诚 信 声 明

我声明，所呈交的毕业论文是本人在老师指导下进行的研究工作及取得的研究成果。据我查证，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得其他教育机构的学位或证书而使用过的材料。我承诺，论文中的所有内容均真实、可信。

毕业论文作者签名： 签名日期： 年 月 日

用于LLM的专业数据集设计

**[摘 要]**

**[关键词]** ； ；

Professional data set design for LLM

**Abstract：**

/\*将中文摘要译成英文，建议在二稿时完成\*/

**Keywords：**  ； ；

/\*关键词是供检索用的主题词条，应采用能覆盖论文主要内容的通用技术词条，一般列3～5个，按词条的外延层次从大到小排列。关键词之间以“；”号间隔 \*/

# 绪论

在深度学习技术高速发展的当下，大模型（Large Model）已然成为了人工智能领域的研究热点。其中，大模型下的一个重要分支——大语言模型（Large Language Models，以下简称LLMs），因其越来越强大的语言能力，正逐渐改变着人们学习、工作方式。大语言模型的快速发展，让人们改变了对人工智能的认知以及使用方式。

## 研究背景及动机

早期对大语言模型的研究主要依赖与规划，但随着2003年神经网络语言模型的提出，深度学习开始在自然语言处理（NLP，以下简称NLP）中发挥作用。2013年Word2Vec引入了此嵌入技术，为文本数据提供了高效的数值表示。2017年，Transformer的提出彻底改变了NLP领域，以其高效处理长序列数据的能力为基础，促进了大预言模型的高速发展。在此之后，越来越多的LLM公开问世，包括非开源模型GPT系列，以及开源模型BERT、LLAMA系列等。

LLM发布的初衷，是为了将其运用到各类垂直领域中，如：教育领域、科研领域和社会各类工作领域等。然而对于开源的LLM，虽然他们的参数量很大，且在一般NLP任务中已经能有不错的表现，但是一旦涉及到专业领域的问题，他们的表现就都差强人意，无法满足在垂直领域部署使用的条件。因此开源LLM的预训练和微调，就显得举足轻重。

然而虽然现在有非常多的文本数据集，可以用作LLM的预训练与微调。但由于制作数据集的人员可能并不是对应专业领域的专业人员，因此这些数据集在对应领域大都难以达到高专业性、高精确的质量要求。训练出来的LLM在对应领域依然缺乏一定的专业性，没法进一步投入到垂直领域进行应用开发。而具有相关垂直领域专业知识的科研人员，却又缺少数据集制作的知识，这导致做出来的数据集会有无法投入训练的情况。而目前LLM的研究领域缺少数据集制作的文章，这导致想要利用LLM做垂直领域开发工作的各领域的科研人员，无法制作出专业的数据集对模型进行训练微调。

## 论文结构概述

为了更好的帮助研究人员对LLM的垂直领域开发，本篇论文将对数据集制作以及模型预测结果评估方式进行总结，并总结出一条切实可行的技术路线供相关科研人员进行参考。以下是文章的结构：

第一章：绪论，概述了LLM的发展背景以及为何要做数据集设计研究。

第二章：数据集设计，概述了数据集设计的基本方法，同时设计了一个用于化学领域LLM训练的数据集。提供具体的数据集设计方法参考。

第三章：数据集评估，概述了数据集质量评估的一些参考指标，并且以第二章中设计的化学数据集为例，来评测该数据集的质量高低，以此来提供一个数据集质量评估的具体做法。

第四章：预测结果评估，对现有的预测指标进行了综述，设计了一套用于评测在该化学数据集上适用的模型预测结果评估系统。并藉由一系列评估数据和人工评估的介入，来比较该评估系统与其他评估指标的优劣。

