书。维基百科的内容丰富多样，涵盖了各种领域的知识，可以作为语言模型训练的数据集。

3. OpenWebText：这是一个从互联网上抓取的开放文本数据集，类似于Common Crawl。它包含了大量的网页文本，可以作为语言模型的训练数据。
4. BookCorpus：这是一个包含了大量图书文本的数据集，用于语言模型的训练。它包括了各种类型的图书，涵盖了广泛的主题和领域。
5. News articles：新闻文章是另一个常用的语言模型训练集。可以通过从新闻网站、新闻API或新闻数据库中收集新闻文章来构建训练集。
6. 其他领域特定数据集：根据具体任务和应用，可以使用特定领域的数据集来训练语言模型。例如，在医学领域，可以使用医学文献或医疗记录作为训练数据；在法律领域，可以使用法律文书或法律条款作为训练数据。

需要注意的是，使用这些数据集时，应该遵守数据的版权和使用规定，确保合法的使用权。此外，还可以通过数据增强技术，如数据合成、数据变换等，来扩充训练集的规模和多样性。

7.进行领域大模型预训练应用哪些数据集比较好?

进行领域大模型预训练时，可以使用以下几种数据集来获得更好的效果：

1. 领域特定文本数据集：收集与目标领域相关的文本数据集，例如专业领域的论文、报告、文档、书籍等。这些数据集可以提供领域内的专业术语、上下文和特定领域的知识。
2. 领域内的网页内容：从目标领域相关的网页抓取文本内容。可以通过爬虫技术从相关网站上获取与目标领域相关的网页文本数据。
3. 领域内的新闻文章：收集与目标领域相关的新闻文章。新闻文章通常包含了领域内的最新信息和事件，可以帮助模型了解领域内的动态和趋势。
4. 行业报告和白皮书：获取与目标领域相关的行业报告、白皮书和研究文献。这些文献通常包含了领域内的专业分析、统计数据和趋势预测，可以帮助模型了解行业背景和发展趋势。
5. 社交媒体数据：收集与目标领域相关的社交媒体数据，如推特、微博、论坛等。社交媒体上的内容通常反映了人们在目标领域中的讨论、观点和问题，可以帮助模型了解领域内的热点和用户需求。
6. 领域内的对话数据：获取与目标领域相关的对话数据，如客服对话、问答平台数据等。这些对话数据可以帮助模型学习领域内的常见问题、解决方案和用户需求。

在选择数据集时，应该确保数据的质量和合法性，并遵守相关的法律和伦理规范。同时，还可以考虑使用数据增强技术，如数据合成、数据变换等，来扩充训练集的规模和多样性。