

## 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]** 随着 ChatGPT 的出现，大语言模型（LLM）成了当今最受关注的技术。构建或者微调 LLM 成为一个重要的发展方向。这其中需要准备高质量的训练数据集，同时还要评估 LLM 输出的结果数据。那么如何准备数据，评估输出结果成为一个实践和理论上需要讨论的问题。本文以构建化学领域数据集为例，给出了一些 LLM 数据集的收集方法、数据集的构建方法数据集的质量评估方法。探讨了包括 BLEU、ROUGE 等传统评估指标，引入了基于语义的评估方法如余弦相似度和 SAS。并且通过对比实验，指出了如何构建 LLM 输出结果的评估方案。

**[关键词]** 大语言模型；数据集设计；预测结果评估

## Professional data set design for LLM

**Abstract:** With the advent of ChatGPT, Large Language models (LLM) have become the most popular technology today. Building or fine-tuning LLMS has become an important development direction. This involves preparing high-quality training datasets and evaluating the resulting LLM output data. Then how to prepare the data and evaluate the output results becomes a practical and theoretical issue that needs to be discussed. In this paper, the methods of collecting LLM data sets, the methods of constructing data sets and the methods of evaluating the quality of data sets are given. Traditional evaluation indexes such as BLEU and ROUGE are discussed, and semantic-based evaluation methods such as cosine similarity and SAS are introduced. And through comparative experiments, it points out how to construct the evaluation scheme of LLM output results.

**Keywords:** Large Language Models; Dataset Design; Prediction Results Assessment

# 目 录

用于 LLM 的专业数据集设计 .....	2
1. 绪论 .....	6
1.1. 研究背景及动机 .....	6
1.2. 论文结构概述 .....	7
2. 数据集设计 .....	9
2.1. 数据集设计方法概述 .....	9
2.1.1. 数据收集处理 .....	9
2.1.2. 数据质量和预测结果评估 .....	10
2.1.3. 数据维护 .....	11
2.2. 语料库设计 .....	12
2.2.1. 什么是语料库 .....	13
2.2.2. 化学语料库数据收集 .....	13
2.2.3. 数据清洗和格式处理 .....	15
2.3. QA 数据集设计 .....	17
2.3.1. QA 数据收集 .....	17
2.3.2. QA 数据处理 .....	19
3. 数据集评估 .....	21
3.1. 数据集质量评估方法综述 .....	21
3.2. 数据集评估设计与结果 .....	22
3.2.1. 数据集评估方案 .....	22
3.2.2. 数据集评估结果 .....	23
4. 预测结果评估 .....	25
4.1. 模型预测结果评估指标综述 .....	25
4.1.1. 精确率、召回率和 F1 分数 .....	25
4.1.2. BLEU、ROUGE 和精确匹配（EM） .....	26
4.1.3. 余弦相似度匹配与 SAS 评估 .....	28
4.2. 预测结果评估方案设计 .....	29
4.3. 多种评估方案与人工评估的对比 .....	31
4.3.1. 各评估方案的评估结果以及分析 .....	32
4.3.2. 与人工评估的对比 .....	33
4.3.3. 评估结果总结 .....	36
5. 总结 .....	37

5.1. 研究结果总结 .....	37
5.2. 局限与未来展望 .....	38
致谢 .....	39
参考文献 .....	40

# 1. 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<sup>40[1]</sup>。

## 1.1. 研究背景及动机

早期对大语言模型的研究主要依赖于规划，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 引入了词嵌入技术，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<sup>[2]</sup>。

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<sup>[3]</sup>。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。因此本文将研究如何制作一个用于 LLM 的专业知识数据集。

## 1.2. 论文结构概述

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，对现有的常用预测指标进行了总结，设计了一系列用于评测在该化学数据集上适用的模型预测结果评估方

案。并藉由评估数据和人工评估的介入，来比较各评估方案与的优劣。

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，即数据集的制作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。



## 2. 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，为了直观地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）。

### 2.1. 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 and 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 [4]。

#### 2.1.1. 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地考量。同时，也需要对潜在的数据来源和获取方法进行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、删除无关值和剔除错误值。目的是除去数据中的噪声，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。

### 2.1.2. 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。

### 2.1.3. 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<sup>[5]</sup>。想要制作一个有效的数据集文档，应当遵循如图 2-1 所示的文档结构。

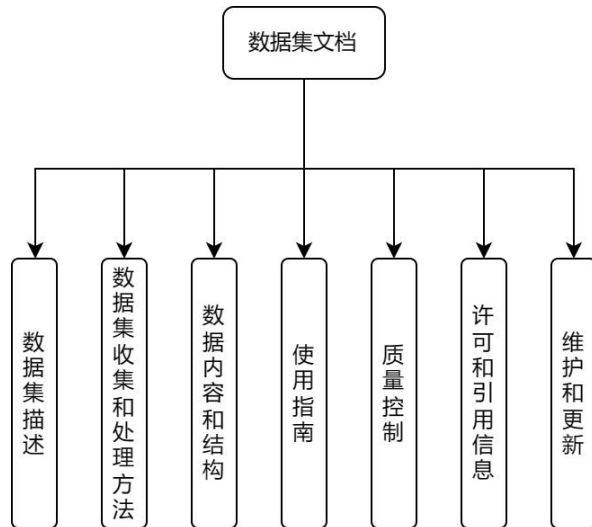


图 2-1 数据集文档结构

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个属性。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。由于本论文主要探讨数据集的收  
集和处理以及预测结果的评估，因此本论文将忽略数据集文档的制作。

## 2.2. 语料库设计

模型预训练的目的在于利用大量的数据资源，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<sup>[6]</sup>。

### 2.2.1. 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地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。

### 2.2.2. 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<sup>[7]</sup>。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 2-2 所示的相关页面。该“化学”关键字的维基百科页面，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。