第五章：总结，对本论文的研究结果进行总结，同时指出不足与局限所在，并对局限不足提出未来改进期望。

由于本文是对LLM的头和尾，即数据集的处理和预测结果的评估进行研究工作，没有足够的条件进行LLM的训练微调，因此预测结果的相关数据，将会使用文心一言来生成。

# 数据集设计

对LLM进行预训练和微调第一步，就是找到合适的数据集，然后进行数据预处理，将数据处理为模型能够接收的格式再进行预训练和微调。由于各企业或者各研究机构对数据集的要求各有不同，且现存的数据集难以很好地满足个性化需求，因此若想训练出能够用于垂直领域应用开发的LLM，优秀的数据集设计就显得举足轻重。本章节将对LLM数据集设计方法进行总结，同时，为了直观的展现数据集的设计流程，本文将设计一个用于化学领域LLM训练的数据集。该数据集包含两部分，分别为预训练数据集——化学专有名词语料库，和监督微调数据集——化学QA数据（其中Q表示问题，A表示答案）。

## 数据集设计方法概述

数据集的设计，不仅需要确保数据的质量和相关性，还需要考虑到数据的多样性和代表性，避免因缺乏多样性和代表性的低质量数据集造成训练后的LLM模型缺乏泛化能力。数据集的设计和构建，需要遵循一系列详尽且严格的标准和步骤，即：明确研究目标和需求分析、数据源的选择和数据收集、数据预处理和数据清洗、数据增强、扩充和标注、质量控制和验证、预测结果评估方式、文档编制和维护。

### 数据收集处理

在数据集设计初步阶段，明确研究领域、问题的定义和研究的目标是至关重要的。这一过程包括对问题域的深入理解，以及确定数据需求的详细规范。需求分析阶段，不仅需要解决所需数据类型、格式、大小等问题，还要对数据的分布情况有细致的考量。同时，也需要对潜在的数据来源和获取方法进行有一定程度的分析了解，确保数据来源的可靠性，能够将其用来支撑后续的工作高质量进行。

当对研究目标和需求分析深入分析透彻后，将进入数据集设计中的一个关键环节——合适数据源的选取。在这一环节中，设计人员要能够评估和选择可以提供高质量、相关性强和代表性好的数据源。优质的数据来源包括但不限于相关书籍、网络百科和专业论坛。数据收集方法则有网络爬虫、公开数据集调用、API调用、社会调查以及数据生成等。在收集数据时，应考虑数据的合法性，遵守隐私保护政策，确保数据收集过程符合相关法律法规和伦理标准。

第一次收集而来的数据，必定包含有不少的脏数据，因此对数据进行清洗，是确保数据集质量的关键步骤。数据清洗主要目标是去除异常值、处理缺失值、删除无关值和排除错误值。目的是为了除去数据中的噪声，以提高数据的准确性和可靠性。同时还需要进行数据的脱敏处理，以去除匿名化敏感信息，保护个人隐私。

### 数据质量和预测结果评估

数据集的质量验证是在数据收集与处理流程完成后进行的一项关键活动。该验证过程的目的是通过一系列既定的指标和细致的检查程序，确保数据集的质量满足预先设定的标准。数据集质量的验证指标广泛覆盖了包括数据量、多样性、代表性及偏见等多个维度。制定一个既合理又适应特定数据集需求的验证方案，是一个值得深入讨论的重要课题。关于数据集质量验证的具体方法，将在第三章中详细展开讨论。

预测结果评估则是提供一套适用于该数据集的评估方式。一个良好的预测结果评估系统，能够最大化地帮助LLM在该数据集上进行高效学习。关于预测结果评估的相关指标以及如何设计一套合适的评估系统，将在第四章中详细展开。

### 数据维护

当数据集建成之后，要保证对数据集的维护。对数据进行维护最好的办法是编写详细的数据集文档。一个详细的数据集文档，能够帮助研究者理解数据集的结构、来源、预处理步骤以及任何潜在的局限性。数据集的维护更新过程是一个动态的、持续的过程，特别对于动态变化的数据源，如：新闻平台、社交媒体平台、金融市场数据、在线论坛和问答网站等。定期对数据集进行更新维护，对于保持数据的相关性和时效性至关重要。在这一过程中，数据集文档发挥着关键作用。

