内蒙古師範大學

硕士研究生学位论文

基于大语言模型的粉笔字规范性书写对话系统研究

Research on Normative Chalk Writing in Dialogue Systems Based on Large Language Models

姓　　名：陈二开

学　　号：20214021012

培养单位：计算机科学技术学院

专业名称：计算机科学与技术

研究方向：自然语言处理

导师姓名：李成城

二〇二四年六月

基于大语言模型的粉笔字规范性书写对话系统研究

中 文 摘 要

师范生对粉笔字练习及获取专业指导有着巨大需求，急需计算机辅助系统对粉笔字练习进行高效率的自动比对、评判、指导。现有的师范生粉笔字书法训练自动评判系统通过计算学生书写粉笔字和教师书写粉笔字特征之间的差异实现自动比对，并使用数据到文本生成技术完成学生书写粉笔字的自动评价。然而目前的评价文本是直接依据数据给予使用者答复，即机械性的回复，没有交流，交互性不强。为进一步提升系统的交互能力，本研究提出了一种基于大语言模型的粉笔字规范性书写对话系统，根据学生的反馈进行个性化的指导，提供更具体的建议，旨在辅助使用者进行粉笔字书写训练。具体工作如下：

（1）粉笔字规范性书写对话数据集构建。首先对粉笔字字帖字典信息库进行整理，形成问答对的结构，其中每个问题对应字典中的一个具体特征，而回答则以自然语言描述该特征的坐标或数字信息。同时将一本粉笔字规范书写教材中的知识点进行整理，以进一步丰富数据集。最后制作了一款 ChatGLM 微调数据集生成工具,以便将数据集整理成大模型微调时指定的数据格式。

（2）在构建的数据集上对开源大语言模型 ChatGLM2-6B 进行参数高效微调训练。在微调时使用了一种更加充分和高效的多轮对话训练方法，并引入了 Prefix-LoRA 联合微调策略。实验结果表明，与原始微调方法相比，BLEU、ROUGE 评分分别提升了 0.07、0.06。为评估微调后的大语言模型对领域知识的“记忆”能力，设计了一种新的评估方式，评估结果表明，微调后的大模型基本掌握了粉笔字规范性书写相关的理论知识和书写技巧。

（3）融合外部知识的检索增强生成。为了解决大语言模型在处理超出训练数据范围或需要最新数据时出现的知识“幻觉”问题，采用了检索增强生成（RAG）技术，此技术结合了检索和生成的优势，可以利用外部知识库回答问题。针对经典 RAG 在构建粉笔字规范性书写对话系统时的缺陷，本研究引入混合检索和重排序策略对经典 RAG 进行改进。使

用命中率（Hit Rate）和平均倒数排名（MRR）来衡量改进 RAG 的检索质量，同时采用开源 RAGAS 框架评估改进 RAG 的生成质量。实验结果表明，改进的 RAG 显著提升了回答问题的准确性和相关性。

关键词：对话系统，ChatGLM，参数高效微调，多轮对话，检索增强生成

Research on Normative Chalk Writing Dialogue System Based on Large Language Models

ABSTRACT

Normal students have a huge demand for chalk writing practice and professional guidance, and urgently need a computer-aided system for high-efficiency automatic comparison, evaluation, and guidance of chalk writing practice. The existing automatic evaluation system for student teachers' chalk calligraphy training achieves automatic comparison by calculating the differences between the characteristics of students' and teachers' chalk writing, and uses data-to-text generation technology to complete the automatic evaluation of students' chalk writing. However, the current evaluation text is a direct reply based on data, that is, a mechanical reply without communication and weak interaction. In order to further improve the interaction of the system, The dialogue system, based on a large language model for standard chalk writing, which provides personalized guidance and more specific suggestions according to students' feedback, aiming to assist users in chalk writing training. The specific work is as follows:

(1)Construction of chalk standard writing question and answer dataset. Firstly, the chalk character model dictionary information database is sorted out to form a question and answer pair structure, in which each question corresponds to a specific feature in the dictionary, and the answer describes the coordinate or numerical information of the feature in natural language. and the knowledge points in a standard chalk writing teaching material are sorted out to further enrich the dataset. Finally, ChatGLM fine-tuning dataset generation tool is developed to organize the dataset into the data format specified by the large model fine-tuning.

(2)Parameter efficient fine-tuning training of the open source large language model ChatGLM2 on the constructed dataset. A more comprehensive and efficient multi-turn dialogue training method than the official one is used

during fine-tuning, and the Prefix-LoRA joint fine-tuning strategy is introduced. Experimental results show that compared with the original method, BLEU and ROUGE have increased by 0.07 and 0.06 respectively. In order to evaluate the “memory” ability of the fine-tuned large language model, a new evaluation method is designed, and the factors affecting the recall rate are analyzed with specific question and answer examples. The evaluation results show that the fine-tuned model has basically mastered the theoretical knowledge and writing skills related to standard chalk writing.

(3)Integration of external knowledge retrieval enhanced generation. In order to solve the knowledge illusion problem of large language models when dealing with situations beyond the training data range or requiring the latest data, retrieval enhanced generation (RAG) technology is adopted. In view of the limitations of traditional RAG methods, mixed retrieval and re-ranking are used to improve it, so as to improve the quality and relevance of the retrieved information. RAG combines the advantages of retrieval and generation, so hit rate (Hit Rate) and mean reciprocal rank (MRR) are used to measure the retrieval quality of RAG, and the open source RAGAS as framework is used to evaluate the generation quality of RAG. Experimental results show that the improved RAG significantly improves the accuracy and relevance of answers.

KEY WORDS： Dialogue System, ChatGLM, Efficient Parameter Fine-Tuning, Multi-Turn Dialogue, RAG

目　录

第 1 章　绪论

1.1 研究背景及意义

粉笔字作为教师的一项基本功，是师范生必备的基本素养和核心技能。师范生对粉笔字练习和获取专业指导有着巨大需求，但高校在粉笔字书法教育面临诸多挑战，例如书法教师资源稀缺，教学配套设施不足，缺乏系统的书法课程等[1]。因此，迫切需要一种高效的计算机辅助系统，能够对粉笔字练习进行高效率的自动比对、评判、指导[2]。课题组研发的师范生粉笔字书法训练自动评判系统经过发展[3]，已经完成了自动比对、自动评分和自动评判。

为构建粉笔字规范书写自动评分模型，李泽瑶[4]从粉笔字书写规范研究入手，提炼影响粉笔字评分的特征，并构建了粉笔字规范书写模板字典，通过计算学生书写粉笔字和教师书写粉笔字特征之间的差异实现自动比对，并利用深度学习技术构建了自动评分模型。

如图 1-1 所示，构建粉笔字模板字典时首先对教师书写粉笔字进行图像预处理，然后进行书写特征提取[5]。整字的特征主要有汉字空间位置，汉字大小、汉字宽高比等，部件的特征主要有部件空间位置、部件大小、部件宽高比等，笔画的特征主要有笔画的空间位置、笔画长度、笔画斜率等。这些特征信息经过量化处理后，存储于 JSON 文件中形成完整的特征描述方式，模板字典的构建为粉笔字规范书写的自动评分和自动评价提供了可靠依据。

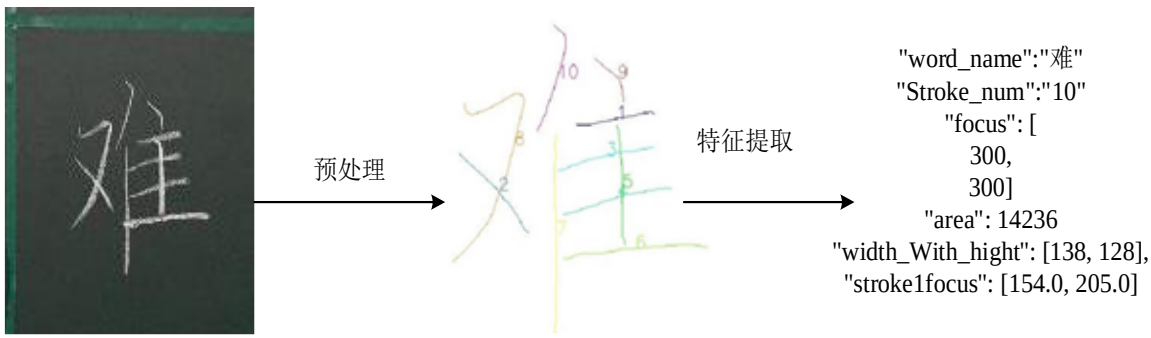

图 1-1 粉笔字模板字典构建过程

提取到学生书写粉笔字的全局特征与局部特征后，范勇峰[6]对提取到的数值型特征与坐标型特征进行转换，结合专家意见与专家字典对特征数据进行推理，完成数值型特征与坐标型特征到文字形式的汉字分级特征，并在此基础上构建了粉笔字规范书写评价数据集，使用基于内容规划的序列到序列文本生成模型，利用手写汉字特征偏差值生成评价文本，如图 1-2 所示：

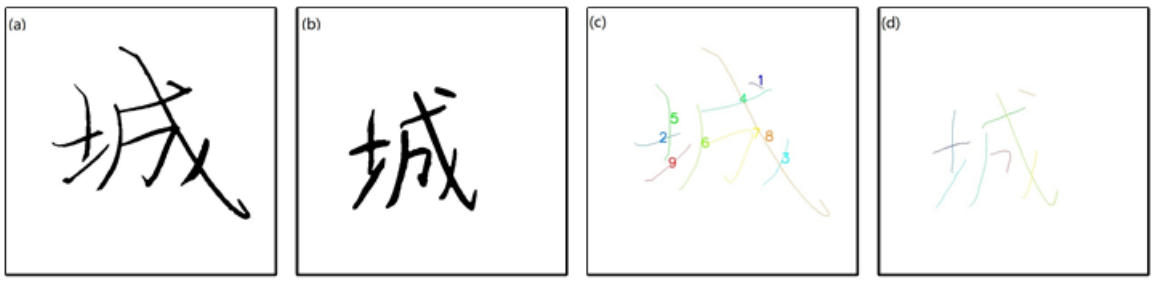

图 1-2 示例图

图 1-2 中，(a)为书法教师书写汉字图像，(b)为学生书写汉字图像，(c)为书法教师书写汉字骨架图，(d)为学生书写汉字骨架图。生成的评价文本中所涉及到的笔画顺序问题依据图 1-2(c)中所标记的笔画顺序为参照（笔画上有对应颜色的数字标注），根据特征提取与特征推理后的汉字分级数据，生成的评价文本如图 1-3 所示：

特征及偏差值	生成文本
笔画1重心：偏上；笔画2长度：偏短；笔画9走势：严重不一致；整字匀称性不匀称；笔画3拐点：略内陷；笔画9斜率：严重倾斜；整字高矮：略矮；笔画8拐点：略内陷；笔画8长度：非常严重偏短；笔画4拐点：略突出；笔画5走势：稍不一致；整字重心：偏下；笔画7走势：严重不一致；整字胖瘦：略瘦；整字面积：严重偏小。	首先从汉字整体结构上看，汉字重心存在着偏下的问题，汉字的大小相较于标准汉字严重偏小，整体在高度上略矮，在宽度上略瘦，且整体的分布较不匀称。然后从笔画层面上看，与标准笔画相比较时，笔画1的重心偏上。笔画2的长度偏短。笔画3的拐点部分存在略内陷的问题。笔画4的拐点部分存在略凸出的问题。笔画5的走势稍不一致。笔画7的走势严重不一致。笔画8的拐点存在略内陷的问题，笔画长度非常严重偏短。笔画9的走势严重不一致，倾斜度上出现了严重倾斜。对以上存在的问题，仍需要您进一步的改进。

图 1-3 评价文本

然而，目前的粉笔字书法训练自动评判系统在提供评价时，只是机械的依据特征偏差生成反馈，这种方式的互动性不足，无法有效的与学生沟通[7]。例如，系统可能会用“重心偏上”或“长度非常偏短”等术语来描述书写问题，这些术语对学生而言

不够具体和直观。此外学生在练习粉笔字时，通常是以一首诗为单元进行练习，如果直接将学生书写的特征信息和粉笔字模板字典信息全部输出，让学生自行对比，这样的信息量对于学生来说是庞大且难以吸收的。因此，在对学生进行规范书写指导时，系统需要针对性的反馈学生书写粉笔字的具体特征信息如坐标或数字，并与粉笔字模板字典中的相关数据进行对比分析，提供具体而精准的指导。除此之外，当前的粉笔字书法训练自动评判系统还缺少关于粉笔字规范书写的基础理论和技巧知识，为了进一步提升指导效果，有必要整合这些理论和技巧知识，确保系统能够基于书法专业知识，通过对话的方式为学生提供更加精确和实用的评价与建议。

2022 年底 OpenAI 发布了对话式 AI 新模型 ChatGPT[8]，它能够进行更为智能化的问答体验，这一里程碑式的技术激发了各行各业对大语言模型在特定领域的深入探索。垂直领域大语言模型[9]是指在特定的领域或行业中经过训练和优化的大型语言模型。与广泛覆盖多领域的通用大语言模型相比，垂直领域大模型更专注于特定领域的知识和技能，因此它们在专业性和实用性方面表现的更出色。

综上所述，鉴于大语言模型在现实场景中的应用潜力，构建基于大语言模型的粉笔字规范性书写对话系统十分有必要。使用微调让大语言模型学习到粉笔字规范性书写的相关知识，让其具备“指导教师”的功能。通过外接本地数据库，将学生书写的详细特征信息引入到大语言模型中，扩展其上下文能力，帮助大语言模型理解学生的问题并给出更加准确和专业的回答，辅助学生更好的掌握粉笔字书写技巧，完成正确而又标准的书写。

1.2 国内外研究现状

1.2.1 对话系统研究现状

对话系统是一种能够理解人类语言意图并与人类形成连贯对话的计算机系统。其研究历史可以追溯到 1966 年，Joseph Weizenbaum 开发了模拟心理治疗师的聊天机器人 ELIZA Chatbot。此类对话系统主要采用符号和规则的方法[10]，通过专家手工编写的规则模板来运作。虽然这种方法搭建简单、具有较好的解释性和修复性，但其人工成本高、鲁棒性和泛化性较差，仅适用于解决预先设定在模板范围内的简单问题。

在 1990 年代至 2000 年代，随着计算能力的增强，研究者开始采用统计机器学习方法构建对话系统，这类对话系统利用大规模语料进行训练，依赖统计模型生成回

复，如基于 n-gram 模型和统计分类器的对话系统。基于统计机器学习的对话系统优点在于相对简单易实现，在特定领域有一定的性能表现[11]。然而，这种方法对于复杂的语境理解能力较弱，往往无法处理含义模糊或语义复杂的对话。此外，这种方法通常需要大量的人工特征工程，对数据质量要求较高。

自 2000 年代以来，随着深度学习和自然语言处理技术的飞速发展，对话系统的研究进入了新的时代。研究者们开始整合规则、统计方法和深度学习技术，以增强对话系统的灵活性和准确性。与此同时，大规模多方对话数据被广泛用作训练语料，并运用深度学习算法学习多方对话模式。深度学习算法的引入使得对话系统在对话学习方面有了更深层次的记忆形式，显著提升了对语言模型中长文本序列的建模能力。基于深度学习的对话系统可以根据是否在输入中利用历史对话信息分为两大类：单轮对话系统和多轮对话系统。此外，根据对话系统的构建方式，可以被进一步细分为检索式、生成式以及融合检索式和生成式的混合方法[12]。具体的分类可参见图 1-4：

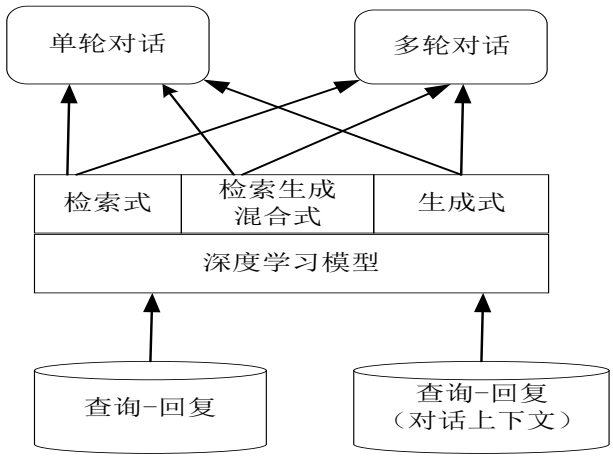

图 1-4 基于深度学习的对话系统

为提高检索效率，大多数检索式对话系统采用了两次排序的策略[13]。主要流程如图 1-5 所示：

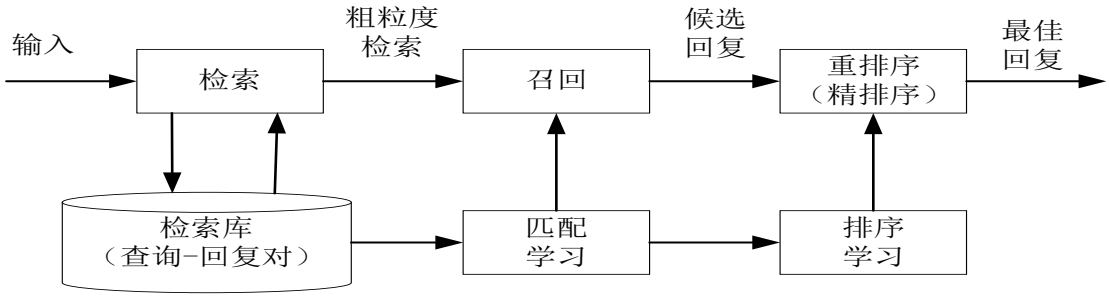

图 1-5 基于检索式方法的对话系统示意图

首先整理好标准的问题-答案对，将其存放在检索库中。当系统收到用户的查询请求时，根据用户的查询与检索库中问题-答案对的匹配程度，召回一批候选回复列表。召回阶段可以采用 TF-IDF、BM25 等经典检索算法[14]，这类基于统计的检索方法相对简单易行，能够快速处理大量文本数据而无需进行复杂的模型调参。接着，在第二阶段的精排中，系统采用更复杂的排序算法或技术，如 BERT 神经网络模型，通过对候选回复进行语义匹配，确保选出最合适的回复。

