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：

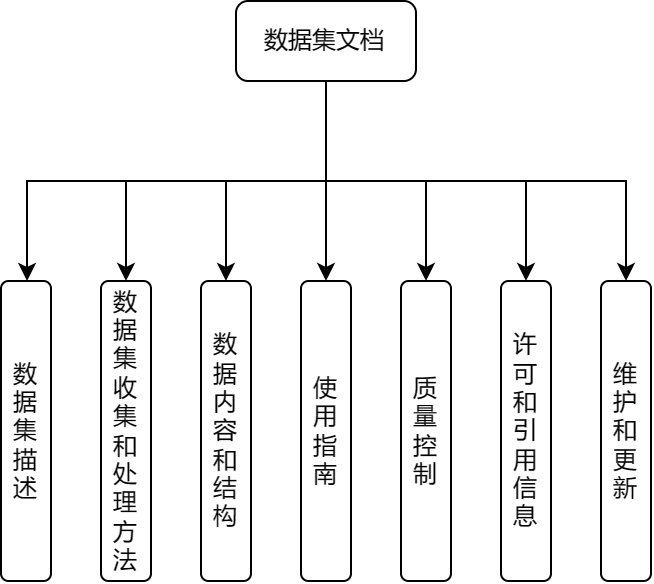


图 1 数据集文档结构

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

### 什么是语料库

语料库是一种语言资源

## QA数据集设计