有效的数据集文档才能帮助研究者进行数据维护工作。想要制作一个有效的数据集文档，应当遵循如图1所示的文档结构：

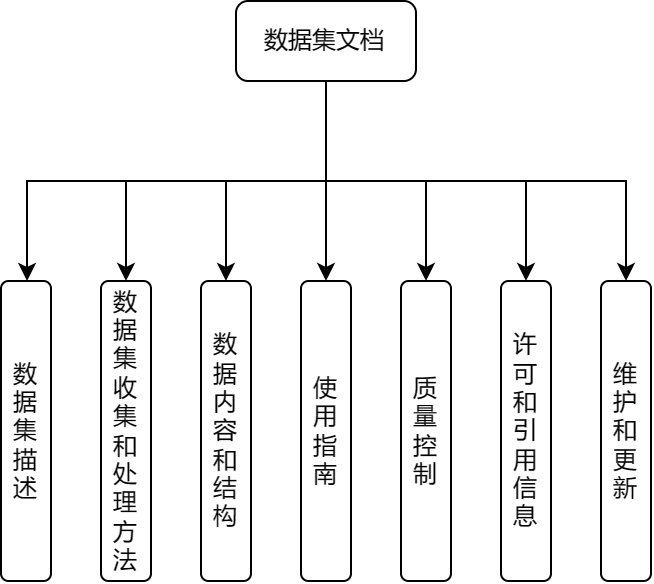


图 1 数据集文档结构

其中，数据集描述应包含概述和数据来源两个属性。数据集收集和处理方法部分，应写明数据的收集方法以及对原始数据进行清洗、筛选、转换等提高数据集质量的步骤。数据的内容和结构，应当描述数据的类型和格式、字段说明，同时还应该提供一定的样本数据，以便于更好地理解数据的结构和内容。数据的获取方式、加载和处理方法，应当在数据集的使用指南中标明。质量控制中，应包含有质量保证措施——数据验证方法等，和已知问题和限制。许可和引用信息包含有使用许可和引用信息两哥属性。维护和更新部分，则应当说明数据的维护情况以及后续更新计划等。

## 语料库设计

模型预训练的目的是利用大量的的数据资源，在没有或仅有少量目标任务数据的情况下，让模型学习到丰富知识，从而在特定任务上的表现更好。而模型预训练通常是在大规模的语料库上进行的。

### 什么是语料库

语料库是一种经过组织的大量的文本或口语数据集，常用于语言研究和自然语言处理（Natural language processing，以下简称NLP）等任务。其中LLM就属于自然语言处理的范畴。语料库的数据来源可以是书籍、报纸等实体语言材料，也可以是网站文章、线上百科全书等虚拟语言材料。语料库中的数据可以是未加工的原始数据，也可以是经过某种程度加工（如关键字标注、语法结构标注等）的数据。 在NLP领域，语料库是语言模型训练中，最基础也是最重要的数据资源。因此为了语言模型能够更好的从语料库中习得知识，语料库的建设要遵循一系列的标准，本论文将以化学语料库为例，展示如何构建一个合格的用于LLM预训练的语料库。

### 化学语料库数据收集

如前文所言，语料库数据可以来自数据和报纸，也可以来自网络文章和在线百科全书。为了在确保数据来源可靠的前提，同时提高数据收集的效率，本论文选取了维基百科（Wikipedia）作为数据源。每个维基百科的页面构成都包含有大量的锚文本（用于指向另一网页或网页内的一个位置），以此来链接各个与当前搜索关键字相关的网页同时维基百科中的数据不涉及敏感数据，因此，它是一个优质的数据来源。

为获取化学领域的专业知识语料库，我们首先在维基百科首页搜索“化学”关键字，可以得到如图2所示的相关页面。其中，蓝色文字为锚文本，这些锚文本大多也是化学领域的专有名词，且它们对应的网页是这些专有名词的对应解释网页。同时该“化学”关键字的维基百科页面，还有对化学领域的二级分类，如：有机化学、无机化学、物理化学、分析化学等，一共二十二个类别，它们同样也是以锚文本的形式在维基百科页面中呈现，因此该化学领域专业数据集的多样性得到一定的保证。

