|  |
| --- |
| 编号： 输入学号 |
|  |
|  |
| 毕业论文（设计） |
|  |
| 题目 输入你的题目，会自动调整为居中，不要自己调整 |
| 院 （系） 人工智能学院 |
| 专 业 机械工程 |
| 学生姓名 |
| 成 绩 |
| 指导教师 ×××（职称） |
| 2022年 5月 |

**诚信声明**

本人郑重声明：所呈交的毕业设计（论文）是我个人在导师指导下, 由我本人独立完成。有关观点、方法、数据和文献等的引用已在文中指出,并与参考文献相对应。

我承诺，论文中的所有内容均真实、可信。如在文中涉及到抄袭或剽窃行为，本人愿承担由此而造成的一切后果及责任。

毕业论文（设计）作者签名：

签名日期： 年 月 日

摘 要

在后面填加摘要正文，然后删掉此句

另有一个模板使用说明，模板中有不明白的事可以看说明。看完说明可以把段删除，凡说明中提及的情况，有不明白者，自行百度，不接受咨询！

毕业论文期间，凡是此模板与黄本规定相矛盾的地方，以黄本为准，凡是黄本中没有提及的地方，以此模板为准。

摘要第一段写研究背景意义。

摘要第二段写研究了那些内容，与文章主要研究内容章节对应，可以分段，每段一章。设计题目请写明最终设计指标，实验研究请给出数值描述的关键实验结果及结论。

摘要最后一段总结展望，不要超过三行。

关键词：关键词1；关键词2；关键词3；关键词3

Abstract

在后面填加摘要正文，然后删掉此句

注意与中文摘要内容对应。

Keywords：关键词1；关键词2；关键词3；关键词3

目 录

[第一章 绪论 1](#_Toc104506215)

[1.1 第一节标题 1](#_Toc104506216)

[1.1.1 第二节标题 1](#_Toc104506217)

[1.1.2 第二节标题 1](#_Toc104506218)

[1.2 本文研究内容 1](#_Toc104506219)

[第二章 章标题 1](#_Toc104506220)

[2.1 第一节标题 1](#_Toc104506221)

[2.2 如何插入公式 1](#_Toc104506222)

[2.3 在第二章插入表格和图片 1](#_Toc104506223)

[2.3.1 第二节标题 1](#_Toc104506224)

[2.4 本章小结 1](#_Toc104506225)

[第三章 章标题 1](#_Toc104506226)

[3.1 第一节标题 1](#_Toc104506227)

[3.2 我要插入公式 1](#_Toc104506228)

[3.2.1 第二节标题 1](#_Toc104506229)

[3.3 本章小结 1](#_Toc104506230)

[第四章 章标题 1](#_Toc104506231)

[4.1 第一节标题 1](#_Toc104506232)

[4.2 第四章插入公式 1](#_Toc104506233)

[4.2.1 第二节标题 1](#_Toc104506234)

[4.3 本章小结 1](#_Toc104506235)

[第五章 章标题 1](#_Toc104506236)

[5.1 第一节标题 1](#_Toc104506237)

[5.1.1 第二节标题 1](#_Toc104506238)

[5.2 我要插入公式 1](#_Toc104506239)

[5.3 本章小结 1](#_Toc104506240)

[第六章 总结与展望 1](#_Toc104506241)

[6.1 第一节标题 1](#_Toc104506242)

[6.1.1 第二节标题 1](#_Toc104506243)

[6.2 展望 1](#_Toc104506244)

[致 谢 1](#_Toc104506245)

[附录A 1](#_Toc104506246)

[附录B 1](#_Toc104506247)

1. 绪论
   1. 研究背景和意义

随着我国经济的快速发展和城市化进程的不断加快，建设项目对环境的影响日益凸显，环境保护已成为全社会共同关注的问题。建筑项目的环境影响评价（简称环评）作为建设项目审批和管理的重要环节，其主要目的是预测和评价建设项目对环境可能产生的影响，制定环境保护措施，确保建设项目与环境保护的协调可持续发展。建筑项目的环境影响评价报告（简称环评报告），旨在全面、客观、公正地反映建设项目实施后可能造成的环境影响、对其影响提出预防或者减轻不良环境影响的对策和措施的报告书，是环评工作的核心成果，其质量直接关系到项目对环境的影响程度和环境保护目标的实现。《中华人民共和国环境影响评价法》、《环境保护法》等法律条令规定：环境影响评价是指导人们规划开发、建设项目等活动的必须行为，是确保项目可行性和环境保护可持续性的重要环节，相关报告需由有环境影响评价资质的单位进行撰写、环保局针对报告进行评估与批复。