基于检索式方法的对话系统的核心是利用匹配模型来减少用户提问和回答之间的语义差距。然而，该类对话系统的性能很大程度上依赖于检索库的规模和质量，使用时可能无法准确反映用户的查询意图或提供与上下文匹配的信息，导致系统生成的回复与用户的实际查询内容不符。

生成式方法是目前倍受欢迎的对话系统构建方式[15]，其不再受限于预定义的问题-答案对，核心理念是通过大规模的问答语料训练神经网络模型，使其能够理解和学习问答对之间的逻辑联系，一旦训练完成，该模型便具备生成全新句子作为回复的能力。2014 年 Vinyals 等人提出的 Sequence-to-Sequence（Seq2Seq）[16]方法成为对话系统研究的一个关键转折点，该方法通过数据驱动的方式挖掘文本之间的规律和特征，将对话系统的研究从基于管道的方法转变为了神经网络联合的端到端方法。这一方法的推出降低了特征工程的不确定性和人工成本。Seq2Seq 模型是一种基于编码器-解码器结构的模型，它通过编码器将输入序列转换为语义表示，然后通过解码器将这些语义表示转换为输出序列，实现了基本的语义理解和生成，使对话系统能够生成符合语义逻辑的回答。

目前，大多数主流的对话系统都基于 Seq2Seq 模型，根据不同任务的需求，采用不同的深度学习技术对模型进行了扩展和优化。如 Transformer[17]模型采用了自注意力机制和多头注意力机制，相比于传统的 Seq2Seq 模型，Transformer 能够更全面的捕捉输入序列中的信息。预训练语言模型如 BERT[18]和 GPT[19]的出现也为生成式对话系统的发展带来了重要的推动力，BERT 采用 Transformer 编码器的结构，通过双向预训练方式丰富了词汇表示，而 GPT 则采用 Transformer 解码器的结构，通过自回归方式训练，使得模型能够生成连贯的文本序列。这些模型通过在大规模语料库上进行预训练，学习到了丰富的语言表示。它们的引入为对话系统提供了更深层次的语义理解能力，使对话系统能够更好的理解问题的语义，生成更加准确和自然的回答。尽管生成式对话系统能够自由生成各种回答，但通常回复的句子存在不通顺的问题，并且需要大量的对话数据进行训练。

1.2.2 生成式大语言模型研究现状

2022 年底，OpenAI 推出了一款专注于对话问答的大语言模型 ChatGPT[20]，标志着对话系统进入到一个全新时代。基于大语言模型的对话系统可以更为高效、准确的理解人类提出的复杂语义问题，并且支持多源异构知识表达和多轮语义交互，可以实现更为智能化的问答体验，对话系统通过不断解决关键技术瓶颈，从简单事实问答到复杂推理问答，其处理能力、覆盖范围与交互智能性不断提升，正朝着模拟人类问答能力的终极目标持续演进。由于大模型对庞大的计算能力和数据资源的高度依赖，研究力量主要来自工业界。但随着清华大学发布了不同参数规模的开源大模型，学术界也得以在有限的计算资源下积极参与进来。

本研究通过文献调研与广泛的资料搜集工作对目前主流的开源大语言模型基本情况进行了归纳，如表 1-1 所示：

表 1-1 主流的开源大模型

名称	发布时间	发布主体	参数规模（B)
Llama[21]	2023.02	Meta	7/13/33/65
Falcon[22]	2023.05	Technology innovation institute	1/7/40
ChatGLM2	2023.06	清华大学	6/12/32/66/130
Bai Chuan	2023.07	百川智能	7/13
Llama2[23]	2023.07	Meta	7/13/70

此类通用大模型一般都是基于已有公开文献与网络数据来训练，面对人群及适用场景十分广泛，由于领域数据积累不足，通用大模型在特定领域的行业针对性与精准度不高[24]。之后部分学者开始以此为切入点构建垂直领域大模型，与通用大模型相比，其主要优势如下：（1）领域专业性：经过专门训练，能够更好的理解和处理特定领域的知识、术语和上下文。（2）高质量输出：由于在特定领域进行了优化，垂直领域大模型的输出质量通常比通用大模型更高。（3）特定任务效果更好：对于特定领域的任务，垂直领域大模型通常比通用大模型表现更好。

本研究通过查阅相关文献和广泛收集各类资料，对当前开源的针对特定领域的大语言模型进行了归纳整理，着重阐述了在构建垂直领域大模型时，相关数据集的构成、使用的基座模型以及训练方法。表 1-2 展示了目前开源的垂直领域大语言模型概览：

表 1-2 垂直领域大模型

领域	大模型名称	数据集构成	基座模型	训练方式
教育领域	Tao li	汉语水平考试试题等	LLaMA-7B	微调
数学领域	ChatGLM-math	加减乘除运算数据集	ChatGLM-6B	微调
科技领域	MaoMozi	科技类数据集	Baichuan-7B	继续预训练+微调
法律领域	LawGPT[25]	法律文书 法律数据	LLaMA-7B	继续预训练+微调
	LeXiLaw	通用领域数据 专业法律数据	ChatGLM-6B	微调
	Lawyer-LLaMA	法律条文裁决文书	LLaMA-13B	继续预训练+微调
医疗领域	ChatDoctor[26]	医疗对话	LLaMA-7B	微调
	ChatGLM-Med	医学文献 医学知识图谱	ChatGLM-6B	微调
	BianQue	医疗问答数据集	ChatYuan-large	微调

1.3 研究内容

本研究首先构建一定规模的粉笔字规范性性书写对话数据集，并在此基础上，使用清华大学开源大语言模型 ChatGLM2-6B 与参数高效微调策略技术和检索增强生成（RAG）技术进行粉笔字规范性书写对话系统研究，通过系统性的研究和反复实验，最终实现基于大语言模型的粉笔字规范性书写对话系统。主要研究内容如下：

（1）对现有的粉笔字字帖字典信息库进行整理，同时整理了一本粉笔字规范书写教材中的知识点，构建了一个包含 5 万对问答的粉笔字规范性书写对话语料。为了满足大模型的微调需求，开发了专用的微调数据集生成工具，以便将语料数据转换为大模型微调时所需的格式。

（2）通过分析 ChatGLM2-6B 官方微调源码，研究了改进的多轮对话训练方式和联合微调策略。

（3）研究了缓解大语言模型“幻觉”问题的检索增强生成（RAG）技术，使用混合检索和重新排序策略对经典 RAG 进行改进，综合大语言模型微调和 RAG 各自的特点，实现了基于大语言模型的粉笔字规范性书写对话系统。

1.4 本文组织结构

本文各个章节组织如下：

第一章为绪论。首先对粉笔字规范性书写对话系统的研究背景与意义进行介绍，接着分别介绍对话系统的研究现状与 ChatGPT 类生成式大语言模型的研究现状，最后说明研究内容及各部分章节安排。

第二章为相关技术及理论基础。对构建对话系统常用的神经网络模型进行介绍，包括循环神经网络、序列到序列模型、Transformer。并详细介绍本研究选用的 ChatGLM2-6B 大语言模型和目前主流的大模型参数高效微调策略技术。

第三章为 ChatGLM2-6B 参数高效微调研究。根据 ChatGLM2-6B 源码分析了 ChatGLM2-6B 的多轮对话数据组织形式和微调训练策略，提出了改进的多轮对话训练方式和联合微调策略。为全面衡量微调后的大模型在特定领域的“记忆”能力，使用 BLEU、ROUGE 和自定义方法进行实验评估，最后给出微调 ChatGLM2-6B 的实验结果。

第四章为大模型检索增强生成（RAG）技术研究。首先说明经典的 RAG 流程，针对经典 RAG 在构建粉笔字规范性书写对话系统时的缺陷，使用混合检索和重排序策略进行改进，最后给出了各种模型的实验结果及分析，并总结了大模型的微调技术和 RAG 技术各自的优缺点。

第五章为总结与展望。首先对本文的研究工作做出总结，进一步分析当前模型存在的缺点和不足，最后对未来工作做出分析及展望。

第 2 章　相关技术及理论基础

本章主要介绍了对话系统所涉及的相关知识，包括循环神经网络、序列到序列模型和 Transformer 模型，并详细介绍了 ChatGLM2-6B 大语言模型和大模型参数高效微调策略。

2.1 循环神经网络及其变体

2.1.1 循环神经网络

循环神经网络（Recurrent Neural Network，RNN）是一种专门用于处理序列数据的神经网络模型[27]。与传统的前馈神经网络不同，RNN 具有循环连接的结构，允许信息在网络中进行循环传递，从而能够对序列数据进行逐步处理，并在处理过程中保持对历史信息的记忆。如图 2-1 所示，RNN 在每个时间步接收输入和前一个时间步的隐藏状态，并产生当前时间步的输出和新的隐藏状态，该机制使得 RNN 能够在不同时间步长之间传递信息，捕捉序列中的长期依赖关系。然而，传统的 RNN 在实际应用中存在一些局限性，特别是在处理长序列时可能会遇到梯度消失或梯度爆炸的问题，这限制了它们捕捉长期依赖关系的能力。为了解决这些问题，研究者们提出了许多改进的 RNN 变体，如长短时记忆网络（LSTM）[28]和门控循环单元（GRU）[29]，这些结构通过引入门控机制来更好地控制和保持长期依赖信息。

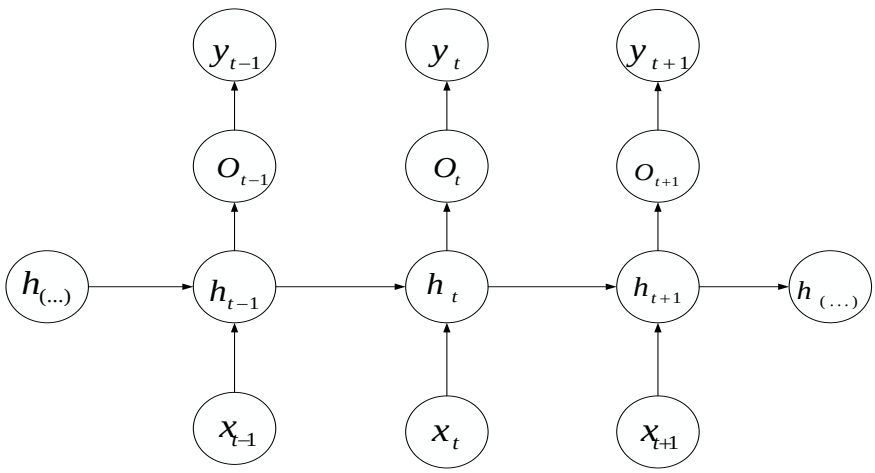

图 2-1 循环神经网络结构图

RNN 的隐藏层结构展开如图 2-2 所示：

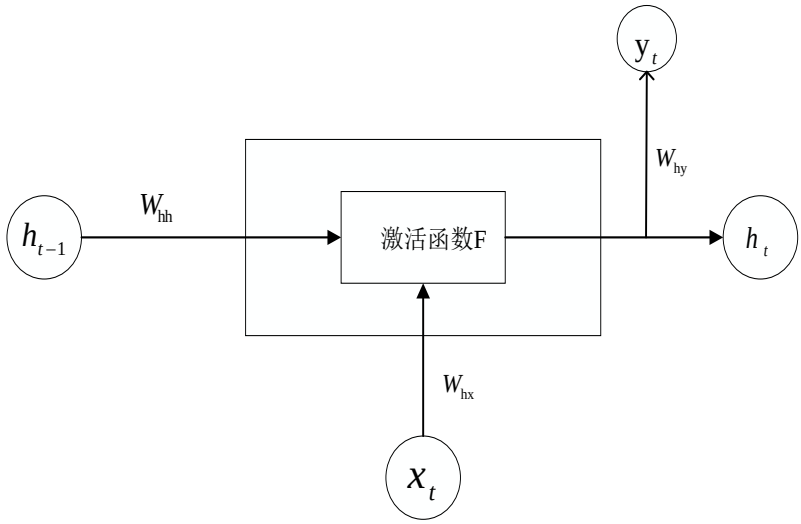

图 2-2 循环神经网络隐藏层结构图

循环神经网络（RNN）的隐藏层结构可以描述为一个或多个隐藏状态组成的层级结构。在时间步 t，隐藏状态 h_t 通过公式（2-1）所计算得到：

$$h_t = \sigma(W_{hh} h_{t-1} + W_{hx} x_t + b_h) \quad (2\text{-}1)$$

其中 W_{hx} 表示输入到隐藏层的权重矩阵，W_{hh} 表示隐藏层到隐藏层的权重矩阵，b_h 是隐藏层的偏置向量。在 RNN 中，每一层都共享参数 W_{hx}、W_{hh}，W_{hy} 和 b_h，参数共享机制降低了网络中需要学习的参数总量，提高了学习效率。

2.1.2 Seq2seq

Sequence-to-Sequence（Seq2Seq）是一种用于序列到序列学习的深度学习模型架构。它最初是由 Google 在机器翻译任务中提出的，并在该任务中取得了显著的成功。如图 2-3 所示，Seq2Seq 模型通常由两个主要部分组成：编码器（Encoder）和解码器（Decoder）。编码器接收输入序列，并将其转换为一个固定长度的向量表示，称为上下文向量 C，如公式（2-2）所示，这个向量表示了输入序列的语义信息。

$$C = q(h_1, h_2, h_3 ... h_t) \quad (2\text{-}2)$$

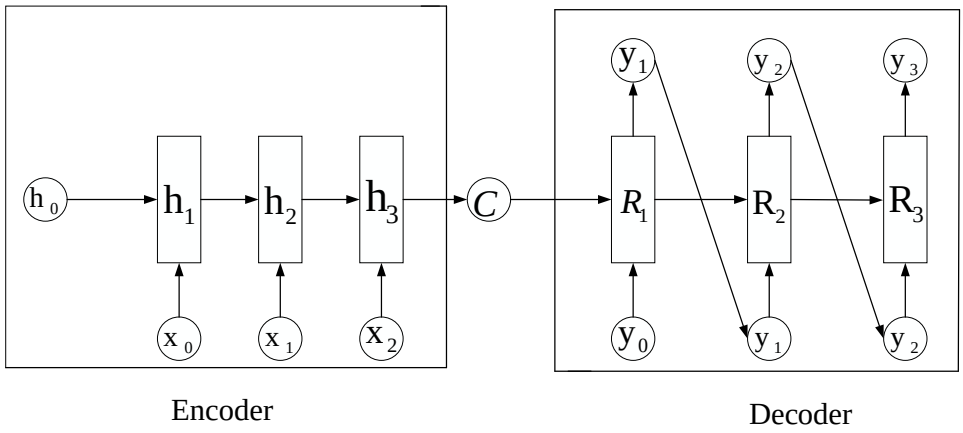

图 2-3 序列到序列模型结构示意图

解码器接受编码器生成的上下文向量，并将其转换为目标序列。使用上下文向量来生成目标序列的输出，一个时间步一个单词或符号，解码过程可以通过公式（2-3）进行描述。

$$y_t = \arg\max P(y_t) = \prod_{t=1}^{T} p(y_t \mid \{y_1, y_2, ...y_{t-1}\}, C) \tag{2-3}$$

Seq2Seq 模型在训练时通常使用教师强制（Teacher Forcing）的方法[30]。教师强制的核心思想是在解码器训练过程中，不是使用上一个时间步的输出作为下一个时间步的输入，而是直接使用上一个时间步的目标值（真实的标签数据）作为输入。此训练方法可以让模型在训练过程中更快的收敛，并且可以减少在训练初期由于模型预测不准确导致的误差累积问题。需要注意的是，虽然教师强制可以加快模型的训练速度，但此训练方式可能导致模型在测试时表现不佳，因为测试时模型无法获得真实的标签数据。

2.1.3 Transformer

Transformer 由 Vaswani 等人在 2017 年提出，与传统的循环神经网络（RNN）和序列到序列学习深度学习模型（Seq2Seq）不同，Transformer 模型摒弃了传统的循环结构[31]，其完全基于注意力机制来处理序列数据，包含两个主要组件：编码器（Encoder）和解码器（Decoder）。

编码器的作用是将输入序列如一个句子中的单词转换为一系列上下文向量表示，

这些向量编码了输入序列中每个元素的全局上下文信息。编码器由多个相同的层堆叠而成，每个层包含两个子层：一个是多头自注意力（Multi-Head Self-Attention）机制，另一个是位置全连接前馈网络。

解码器也由多个相同的层堆叠而成，每个层包含三个子层：一个是多头自注意力机制，用于关注输入序列的各个元素；一个是多头注意力（Multi-Head Attention）机制，用于关注编码器的输出；最后是一个位置全连接前馈网络。解码器的作用是基于编码器提供的上下文向量生成输出序列（如翻译后的句子）。Transformer 模型的整体结构如图 2-4 所示。