**化学**是在原子、分子层次上研究物质的组成、结构、性质以及变化规律的科学。化学研究的对象涉及物质之间的相互关系，或物质和能量之间的关联。传统的化学常常都是关于两种或以上的物质之间的接触和其后的变化，即**化学反应**<sup>[1]</sup>，又或者是一种物质变成另一种物质的过程。这些变化有时会需要使用**电磁波**，当中电磁波负责激发化学作用。不过有时化学并不一定要关于物质之间的反应。**光谱学**研究物质与光之间的关系，而这些关系并不涉及化学反应。准确的说，化学的研究范围是包括分子、电子、离子、原子、原子团在内的核-电子体系。<sup>[2]</sup>

“化学”一词，若单从字面解释就是“变化的学问”之意。化学主要研究的是**化学物质**<sup>[3]</sup>互相作用的科学。化学如同更广义的**物理**皆为**自然科学之基础科学**。很多人称化学为“**中心科学**”，因为化学为部分科学学门的核心，连接物理概念及其他科学，如**材料科学**、**纳米技术**、**生物化学**等。研究化学的学者称为**化学家**。在化学家的概念中一切物质都是由原子或比原子更细小的物质组成，如**电子**、**中子**和**质子**。<sup>[4]</sup>但化学反应都是以**原子**或**原子团**为最小结构进行的。若干**原子**通过某种方式结合起来可构成更复杂的结构，例如**分子**、**离子**或者**晶体**。

图 2-2 部分维基百科页面，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。

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 2-1 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。

表 2-1 锚文本-链接对示例

锚文本	链接
化学元素	<a href="https://zh.wikipedia.org/wiki/%E5%8C%96%E5%A">https://zh.wikipedia.org/wiki/%E5%8C%96%E5%A</a>
中子	<a href="https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%A">https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%A</a>
氧化还原反应	<a href="https://zh.wikipedia.org/wiki/%E6%B0%A7%E5%8C">https://zh.wikipedia.org/wiki/%E6%B0%A7%E5%8C</a>

### 2.2.3. 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 2-2 所示。

表 2-2 脏数据示例

锚文本	链接
编辑	<a href="/w/index.php?title=%E7%94%9F%E7%89%A9%E6">/w/index.php?title=%E7%94%9F%E7%89%A9%E6</a>
[13]	#cite_note-13
Null	<a href="https://zh.wikipedia.org/wiki/File:Norbornenecycload">https://zh.wikipedia.org/wiki/File:Norbornenecycload</a>

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。但对于其他与化学领域无关但又无

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地增长，加大了数据清洗的难度。最终两次迭代获得的数据量为9639对。

表 2-3 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 2-3 所示，其中各字段分别为：**id**（序号）、**url**（内容所在网址）、**title**（化学专有名词）、**text**（解释文本，其中不同文



段用\n\n分割开）。

```
{
  "id": 2,
  "url": "https://zh.wikipedia.org/wiki/%E5%88%86%E5%AD%90%E7%89%A9%E7%90%86%E5",
  "title": "分子物理学",
  "text": "分子物理学是研究分子的物理性质以及将原子结合为分子的化学键性质的学科，与化学学科紧密相连，同时和原子物理学密切相关。分子物理学中最重要的实验手段是光谱分析。分子谱和原子谱的最大区别是，除了组成原子的原子能级之外，还有分子本身的转动和振动能级。原子物理学的原子轨道理论，在分子物理学里，扩展为分子轨道理论。"
```

图 2-3 JSON 格式存储示例

## 2.3. QA 数据集设计

模型微调，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<sup>[9]</sup>。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地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<sup>[10]</sup>。正因为 QA 数据集有着这些优点，因此 QA 数据集非常适合用于 LLM 的特定任务微调训练。

### 2.3.1. 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（如图 2-4 所示），如此一来，就可以确定数据选取的目标和范围。

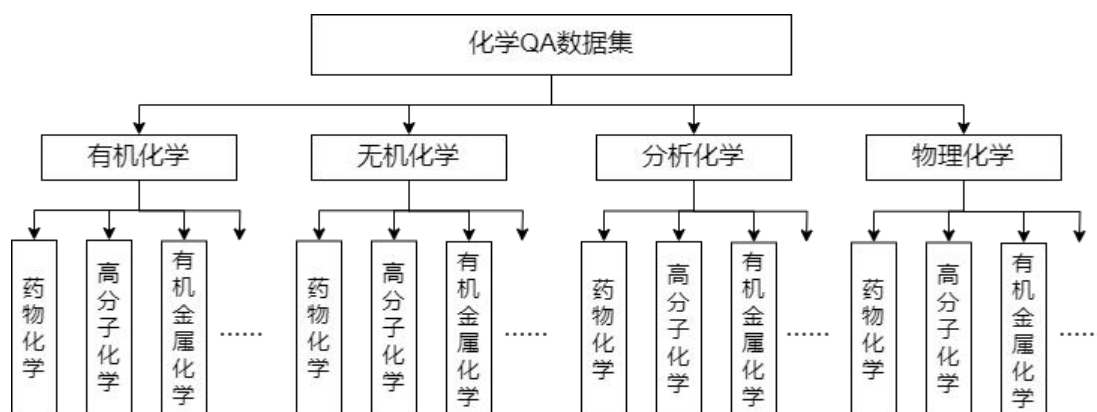


图 2-4 QA 数据集分类结构

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，因此对于问题的收集，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，便是答案收集阶段，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，部分收集结果如表 2-4 所示（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）。

表 2-4 问题数据示例

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...
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...
如何通过有机合成方法合成具有生物活性的多肽和蛋白质？	合成生物活性的多肽和蛋白质通常采用固相肽合成技术，通过逐...