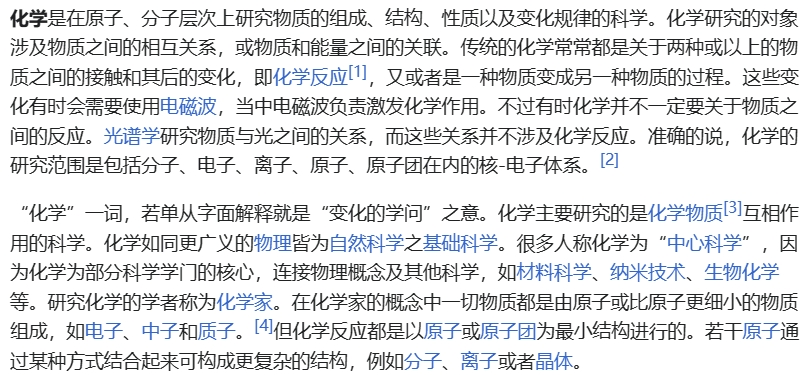


图 2 部分维基百科页面

在分析完页面数据构成后，可以以此来制定如下的数据收集方案：

1. 编写爬虫脚本，爬取“化学”维基百科页面中的锚文本及其包含的网页链接。
2. 进行锚文本的数据清洗，去除脏数据以及隐私数据。
3. 根据所爬取的文本——链接数据对，爬取锚文本对应的网页中，对该关键字的解释描述。
4. 进行所得语料库的数据清洗，去除脏数据以及隐私数据。

按照上述步骤进行第一步锚文本爬取，分析维基百科页面的HTML代码，可以确定锚文本及其链接在HTML代码中属于<a></a>标签中的内容。由于网页中存在无效链接，即链接对应网页为空，这类链接所在的a标签中包含有类名为new的css类，因此还需要在爬虫脚本中忽略对该类无效链接的爬取。

在第一次爬取后，可以得到1215对锚文本-链接对，如表1所示。经过步骤二数据清洗后，可以得到321对干净的锚文本-链接对。由于所得数据较少，因此本论文在第一次爬取数据的基础上进行了二轮迭代，得到26729对数据，清洗后共9639对数据。

|  |  |
| --- | --- |
| 锚文本 | 链接 |
| 化学元素 | https://zh.wikipedia.org/wiki/%E5%8C%96%E5%AD%B8%E5%85%83%E7%B4%A0 |
| 中子 | https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%AD%90中子 |
| …… | …… |

表 1 锚文本-链接对

### 数据清洗和格式处理

本节将讨论在上一节收集数据的过程中，如何进行数据清理。由于是对维基百科页面的锚文本进行爬取，因此不可避免会爬取下来非常多的脏数据，如表2所示。

|  |  |
| --- | --- |
| 锚文本 | 链接 |
| 编辑 | /w/index.php?title=%E7%94%9F%E7%89%A9%E6%AD%A3%E4%BA%A4%E5%8C%96%E5%AD%A6&action=edit&section=2 |
| [13] | #cite\_note-13 |
| Null | https://zh.wikipedia.org/wiki/File:Norbornenecycloaddnmech.png |
| …… | …… |

表 2 脏数据示例

数据清洗的标准为清除与化学领域无关的数据，对于特征明显的脏数据，比如冗余数据、链接格式不正确、非中文维基百科的网页链接、空数据和如“编辑”、[13]等字样的脏数据，可以直接通过编写正则匹配脚本来进行数据的清洗。但对于其他明显与化学领域无关但又无法总结出正则匹配规则的数据，只能通过人工进行数据筛选清洗。

值得一提的是，在第二次迭代爬取数据前，必须先进行一次数据清理，否则在第二次迭代爬取数据的过程中，爬取出来的脏数据会成倍的增长，加大了数据清洗的难度。最终两次迭代获得的数据量为9639。