环评报告评审是环境影响评价中的关键环节，目前的环评报告评审过程完全依赖于人工评审，评审人员需要对冗长的报告进行报告解读、完备性审查、意见书生成等工作，这使得报告评审工作仍面临着一系列挑战。首先，环评报告中包含大量的数据、图表和专业术语，涉及多个学科领域交叉，这需要评审人员具备一定的跨学科能力和专业素质。其次，环评报告的评审往往需要在较短的时间内完成，然而大量的环评报告，机械重复、高度消耗的报告阅读工作占据了评审人员的大量精力，使得评审过程容易出现疏漏和错误；同时，人工评审的方法在理解报告时容易受到主观因素和个人能力的影响，导致评审结果的准确性与一致性受到影响，进而造成经济损失与环境污染。

近年来，大型语言模型（Large Language Models, LLM）飞速发展，其在解决复杂的自然语言处理及其衍生任务方面表现出了前所未有的强大能力。为了使基于LLM的系统可以在更侧重于客观准确描述及专业性决策的垂直领域中落地，许多研究在一定程度上突破了LLM所具有的幻觉、知识点截止、上下文限制等局限性，这使得LLM能够完成更复杂的专业领域任务，如科研领域论文研读、法律领域合同判读、金融领域财报分析、医疗领域疾病问诊和教育领域教案制作等。

这为解决环评报告评审工作中的痛点问题提供了新的思路和方法，在概念层面，本文旨在研究并实现一系列方法，用以提高大模型完成环评报告分析领域完成相关任务的能力；在实践层面，基于环保局的实际需求，我们将打造环评报告智能分析工作流，构建基于大语言模型的环评报告智能分析系统，构造文档智能领域的实现范本，实现人工智能技术在环评领域的落地应用，帮助报告评估人员高效准确地完成报告评审分析，提高环评报告评审工作的质量和效率。

* 1. 研究现状

要打造能胜任环评报告分析领域相关任务的专用大模型，需要对大语言模型(简称LLM)进行能力增强（LLM augmentation），以突破通用预训练语言模型固有的局限性，使其能够进行报告分析、文档问答、推理决策等任务。本节将会介绍构造专用大模型过程中需要解决的问题（即进行大语言模型能力增强的原因）以及相关的能力增强方法。

* + 1. 构造专用大模型需要解决的问题

虽然通用（或基础）LLM在解决通识问题上表现良好，但鉴于通用LLM的固有局限性，决定了其无法在更侧重于客观准确描述及专业性决策的垂直领域中落地。LLM能力增强的主要目的是解决以下问题：

幻觉(Hallucinations)：幻觉的产生归根于机器学习领域的过拟合现象，当模型过拟合时，它在训练集上可以达到极高的准确率，但是在测试集上性能会显著下降，这是因为模型过分地记住了训练集的特定特征而不是数据的内在规律。幻觉是过拟合的一种极端表现形式，在使用大语言模型生成文本时，模型会产生不准确甚至错误的陈述，生成看似合理但实际上是虚构的信息，即编造答案或歪曲事实。大模型的幻觉导致用户在使用大模型时需要保持警惕，必要时还需要对模型提供的输出进行验证。要达成模型在侧重于客观准确描述及专业性决策的垂直领域中落地的目标，需要首要解决幻觉问题。

知识点截止(Knowledge Cut-offs)：通常指大型语言模型在训练过程中所使用的数据都有一个时间截止点，即模型只学习了该时间点之前的知识，而不具有对训练数据之外的认知。由于时效性信息和领域数据的缺失，知识点截止导致模型无法准确理解或处理私有领域数据，也间接地导致了模型幻觉地产生。