2.3.2. QA 数据处理

QA 数据的清洗处理，与上文中语料库的处理方式大同小异，都是需要根据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，和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 2-5 所示。

```
def contains_chinese(s):  
    return any('\u4e00' <= c <= '\u9fff' for c in s)  
  
file_path = './Chemistry_QADATA.xlsx'  
data = pd.read_excel(file_path)  
  
# 删除答案列中的“*”，*被原始数据用来做加粗符号。  
data['答案'] = data['答案'].astype(str).str.replace('*', '', regex=False)  
  
# 删除答案列中的乱码数据以及无用字符  
data = data[data['问题'].apply(contains_chinese) & data['答案'].apply(contains_chinese)]
```

图 2-5 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 2-6 所示。

```
{  
  "id": 822,  
  "firsttitle": "分析化学",  
  "secondtitle": "食品分析化学",  
  "question": "食品中防腐剂的分析方法有哪些？",  
  "answer": "食品中防腐剂的分析方法包括高效液相色谱（HPLC）、气相色谱（GC）和毛细管电泳（CE）。这些技术可用于分析和定量食品中添加的防腐剂，如苯甲酸、山梨酸和硝酸盐。"  
},  
{  
  "id": 823,  
  "firsttitle": "分析化学",  
  "secondtitle": "食品分析化学",  
  "question": "如何鉴定食品样品中的人工添加剂？",  
  "answer": "人工添加剂的鉴定可以通过高效液相色谱（HPLC）、气相色谱-质谱联用（GC-MS）和液相色谱-质谱联用（LC-MS）进行。这些技术能够鉴定和量化食品中的人工色素、甜味剂、防腐剂和其他添加剂。"  
},
```

图 2-6 QA 数据保存格式示例

### 3. 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因素影响，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，需要在将数据投入使用前，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。

#### 3.1. 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<sup>[11-12]</sup>。

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。

## 3.2. 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。

### 3.2.1. 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，还要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，尽可能地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。

### 3.2.2. 数据集评估结果

语料库和 QA 数据集的数据量加起来有共 11424 条数据，且经过严格制定的清洗规则对数据进行清洗后，数据的一致性和完整性也有一定的保障，不存在格式不一致和脏数据清理不完全的情况。同语料库和 QA 数据集的知识构成不单单只有基础化学，对于语料库来说，两次数据迭代获取，让其有了广泛的化学知识积累；对于 QA 数据集来说，由有机化学、无机化学、物理化学和分析化学这四个主类为起点，并在这四个主类下依然分了 25 个子领域，因此 QA 数据有着一一定的全面性。而对于数据的时效性，维基百科和 Quora 论坛中的数据均是持续更新的，因此在时效性方面绝对有保障。对于数据的准确性，按照前文制定的检测方法，可以得到如表 3-1 的检测结果（答案正确性检测不仅局限于文本比较，而是将语义纳入检测范围）。其中错误数据的案例如图 3-1 所示。

表 3-1 数据准确率抽样检测结果

检测对象	检测数量	合格数量	合格率
语料库	100 条	100 条	96%
QA 数据集	50 对	48 对	100%

问题：

在有机金属化学中，过渡金属与碳之间的键有哪些类型？

QA数据集中答案：

过渡金属与碳之间的键包括 $\sigma$ 键和 $\pi$ 键。

维基百科中的答案：

在有机金属化学中，过渡金属与碳之间的键主要有三种类型：

$\sigma$ 键（**sigma**键）：这是最常见的键类型，类似于碳与碳之间的单键，其中过渡金属的 $d$ 轨道与碳的 $sp^2$ 或 $sp^3$ 杂化轨道形成共价 $\sigma$ 键。这种键在大多数有机金属化合物中都很常见。

$\pi$ 键（**pi**键）：这种键是由过渡金属 $d$ 轨道与碳的 $p$ 轨道重叠形成的。 $\pi$ 键通常是额外的键，可以增加化合物的稳定性，尤其在芳香化合物和烯烃配合物中很常见。

$\delta$ 键（**delta**键）：这是一种相对较弱的键，是过渡金属的 $d$ 轨道与碳的 $d$ 轨道之间的相互作用形成的。这种键在一些高度配合的过渡金属化合物中出现，通常对于反应活性和反应机理有重要影响。

图 3-1 QA 数据集中检测出的错误数据示例

综上所述，本文所做的语料库和 QA 数据集，在数据的完整性、一致性、多样性、准确性和时效性，都有着较高的质量，其中虽然 QA 数据集中存在着微量错误数据，但出现的错误也并不严重，只是缺失了一点信息。同时，为了保证数据集的质量，保持对数据集维护检查的习惯。



## 4. 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地反馈给模型，有利于模型的进一步学习改进。

### 4.1. 模型预测结果评估指标综述

大语言模型的预测结果评估指标有非常多种，按类别分，有用于分类任务的评估指标、用于生成任务的评估指标和用于语义和语法评估的指标等<sup>[13]</sup>。接下来将分别对上述三种评估指标进行举例介绍说明。

#### 4.1.1. 精确率、召回率和 F1 分数

精确率、召回率和 F1 分数，是用于分类任务模型的常见评估指标，特别是在二分类任务中。它们通过比较模型预测结果与实际标签之间的差异来对模型的准确性和完整性进行评估。

其中精确率是指模型在预测结果为正类的样本中，有多少是真正的正类。也就是说，精确率衡量的是模型在所有预测为正类的样本中，预测的准确性如何。精确率的计算方式如公式 4-1 所示，其中 TP (True Positives) 表示真正例，FP (False Positives) 表示假正例<sup>[14]</sup>。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4-1)$$

召回率是指模型所有正确预测为正类的样本占实际正类样本的比例。也就是说，召回率衡量的是模型找出所有正类样本的能力。召回率的计算方式如公式 4-2 所示，其中 TP 与精确率计算公式中的含义一样，FN (False Negatives) 表示假负例，即模型没有正确找出的正

类样例<sup>[14]</sup>。

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4-2)$$

F1 分数是对精确率和召回率的综合考量，因为其综合考虑了模型精确率和召回率，因此能够提供一个相对更加全面的评估模型性能的指标，其计算方式如公式 4-3 所示。但由于对不同的预测任务，对召回率和精确率的权重有时不同，因此在某些特定情况下，F1 分数由于其无法分辨精确率和召回率的差异的特点，而难以适用，因此可能需要结合其他指标来更全面地评估模型性能<sup>[15]</sup>。

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4-3)$$

#### 4.1.2. BLEU、ROUGE 和精确匹配（EM）

BLEU（Bilingual Evaluation Understudy Score）、ROUGE（Recall-Oriented Understudy for Gisting Evaluation）和精确匹配（Exact Match，简称 EM），常用于评估 NLP 任务性能的指标，但它们分别又适用于不同的任务场景。