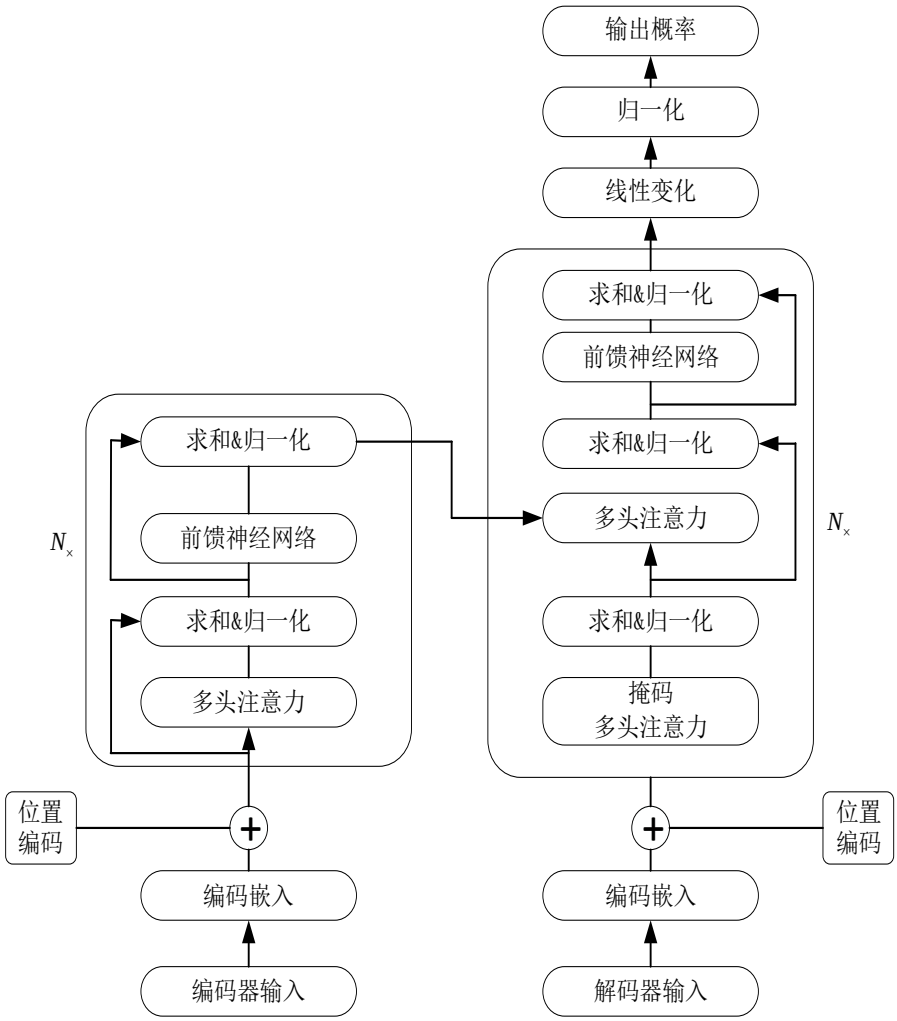

图 2-4 Transformer 的编码器-解码器的示意图

注意力机制是 Transformer 模型的核心部分[32]，包括自注意力和多头注意力。自注意力（Self-Attention）允许模型在处理序列中的每个元素时同时考虑到序列中的所

有其他元素。自注意力机制通过学习一个权重矩阵来计算序列中每个元素与其他元素之间的关联度，这些权重表示了在生成当前元素的表示时，其他元素应该给予多少“注意力”。

多头注意力（Multi-Head Attention）则是在自注意力的基础上进一步扩展。它将输入序列分割成多个“头”，每个头有自己的权重矩阵，这样可以并行地在不同的表示空间中计算注意力。每个头关注输入序列的不同部分，这样模型就能够捕捉到更加丰富的信息。最后，将这些头的输出拼接起来，并通过一个线性层进行处理，得到最终的输出。多头注意力机制流程如图 2-5 所示

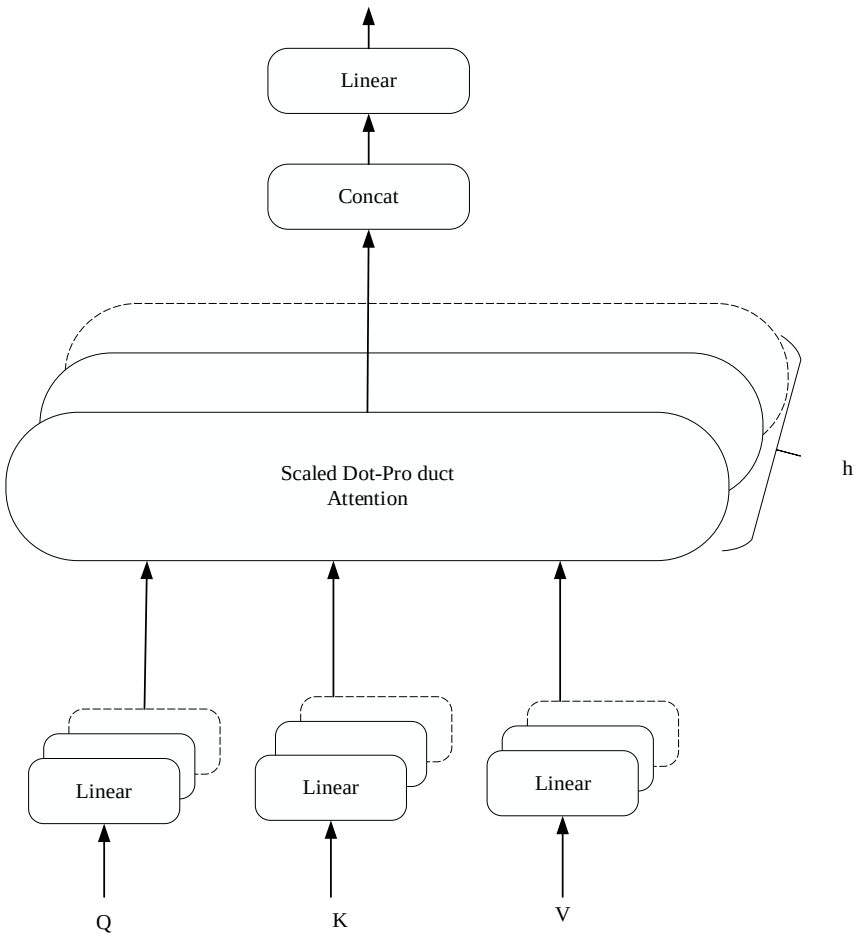

图 2-5 多头注意力机制流程图

多头注意力计算方式如公式（2-4）、（2-5）、（2-6）表示：

$$Attention(Q,K,V) = Soft\max(\frac{Qk^T}{\sqrt{d_k}})V \tag{2-4}$$

$$MultiHead(Q,K,V) = concat(head_1,...head_h)w^o \quad (2\text{-}5)$$

$$head_i = Attention(Qw_i^Q, Kw_i^K, vw_i^v) \quad (2\text{-}6)$$

式中的 Q、K 和 V 表示查询向量、键向量和值向量。h 表示头的个数，w_i^Q, w_i^K，w_i^v 是与第 i 个注意力头相关联的线性变换矩阵。w^o 是用于拼接后的线性变换矩阵，$concat$ 表示拼接操作。缩放因子 $\sqrt{d_k}$ 将点积的结果进行缩放，使得点积的值落在一个相对较小的范围内，避免了由于点积值过大或过小导致数值计算不稳定的情况，有助于提高训练过程中 Softmax 函数的稳定性[33]。

2.2 生成式大语言模型

2.2.1 ChatGLM

随着算力的不断发展，国内外开发出了一系列大语言模型，由清华大学提出的通用大语言模型 ChatGLM[34]正在被国内各个组织和企业广泛使用。结合模型量化技术，用户可以在消费级的显卡上进行本地部署，ChatGLM 各量化等级和推理微调时对 GPU 显存的要求如表 2-1 所示：

表 2-1 ChatGLM 版本

量化等级	最低 GPU 显存（推理）	最低 GPU 显存（高效参数微调）
FP16(无量化）	13GB	14GB
INT8(量化）	8GB	9GB
INT4(量化）	6GB	7GB

ChatGLM-6B 是一个开源的、支持中英双语的对话语言模型，ChatGLM-6B 使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 虽然规模不及千亿模型，但大大降低了推理成本，提升了效率，并且已经能生成相当符合人类偏好的回答[35]。

ChatGLM2-6B 是开源中英双语对话模型 ChatGLM-6B 的第二代版本，在保留了初代模型对话流畅、部署门槛较低等众多优秀特性的基础之上，ChatGLM2-6B 引入

了如下新特性：

（1）更强大的性能：基于 ChatGLM 初代模型的开发经验，全面升级了 ChatGLM2-6B 的基座模型，模型结构改从 Prefix-LM 回归纯粹的 Decoder-Only 结构。ChatGLM2-6B 使用了 GLM 的混合目标函数，经过了 1.4T 中英标识符的预训练与人类偏好对齐训练，评测结果显示，相比于初代模型：ChatGLM2-6B 在 MMLU（+23%）、CEval（+33%）、GSM8K（+57%）、BBH（+60%）等数据集上的性能取得了大幅度的提升，在同尺寸开源模型中具有较强的竞争力。

（2）更长的上下文：基于 Flash Attention 技术，将基座模型的上下文长度（Context Length）由 ChatGLM-6B 的 2K 扩展到了 32K，并在对话阶段使用 8K 的上下文长度训练。

（3）更高效的推理：基于 Multi-Query Attention 技术，ChatGLM2-6B 有更高效的推理速度和更低的显存占用：在官方的模型实现下，推理速度相比初代提升了 42%，INT4 量化下，6G 显存支持的对话长度由 1K 提升到了 8K。

（4）更开放的协议：ChatGLM2-6B 权重对学术研究完全开放，在填写问卷进行登记后亦允许免费商业使用。

由于目前很多大模型需要强大的算力支持，但在计算资源只有单卡的环境下，ChatGLM2-6B 也能够进行部署和微调，并取得不错的效果，这也是本研究选择该模型的主要原因。

2.2.2 参数高效微调策略

大语言模型在海量文本数据上训练而来，天然适应各种自然语言处理任务。然而直接将原始训练而来的大语言模型用于特定领域，效果往往差强人意。为了更好的适应特定领域的任务，需要进一步在领域知识中进行微调训练，将外部知识内化融合到大语言模型参数中，使其进一步提升对话能力。与之前微调预训练语言模型（如 BERT）不同的是，微调大语言模型需要使用更为海量、高质量的数据，以及更高的计算资源。

重新开始训练大模型需要大量高性能 GPU，这对没有充足计算资源的研究人员来说是个挑战。为了应对上述问题，研究人员开始探索参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）技术[36]。PEFT 技术的目标是通过最小化微调参数和计算复杂度，提高预训练模型在新任务上性能的同时，也不出现严重的灾难性遗忘问题，在提升模型效果的同时，也能显著缩短模型训练时间和计算成本。

当前，主流的参数高效微调策略技术可大致分为三类：Adapter[37]、Prefix Tuning[38]以及 LoRA[39]（Low-Rank Adaptation）。它们在模型结构中的嵌入位置和特点各异，具体细节请参考图 2-6：

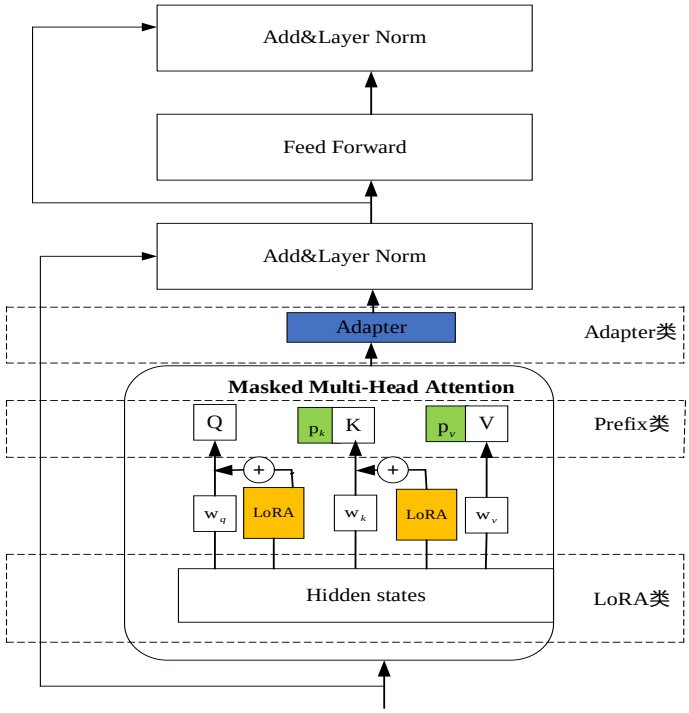

图 2-6 三类参数高效微调策略图

在传统的微调方法中，整个模型的所有参数都会被调整以适应新的任务，这通常需要大量的计算资源和时间。Adapter 类微调技术则通过引入一些小的、可学习的模块来降低参数。Adapter 模块被插入在前馈神经网络（FFN）和自注意力模块之间，Adapter 模块通常由两个或多个全连接层组成，这些层对输入进行降维和升维操作，在微调时只有这些模块的参数需要被调整，而预训练模型的主干部分保持不变。Adapter 类微调技术的优点包括：（1）参数效率：由于只有 Adapter 的参数需要被调整，因此这种方法大大减少了需要存储和传输的参数数量，提高了参数效率。（2）计算效率：由于大部分预训练模型的参数保持不变，因此 Adapter 类微调技术可以在更短的时间内完成训练，减少了计算资源的消耗。Adapter 方法在预训练模型的层中插入可训练模块的形式简单，但增加了模型层数，引入了额外的推理延迟。

Prefix Tuning 是另一种高效的微调策略，它通过在模型输入前附加一个可学习的连续前缀向量序列来引导模型生成目标输出。此方法的基本理念是通过优化前缀向量的内容，来调控模型的注意力机制和隐藏层状态，进而使模型生成的输出更加符合

特定任务的要求。在 Prefix Tuning 中，可学习的前缀被添加到输入序列的起始位置，并且具有固定的长度。可学习的前缀是专门针对每个任务进行训练的，它与模型的原始输入一起进入注意力层，通过微调训练影响模型对输入序列中各个元素的注意力分配。可学习的前缀充当特定任务的“指令”，指导模型生成更加符合任务需求的输出。与 Adapter 类方法相似，Prefix Tuning 仅对前缀参数进行更新而保持预训练模型的核心参数不变。这种方法使模型能够针对不同任务学习独立的前缀，从而支持多任务学习，并且能够轻松的为新任务添加新的前缀。Prefix Tuning 适用于多种任务类型，包括文本生成、文本分类、机器翻译等。然而，由于前缀参数的学习涉及到对模型内部注意力机制的精细调整，这可能会使训练过程变得复杂，需要更细致的参数调整和更长的收敛时间。此外，由于前缀的添加可能会占据模型能够处理的最大序列长度的一部分，可能会减少下游任务输入的可用长度，从而对模型性能产生影响。

LoRA 类技术在微调过程中冻结预训练模型权重，并在每个 Transformer 块中注入可训练的低秩分解矩阵以近似模型权重矩阵的更新，通过添加一个旁路并利用低秩分解来模拟参数的更新，实现了轻量微调的目的。相较于其他微调方法，LoRA 不会因增加参数量而降低性能，也不会增加推理耗时，更易于优化。LoRA 支持可插拔式的任务切换，能够节省显存，并且能将现有大模型转化为不同领域的专业模型。此外，低秩适配法与大多数高效参数微调算法兼容，可以联合使用，进一步提升大模型在新任务上的性能。具体细节如图 2-7 所示：

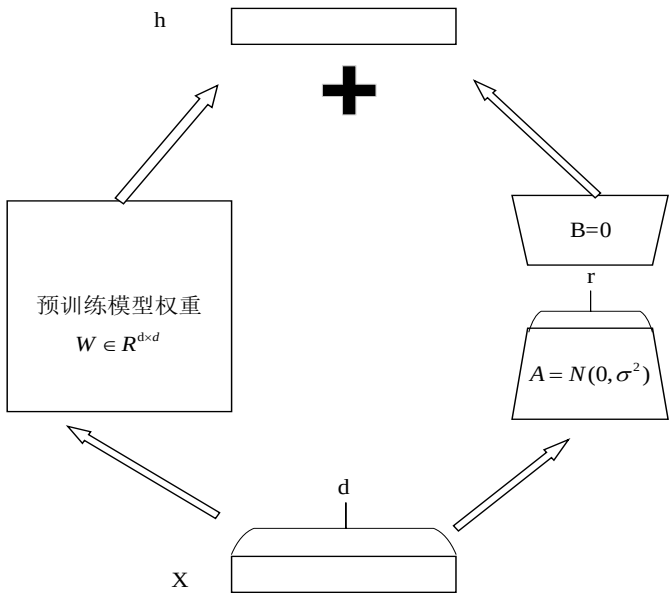

图 2-7 LORA 微调策略图

以单个 Linear 层举例，假设原始预训练模型的权重为$W_0 \in R^{d\times k}$，定义两个低秩矩阵$B \in R^{d\times r}$和$A \in R^{r\times k}$,其中 r<<min(d,k)。Linear 层原始微调的计算方式如公式（2-7）所示：

$$h = W_0 X + \Delta W X \tag{2-7}$$

LoRA 微调时的计算方式如公式（2-8）所示：

$$h = W_0 X + BAX \tag{2-8}$$

对矩阵 A 使用随机高斯初始化，对 B 使用零初始化,可以保证在微调开始之前ΔWX=0。也可以使用$\frac{\partial}{r}$作为缩放参数来调节ΔW,通过调节缩放比例可以调节预训练模型与 LoRA 的加权占比。

以单个 Linear 层为例探究 LoRA 微调时反向传播阶段与梯度下降阶段的复杂度，不使用 LoRA 微调时 Linear 反向传播的梯度计算公式如（2-9）：

$$\begin{cases} \dfrac{\partial L}{\partial w} = \dfrac{\partial L}{\partial h} x^T \\ \dfrac{\partial L}{\partial x} = \dfrac{\partial L}{\partial h} w^T \end{cases} \tag{2-9}$$