在锚文本-链接对数据清理完毕后，下一步为通过这些锚文本-链接数据，根据其中的网页链接，爬取对应化学关键字的描述。在维基百科中，通过锚文本跳转的网页H1标题（HTML中的H1标题）即为该锚文本的内容，而该关键词的解释描述，即在该标签下，因此便可以设定对应的规则，进行数据爬取。爬取的数据如表3所示。之后进一步对获取的数据进行与锚文本-链接数据相同的数据清洗流程。即可获得8924条数据。asdads

|  |  |  |
| --- | --- | --- |
| 关键字 | 网址 | 解释文本 |
| 晶体 | https://zh.wikipe...a.org/wiki/...%E6%99%B6%E4%BD%93 | 晶体是原子、离子或分子按照一定...周期性，在结晶过程中，在空间排列形成具有一定规则的几何外形的固体。  晶体的分布非常广泛，自然界的固体物质中，绝大多数是晶体。气体、液体和非晶物质在一定的合适条件下也可以转变成晶体。  晶体内部原子或分子排列的三维空间周期性结构，是晶体最基本的、最本质的特征，并使晶体具有下面的通性： |
| 单质 | https://zh.wikipe... | 单质又称单纯物质、元素态物... |
| …… | …… | …… |

表 3 语料库数据示例

对于语料库的格式，本论文决定采用JSON格式进行保存。JSON格式相对于其他格式的文件，它具有轻量级、数据交换效率高、具有良好的跨平台兼容性等优点，因此JSON是存储语料库数据的理想选择。保存结果如图3所示，其中各字段分别为：id（序号）、url（内容所在网址）、title（化学专有词）、text（解释文本，其中不同文段用\n\n分割开）。

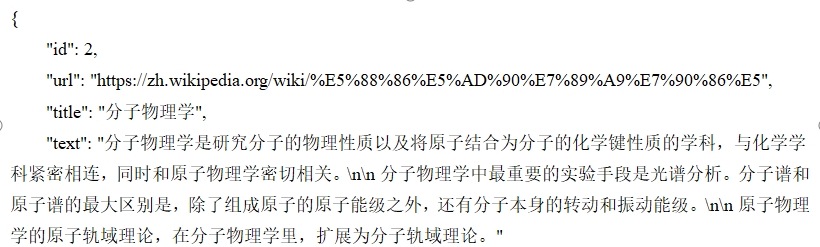


图 3 JSON格式存储示例

## QA数据集设计

模型微调，是指在一个已经在大规模数据集上预训练过的模型，进一步在特定任务的较小数据集上进行训练，以适应特定任务的需求。QA数据集通常包含一系列的问题以及答案，这些数据集可以有效的帮助模型学习理解问题的上下文意图，以及如何从给定的信息源中检索或推理出正确的答案。正因为QA数据集有着这些有点，因此QA数据集非常适合用于LLM的特定任务微调训练。

### QA数据收集

由于需要收集化学领域的专业知识问答数据，而维基百科中提供的更多为名词解释性数据，因此化学领域QA问答数据无法照搬上文中语料库的做法。权威的专业知识论坛中，有非常多高质量的问答记录，因此将论坛选为QA数据源。

目前比较流行的专业知识论坛网站有：知乎、CSDN、Stackoverflow和Quora等。其中，CSDN和Stackoverflow是专注于计算机领域的论坛，而知乎中的数据残次率又十分高，因此上述三个论坛网站并不适用于做数据集的数据来源。Quora论坛中，有明确分出不同的知识领域，其中就有化学的专业知识领域，且其中的问答数据均有实时更新，具有良好的时效性，同时高赞回答的数据，均有较高的正确率，因此对比之下，Quora论坛更适合作为数据源。

确定数据源之后，就应当考虑如何确定问答数据的范围和类别。在上文的语料库制作中，有提到维基百科上有对化学领域的二级分类，其中有4个大类，即有机化学、无机化学、物理化学、分析化学。至于其他领域，如：电化学、环境化学等，均为上述4个类别的衍生类。由于仅仅分为4个类别便开始收集数据的话，难免会造成一定的数据偏好，进而导致数据的多样性与覆盖范围下降，因此本文根据化学研究领域的子任务（源自维基百科），在4个主类下面各分出了25个子类别，如此一来，就可以确定数据选取的目标和范围了。