BLEU 主要用于评估机器翻译系统质量，它主要通过比较机器生成的翻译与人类专家的参考翻译之间的相似度来量化翻译的准确性。其计算方式如公式 4-4 和公式 4-5 所示。其中对于公式 4 中，BP 是短文惩罚，用于惩罚生成文本过短的情况，它是参考文本与生成文本的长度之比的函数；N 是 n-gram 的最大长度； $P_n$  是 n-gram 的精度，他是生成文本中与参考文本匹配的 n-gram 数量与参考文本中 n-gram 的总数之比； $w_n$  是各个 n-gram 精度的权重，通常经过几何平均计算得到。对于公式 4-55 中，c 为生成文本，r 为参考文本，用于计算惩罚

因子 BP 的大小<sup>[16]</sup>。

$$\text{BLEU} = \text{BP} \times \exp(\sum_{n=1}^N w_n \cdot \log P_n) \quad (4-4)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \quad (4-5)$$

ROUGE 指标和 BLEU 指标适用任务类型相同，也是通过对生成文本和参考文本之间的重合度，来对生成文本进行质量评估。ROUGE 的具体计算公式因指标类型而异，主要有 ROUGE-N、ROOUGE-L 和 ROUGE-W。其中 ROUGE-N 主要关注 n-gram 的重叠度情况；ROUGE-L 关注最长公共子序列的长度，能够更好地反映生成文本和参考文本之间的整体结构相似性；ROUGE-W 通过引入词级别的权重，更加注重词级别的重叠度情况，能够更准确地反映系统生成文本与参考文本之间的语义相似度<sup>[17]</sup>。上述三种 ROUGE 指标计算方式分别如公式 4-6、公式 4-7 和公式 4-8 所示。其中公式 4-8 中，Weighted LCS 是生成文本与参考文本之间的加权最长公共子序列，Weight of all words in reference 是参考文本中所有词的总权重<sup>[18]</sup>。

$$\text{ROUGE} - \text{N} = \frac{\text{n-gram 重叠总数}}{\text{参考文本本地 n-gram 总数}} \quad (4-6)$$

$$\text{ROUGE} - \text{L} = \frac{\text{生成文本与参考文本的最长公共子序列长度}}{\text{参考文本的字数}} \quad (4-7)$$

$$\text{ROUGE} - \text{W} = \frac{\sum \text{Weighted LCS}}{\sum \text{Weight of all words in reference}} \quad (4-8)$$

精确匹配常在问答系统等任务中被使用。精确匹配指标计算简单，只需要比较生成文本与参考文本之间是否完全相同，若完全相同，才会给出满分，否则得分为 0<sup>[19]</sup>。精确匹配指标在一些要求精确匹配特定答案的情况下，比较适用。但是它也有一些限制，精确匹配只考虑答案文本是否完全一致，而不考虑答案的语义相似度或近似匹配的情

况，这可能会忽略了部分正确但不完全一致的答案，造成大语言模型的训练效率偏低。

#### 4.1.3. 余弦相似度匹配与 SAS 评估

上文提到的评估指标，都是基于字符匹配来进行模型预测结果评估，但是中文文本的意思不仅局限于文本，而是还有着一层语义，因此两个不相同的文本，它们的语义可能相同，因而能表达出相同的意思。所以对于中文 NLP 任务中需要考虑语义相似度的预测评估，上文中提到的预测结果评估方法便不完全适用，因此本论文将提出能考虑语义相似度的评估指标——余弦相似度匹配和 SAS。

余弦相似度匹配，适用于各类如：信息检索、推荐系统、文本分类等文本相似度度量的任务，它能够很好地捕捉文本之间语义的相似性<sup>[20]</sup>。其计算方式如公式 4-9 所示。其中，A、B 表示两个文本的词向量（通常是词袋模型或者词嵌入模型生成的表示），“·”表示向量的内积“ $\|A\|$ ”表示向量的模。

$$\text{similarity}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (4-9)$$

语义答案相似性（SAS）评估指标，旨在解决传统评估指标的局限性<sup>40</sup>。如精确匹配（EM）和 F1 分数，这些传统主要关注于问答（QA）模型中的词汇相似性，而 SAS 比较生成文本和预测答案的语义相似性。这种方法可以更准确地评估模型地实际性能，特别是在生成文本有相同意义但表述不同地情况下。SAS 的计算过程如图 4-1 所示。其中文本输入和预处理是将两个文本进行如分词、去除停用词等预处理；文本合并，是将处理后的两个文本通过一个特殊的分隔符连

接起来，通过特殊符号来告诉模型这两部分文本是不同的输入；跨编码器模型，是用以处理输入的文本，理解它们之间的语义联系，其中跨编码器模型通常基于 BERT 或其他 Transformer 架构；语义相似度评分，是模型对这两段输入文本的语义相似度进行评分，得分通常是 0 到 1 之间，分数越高，表示语义相似度越高<sup>[21]</sup>。

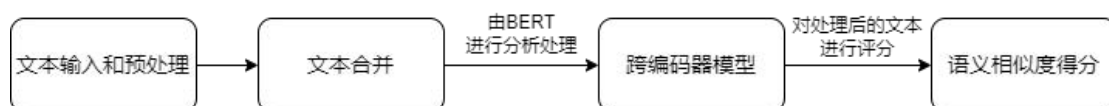


图 4-1 SAS 评估流程

## 4.2. 预测结果评估方案设计

经过上文对预测结果评估指标的概述，我们可以知道在 NLP 领域，对模型预测结果的评估大致分为基于文本比对的评估指标和对文本语义相似度进行评估的指标。对于评估方案的设计，应将具体的任务和数据的类型纳入考量范围。本文的数据集设计适用于 LLM 化学领域的预训练和微调，旨在经过训练，能够完成对 LLM 在化学领域的垂直应用开发。而化学领域的一些问答，可能会存在表述不同但意思相同的情况，因此预测结果评估指标应将语义相似度纳入考量范围。但倘若只考虑语义相似度会发生什么情况呢？图 4-2 展示的结果为 SAS 和余弦相似度评估，在对两个文本长度相差过大时的评估结果。