使用 LoRA 微调时 Linear 的反向传播梯度计算公式如（2-10）：

$$\begin{cases} \dfrac{\partial L}{\partial B} = \dfrac{\partial L}{\partial h} x^T A^T \\ \dfrac{\partial L}{\partial A} = \dfrac{\partial L}{\partial h} x^T B^T \\ \dfrac{\partial L}{\partial x} = (w + BA)^T \dfrac{\partial L}{\partial h} \end{cases} \tag{2-10}$$

不难发现在反向传播阶段，同一层使用 LoRA 相对于不使用 LoRA 所需要求解的梯度计算量还要多一些，因为 r<<min(d,k)的缘故，所以(d*r+k*r)<<d*k。事实上并不是所有可学习的层都在微调时都要使用 LoRA，原论文中只对生成 QKV 的三个

Linear 使用了 LoRA，不使用 LoRA 微调的层只需要求输入梯度$\frac{\partial L}{\partial x}$即可。同理，在梯度下降阶段，因为可训练的参数量少，所以需要梯度下降的参数也少，只需要对 LoRA 的两个低秩矩阵梯度下降即可。因此可以得出：LoRA 微调在反向传播阶段计算复杂度还要多一些，只是需要梯度下降的参数少，所以节省显存，梯度下降的也快。

关于 B 和 A 这两个低秩矩阵的初始化问题，首先是需要 BA 的结果初始是 0，这样能保证微调开始时新引入的低秩矩阵不会对最终结果造成影响，最直接的方式是令其中一个低秩矩阵初始阶段为全零，另一个为非全零即可。但是两者不能都为全零，通过 LoRA 的梯度计算公式（2-10）可知，如果 B 和 A 这两个低秩矩阵初始化都是 0，则两个矩阵的梯度都是 0，模型在微调时无法进行有效训练。

2.3 本章小结

本章主要介绍本文研究方法所用到的相关理论，详细介绍了循环神经网络及其变体的相关知识，包括循环神经网络 RNN、序列到序列模型、Transformer，并对模型的结构与参数进行了详细解释。此外还重点介绍了清华开源大语言模型 ChatGLM 和三类主流大语言模型参数高效微调策略技术：Adapter，Prefix-Tuning，LoRA，并针对每类技术进行了详细分析，为后续实现基于大语言模型的粉笔字规范性书写对话系统提供理论指导。

第 3 章　ChatGLM2-6B 参数高效微调研究

3.1 引言

与通用大语言模型相比，垂直领域大语言模型更专注于某个特定领域的知识和技能，具备更高的领域专业性和实用性。本章通过构建的粉笔字规范性书写对话数据集对 ChatGLM2-6B 进行微调。针对 ChatGLM2-6B 的模型结构提出了一种更加充分和高效的多轮对话训练方式，在微调大模型时使用了联合微调训练方式，并通过多组对比实验进行了研究。

3.2 改进的多轮对话训练方式

在分析 ChatGLM2-6B 官方微调源码时[40]，发现只有最后一轮的对话内容参与计算 loss，其他轮次的回复内容不参与计算 loss。首先分析 ChatGLM2-6B 的多轮对话数据组织形式，具体内容如图 3-1 所示：

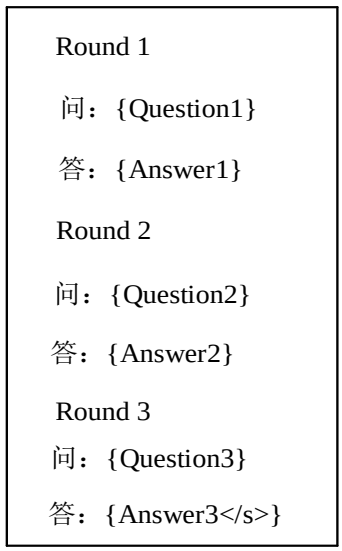

图 3-1 多轮对话组织形式

从图 3-1 可以看出 ChatGLM2-6B 多轮对话数据组织形式，其中[Round]表示多轮对话的轮次，</s>表示模型的生成结束符。

接下来探究 ChatGLM2-6B 采用何种方式训练多轮对话，通过分析源码发现模型

最终的输入由 prompt、answer 和结束符</s>组成，其中 prompt 是将历史对话和当前轮次的用户输入进行拼接，answer 是当前轮次的回复，具体内容如图 3-2 所示：

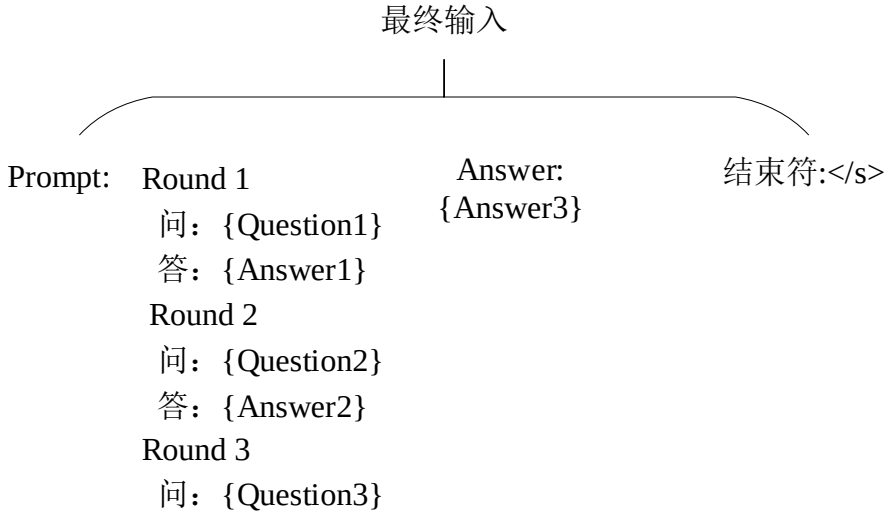

图 3-2 微调输入组织形式

通过分析 ChatGLM2-6B 微调源码可以得出：在微调时模型的最终输入部分除了最后一个轮次的回复内容外，其它所有位置的 tokens 都被置为 pad_token_id。说明只有最后一轮的回复内容参与计算 loss，其他轮次的回复内容不参与计算 loss，训练数据没有充分利用，被浪费了。具体过程如图 3-3 所示：

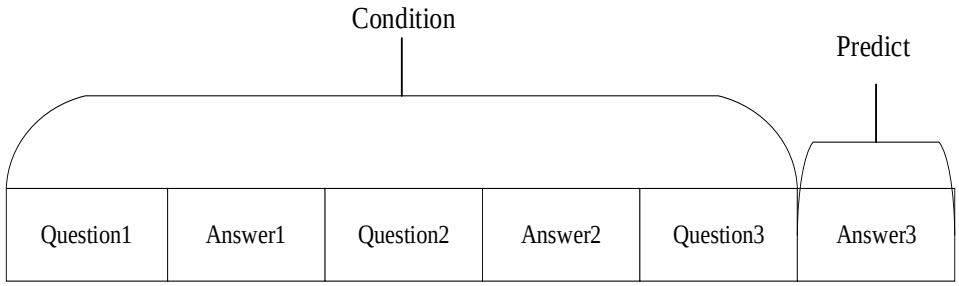

图 3-3 原始多轮对话训练方式

事实上在大模型的指令微调阶段，一般只有 Answer 部分的 loss 会用于梯度回传来更新权重，而 Question 部分的 loss 不会用于更新权重。ChatGLM2-6B 在进行多轮对话微调训练时将 Question1、Answer1、Question2、Answer2、Question3 的文本都视为模型的输入部分，将 Answer3 的文本视为模型的预测部分，因此只有 Answer3 部分的 loss 参与权重更新。该类方法的弊端在于，模型在训练时并没有充分利用多轮

对话的训练数据，Answer1 和 Answer2 的内容没有参与模型训练，这部分数据在训练时被浪费了。事实上对于多轮对话数据而言，往往是中间的回复部分的信息量更丰富详细，最后一个 Answer 回复部分往往是“谢谢”、“不客气”等诸如此类的较为简短的文本[41]。如果只使用这部分文本训练模型，会严重影响模型的训练效果。为充分利用多轮对话的训练数据，可以使用方式二改进原有的训练方式，方式二如图 3-4 所示。具体而言将一条多轮对话数据拆分成多条数据[42]，例如将以上示例拆分成以下三条数据。

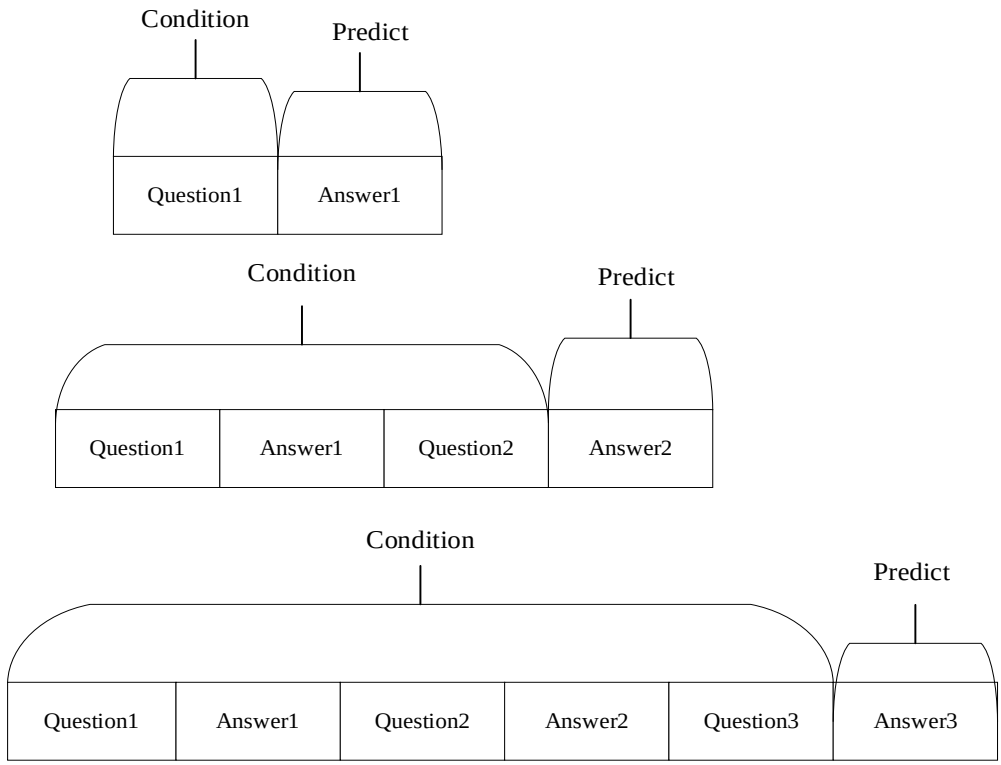

图 3-4 拆分的多轮对话训练方式

方式二能够更加充分利用多轮对话中每一个Answer的回复内容。但是弊端在于，需要将一个包含 n 轮对话的数据，拆分成 n 条数据，训练效率降低了 n 倍，训练方法不高效，需要更多的 GPU 显存来存储和处理数据，对没有充足计算资源的使用者来说是一个挑战。

针对方式一和方式二各自的优缺点，本研究使用了一种更加充分和高效的多轮对话训练方式。具体流程如图 3-5 所示，将一条多轮对话数据拼接之后输入模型，并行计算每个位置的 loss，只有 Answer 部分的 loss 参与权重更新。

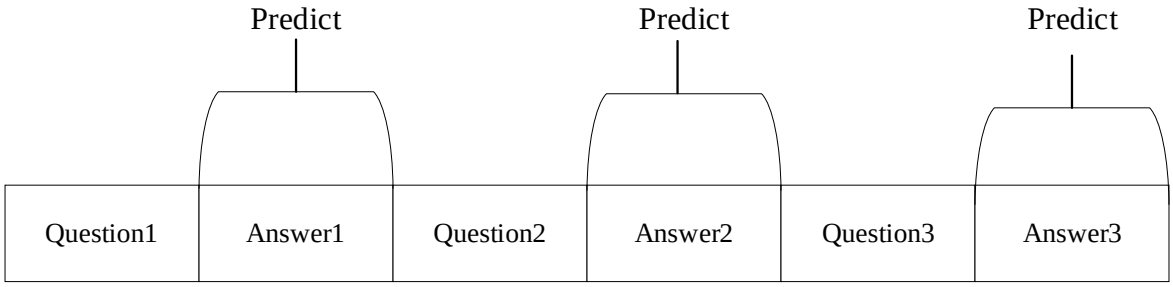

图 3-5 改进的多轮对话训练方式

能进行并行计算的主要原因是以 GPT 为代表的因果语言模型（Causal Language）的 attention mask 是一个对角掩码矩阵[43]，如图 3-6 所示，每个 token 在编码的时候，只能看到它之前的 token，看不到它之后的 token。所以 Answer1 部分的编码输出，只能感知到 Question1 的内容，无法感知到它之后的文本，可以用来预测 Answer1 的内容。而 Answer2 部分的编码输出，只能看到 Question1、Answer1、Question2 的内容，可以用来预测 Answer2 的内容，依此类推。对于整个序列，只需要输入模型一次，便可并行计算每个回复部分的 loss。

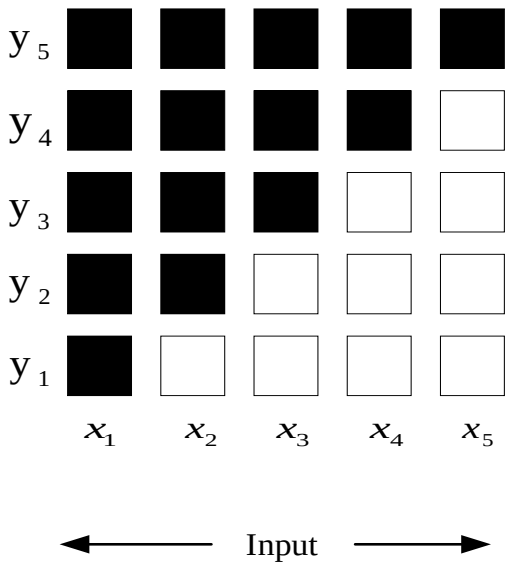

图 3-6 注意力掩码矩阵

值得注意的是，GLM 类大语言模型不属于严格意义上的因果语言模型，因为它们存在 prefix attention mask 的设计。例如对于 prefix 而言，它的 attention 是双向的，而预测部分的 attention 是单向的[44]。ChatGLM2-6B 是开源中英双语对话模型

ChatGLM-6B 的第二代版本，在保留了初代模型对话流畅、部署门槛较低等众多优秀特性的基础之上，ChatGLM2-6B 的模型结构做了改变，从 ChatGLM-6B 的 Prefix-LM 改为了纯粹的 Decoder-Only 结构，因此本研究使用的多轮对话训练方式适配 ChatGLM2-6B 大模型。

ChatGLM2-6B 是一个经过指令微调的模型，在微调时遵从官方的数据组织格式，才能达到最优效果。为防止模型在推理时出现“自问自答”和“不停止”的情况，训练的时候，在每个 Answer 的回复后面都添加</s>，作为此轮对话生成结束的标识符，否则推理的时候，模型很难采样到</s>，无法结束生成。

改进的多轮对话训练方式具体实现过程如图 3-7 所示：在模型训练生成 token_id 的时候，还会另外生成一个 target_mask，取值为 0 或 1，用来标记每个 token 是否属于 target 部分，即是否需要模型进行预测。

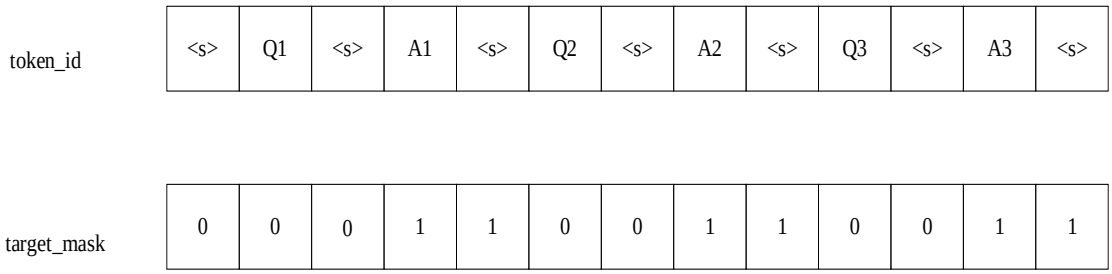

图 3-7 微调过程编码图

如图 3-7 所示，Answer 和之后</s>的 target_mask 均为 1，其他部分为 0。只有 target_mask=1 的部分位置的 loss 才会参与权重更新,以此来并行计算每个 Answer 位置的 loss。这种方式利用了模型并行计算的优势，并且多轮对话中的每个回复部分都参与了训练，更加充分利用了数据。与方式二将一条多轮对话数据拆分成多条数据相比，不需要更多的显存，计算更加高效。

3.3 联合微调

将预训练好的语言模型在下游任务上进行微调已成为处理 NLP 任务的一种范式。与开箱即用的预训练 LLM(例如：零样本推理)相比，在下游数据集上微调这些预训练 LLM 会带来巨大的性能提升。但是，随着模型变得越来越大，在消费级硬件上对模型进行全部参数的微调（full fine-tuning）变得不可行。此外，为每个下游任务独立存储和部署微调模型变得越来越昂贵，因为微调模型（调整模型的所有参数）与原始

