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。

文章将从三个部分进行专业领域数据集制作的介绍。第一部分，即文章的第二章节，将介绍数据集制作的一些基本方式，并且以化学领域专业数据集制作作为实例。第二部分，即文章的第三章节，将介绍一些数据集质量评估的常用指标，然后再根据前面的实例，制定一套评估标准来评测数据集质量的好坏。第三部分，即文章的第四章节，将对现有的模型预测结果评估指标进行概述总结，然后再根据上文的实例，制定一套合理的预测结果评估规则，来对预测结果进行评估，同时与其他指标进行对比分析。

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。