（其中 SAS 评估所用的跨编码器模型为 Huggingface.co 上公开的编码器模型，语义相似度评估所用的词向量嵌入模型也为 Huggingface.co 上公开的模型。）其中 sentence1 为参考文本，sentence2 为人为删减的生成文本，目的是检验在文本长度差异过大时，评测指标的评估效果如何。

```
sentence1 = "在无机光电材料中，载流子主要通过光激发过程产生。当光照射到材料上时，光子的能量被电子吸收，使电子从价带激发到导带，形成电子和空穴。这些电子和空穴是电荷载流子，它们在材料内部的传输过程受到材料的电子结构和缺陷状态的影响。有效的载流子传输需要良好的材料导电性和最小化的载流子复合，以提高光电转换效率。"  
sentence2 = "无机光电材料中的载流子主要通过光激发生成"  
  
result:  
SAS: 0.4993584156036377  
余弦相似度: 0.600745677947998
```

图 4-2 SAS 和余弦相似度对于长短差距过大的文本评估结果

```
sentence1 = "在无机光电材料中，载流子主要通过光激发过程产生。当光照射到材料上时，光子的能量被电子吸收，使电子从价带激发到导带，形成电子和空穴。这些电子和空穴是电荷载流子，它们在材料内部的传输过程受到材料的电子结构和缺陷状态的影响。有效的载流子传输需要良好的材料导电性和最小化的载流子复合，以提高光电转换效率。"  
sentence2 = "无机光电材料中的载流子主要通过光激发生成"  
  
ROUGE scores: {'rouge-1': {'r': 0.192307, 'p': 0.909090, 'f': 0.317460},  
                'rouge-2': {'r': 0.097560, 'p': 0.8, 'f': 0.173913},  
                'rouge-l': {'r': 0.173076, 'p': 0.8181812, 'f': 0.285714}}  
BLEU score: 0.0006823626951308315
```

图 4-3 ROUGE 和 BLEU 对于长短差距过大的文本评估结果，其中对于 ROUGE 指标中，r 代表召回率，p 代表精确率，f 代表 F1 分数。

从结果可以看出，对于生成文本与参考文本长度差异较大时，SAS 和余弦相似度的评估效果反而不理想，因此若只将语义相似度当作唯一的评估指标，反而会不利于模型对参考文本的学习。相比之下，BLEU 和 ROUGE 这两种基于文本匹配的指标在处理长短差异过大的时的效果反而更好，如图 4-3 所示，。

因此对于以本文所建立的数据集来进行 LLM 的预训练和微调，预测结果的评估方案应该将语义相似度和文本比较均纳入考量，其中为了找出较优的方案，本文举例出了 4 个方案，如下：

1. 单独采用 ROUGE 指标作为评估标准。
2. 单独采用 BLEU 指标作为评估标准。
3. 分别单独采用 SAS 和余弦相似度计算，作为评估标准。

4. 采用 BLEU 分别和 SAS、余弦相似度组合，作为评估标准。

其中方案 4，因为 BLEU 引入了短文惩罚机制，因此其评分占比因素应当考虑生成文本长短的情况。在生成文本长度与参考文本长度的差值，大于参考文本长度的三分之一时，BLEU 评估的占比为 70%；小于等于三分之一时，BLEU 的评估占比为 30%。

对于预测结果的评估方式，需要结合多种评估指标一起考虑，因此需要做大量的实验来检验评估方案是否合理，本文列举出上述 4 个方案，意在对各评估指标的效果进行比对，最优的方案并不局限于上述 4 个方案之中。

### 4.3. 多种评估方案与人工评估的对比

为了实现上述评估方案，本文采取随机抽样的方式，从 QA 数据集中，每个类别下各抽取一个 QA 对，共 100 对，来作为实验数据。预测答案来源于文心一言，其中为了展现过短的生成文本在上述各评测方案中的评估效果如何，部分文心一言的回答被限制在了 50 字左右（如图 4-4 所示）。其中 ROUGE 和 BLEU 的评估系统，均调用自 python 中的第三方轮子。余弦相似度计算所需的词嵌入模型，调用自 huggingface.co 中开源的中文 Sentence-BERT 模型。SAS 评估所需的跨编码器模型调用自 huggingface.co 中的 “sentence-transformers/paraphrase-xlm-r-multilingual-v1” 开源模型。

文心一言50字限制回答示例：

Q：配位键是如何形成的？

A（参考答案）：

配位键是由一个原子单方面提供一对电子与另一个接受电子的原子共用而形成共价键。具体而言，当共价键中共用的电子对由其中一个原子独自供应，而另一个原子提供空轨道时，就形成配位键。成键的两原子间共享的两个电子不是由两原子各提供一个，而是来自一个原子。

A（文心一言）：

配位键是一种化学键，由配体的孤对电子与中心金属离子的空轨道共享形成。这种键具有典型的共价键特性，但电子主要来源于配体。

文心一言不加限制的回答示例：

Q：什么是立体化学，它如何影响分子的化学性质？

A（参考答案）：立体化学是研究分子中原子的三维排列及其对化学反应性和物理性质的影响的科学。分子的空间构型可以影响分子间的相互作用、光学活性和反应途径，因此立体化学对药物设计、合成化学和生物分子的功能研究至关重要。分子的立体构型决定了其与其他分子相互作用的特异性，进而影响其化学反应的选择性和速率。

A（文心一言）：立体化学是研究分子三维结构及其与性质、反应行为关系的学科。分子中原子的空间排列方式不同，其化学性质也会不同。立体异构体、空间位阻效应和反应立体化学等因素均影响分子的反应速率、产物的构型和产率等。因此，立体化学对理解分子行为和化学反应至关重要。

图 4-4 文心一言预测结果示例

### 4.3.1. 各评估方案的评估结果以及分析

方案 1 至方案 4 的评估结果如表 4-1 所示。其中方案 1 的 ROUGE 指标包含 ROUGE-1.f、ROUGE-2.f 和 ROUGE-L.f 标准。