预训练模型的大小相同。

第二章节介绍了三种主流的参数高效微调技术，众多学者已经针对三种高效微调技术分别做了改进，例如在 LoRA 基础上衍生出了 AdaLoRA[45]，QLoRA[46]等技术。但很少有学者针对三类高效微调技术的联合微调进行探究，本研究针对 LoRA 与大多数高效参数微调算法兼容，可以联合使用的特点，提出 Prefix-LoRA 联合微调[47]方法，具体步骤如下：先使用 Prefix 微调 ChatGLM2-6B，将得到的最优参数权重与 ChatGLM2-6B 组合，保存为 ChatGLM2-Prefix 模型。基于相同的训练集，使用 LoRA 方法进一步微调 ChatGLM2-Prefix 模型。将 ChatGLM2-Prefix 模型与 LoRA 参数权重相结合，得到 ChatGLM2-Prefix-LoRA 模型。

3.4 实验设计及结果分析

为验证本研究提出的改进的多轮对话训练方式和联合微调，使用构建的数据集对 ChatGLM2-6B 进行微调，设计了五组实验，通过 Prefix 微调和 LoRA 微调并结合改进的多轮对话训练方式进行实验，再此基础上，使用联合微调进行对比分析。

3.4.1 实验环境及数据

本研究的实验环境具体配置如表 3-1 所示。

表 3-1 实验环境配置

名称	配置
处理器	NVIDIA GeForce RTX 4090
操作系统	Linux
单卡显存大小	24G
编程语言	Python 3.10
深度学习框架	Pytorch 2.0

3.4.2 数据集构建

实验使用的数据集由两部分组成，第一部分对现有的粉笔字模板字典信息进行整理。第二部分选取了一本粉笔字规范书写教材，下面介绍数据集构建方法：

（1）粉笔字模板字典信息整理。粉笔字模板字典信息记录了书法教师书写 800

个粉笔字的特征信息，特征信息主要包括整字特征信息、部件特征信息、笔画特征信息。整字的特征主要有汉字空间位置，汉字大小、汉字宽高比等，部件的特征主要有部件空间位置、部件大小、部件宽高比等，笔画的特征主要有笔画的空间位置、笔画长度、笔画斜率等。这些特征信息经过量化处理后存储于 JSON 文件中，在大模型微调阶段，需要把 JSON 文件中的结构化信息转为文本信息。在转换时使用 Python 的 JSON 模块读取此 JSON 文件，将 JSON 文件中的关键节点转换为自然语言描述。例如"word_name"转换为“汉字名”，"Stroke_num"转换为“笔画数目”，"focus"转换为“整字重心”，"area"转换为“文字大小”。转换结果如图 3-8 所示：

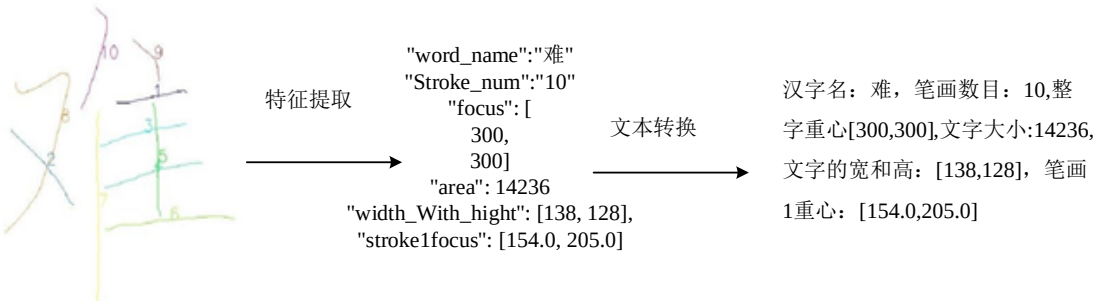

图 3-8 粉笔字模板字典信息整理

将粉笔字模板字典信息转换为文本后，需要针对此文本设置配套的问题，以便构建 ChatGLM2-6B 微调的问答对。本研究中采用直接问答的方式进行构建，即针对书法教师书写 800 个粉笔字的特征信息批量生成问题。

（2）对于粉笔字规范书写的 pdf 格式的教材，使用 Python 工具包自动提取文本，并对文本进行初步预处理，只保留每个段落的纯文本，选择高质量、合适的段落文本作为语料库数据。

（3）与以往 BERT 模型微调不同[48]，大模型微调时需要遵守其指定的数据格式，因此制作了一款 ChatGLM 微调数据集生成工具,其主要功能是读取用户保存在文本文件的问题。在本地服务器上以 API 方式部署 ChatGPT。将每条处理后的数据逐一输入到 ChatGPT 中，通过编写适当的提示，引导 ChatGPT 根据提示和给定的文本，以自我指导的方式自动生成并返回若干问答数据，或者用户自行撰写回答，最后自动保存为微调大模型所需要的 json 格式文件。具体内容如图 3-9 所示：

图 3-9 数据集生成工具

3.4.3 模型评价指标

目前对智能对话系统评价基本有两条路径，一是借鉴一些已经在其他类型的自然语言处理任务取得卓越评价效果的指标，将其应用到对话系统评价中。代表性指标包括 BLEU、METEOR 和 ROUGE。其中，BLEU 和 METEOR 广泛用于机器翻译评价，而 ROUGE 则更多用于度量自动生成摘要的准确性[49]。另一种是人工评分，即主观性评价或者进行模拟人工评分，使用人工评测方法的优势在于其能够提供真实且相对准确的评估数据，然而，人工评测方法通常需要巨大的资源投入和大量的时间，常见于资源充裕的实验环境中。

为使实验结果更加直观和全面，本研究采用多种评价指标来评估微调后的模型的性能。首先选择使用 BLEU、ROUGE 两种评价指标对微调后的模型进行初步的评价和分析，这两种指标可以帮助了解大语言模型在生成文本和回答问题方面的表现。此外，为评估微调后的大模型“记忆”特定领域知识的能力，设计了一种新的评估策略，下面分别介绍这三种评价指标：

（1）BLEU

BLEU 是由 Kishore Papineni 等人在 2002 年提出的一种常用于自动化评估机器翻译质量的指标。它通过比较机器翻译的结果和人工翻译的结果之间的相似度来衡

量翻译的准确性。BLEU 的计算基于 n-gram 的精度以及一个惩罚因子，用于惩罚较短的翻译。BLEU 使用 n-gram 精度来衡量翻译结果与参考翻译之间的匹配程度，通过计算翻译结果中的 n-gram（连续 n 个词）与参考翻译中的 n-gram 的重叠程度，这样就能够捕捉到翻译的准确性，通常的 n 值设置为 1 到 4。BLEU 还引入了一个惩罚因子，用于惩罚较短的翻译结果。这是为了解决机器翻译系统倾向于生成过短翻译的问题。如果翻译结果较短，则会受到惩罚，从而降低 BLEU 分数，其公式如下：

$$BLEU = BP \times \exp(\sum_{n=1}^{N} W_n \times \log P_n) \tag{3-1}$$

$$BP = \begin{cases} 1 & if \quad l_c > l_s \\ e^{1-\frac{l_s}{l_c}} & if \quad l_c \le l_s \end{cases} \tag{3-2}$$

其中W_n为 n-gram 的权重，P_n为 n-gram 的精确率，BP 为惩罚因子，l_c是答句的长度，l_s是最短的参考答句的长度。BLEU 的 1-gram 精确率表示了答句中与原始答句的匹配程度，而其他 n-gram 则反映了答句生成的流畅程度。

然而，需要注意的是，BLEU 并不总是与人类判断一致，因此在使用时需要结合其他评价指标来综合评估系统的性能。

（2）ROUGE

ROUGE 是一种常用于自动评估文本摘要生成质量的指标。与 BLEU 类似，ROUGE 也是一种自动化的评价方法，主要关注于召回率，即系统生成的摘要是否能够涵盖到参考摘要中的重要信息。

ROUGE-N 衡量的是模型生成的文本和参考文本中 n-gram 的重叠。ROUGE-1 是比较一元组的重叠，一元组指的是文本中的单个词或字符。例如，在句子“我喜欢吃苹果”中，“我”、“喜欢”、“吃”和“苹果”都是一元组。ROUGE-2 则是比较二元组的重叠，例如上面的例子，“我 喜欢”、“喜欢 吃”和“吃 苹果”都是二元组。ROUGE-N 的值越高，表示模型生成的文本和参考文本在 n-gram 级别上的相似度越高。

ROUGE-L 专门用于衡量生成文本与参考文本之间长序列匹配的程度。ROUGE-L 的核心思想是比较生成文本与参考文本之间的最长公共子序列（LCS）。LCS 指两个文本序列中从开始到结束的最长、完全相同的子序列。在计算 ROUGE-L 时，会考

虑生成文本中每个 n-gram（从 1-gram 到最大 n-gram）与参考文本中相应 n-gram 的匹配情况，然后根据匹配的 n-gram 数量来计算召回率和 F1 分数。

（3）在本研究中，采用的数据集包含从粉笔字模板字典中提取出的数字和坐标信息，研究的目标是确保微调后的大模型能够准确的记忆这些关键信息。因此，评估不仅关注文本生成质量，而且强调模型对特定领域知识的掌握和应用能力。所以引入了一种新的评估方式，其主要思想是提供测评相关的成对问题和答案，使用微调后的模型输出预测答案，通过编写适当的提示词借助 ChatGPT 来评估预测答案与标准答案的匹配程度。具体定义如下：

TP：模型正确的回答了问题，即模型的输出与真实答案相匹配。

FN：模型未能提供一个正确的答案，即模型没有输出真实答案。

使用公式（3-3）计算召回率：

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{3-3}$$

召回率又被成为查全率，指模型正确预测的正样本数量占所有真实正样本的比例，即模型正确回答的问题数量占所有问题的数量的比例。召回率衡量了模型找出所有真实正例的能力，通过这一指标，能够对微调后的大模型在学习数字和坐标信息方面的性能进行深入的评估，为模型的进一步改进提供有价值的参考。

3.4.4 改进的多轮对话微调实验

使用LoRA和Prefix方式进行参数高效微调实验，具体步骤如下：

（1）将构建的数据集分为 45000 个训练样本和 5000 个测试样本。

（2）使用训练样本在ChatGLM2-6B上进行Prefix微调训练，微调后的模型记为ChatGLM2-Prefix1，使用同样的训练样本通过LoRA微调ChatGLM2-6B,微调后的模型记为ChatGLM2-LoRA1。

（3）使用改进的多轮对话训练方式，在ChatGLM2-6B上进行Prefix微调训练，微调后的模型记为ChatGLM2-Prefix2。使用改进的多轮对话训练方式，在ChatGLM2 上进行LoRA微调训练，微调后的模型记为ChatGLM-LoRA2。

（4）使用测试样本对上述所有微调后的模型进行性能评估，以比较不同微调方法的效果。

实验中 Prefix 的具体参数配置如表 3-2 所示：

表 3-2 Prefix 微调超参数设置

参数	参数意义	参数值
pre_seq_len	前缀序列长度	128
learning_rate	学习率	2e-2
batchsize	批大小	20
epoch	周期	10
max_source_length	最大源长度	128
max_target_len	最大目标长度	256

实验中 LoRA 的具体参数配置如表 3-3 所示：

表 3-3 LoRA 微调超参数设置

参数	参数意义	参数值
lora_r	低秩矩阵秩	128
lora_alpha	缩放因子	2e-2
lora_dropout	随机忽略连接比例	20
learning_rate	学习率	10
batchsize	批大小	128
epochs	周期	256

微调后的实验结果如表 3-4 所示：

表 3-4 微调实验结果

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
ChatGLM2-6B	26.3%	35.6%	28.3%	31.3%
ChatGLM2-LoRA1	38.1%	49.4%	40.7%	42.5%
ChatGLM2-LoRA2	42.9%	54.3%	43.8%	43.7%
ChatGLM2-Prefix1	40.7%	45.2%	38.1%	33.8%
ChatGLM2-Prefix2	44.3%	49.1%	40.4%	34.2%

通过比较表 3-4，可以发现与 ChatGLM2-6B 相比，不管采用哪种微调方式都可以显著提高测试样本的 BLEU 和 ROUGE 评分。并且在微调过程中结合改进的多轮对话训练方式后，BLEU 和 ROUGE 均有进一步提升。然而各类指标得分普遍较低，主要是因为大语言模型生成的答案与原始答案之间存在较大差异，也从侧面反映出生成式大语言模型具有较高的创造能力。

具体而言，与 ChatGLM2-6B 模型相比，ChatGLM2-Prefix1 可分别将 BLEU-4 评分增加约 13%，Rough-1 评分提升约 9%。ChatGLM2-LoRA1 将 Bleu-4 评分增加约 12%，Rough-1 评分提升约 14%。LoRA 微调的 ROUGE-L 分数高于 Prefix 微调，主要原因是 LoRA 微调通过引入少量特定任务的参数，能够在保持模型通用语言能力的同时，提高对特定任务的适应性。相比之下，Prefix 微调需要更多的数据来调整模型权重，且容易受到原始模型已学到的通用语言知识的影响，因此在 ROUGE-L 上的表现不如 LoRA 微调。

在微调过程中使用改进的多轮对话训练方式后，与 ChatGLM2-LoRA1 相比，ChatGLM2-LoRA2 将 BLEU-4 评分增加约 5%，与 ChatGLM2-Prefix1 相比，ChatGLM2-Prefix2 的 ROUGE-1 评分也有 4%左右的提升，说明在微调过程中，使用改进的多轮对话训练方式能使模型更好的捕捉上下文信息，理解对话的连贯性，从而生成更加准确和连贯的回复。实验结果验证了本研究提出的改进的多轮对话训练方式的有效性，同时也意味着，通过持续优化训练策略，可以进一步改善语言模型在复杂对话场景中的表现，使其更加接近人类对话的水平。

3.4.5 联合微调实验

为验证本研究提出的 Prefix-LoRA 联合微调方法，进行了对比实验，将 3.4.3 节中描述的 ChatGLM2-Prefix2 作为基线模型,基于相同的训练集，使用 LoRA 方法进一步微调 ChatGLM2-Prefix2 模型。将模型超参数设定为与第 3.4.3 节相同的值，以确保可比性,联合微调后得到 ChatGLM2-Prefix-LoRA 模型。

表 3-5 联合微调实验结果

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
ChatGLM2-Prefix2	44.3%	49.1%	40.4%	34.2%
ChatGLM2-Prefix-LoRA	48.1%	52.4%	42.1%	38.7%

通过表 3-5，可以发现 ChatGLM2-Prefix-LoRA 模型经过联合微调训练后，可以进一步提高测试样本的 BLEU 和 ROUGE 评分。与 ChatGLM2-Prefix2 模型相比，该模型可以将 BLEU-4 评分提高约 4%，将 ROUGE-1 评分提高约 3%，实验结果验证了 Prefix-LoRA 联合微调方法在粉笔字规范性书写对话数据集上的有效性。

基于以上实验结果，联合微调能提升实验效果的主要原因是 Prefix 微调通过在输入序列前添加特定任务的指令或提示，引导模型生成与任务相关的输出，这种方法可以让模型在不改变预训练参数的情况下，快速适应新任务。在 Prefix 微调之后，模型已经对任务有了初步的理解和适应，使得 LoRA 再继续微调时更加有效，模型可以从不同的角度看待问题，从而提高泛化能力。联合使用这两种方法进行微调训练，可以充分发挥各自的优势。Prefix 方法可以利用原始文本中的上下文信息，有助于模型生成与参考答案更一致的文本，LoRA 方法使模型能够在无监督条件下通过自监督学习生成更高质量的文本。Prefix 微调主要影响模型的前几层，而 LoRA 微调则在这些层的基础上进一步调整。这种多层次优化有助于模型更好的理解和处理新任务。

3.4.6 领域知识问答性能评估

与传统的命名实体识别（NER）[50]或分类问题不同，评估微调后的大语言模型在“记忆”粉笔字模板字典中的数字和坐标信息能力时，无法按照划分训练集和测试集方法来进行评分。因为每个汉字提取出的特征信息都是唯一的，不存在可以划分的独立标签，因此采取 3.4.2 中提出的新的评估方式来检验模型的“记忆”能力，在测评时，选择联合微调后的 ChatGLM2-Prefix-LoRA 模型进行评测，随机选择训练语料中的 1000 条语料进行验证，实验结果如表 3-6 所示：

表 3-6 模型记忆能力评估结果

评估指标	数值
TP（真正例）	533
FN（假负例）	467
召回率	53.3%

可以发现，召回率不足百分之六十，下面结合真实的问答案例进行具体分析，选择了六条对话数据进行展示，主要包含微调前数据集中的标准答案和微调后模型给出的回答，具体内容如表 3-7 所示：

表 3-7 真实问答案例

问题	标准答案	微调后的答案
在粉笔字评判体系中标准的读字大小是多大？	在粉笔字评判体系中标准的读字大小为 14002	我了解到在粉笔字评判体系中，标准的字体尺寸设定为 14002。
那这个字宽是多宽？	在粉笔字评判体系中标准的读字宽为 331	我确认在粉笔字评分标准中，标准的字体宽度为 405。
字高是多高？	在粉笔字评判体系中标准的读字高为 405	我了解到在粉笔字评判体系中，标准的字体高度为 405。
整字的重心坐标是多少？	在粉笔字评判体系中标准的读字重心坐标为(311,423)	我得知在粉笔字评判体系中，标准的字体重心坐标定位在 405。
是什么结构？	在粉笔字评判体系中标准的读字的结构为是 LR 结构	我找到的信息表明，在粉笔字评判体系中，标准的字体结构属于 LR（左右）结构。
书写技巧呢？	“读”，左窄右宽，左占 1/3，右占 2/3。言字旁点和折相对不相连，折要短而直。右部“卖”的上竖和竖撇对正，最后一笔是右点。整个字左收右放。	我了解到“读”字在粉笔字评判体系中的书写规范如下：该字采用左窄右宽的结构，左侧占据字宽的 1/3，右侧占据 2/3。言字旁的点和折笔不相连接，且折笔应短而直。右侧的“卖”字部分，上竖和竖撇需对齐，最后一笔为右点。整体上，字形的左侧紧凑，右侧舒展。