由于存在化学领域专业知识的限制，无法直接分辨出问题所在领域，因此对于问题的搜集，本文采用GPT根据上文定好的数据范围进行辅助问题生成，问题规模为每个子类下各25个问题，一共2500个问题。在确定问题后，便是答案搜集阶段，由于Quora论坛的反爬虫机制，导致无法直接使用爬虫自动获取答案，因此只能采取人工收集的方式进行数据收集。收集规则为，选取每个问题下的高赞且具有一定文本长度的回答（因为高赞回答一般即为正确有用的回答，对于数据正确性的研究，将于第三章进行探讨）。收集完的数据如表4、表5所示。

|  |  |
| --- | --- |
| 主类 | 子类 |
| 有机化学 | 合成有机化学 |
|  | 高分子化学 |
|  | …… |
| 无机化学 | 晶体化学 |
|  | 无机合成化学 |
|  | …… |
| …… | …… |

表 4 QA数据分类结构

|  |  |
| --- | --- |
| 问题 | 答案 |
| 如何设计一个高效的多步合成路线来合成复杂的天然产物？ | 分析目标分子结构，识别关键构建块和合成前体。确定合成……择高效反应类型和条件。利用文献调研和计算化学预测反应路径。采用串联反应或多组分反应减少步骤，提高收率。 |
| 对于具有多个手性中心的分子，如何实现立体选择性合成？ | 使用手性起始材料或辅助剂，选择立体选择性催化剂，采用……称合成策略。优化反应条件以控制立体化学，利用手性保护基团和解除策略。 |
| …… | …… |

表 5 问题数据示例

### QA数据处理

QA数据的清洗处理，于上文中语料库的处理方式大同小异，都是需要更具脏数据的特点来制定清洗规则。由于问答数据均为人工收集，因此不存在出现空数据、无关数据等脏数据。但可能会出现冗余数据的情况，以及答案文本中出现乱码字符等，因此对QA数据的清理主要在于乱码字符的清理和冗余数据的清理。相关清理代码如图4所示。



图 4 QA数据清洗部分代码

数据清洗完后的保存方式，参考上文语料库的保存方式，采用JSON格式进行保存，相关字段为：id（序号）、firsttitle（一级标题）、secondtitle（二级标题）、question（问题）、answer（答案）。保存后的结果如图5所示。