表 4-1 各评估方案的评分分布表，分为 4 个区间来展示评分的分布。

评估方案	评分 0-0.25	评分 0.25-0.5	评分 0.5-0.75	评分 0.75-1
方案 1 ROUGE-1.f	28.0%	13.0%	20.0%	39.0%
方案 1 ROUGE-2.f	27.0%	24.0%	25.0%	24.0%
方案 1 ROUGE-L.f	34.0%	16.0%	26.0%	24.0%
方案 2 BLEU	99.0%	1.0%	0.0%	0.0%
方案 3 余弦相似度	0.0%	68.0%	32.0%	0.0%
方案 3 SAS	1.0%	5.0%	55.0%	39.0%
方案 4 余弦相似度 &BLEU	15.0%	25.0%	60.0%	0.0%
方案 4 SAS&BLEU	21.0%	79.0%	0.0%	0.0%



从上述结果中可以看出，仅使用 BLEU 作为评估本文数据集进行预训练和微调的预测结果的指标是非常不合适的。因为在这次的随机抽样实验中，BLEU 无一例外给了不超过 0.5 的低分，甚至 99% 的预测结果给到了 0.25 分以下，如此严格地要求预测结果的文本与参考答案的文本一致，会导致模型的泛化能力下降，无法理解中文的语义，不利于模型的训练，仅适用于翻译任务和文本摘要等对文本还原要求高的任务。而其他三个方案如果只从评分分布的角度来看，难以区分优劣，因此还应当抽出其中具体的评估案例，加以人工比对，来判断哪一种评估方案更加优秀。

#### 4.3.2. 与人工评估的对比

对于某些对语义评估有要求的大语言模型，人工评估的优势在于，人能够更好地理解两个文本之间的语义关系，进而判断两个文本的相似程度如何。但人工评估的缺点也非常明显，就是效率低，且容易发生疲劳，进而导致评估精确度下降。

为了比对除方案 2 外，何种评估方式更加贴近于人工评估、更加优秀，下面将从各评估方案的实验中，抽出最高分的预测结果和最低分的预测结果，通过人工来进行判断其给分是否合理、是否更加贴近人工评估。其中 ROUGE 系列指标以 ROUGE-2 为例进行讨论。方案一、方案 3 和方案 4 的最高分最低分预测结果如图 4-5 至图 4-7 所示。

Highest rouge-2.f:

Score: 0.9829015904089162

Q: 光谱分析的基本原理是什么?

Reference A: 光谱分析基于物质对电磁辐射（包括可见光、紫外光、红外光等）的吸收、发射或散射特性。当物质与电磁辐射相互作用时，其内部电子、原子核或分子间的能级会发生变化，从而产生特定的光谱信号。通过测量这些信号，可以获得物质的组成、结构和其他特性的信息。例如，不同的化合物会吸收或发射不同波长的光，通过分析这些光谱数据，可以识别和量化样品中的各种成分。

Predicted A: 光谱分析的基本原理是利用物质与光的相互作用产生特定的光谱现象，通过对光谱的特征进行观察和分析，获得关于物质的物理性质和化学组成的重要信息。光谱分析包括吸收光谱分析、发射光谱分析和散射光谱分析等多种方法。

Lowest rouge-2.f:

Score: 0.01218642895691291

Q: 什么是密度泛函理论（DFT）?

Reference A: 密度泛函理论（DFT）是一种量子力学方法，用于研究多电子系统（如分子和凝聚态系统）的电子结构。它基于密度泛函假设，即系统的所有性质都可以通过其电子密度来描述。DFT将复杂的多体问题简化为以电子密度为主要变量的泛函问题，使得计算成为可能，广泛应用于物理、化学和材料科学领域。

Predicted A: 密度泛函理论（DFT）是一种基于量子力学的方法，用于计算原子、分子和固体电子结构。它将电子密度视为基本变量，通过最小化系统总能量得到基态电子密度分布，进而获得系统基态能量和其他性质。DFT在物理和化学领域具有广泛应用。

图 4-5 方案 1 的最高分和最低分

Highest SAS Similarity Score:

Score: 0.8829872608184814

Q: 水处理中的高级氧化过程（AOPs）是什么?

Reference A: 高级氧化过程（AOPs）是水处理技术，通过生成高反应性的氧化剂，如羟基自由基，来降解难以去除的有机污染物。这些氧化剂能有效分解污染物至无害物质。AOPs通常涉及紫外光、臭氧、过氧化氢等，用于处理工业废水、城市污水和饮用水，尤其针对持久性有机污染物。这种方法能显著提高水质，减少环境风险。

Predicted A: 高级氧化过程（AOPs）在水处理中是一种通过产生强氧化剂（如羟基自由基）来处理有机和有害污染物的技术。它利用这些强氧化剂与污染物发生反应，将其降解为无害物质，具有高效、环保的特点，广泛应用于废水处理和饮用水处理等领域。

Lowest SAS Similarity Score:

Score: 0.22662393748760223

Q: 无机化学中的数据挖掘技术有哪些，它们如何应用?

Reference A: 稀土基超导材料，如含稀土元素的铜氧化物高温超导体，目前是研究的热点。这些材料在相对较高的温度下表现出超导性，减少了冷却成本。研究正在寻求提高这些材料的临界温度和临界磁场，扩大其在电力输送、磁共振成像（MRI）和量子计算等领域的应用。

Predicted A: 无机化学中的数据挖掘技术包括聚类分析、关联规则挖掘、分类与回归等。这些技术可应用于化合物性质预测、反应路径优化及材料设计等领域，通过分析大量数据揭示规律，提高研发效率，推动无机化学领域的创新发展。

Highest Cross Encoder Similarity Score:

Score: 0.5086862444877625

Q: 离子迁移谱与质谱的区别和联系是什么?