召回率低的原因主要是由于数据集中的数字和坐标信息在模型微调阶段未能被充分记忆。这一现象可能由以下几个因素造成：

（1）在语言模型中，词通常被表示为词向量[51]，这些向量捕捉了词的语义和上下文信息。然而，在语言模型中数字和坐标通常被视为离散的符号，这意味着模型在处理数字时，可能不会考虑到数字的数值属性，如大小、顺序或与其他数字的关系。

（2）构建的数据集语境不足，微调过程中可能未能提供足够的语境信息，使模型无法正确理解数字和坐标信息所处的语境。数字和坐标通常需要周围文本的上下文来正确理解其含义，语境信息不足，模型可能会记忆不准确。

虽然微调后的模型对数据集中的数字和坐标信息的“记忆”能力没有达到预期，但通过问答案例可以发现，模型对数据集中粉笔字规范书写的理论知识和书写技巧的学习是成功的，因此，微调仍然具有实际意义。针对微调后的大语言模型无法精准记忆数字和坐标的情况，本研究拟采用检索问答的方式来弥补这个缺点，具体内容将在第四章进行详细阐述。

3.5 本章小结

本章主要介绍了粉笔字规范性书写的领域知识注入 ChatGLM2-6B 大语言模型的过程，着重介绍了微调过程中改进的多轮对话训练方式和联合微调方法，通过对比实验验证了改进方法的有效性。为评估微调后的大语言模型的“记忆”能力，使用了一种新的评估方式，并结合具体的问答实例分析影响召回率的原因，实验结果表明，微调后的模型基本掌握粉笔字规范性书写的理论知识和书写技巧。

第 4 章　外部知识检索增强生成研究

4.1 引言

微调固然效果好，可以让模型真正的“学会”一些私域知识。但是微调也会带来几个问题：首先，由于生成模型依赖于内在知识（权重），因此模型还是无法摆脱幻觉的产生，在对理解门槛高且准确性要求严格的场景下，这是完全无法接受的，因为用户很难从回答的表面看出模型是否在胡说八道。其次，在真实场景中，每时每刻都在产生大量数据，对一个事物的概念会迭代的飞快，如某个指标的调整等。而模型微调并不是一个简单的工作，无论是从数据准备、算力资源、微调效果、训练时间等各个角度来看，随时用新产生的数据来进行微调都是不现实的，能够做到每月更新一次都已经是很理想的状态，且最终微调的效果也无法保证。

另一种解决方案是外部知识检索增强生成(RAG)技术。它为生成式大语言模型与外部世界的互动提供了一个很有前景的方法[52]。RAG 的主要作用类似搜索引擎，找到用户提问最相关的知识或者是相关的对话历史，并结合原始提问，创造信息丰富的 prompt，指导模型生成准确输出。其本质上应用了情境学习的原理。

4.2 RAG 经典流程

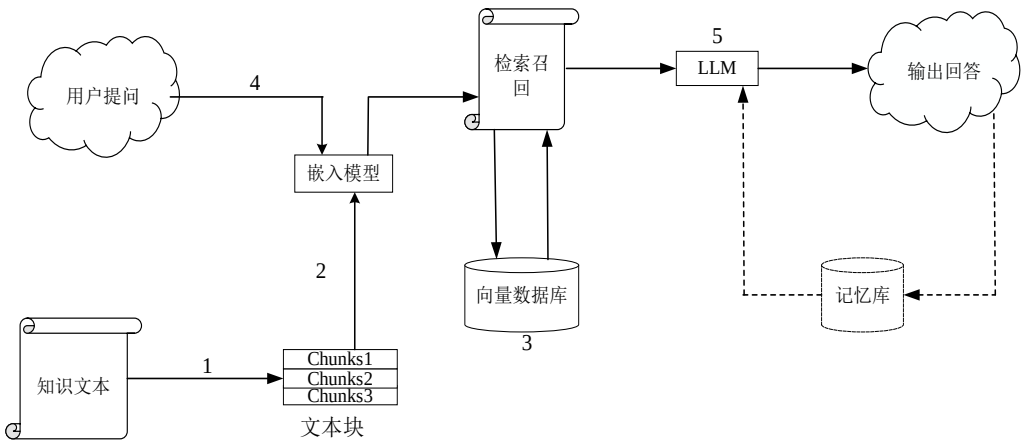

图 4-1 RAG 流程图

图 4-1 所示为 RAG 的经典流程，可分为 5 个基本流程：知识文档的准备，嵌入模型（embedding model），向量数据库，查询检索和生成回答。

在构建一个高效的 RAG 系统时，首要步骤是准备知识文档。现实场景中，面对的知识源可能包括多种格式，如 Word 文档、CSV 数据表、图片和视频等。因此，第一步需要将这些丰富的知识源转换为大语言模型可理解的纯文本数据。此外，鉴于文档可能存在过长的问题，需要将长篇文档分割成多个文本块（chunks），以便更高效的处理和检索信息。这不仅有助于减轻模型的负担，还能提高信息检索的准确性。

嵌入模型作为连接用户查询和知识库的桥梁，核心任务是将文本转换为向量形式，这些向量之间的相似度用于衡量它们的关联程度，嵌入模型通过复杂的网络结构捕捉更深层次的语义关系，确保系统回答的准确性和相关性。在 RAG 系统中，通过嵌入模型生成的所有向量都会被存储在向量数据库中。向量数据库优化了处理和存储大规模向量数据的效率，使得在面对海量知识向量时，能够迅速检索出与用户查询最相关的信息。用户的问题会输入到嵌入模型中进行向量化处理。然后，系统会在向量数据库中搜索与该问题向量语义上相似的知识文本或历史对话记录并返回。

获取到检索的结果后，在进入大模型之前，还需要通过 prompt 工程把检索结果、用户问题以及其他辅助信息进行有机拼接送入模型以获取最终答案。整体流程如下所示：

RAG_PROMPT=“你是一名出色的粉笔字规范书写教师，请根据用户提问和已知的参考资料进行回复，不要胡编乱造。

用户提问：{}

参考材料：{}”

从结果来看，prompt 工程是一个字符串拼接工作，但已经有研究证明，合理的 prompt[53]，能让大模型返回的结果更加优秀且符合预期。

4.3 改进 RAG

使用经典的RAG流程在构建粉笔字规范性书写对话系统时，主要存在两个问题：（1）不能正确检索到所需要的信息，例如询问“城”字笔画 1 为什么偏长？向量检索（Vector search）无法精准检索学生书写笔画 1 的相关长度信息，返回的是其他笔画的信息。（2）检索到的信息能够存到文档中，但由于排名不高，无法有效提供给大型模型进行后续处理。

虽然向量检索在语义搜索能力中具有明显优势，但在某些情况中效果不佳。比如搜索缩写词或短语（例如 RAG、RLHF），搜索 ID（例如笔画 1、笔画 2），此外，

向量检索取决于生成的向量嵌入质量，并对领域外术语敏感。而上述缺点正好是传统关键词检索（Keyword search）的优势所在。传统关键词检索擅长精确匹配（如汉字名称）。对于大多数文本检索的情境，需要精确的关键字匹配，在检索时首要是确保潜在最相关的结果能够出现在候选结果中。

针对传统RAG面临的两个问题，本研究采用了混合检索加重排序策略进行优化。具体流程如图 4-2 所示，将基于关键字的检索和向量检索结合为混合检索，以此提高检索结果的相关性。混合检索[54]能够检索大量上下文，但并非所有上下文都与问题相关。因此，通过重新排序流程对文档进行重新排序和过滤，将相关文档放在最前面，提高 RAG 系统的有效性。

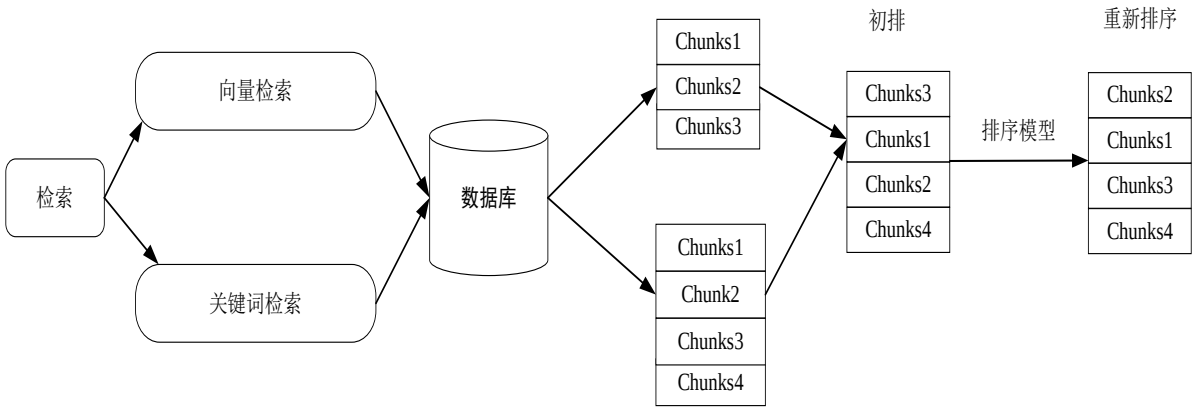

图 4-2 混合检索加重排序流程图

4.3.1 混合检索

向量检索和关键词检索在检索领域各有其优势，混合检索正是结合了这两种搜索技术的优点，同时弥补了两者的缺陷。在用户输入问题时，通过两种检索模式分别在文档中检索出最相关的内容（见图 4-2），不同的检索系统各自擅长寻找文本（段落、语句、词汇）之间不同的细微联系，包括精确关系、语义关系等。

在混合搜索中，基于关键词的搜索通常使用一种叫做稀疏嵌入的表示法，因此也被称为稀疏向量搜索。稀疏嵌入是指大部分值为零，只有少量非零值的向量。最常用的稀疏嵌入算法是 BM25（Best match25），它建立在 TF-IDF（词频-逆文档频率）方法的基础上，通过引入额外的参数来提高搜索相关性。BM25 算法的核心思想是为查询中的每个术语赋予一个权重，该权重取决于术语在文档中的频率和在语料库中的分布。BM25 算法的具体公式如（4-1）、（4-2）所示：

$$IDF(q_i)=\log(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}) \tag{4-1}$$

$$BM25(D,Q)=\sum_{i=1}^{n}IDF(q_i)\frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{\text{avg}DL})} \tag{4-2}$$

其中，$IDF(q_i)$是词q_i的逆文档频率，N是语料库中文档的总数，$n(q_i)$是包含词q_i的文档数，D是待评分的文档，Q是查询，q_i是查询中的第i个词，$f(q_i,D)$是词q_i在文档D的频率，|D|是文档D的长度，avgDL是语料库中文档的平均长度。k_1和b可是调节参数。

向量检索是随着机器学习的发展而出现的一种现代搜索技术。机器学习算法如 Transformer 可以生成数据对象在各种模式（文本、图像等）中的数字表示，向量嵌入通常信息密集，且大多由非零值（密集向量）组成[55]。因此，向量检索也被称为密集向量搜索。检索查询被嵌入到与数据对象相同的向量空间中，为了找到与查询最相似的数据对象，通常会计算查询向量与数据对象向量之间的余弦相似度。假设用户输入表示为二维向量 a，数据对象的向量表示为 b,具体计算方式如图 4-3 所示：

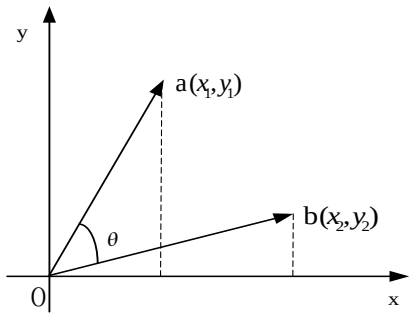

图 4-3 余弦相似度

余弦距离相似度计算公式如（4-3）所示：

$$\cos\theta=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^{\,2}+y_1^{\,2}}\sqrt{x_2^{\,2}+y_2^{\,2}}} \tag{4-3}$$

若向量维度为 n，则向量 A 与 B 之间的余弦距离计算方式如公式(4-4)所示：

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2 * \sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| * |B|} \quad (4\text{-}4)$$

余弦相似度通过计算两个向量之间的夹角来评估它们的相似性，夹角越小，相似度越高。数据对象根据与查询向量的相似度得分进行排序，最相似的排在最前面，从而提供给用户最相关的搜索结果。

4.3.2 初排

混合检索会返回更多更好的结果，但是不同检索方式返回的查询结果需要合并和归一化，以便后续的重新排序阶段进行处理。使用公式（4-5）对检索结果进行混合评分：

$$hybrid\_score = (1 - alpha) * sparse\_score + alpha * dense\_score \quad (4\text{-}5)$$

sparse_score 是关键词检索结果的评分，dense_score 是向量检索结果的评分。参数 alpha 的取值范围在 0 到 1 之间。alpha=0.5 时计算过程如图 4-4 所示：

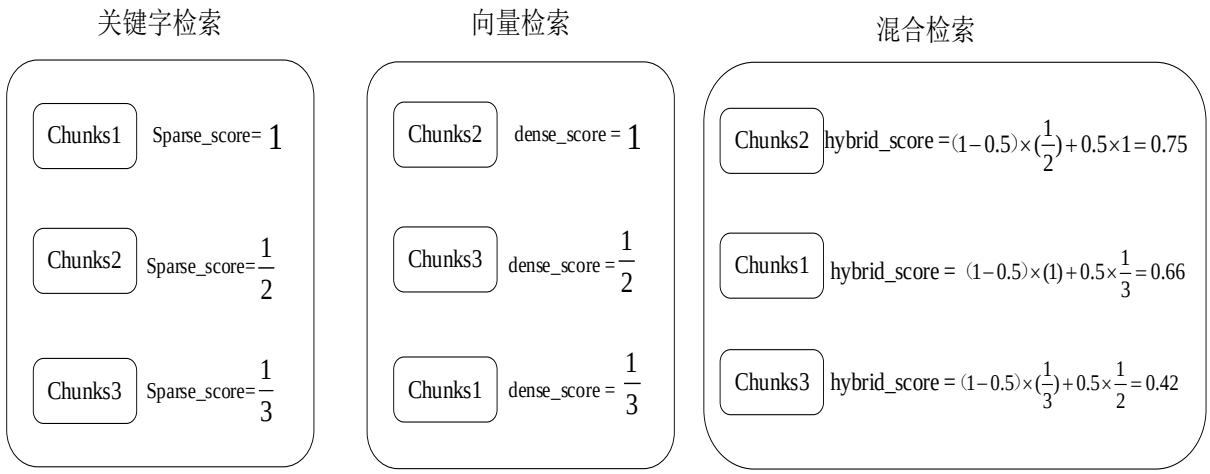

图 4-4 混合检索结果合并图

4.3.3 重排序

重排序模型通过将候选文档列表与用户问题的语义匹配度进行重新排序，从而改进混合排序的结果。当检索器从索引集合中检索出多个上下文时，这些上下文可能与用户的查询具有不同的相关性。重新排序的任务是评估这些上下文的相关性，并优先选择最有可能提供准确和相关答案的上下文。这样，LLM 就能在生成答案时优先考虑这些排名靠前的上下文，提高答案的准确性和质量。

在经典 RAG 系统中，基于向量模型检索的结果，得分最高的文档并不总是意味着它是最相关的。这是因为检索阶段采用的向量模型都是双向编码器[56]，如图 4-5 所示：

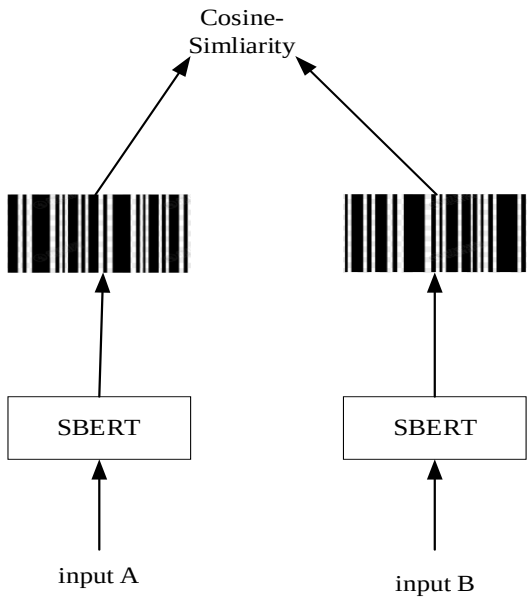

图 4-5 双向编码器

双向编码器将输入文本编码成固定长度的向量，当计算两个句子的相似性时，将两个句子编码成两个向量，然后计算它们之间的相似性。双向编码器在训练时通常会使用相似性损失函数，如公式（4-6）所示：

$$L(\mathrm{x}_1,\mathrm{x}_2,y)=\frac{1}{2N}\sum_{i=1}^{N}y_i d^2+(1-y_i)\max(m-d^2,0) \qquad (4\text{-}6)$$

其中，x_1和x_2是一对输入样本的编码向量，y_i是样本对的标签，1 表示相似，0

表示不相似，d 是 x_1 和 x_2 之间的欧氏距离，m 是一个超参数，表示损失函数的间隔（margin）,如果 y_i=1，则 d 越小越好；如果 y_i=0，则希望 d 大于间隔 m.