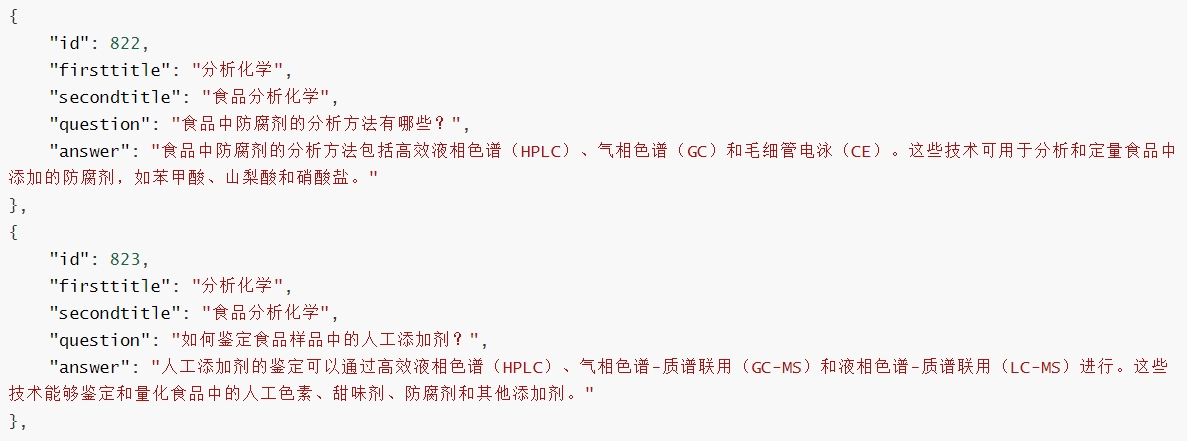


图 5 QA数据保存格式示例

# 数据集评估

当数据收集清洗完毕后，应当对数据集的质量进行评估，来判断该数据是否是合格的高质量数据。由于模型训练得到的预测结果受多方面要素影响，例如模型本身的质量高低、训练方案设计的合理与否等。因此，想要通过将数据处理完后给模型进行训练，再根据模型训练的预测结果来对数据集的质量进行判断是不切实际的。因此，需要再将数据投入使用前，对所收集处理后的数据进行一个质量评估，来判断该数据是否有一定的训练价值。

## 数据集质量评估方法综述

对一个数据集质量的评估，需要从数据质量、数据多样性、数据代表性、数据可解释性和透明度、数据合法性和伦理性、数据的可接入性等多方面进行评估。下面本文将一一介绍如何从这些角度进行数据集的质量评估。

对数据质量的评估，包含对数据体量、数据完整性、数据一致性、数据准确性和数据新鲜性的评估。数据量的大小，包含样本数据量和特征数据量，通常优质的大语言模型数据集都有着较大的体量，能为模型提供更多丰富的信息，但冗余的数据只会加大模型的处理时间和计算成本，因此在提高数据体量的同时，应当注意冗余数据的清洗。数据完整性即检查数据是否存在缺失值。数据一致性要求对数据格式、类型的处理要保持一致，例如日期格式、数字格式等。数据准确性即检查数据的准确度，是否存在错误或异常值。数据的新鲜性，则是要求数据要求一定的时效性，避免因过时数据而导致对模型的误导。

对数据多样性的评估，则是为了消除数据的偏见性，导致模型泛化能力低下。对数据的代表性评估，则是为了保证数据能够代表真实世界的分布，包括不同的群体、条件和情境，让模型能透过尽可能少的数据，学到尽可能多的知识。数据的可解释性和透明度的评估，要求对数据集的来源、收集方法和预处理的步骤做详细记录，如此一来对于理解数据的限制和潜在的偏见有很大作用。数据的合法性和伦理性则是对数据的收集和使用做是否符合法律法规、是否尊重个人隐私等相关方面的评估。数据的可接入性评估，则是要对数据的存储格式、大小和访问权限方面做出评估，以此来判断该数据集使用容易获取和使用。

## 数据集评估设计与结果

有了前文介绍的数据集评估指标，再根据第二章收集的化学语料库和QA数据集，便可以制定如下的数据集评估方案。

### 数据集评估方案

对于语料库的评估，将从数据体量、数据完整性、数据一致性、数据准确性、数据多样性等方面进行评估。对于QA数据集的评估，因为Quora论坛中问答数据涉及一定的时效性和隐私合法性，所以，除了对它做上述语料库需要做的评估，还需要对其时效性和合法性进行评估。

鉴于语料库有近9000条数据，QA数据集有2500对问答数据，而上述评估方式中，数据的一致性准确性和时效性指标都需要人工进行评估，因此为了提高评估效率的同时，还保证评估结果的可靠性，本文采用随机抽样评估的方式。随机抽样评估能在提高评估效率的同时，尽可能的提高评估结果的客观性、可靠性。随机抽样方式，将对语料库随机抽取100条数据，对QA数据随机抽取50个问题。对这些数据的准确性评估，将采用通过对比其他权威知识网站中的数据来判断合规与否。

### 数据集评估结果

# 预测结果评估

模型的预测结果是否正确需要有合理的预测结果评估指标，而合理的评估指标可以将模型的学习结果真实的反馈给模型，有利于模型的进一步学习改进。

## 模型预测结果评估指标综述

### BLEU与ROUGE

### F1分数与精确匹配（EM）

### SAS评估与余弦相似度匹配

## 预测结果评估方案设计

## 与人工评估和其他多种评估指标的对比

# 总结

## 研究结果总结

## 局限与未来展望