Reference A: 离子迁移谱（IMS）和质谱（MS）都是基于离子检测的技术，但IMS通过离子在电场中的迁移速度进行分析，而MS则是通过测量离子的质荷比（m/z）。两者可以联用，IMS提供快速的离子分离能力，而MS提供精确的质量分析，二者联用可以提高分析的灵敏度和选择性。

Predicted A: 离子迁移谱与质谱在原理上有相似之处，都需先对目标物离子化并检测离子。然而，它们的离子分离原理不同：离子迁移谱利用离子的淌度分离，而质谱则基于离子的质荷比。此外，离子迁移谱在工作环境和检测速度等方面具有独特优势。两者可结合使用，提高分析的灵敏度和选择性。

Lowest Cross Encoder Similarity Score:

Score: 0.47932398319244385

Q: 如何利用动态共价化学构建自修复材料?

Reference A: 利用动态共价化学构建自修复材料，关键在于引入可逆共价键，使材料在受损时能够通过化学键的动态变化实现自我修复。这种方法不仅提高了材料的耐用性，还拓展了其在各种环境下的应用潜力。

Predicted A: 自修复材料可以通过动态共价化学反应实现，选用可逆的化学键如二硫键或烯烃亚胺键等，当材料损伤时，这些键可以断裂再连接，从而实现材料的自我修复。

图 4-6 方案 3 的最高分和最低分

Highest Cosine & BLEU Similarity Score:

Score: 0.6531316918525368

Q: 什么是密度泛函理论 (DFT) ?

Reference A: 密度泛函理论 (DFT) 是一种量子力学方法, 用于研究多电子系统 (如分子和凝聚态系统) 的电子结构。它基于密度泛函假设, 即系统的所有性质都可以通过其电子密度来描述。DFT将复杂的多体问题简化为以电子密度为主要变量的泛函问题, 使得计算成为可能, 广泛应用于物理、化学和材料科学领域。

Predicted A: 密度泛函理论 (DFT) 是一种基于量子力学的方法, 用于计算原子、分子和固体电子结构。它将电子密度视为基本变量, 通过最小化系统总能量得到基态电子密度分布, 进而获得系统基态能量和其他性质。DFT在物理和化学领域具有广泛应用。

Lowest Cosine & BLEU Similarity Score:

Score: 0.1566218723296275

Q: 环境有机化学中的生物富集是什么意思, 它如何影响生态系统?

Reference A: 生物富集是指在食物链中, 有机污染物在生物体内的浓度逐级增加的现象。这种现象通常发生在水生生态系统中, 其中污染物在水中的浓度相对较低, 但在食物链的高级消费者中可以达到更高的浓度。生物富集会导致捕食者, 特别是位于食物链顶端的生物, 如鸟类、鱼类和哺乳动物, 体内积累高浓度的有害物质, 对其生存和繁殖能力造成影响。

Predicted A: 生物富集在环境有机化学中, 指生物体从环境中吸收并积累低浓度有害物质的过程。这种积累可能破坏生态平衡, 影响其他生物健康, 甚至间接威胁人类健康。因此, 对生物富集的研究和控制对维护生态系统至关重要。

Highest SAS & BLEU Similarity Score:

Score: 0.43720187222426393

Q: 什么是光电化学分解?

Reference A: 光电化学分解是一种利用光能将水分解成氢气和氧气的技术。这一过程主要通过半导体材料制成的光电极在光照条件下实现。当光照射到光电极上时, 产生的光生电荷能够驱动水的氧化还原反应, 从而生成氢气和氧气。这种技术是将太阳能直接转化为化学能的有效方式, 被视为可再生能源技术中的一种重要发展方向。

Predicted A: 光电化学分解是一种利用光能将水分解成氢气和氧气的过程。在此过程中, 光能被吸收并激发出电子, 这些电子被输送到催化剂表面, 在外部电场的作用下, 电子传输到气体室中, 与水反应生成氢气和氧气。这种方法可用于生产清洁的氢能源。

Lowest SAS & BLEU Similarity Score:

Score: 0.16132763324754953

Q: 配位键是如何形成的?

Reference A: 配位键是由一个原子单方面提供一对电子与另一个接受电子的原子共用而形成共价键。具体而言, 当共价键中共用的电子对由其中一个原子独自供应, 而另一个原子提供空轨道时, 就形成配位键。成键的两原子间共享的两个电子不是由两原子各提供一个, 而是来自一个原子。

Predicted A: 配位键是一种化学键, 由配体的孤对电子与中心金属离子的空轨道共享形成。这种键具有典型的共价键特性, 但电子主要来源于配体。

图 4-7 方案 4 的最高分和最低分

从方案 1 开始分析。ROUGE-2 评出的最高分和最低分示例中, 可以看见, 最高分的预测结果在语义上虽然与参考文本十分相似, 但文本长度上却相差较多, 难免会失去一些信息导致意思表达得不够精准, 而最低分的预测结果虽然与参考文本在字符匹配上表现较差, 但该预测结果的语义却和参考文本表达得比较相似。因此可以确定, ROUGE-2 在预测结果评估上与人工评估相差较大。

而虽然方案 3 的最高分和最低分的预测结果与参考文本之间的语义匹配虽然较为准确, 但就如前文所提到的, 在处理文本长度相差较多的情况下, 依然会给出较高的评分。我们当然希望预测结果能够表达出与参考文本相同的意思的前提下, 还能尽可能地完善细节信息。因此方案 3 的评估效果虽然比方案一的结果要强一点, 但还是不够贴

近人工评估标准。

从方案 4 的评估的最高分和最低分的评估结果来看，即使预测结果表达的语义与参考文本相似，但如果预测结果的文本长度与参考结果相比过短，还是会被给出低分。因为加入了 BLEU 指标来对短文本进行惩罚，因此在评估效果方面，会比较贴近人工评估的标准。

#### **4.3.3. 评估结果总结**

从上文的评估结果中我们可以知道，方案 4 是最适用于对在该数据集上的预测结果进行评估的方案。同时也可以直观地体会到，仅仅使用单个评估指标来作为预测结果评估方案是不够全面的，无法贴合人工评估的标准。因此，在制定评估方案的时候，应当综合考虑数据集的情况与模型训练的要求，综合多个指标进行评估方案的设计，才能够设计出有利于大语言模型训练的评估方案，加强大语言模型的性能。

## 5. 总结

若想用自己的数据去训练能完成特定任务的大语言模型，学会如何制作一个能用于大语言模型训练的数据集是必不可少的。本文对大语言模型数据集的设计做了总结，同时通过具体实例来辅助说明了数据集制作的具体流程。