该损失函数会鼓励模型将正样本（相似的句子对）的嵌入向量拉近，同时将负样本（不相似的句子对）的嵌入向量推远，模型通常是针对单个句子或固定长度的文本片段进行训练的，由于每个句子的嵌入都是相互独立的，双向编码器关注的是单个句子的内部结构和语义，并没有明确的学习如何处理句子之间的关联。它能够很好的理解句子中的词语如何相互作用，但对于不同句子之间的逻辑关系等信息，不会有显式的表示。

交叉编码器是同时编码两个句子，并输出一个分类分数，表示这两个句子是否相关。图 4-6 展示了交叉编码器和双向编码器之间的区别：

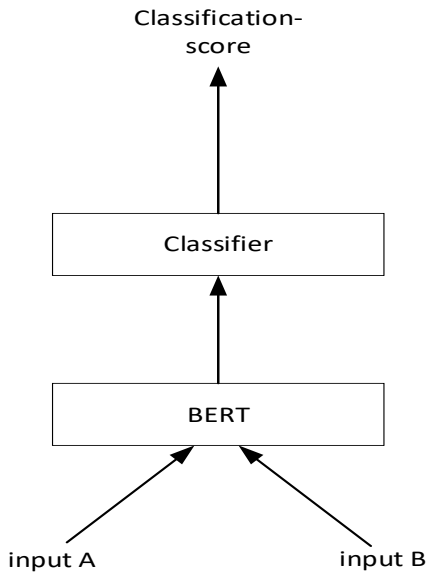

图 4-6 交叉编码器

交叉编码器在训练时使用交叉熵损失函数（Cross Entropy Loss），具体计算方式如公式（4-7）所示：

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \quad (4\text{-}7)$$

其中，y 是实际的标签，$\hat{y}_i$ 是模型的预测概率分布，N 是样本的数量，对于多分类问题，交叉熵损失函数的表达式略有不同，但思想是相似的。

交叉编码器专门针对句子对之间的关联进行设计和训练。在处理句子对时，交叉

编码器不仅考虑每个句子内部的词语相互作用，而且还关注两个句子之间的相互作用和关系。

综合以上分析可知，双向编码器能够独立的对每个输入句子进行编码，在处理大量句子时可以实现并行化处理，这种特性对于搜索引擎等需要快速索引和检索大量文本的应用场景非常有利，由于可以同时处理多个句子，双向编码器能够提高处理速度和效率。相对而言，交叉编码器在处理句子对时速度较慢，且需要更多的内存资源。这是因为交叉编码器需要同时考虑两个句子之间的相互作用，通常涉及到更复杂的计算。然而，这种细粒度的交互建模使得交叉编码器在一些需要高精度分类和排序的任务上表现出色。

因此，在研究中，先使用双向编码器进行初步筛选，以快速缩小候选集。随后，再利用交叉编码器对这些筛选后的句子进行更精确的重排序，以提升最终结果的准确性和相关性。这种两阶段的方法结合了双向编码器的效率和交叉编码器的精确度，能够在实际应用中取得较好的平衡。

4.4 实验及结果分析

4.4.1 实验环境及数据

本研究的实验环境分为硬件环境与软件环境，具体配置如表 4-6 所示：

表 4-6 软硬件配置

名称	配置
处理器	NVIDIA GeForce RTX 4080
操作系统	Windows
显存大小	16G
编程语言	Python 3.8
深度学习框架	Pytorch2.0

RAG 对显存的需求低于大模型微调，主要原因是 RAG 中的检索组件从文档库中检索相关信息时，这些信息通常存储在硬盘上，而不需要加载到显存中。生成组件只需要处理检索到的信息，而不是整个文档库，RAG 的显存需求主要体现在大模型的加载和推理上。相比之下，大模型的微调通常需要将整个模型和数据集加载到显存

中以便计算梯度并进行参数更新，导致显存需求显著增加。

在构建 RAG 检索系统时，原始的知识文本通常会被分割成更小的片段。固定长度的分割策略可以提高处理效率，因为模型可以并行处理多个片段，同时减少了对内存和计算资源的需求。但分割可能会导致文本的语义断裂，特别是在分割点位于重要短语或句子中间时，会导致模型生成的内容缺乏必要的上下文信息，影响模型的生成质量。因此在本研究中将图片中的特征信息转为文本时[57]，使用了关键信息前缀策略，以此来避免在切割时丢掉信息，具体如图 4-7 所示：

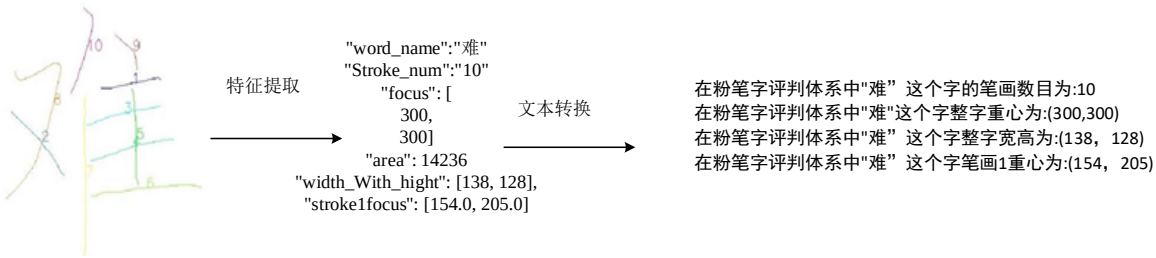

图 4-7 RAG 文本信息转换过程

4.4.2 评价指标

RAG 在自然语言处理领域的快速发展和日益普及已经将 RAG 模型的评估推到了 LLM 社区研究的前沿。评估的主要目标是理解和优化 RAG 模型在不同应用场景下的性能。对于 RAG 模型的评估主要围绕着两个关键组成部分：检索模块和生成模块。这种分工确保了对提供的上下文质量和生成的内容质量进行全面评估。

（1）评估检索质量对于确定检索器组件提供的上下文的有效性至关重要。来自搜索引擎、推荐系统和信息检索系统领域的标准度量指标被用来衡量RAG检索模块的性能。常用的指标包括Hit Rate（命中率）和MRR（Mean Reciprocal Rank）。本研究使用大模型应用开发框架Llama Index—Retriever Evaluator接口进行检索质量评估，为其提供测试数据集，向量数据库，向量化模型，就能自动的进行评估，输出命中率和平均倒数排名。

Hit Rate 衡量的是系统返回的结果中，有多少比例是相关的或者用户认为是有用的。Hit Rate 的计算公式如（4-8）所示：

$$\text{Hit R}ate = \frac{\text{命中次数}}{\text{总查询次数}} \tag{4-8}$$

其中，命中次数是指系统返回的结果中被用户认为是相关或有用结果的次数，总查询次数是指用户发起的总查询次数。

MRR 衡量的是系统返回的最相关的结果在整个结果列表中的位置的倒数平均值，它更关注最相关结果的位置。MRR 的计算公式如（4-9）所示：

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{4-9}$$

其中，Q是查询的总数，$rank_i$ 是第i个查询的最相关结果的排名位置。MRR的值范围从 0 到 1，其中 1 表示每个查询的最相关结果都排在第一个位置，而 0 则表示没有查询的相关结果排在前面的位置。MRR越接近 1，说明系统的性能越好。

（2）使用开源的RAGAS框架[58]评估RAG系统的生成质量。该框架考虑LLM以忠实方式利用检索到的上下文的能力，以及生成文本的质量。RAGAS需要以下信息：

Question:用户输入的问题。

Answer:从RAG系统生成的答案(由LLM给出)。

Contexts:根据用户的问题从外部知识源检索的上下文即与问题相关的文档。

ground_truths:人类提供的基于问题的正确答案，这是唯一的需要提供的信息。

评估改进RAG的生成质量时主要使用RAGAS框架中的忠实度和答案相关性指标，下面进行详细介绍：

忠实度衡量了生成的答案与给定上下文的事实一致性。它是根据answer和检索到的context计算得出的。并将计算结果缩放到（0,1）范围，计算结果越高，表示在生成答案时，大模型更多的使用了检索到的内容。

如果答案中提出的所有基本事实都可以从给定的上下文中推断出来，则生成的答案被认为是忠实的。为了计算这一点，首先从生成的答案中识别一组事实，然后，将这些事实中的每一项与给定的上下文进行交叉检查，以确定是否可以从给定的上下文中推断出它。忠实度分数由公式（4-10）得出：

$$忠实度分数=\frac{生成答案事实总数}{给定上下文推断出的事实数量} \tag{4-10}$$

答案相关性重点评估生成的答案与用户问题之间相关程度，不完整或包含冗余信息的答案将获得较低分数。该指标是通过计算 question 和 answer 获得的，它的取值范围在 0 到 1 之间，其中分数越高表示相关性越好。在原论文中作者给出的具体计算步骤如下：

（1）使用text-embedding模型为所有问题获得嵌入（embeddings）。

（2）对于每个问题 q_i 计算它与原始问题q之间的相似度sim(q, q_i)，即对应嵌入之间的余弦值。

（3）计算原始问题 q 的答案相关性分数 AR，计算方式如公式（4-11）所示：

$$AR(\mathrm{q})=\frac{1}{n}\sum_{i=1}^{n} sim(q,q_i) \tag{4-11}$$

当答案直接且适当地解决原始问题时，该答案被视为相关。评估的重点不是答案是否反映了真实情况，而是答案是否完整的回答了原始问题，并且是否避免了包含不必要或多余的细节。为了计算这个分数，LLM会被提示多次为生成的答案生成适当的问题，并测量这些生成的问题与原始问题之间的平均余弦相似度。基本思想是，如果生成的答案准确的解决了最初的问题，LLM应该能够从答案中生成与原始问题相符的问题。

4.4.3 检索阶段实验

（1）单一模型召回实验

在当前的RAG流程中，众多先进的向量嵌入模型引起了关注，排名靠前的是OpenAI 的 "text-embedding-ada-002" 以及 BAAI 的 "BGE Embedding" 系列中的 "BAAI/BGE-base"和"BAAI/BGE-large"。这些模型在处理文本数据时各有所长。实验中需要根据具体任务需求，挑选合适的向量模型。本研究基于构建的数据集，在RAG框架的Retrieve阶段中，首先对以上三款Embedding模型算法进行了评估，为方便评估结果能够直观展示，在实验中使用Plotly模块绘制指标的条形图，结果如图 4-8 所示：

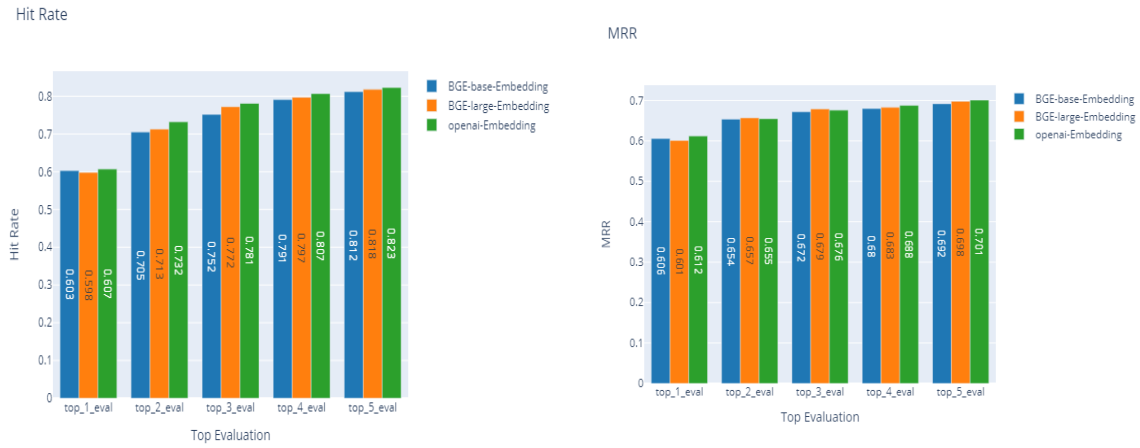

图 4-8 单一模型实验结果图

从图 4-8 中可以看出：对于三种Embedding模型，随着k的增加，top_k的Hit Rate指标和MRR指标都有所增加，BGE-large-Embedding模型top_5 与top_1 相比，Hit Rate提升约 18%左右，MRR指标提升 10%左右。BGE-large-Embedding模型的召回效果比OpenAI-Embedding模型会好一些，但提升不大,同等情况下，二者的Hit Rate指标和MRR指标相差在 0.07 左右。综合考虑之下，后续的实验选择BGE-base-Embedding模型作为基线向量检索模型。

（2）混合检索实验

为验证混合检索对RAG检索性能的影响，进行了消融实验，分别使用关键词检索即BM25 算法，BGE-base-Embedding向量检索和两者的混合检索进行对比，混合检索时参数alpha的取值为 0.5。混合检索的消融实验结果如图 4-9 所示：

图 4-9 混合检索实验结果图

从图 4-9 可以看出：就单独的检索算法而言，本研究中BM25 的检索效果比向量检索好，这可能与生成的问答来源于文档有关，但这不是普遍结论，两种算法各有合适的场景。就总体检索效果而言，混合检索要优于单独检索算法，混合检索与BM25相比，在top_1 时，HitRate提升约 3%，MRR指标提升了 2%左右，上述实验结果可以验证在RAG流程中使用混合检索测量的有效性。

（3）重排序实验

目前，没有太多可用的重新排序模型。Cohere Rerank 模型目前闭源，对外提供API，普通账号提供免费使用额度。bge-reranker 是 BAAI（北京智源人工智能研究院）发布的系列模型之一，bge-reranker 模型在 Hugging Face 上开源，有 base、large 两个版本模型。重排序实验中选取 BM25+bge-base-Embedding 作为基线模型，分别与三种排序模型组合，进行对比试验。实验结果如图 4-10 所示：

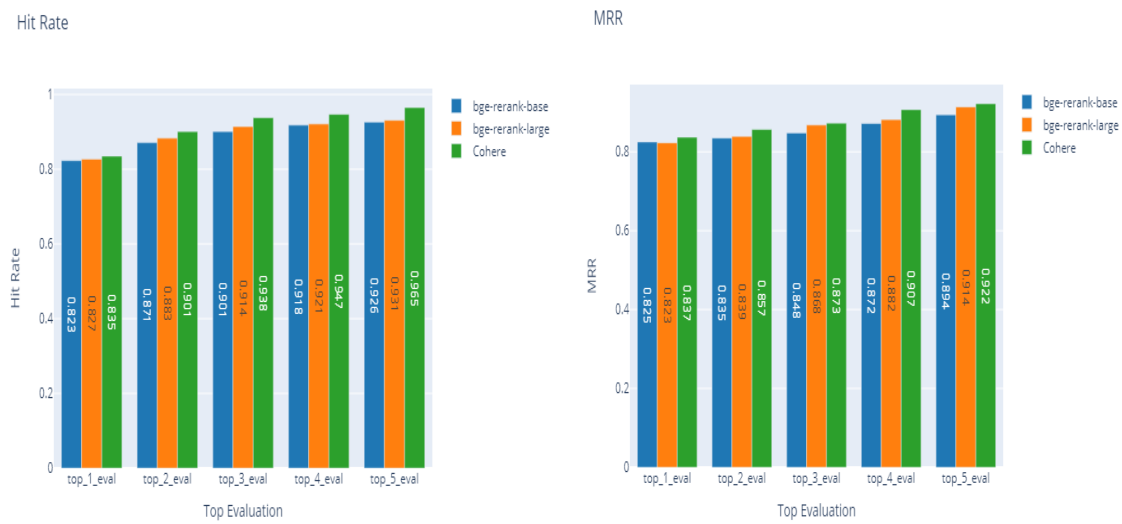

图 4-10 重排序实验结果图

从图 4-10 可以看出：在混合检索阶段后加入重排序模型后检索效果会有明显提升，就检索效果而言，重排序模型的结果为：Cohere>bge-rerank-large>bge-rerank-base，在 top_5 时，最好情况的 Cohere 的 Hit Rate 已经提升至 0.965，MRR 指标提升至 0.922。与经典 RAG 中仅使用单一向量检索相比，采用混合检索加重排序的改进 RAG 有明显的提升。这说明在 RAG 流程中，结合多种检索策略并使用重排序模型可以有效提高检索信息的准确性和相关性。

此外，为进一步验证在 RAG 中引入重新排序流程的重要性，还研究了在单一检

索模式下进行重排序的对比试验。分别选取 BM25 和 bge-base-Embedding 作为单一检索模型，重排序模型使用 bge-rerank-base 模型。实验结果如图 4-11 所示：

图 4-11　单一检索重排序实验结果图

从图 4-11 可以看出：即使在单一检索模式下，引入重新排序步骤也能提升 RAG 的检索效果。例如在进行关键词检索之后再加入重排序，与单一的 BM25 算法相比，在 top_1 时，Hit Rate 提升约 5%，MRR 指标提升约 4%。

4.4.4 生成阶段实验

使用 RAGAS 进行测评时，选取了 100 条问答数据，对于每条问答的上下文信息，选择 RAG 检索结果排名的前五条。进行了三组对比试验，分别为 RAG1（单一向量检索+ChatGLM2-6B）、RAG2（单一向量检索+ChatGLM2-Prefix-LoRA），改进的 RAG（混合检索+重排序+ChatGLM2-Prefix-LoRA），向量检索模型为 BGE-base-Embedding,排序模型为 bge-rerank-base，测评结果如表 4-7 所示：

表 4-7 RAGAS 测评结果

名称	忠实度	答案相关性
RAG1	0.817	0.752
RAG2	0.835	0.754
改进的 RAG	0.841	0.811

从表 4-7 的测评结果可以看出：三者的忠实度分数普遍得分较高，主要原因是 RAG 在回答问题时，有提示语句“根据用户提问和已知的参考资料进行回复，不要胡编乱造”，大模型在有“参考资料”作为参考时基本会遵循指令提示。RAG2 的忠实度分数要比 RAG1 的忠实度分数高 2%左右，说明在构建 RAG 系统时，大语言模型的微调也是有必要的。改进的 RAG 答案相关性的分数要比 RAG1 高 6%左右，说明改进的 RAG 通过提升检索质量，生成的答案基本可以完整的回答原始问题。