### 5.1. 研究结果总结

在第二章，本文对数据集的制作流程做了总结，同时完成了用于化学领域大语言模型训练的语料库和 QA 数据集；在第三章，本文总结了评估一个数据集质量高低的指标有哪些，要从哪些方面去看待一个数据集是否是高质量数据集，同时还对第二章做的数据集进行了质量评估，以实操来体现如何对数据集质量进行评估；第四章，本文对现有的常用预测结果评估指标进行了总结，并根据第二章所制作的数据集的特点，制定了一系列预测结果评估的方案。通过这些评估方案对预测结果的评估结果对比，并在人工评估的角度下，对这些方案进行了对比，能更加直观地看出各种评估指标的特点和优劣。同时指出预测结果评估方案的设计，特别是对于中文预测数据的评估，不能单单只将一种评估指标作为评估标准，而是应该针对不同的情况，综合使用不同的评测指标才能更好地完成预测结果评估任务。

本文最终设计了一个用于化学领域 LLM 训练的数据集，其中包含一个有 8924 条数据的语料库和一个有 2500 对 QA 问答数据的 QA 数据集。同时还设计了一套用于评估在该数据集上进行结果预测的评估指标，并通过实例证明了方案的可用性。

## 5.2. 局限与未来展望

本文对于用于 LLM 的专业数据集设计的局限性,在于对化学专业知识了解的限制,导致对数据进行人工筛检以及对数据集质量进行评估的时候,存在专业方面的局限性,导致数据筛检以及质量评估的结果也失去了一定的可信度。同时对于预测结果评估方案的设计,只设计了 4 个方案进行对比,没有深入探讨混合指标方案中,各个指标的占比究竟分别占多少才能达到最优效果。而且在引入人工评估对比的时候,也会由于专业知识的限制,导致对预测结果的判断会掺入一定的主观意识。

虽然本文的研究存在着如上的局限性。但本文的研究目的在于对目前的数据集制作方法做一个综述性文章,并通过实际数据集制作的实验来佐证本文总结的数据集制作路线是切实可行的。在大语言模型高速发展的现在,对高质量数据集的要求也越来越多,本文相信,在日后的大语言模型数据集研究领域,将会有更多更成熟的研究成果面世。希望本文能对数据集制作研究提供一定的借鉴。

## 致谢

四年的读书生活在这个季节即将划上一个句号，而于我的人生却只是一个逗号，我将面对又一次征程地开始。在论文即将付梓之际，思绪万千，心情久久不能平静，四年的大学生涯转瞬即逝，回顾这四年的大学生活，我得到了许多人的帮助，倘若没有他们的帮助，我的大学生活不可能如此的顺利、丰富。在此，我有许多感谢的话，写给曾今帮助过我的人。

感谢我的父母，养育之恩，无以回报，你们永远健康快乐是我最大的心愿。

感谢我的导师——张庆丰老师。当我在科研和学习生活中遇到困难时，总会给予我不可或缺的帮助，助我解决困难，继续前行。

其次，我还要谢谢我身边所有的朋友和同学，你们是我大学生活中最重要的组成部分。在这几年的大学生活里，不管是学习还是生活，你们都给了我最多的帮助和支持。

在论文即将完成之际，我的心情无法平静，从开始进入课题到论文的顺利完成，有多少可敬的师长、同学、朋友给了我无言的帮助，在这里请接受我诚挚谢意！

最后再一次感谢张庆丰老师，以及其他帮助过我的同学们。

## 参考文献

- [1] 刘学博,户保田,陈科海 & 张民.(2023).大模型关键技术与未来发展方向——从 ChatGPT 谈起. 中国科学基金 (05),758-766. doi:10.16262/j.cnki.1000-8217.20231026.004.
- [2] Singh, Y., Bhatia, P. K., & Sangwan, O. (2007). A review of studies on machine learning techniques. *International Journal of Computer Science and Security*, 1(1), 70-84.
- [3] Becker, R., Gilz, L., & Shrestha, M. (2021). Development of a Language model for the medical domain (Doctoral dissertation).
- [4] Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefèvre, B., & Soumarmon, T. (2020, October). Ensuring dataset quality for machine learning certification. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)* (pp. 275-282). IEEE.
- [5] Combi, C., & Shahr, Y. (1997). Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Computers in biology and medicine*, 27(5), 353-368.
- [6] Edwards, A., Camacho-Collados, J., De Ribaupierre, H., & Preece, A. (2020, December). Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th international conference on computational linguistics* (pp. 5522-5529).



- [7] Zhu, H., Tiwari, P., Ghoneim, A., & Hossain, M. S. (2021). A collaborative AI-enabled pretrained language model for AIoT domain question answering. *IEEE Transactions on Industrial Informatics*, 18(5), 3387-3396.
- [8] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data* (pp. 2201-2206).
- [9] Rogers, A., Gardner, M., & Augenstein, I. (2023). Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10), 1-45.
- [10] Zhao, H., Ling, Q., Pan, Y., Zhong, T., Hu, J. Y., Yao, J., ... & Shao, Y. (2023). Ophtha-llama2: A large language model for ophthalmology. *arXiv preprint arXiv:2312.04906*.
- [11] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1), 63-93.
- [12] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3), 1-52.
- [13] Li, Z., Xu, X., Shen, T., Xu, C., Gu, J. C., & Tao, C. (2024). Leveraging large language models for nlg evaluation: A survey. *arXiv*

preprint arXiv:2401.07103.

- [14] Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1), 12-19.
- [15] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).
- [16] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
- [17] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).
- [18] Barbella, M., & Tortora, G. (2022). Rouge metric evaluation for text summarization techniques. Available at SSRN 4120317.
- [19] Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *The Stata Journal*, 9(4), 524-546.
- [20] Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012, October). Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST* (Vol. 4, No. 1,

p. 1). South Korea: University of Seoul.

- [21] Risch, J., Möller, T., Gutsch, J., & Pietsch, M. Semantic answer similarity for evaluating question answering models. 2021.