从上述的实验可以看出，大模型的 RAG 技术和微调并不冲突[59]。RAG 的优势在于能够快速更新知识库，仅通过更新数据库来反映最新信息，无需重新训练模型，在稳定性和解释性方面表现较好，因为其生成的回答基于检索到的具体事实。微调的优势在于能够学习特定领域的深入知识，对于简单任务，可能达到比 RAG 更高的性能。RAG 通过更新数据库来更新知识，微调则是通过重新训练来吸收新知识。RAG 通常在生成回答时更稳定，而微调能达到更高的性能上限。微调在训练时消耗资源较多，RAG 在推理时增加额外的检索成本。

RAG 和微调各自有优势，在构建粉笔字规范性书写对话系统时，首先使用微调，让模型适应特定领域知识，然后使用 RAG 来补充微调未覆盖的知识。通过这样的结合使用，可以充分利用 RAG 的快速知识检索能力和微调的深度知识学习能力，提升模型在复杂任务上的表现。同时，也可以帮助平衡实时性、准确性和资源消耗等多方面的需求。

4.4.5 RAG 问答案例

在部署 RAG 时，使用 LangChain 框架。LangChain 框架是一个开源工具，它的目标是为各种大型语言模型提供通用接口，便于开发者根据需要选择合适的模型，从而简化应用程序的开发流程。LangChain 框架能够连接大语言模型与其他数据源，比如知识库或数据库，这样 RAG 模型在生成回答时可以更加准确的引用外部信息。此外 LangChain 框架支持语言模型与外部环境交互，例如，在对话系统中，模型可以根据用户的反馈来调整自己的回答，更好地实现人机对话的连贯性和自然性。

下面展示部署的 RAG 回答学生问题的过程，在本研究中，学生练习粉笔字是以诗为单位进行练习，图 4-12 为真实书写的粉笔字，左侧为教师书写的粉笔字，右侧为学生练习时书写的粉笔字。图 4-13 为现有的师范生粉笔字书法训练自动评判系统对学生书写的“盖”字生成的部分评语。

图 4-12 粉笔字

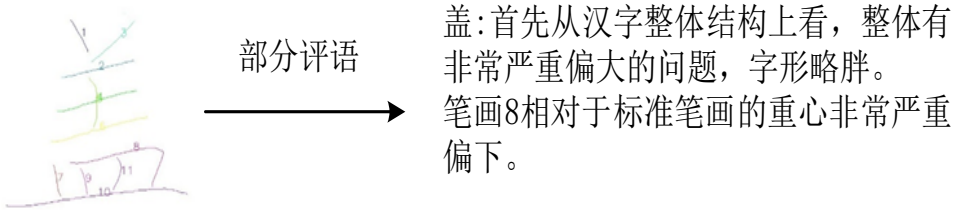

图 4-13 评语

改进的 RAG 系统会将教师书写粉笔字和学生书写粉笔字的特征信息即 JSON 文件进行文本信息转换，转换时为防止后续切割文本丢失关键信息，采用上文提到的关键信息前缀策略进行转换，转换后的文本信息写入文本文件中，此文本文件为 RAG 后续问答的“知识库”。学生提出问题后，改进的 RAG 会依据学生的问题在文本文件中检索相关信息，经过混合检索和重排序流程后，RAG 会将检索到的有用信息传给大模型充当上下文“参考资料”，通过适当的提示语句最终让大模型给出合理的回复。

表 4-8 为引入 RAG 后，对话系统对当前评语做出的解释。可以看出，与微调相比，RAG 会通过检索相关信息进行回答，检索增强生成问答能有效缓解大模型的知识“幻觉”现象。原始的 RAG 系统并没有依据真实数据回答，而是简单的回复了评判的标准，主要原因是检索时并没有正确检索到学生书写粉笔字的特征信息，导致后续的回复阶段没有相关依据。在使用改进的 RAG 后，回答问题时利用检索到的学生书写粉笔字特征信息和教师书写粉笔字的特征信息进行分析和回复。

表 4-8 RAG 问答案例

模型	答案
原 始 RAG	根据文档内容，学生书写“盖”字时，存在以下问题： 1.整体结构：整体字形有严重偏大的问题，具体表现为整体面积明显偏大。 2.高矮胖瘦特征：整体字形略显肥胖，具体表现为外接矩形宽度差大于标准整字胖瘦值的+120，被归类为非常严重偏胖。 3.笔画 8 重心：相对于标准笔画的重心非常严重偏下，被归类为非常严重偏下问题。 综上所述，学生书写的“盖”字整体结构偏大、略显肥胖，且笔画 8 的重心位置与标准存在严重偏移。
改 进 RAG	根据文档内容，学生写的整字面积和整字胖瘦与标准整字存在较大偏差： 1.学生写的整字面积为 201422，标准整字面积为 151285。面积偏移百分比超过 20%，属于“非常严重偏大” 2.学生写的整字胖瘦为 423，标准整字胖瘦为 383。外接矩形宽度差超过 120，属于“非常严重偏胖”。 3.学生写的笔画 8 重心值为[262,270]，标准笔画 8 重心为[318,346]。坐标偏移值超过 60，属于“非常严重偏下”。 4.整字面积和胖瘦的偏差导致字形看起来偏大和偏胖。 综上所述，从汉字整体结构上看，学生写的整字整体有非常严重偏大的问题，字形略胖，笔画 8 相对于标准笔画的重心非常严重偏下。这些因素共同导致字形整体偏大和偏下。

4.5 本章小结

本章主要介绍了利用 RAG 技术来缓解大模型知识“幻觉”的问题，首先介绍了经典的 RAG 流程，接着给出经典的 RAG 流程在本研究中的局限性，并提出混合检索加重排序的改进 RAG。在评估改进 RAG 性能时，分别从检索性能和生成性能进行评估，通过多组对比试验证明改进的 RAG 明显优于经典 RAG，并总结了 RAG 和微调各自的优缺点，最后给出了改进的 RAG 在本研究中的真实问答案例。

第 5 章 总结与展望

5.1 总结

近年来，大语言模型在人工智能领域迅速崛起，成为了研究者和企业争相追逐的热点技术。大语言模型在垂直领域中展现出了巨大的潜力，特别是在对话系统领域。大语言模型能够为用户提供更加智能、个性化的服务，通过训练和优化，这些对话系统能够更好地理解用户的需求，提供准确的回答和解决方案。基于大语言模型的粉笔字规范性书写对话系统研究处于起步阶段，由于领域自身的复杂性和对话语料的稀缺导致研究难以开展。针对以上问题本文展开了如下研究：

首先，对现有的粉笔字字帖字典信息库进行了整理，并整理了一本粉笔字规范性书写教材中的知识点，构建了一个包含 5 万对问答的粉笔字规范性书写对话语料。然后开发了专用的微调数据集生成工具，将语料数据转换为大模型微调所需的格式。

其次，使用参数高效微调技术将粉笔字规范性书写的领域知识注入 ChatGLM2-6B 大语言模型。在微调时，使用了一种比官方更加充分和高效的多轮对话训练方式，并结合微调技术的特点，引入了联合微调策略。为全面衡量微调后的大模型在特定领域的性能，除了使用 BLEU、ROUGE 等常用的评价指标，专门设计了一种新的评估方式，通过借助 ChatGPT 来评估微调后的大模型对领域知识的“记忆”能力。实验结果表明：使用改进的多轮对话训练方式和联合微调策略后，BLEU 和 ROUGE 分数有明显提升，虽然大语言模型在记忆坐标和数值的能力没有达到预期，但微调后的模型基本掌握粉笔字规范性书写的理论知识和书写技巧。

最后，使用了 RAG 技术来缓解大语言模型的知识“幻觉”问题。针对经典的 RAG 在检索过程中遇到的问题，如不能正确检索到所需要的信息以及检索到的信息因排名较低而无法有效提供给大语言模型进一步处理。本研究使用了混合检索加重排序策略来优化 RAG，RAG 结合了检索和生成的特点，评估时分别使用 Hit Rate 指标和 MRR 指标来评估检索质量，使用开源的 RAGAS 框架评估生成质量，并给出了具体的问答案例。实验结果表明，优化后的 RAG 检索质量和生成质量都优于经典的 RAG。

5.2 展望

本研究在构建基于大语言模型的粉笔字规范性对话系统时取得了具体成果，可目前仍然存在缺陷，还有很大的发展空间，主要有以下考虑的方面：

（1）数据集质量有待进一步增加：本研究构建的粉笔字规范性书写数据集相对较小，且是采用直问直答的形式进行构建，没有提供充足的语境信息，这可能是导致微调后的模型在记忆数字和坐标不够准确的一个重要原因，在构建数据集时，可以尝试提供更多的上下文信息，让模型能够更好地理解问题的背景和含义。

（2）微调方法：本研究探讨了 Prefix、LoRA 和联合微调方法在构建基于大语言模型的粉笔字规范性对话系统。将来，还可以探索 LoRA+Prefix 的联合微调方法来分析不同的微调序列对模型性能的影响。同时，通过比较其他有效的微调方法，如 P-Tuningv2[60]、Adapter 等，可以进一步证明联合微调方法的优越性。

（3）大语言模型的选择：本研究利用 ChatGLM2-6B 构建领域对话系统。未来，随着越来越多的中文大语言模型开源出来，应使用更多的开源大模型进行实验。

（4）在使用 RAG 缓解大语言模型的幻觉问题时，检索的知识是非结构化知识，非结构化知识在分割时，容易丢失信息，而知识图谱是一种结构化的知识表示形式，这些结构化的信息可以帮助 RAG 在生成文本时提供准确的事实依据。通过检索知识图谱，RAG 模型可以在生成文本时引用图谱中的知识，这样可以提高生成文本的准确性和可靠性，进而给予使用者完整的指导意见。

参考文献

[1] 张墨涵,涂聪,辛春薇.师范生粉笔字书写现状调查——以南宁师范大学师园学院为例[J].西部素质教育,2023,9(10):165-170.

[2] 李泽瑶,李成城.基于结构知识的手写体汉字部件提取算法[J].计算机工程与设计,2023,44(05):1479-1486.

[3] 肖雪,李成城.手写汉字评价方法研究进展[J].计算机工程与应用,2022,58(02):27-42.

[4] 李泽瑶. 基于模板匹配的粉笔字书法自动评分方法研究[D].内蒙古师范大学,2023.

[5] 范勇峰,李成城,林民.基于 K-P 算法优化的手写汉字细化算法[J].计算机工程与设计,2023,44(10):3076-3083.

[6] 范勇峰,李成城,林民等.基于局部信息的手写汉字笔画提取[J].内蒙古师范大学学报(自然科学汉文版),2023,52(02):181-188.

[7] 肖雪. 基于模板提取的手写汉字评价自动生成方法[D].内蒙古师范大学,2023.

[8] Philip Welsby, Bernard M Y Cheung.ChatGPT[J].Postgraduate medical journal,2023,Vol.99(1176): qgad056

[9] 竹倩叶，鄂海红.基于大语言模型的垂直领域问答系统研究[J].新一代信息技术,2023,(17)

[10] 陈跃鹤.基于信息检索的知识图谱问答技术研究[D].苏州大学,2023.

[11] 张晨曦.面向复杂问题的智能问答技术研究与实现[D].电子科技大学,2023.

[12] 特日格勒呼.检索与生成相融合的蒙古文自动问答研究[D].内蒙古师范大学,2022.

[13] 顾佳宸.基于深度学习的检索式对话系统研究[D].中国科学技术大学,2022.

[14] 梁中阁,陈孝如.基于关键词分级检索的 Web 信息访问监控算法[J].计算机仿真,2021,38(11):433-437.

[15] 张伟.基于 GPT-2 模型的生成式对话系统应用研究[D].华东师范大学,2022.

[16] Sutskever I , Vinyals O , Le Q V .Sequence to Sequence Learning with Neural Networks[J].Advances in neural information processing systems, 2014.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[18] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[19] 冯志伟,张灯柯.GPT 与语言研究[J].外语电化教学,2023,(02):3-11+105.

[20] Mallio C A , Sertorio A C , Bernetti C ,et al.Large language models for structured reporting in radiology:

performance of GPT-4, ChatGPT-3.5, Perplexity and Bing[J].La radiologia medica, 2023, 128(7):808-812.

[21] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[22] Penedo G, Malartic Q, Hesslow D, et al. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only[OL].arXiv preprint arXiv:2306.01116, 2023.

[23] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[OL]. arXiv preprint arXiv:2307.09288, 2023.

[24] 赵月,何锦雯,朱申辰,等.大语言模型安全现状与挑战[J].计算机科学,2024,51(01):68-71.

[25] Cui J, Li Z, Yan Y, et al. Chatlaw: Open-source legal large language model with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092, 2023.

[26] Zhang H, Chen J, Jiang F, et al. Huatuogpt, towards taming language model to be a doctor[J]. arXiv preprint arXiv:2305.15075, 2023.

[27] 刘建伟,宋志妍.循环神经网络研究综述[J].控制与决策,2022,37(11):2753-2768.

[28] Wang J , Zhang J , Wang X .Bilateral LSTM: A Two-Dimensional Long Short-Term Memory Model With Multiply Memory Units for Short-Term Cycle Time Forecasting in Re-entrant Manufacturing Systems[J].IEEE Transactions on Industrial Informatics, 2018, 14(2):748-758.

[29] Cho K , Van Merrienboer B , Gulcehre C ,et al.Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J].Computer Science, 2014.

[30] 常锴文.基于 Seq2Seq 架构的问题生成方法研究[D].山西大学,2023.

[31] 陈健威,俞璐,韩昌芝,等.Transformer 在域适应中的应用研究综述[J/OL].计算机工程与应用,1-18[2024-04-02].

[32] 祁宣豪,智敏.图像处理中注意力机制综述[J].计算机科学与探索,2024,18(02):345-362.

[33] 曾亚竹,孙静宇,何倩倩.融合 BiGRU 和记忆网络的会话推荐算法[J].计算机工程与设计,2023,44(02):335-342.

[34] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[35] 文森,钱力,胡懋地等.基于大语言模型的问答技术研究进展综述[J/OL].数据分析与知识发现,1-17[2024-03-26].

[36] 丁鑫,邹荣金,潘志庚.基于高效参数微调的生成式大模型领域适配技术[J].人工智能,2023,(04):1-9.DOI:10.16453/j.2096-5036.2023.04.001.

[37] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]//International conference on machine learning. PMLR, 2019: 2790-2799.

[38] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

[39] Hu E, Shen Y, Wallis P, et al. Low-rank adaptation of large language models[J]. arXiv, 2021.

[40] Wang Z, Li K, Ren Q, et al. Traditional Chinese Medicine Formula Classification Using Large Language Models[C]//2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2023: 4647-4654.

[41] 王婷,王娜,崔运鹏等.基于人工智能大模型技术的果蔬农技知识智能问答系统[J].智慧农业(中英文),2023,5(04):105-116.

[42] 唐婧尧.基于外部知识的多轮对话系统的研究与实现[D].北京邮电大学,2023

[43] 陈锡,陈奥博.基于掩码矩阵-BERT 注意力机制的神经机器翻译[J].现代电子技术,2023,46(21):111-116

[44] 吴娜,刘畅,刘江峰,等.AIGC 驱动古籍自动摘要研究：从自然语言理解到生成[J/OL].图书馆论坛,1-14[2024-04-02].

[45] Zhang Q, Chen M, Bukharin A, et al. Adaptive budget allocation for parameter-efficient fine-tuning[C]//The Eleventh International Conference on Learning Representations. 2023.

[46] Dettmers T, Pagnoni A, Holtzman A, et al.Qlora: Efficient finetuning of quantized llms[J]. Advances in Neural Information Processing Systems, 2024, 36.

[47] Wang X, Ling X, Zhang T, et al. Optimizing and Fine-tuning Large Language Model for Urban Renewal[J]. arXiv preprint arXiv:2311.15490, 2023.

[48] 罗亮.基于 BERT 和外部知识的答案选择模型研究[D].南京邮电大学,2023.

[49] 王乃钰.面向教育领域的对话系统研究[D].吉林大学,2021.

[50] 张天宇,孙媛媛,杜文玉,等.基于语义边界增强的司法命名实体识别[J/OL].清华大学学报(自然科学版):1-11[2024-03-31].

[51] Cambria E, White B. Jumping NLP curves: A review of natural language processing research[J]. IEEE Computational intelligence magazine, 2014, 9(2): 48-57.

[52] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[53] 王方方，AI 提示工程师引发的思考[J].科技与金融,2023,(06):51-52.

[54] 陈承.面向数学公式及上下文的混合检索模型研究[D].华东师范大学,2022.

[55] Ram P, Gray A G. Maximum inner-product search using cone trees[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 931-939.

[56] 郑少帅,翁境鸿,蒋小洋.基于 BM25、文本 Embeddings 与交叉编码器的民航客服知识库检索研究[J].无线互联科技,2023,20(24):122-125.

[57] 李雅红,周海英,徐少伟.基于对象关系网状转换器的图像描述模型[J].计算机工程,2021,47(05):197-204.

[58] Es S, James J, Espinosa-Anke L, et al. Ragas: Automated evaluation of retrieval augmented generation[J]. arXiv preprint arXiv:2309.15217, 2023.

[59] 姜雨杉,张仰森.大语言模型驱动的立场感知事实核查[J/OL].计算机应用:1-9[2024-03-31].

[60] Liu X, Ji K, Fu Y, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.