上下文限制(Context Limited)：大多数语言模型都是基于Transformer的预训练模型，但Transformer在计算上的时间复杂度与空间复杂度随着序列长度呈二次幂增长，由于显存、内存等计算资源的限制，大语言模型不能同时处理无限长的序列，所以LLM都有一个上下文长度限制，也被称为上下文窗口限制，约束了LLM输入与输出的范围。在文档处理相关任务中，模型时常需要理解文档片段、捕捉其他段落的长距离信息甚至整个文档的语义，然而我们无法将整篇文档和所有必要信息输入给模型，这导致模型对超出其上下文窗口限制的任务无能为力。

* + 1. 大语言模型能力增强方法

近年来，LLM能力增强的通用解决方案是使用预训练(Pre-train)、微调(Fine-tuning)、检索增强生成(Retrieval Augmented Generation, RAG)、提示工程(Prompt Engineering)等生成式AI构造方法，配合大模型应用技术，定制能完成目标领域特定任务的专用LLM。

预训练(Pre-train)：即针对最终目标从头开始用大规模的数据集训练一个语言模型，以便模型能学习到期许的知识。这通常需要大量的计算资源（如GPU或TPU）和大规模的数据集，还需要考虑文本数据的预处理、标记化和训练过程的调优等方面，通常仅在面对领域跨度极大的业务时才采用预训练的方式从头构建一个语言模型，例如CodeGeeX和Sora是分别针对代码和视频生成领域所训练的专用模型。针对自然语言处理相关任务，一种普适方法是选择开源的预训练模型作为基座模型并以此作为基础针对特定任务构建解决方案，基座模型已经通过大规模的文本数据学会了语言的通用知识和规则，因此可以用于处理各种简单的文本相关任务，例如文本生成、机器翻译、自动摘要等。这样使得专用LLM的开发无需经历预训练阶段，免除了海量成本和计算压力。通过对基座模型进行指令微调、领域适配来增强模型完成特定领域自然语言处理相关任务的能力，再通过数据整合、大模型开发框架应用等方式构建解决方案，打造完整的面向垂直领域特定任务的LLM构建及部署流程。截止本论文撰写时，在自然语言处理任务上，最新一代的大型语言模型，如清华大学KEG实验室的GLM-4、OpenAI公司的GPT-4、Anthropic公司的Claude-3等，在文本生成、逻辑推理、中文支持和知识广度等方面都有着出色的表现，并且都经过微调使其具备了函数/工具调用的能力，通过自定义模型的行为，它们能够执行更加复杂的任务。

微调(Fine-tuning)：预训练LLM已经获得了解决自然语言任务的一般能力，其能力还可以根据应用领域和具体目标进行进一步调整，这个过程被称为LLM的“适配”(也称“适应”)过程。进行适配的主流方法是微调，即使用任务特定的较小数据集对模型参数进行调整。针对LLM适配，主要使用参数高效微调(Parameter-Efficient Fine-Tuning, PEFT)方法进行指令微调和对齐微调。参数高效微调一直是高效适配预训练语言模型到下游任务中的重要主题，由于LLM由大量模型参数组成，因此执行全量微调的成本相当巨大，PEFT旨在保持良好性能的同时，减少训练参数的数量。Transformer模型的四种参数高效微调方法主要有适配器微调(Adapter Tuning)、前缀微调(Prefix Tuning)、提示微调(Prompt Tuning)和LoRA(Low-Rank Adaptation)。指令微调(Instruction Tuning)旨在增强LLM在特定领域的能力，是在自然语言形式的格式化实例集合上对预训练LLM进行微调的方法，即通过由（指令，输出）对组成的数据集上进一步训练LLM的过程，指令微调有助于弥合LLM词预测目标与人类期望输出之间的差距，帮助LLM拥有更好的推理能力，从而展现出泛化到未见过任务的卓越能力，这种技术已经在LLM中得到广泛应用，例如InstructGPT和GPT-4。对齐微调(Alignment Tuning)旨在使LLM的输出与人类价值观或偏好保持一致，它令LLM输出业务特定下用户更加偏好的回答。一种高效的对齐微调方法是基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)。通过收集人类反馈数据对LLM进行微调，RLHF采用强化学习算法，通过学习奖励模型让LLM适应人类反馈。

检索增强生成(Retrieval Augmented Generation, RAG)：通过从外部知识库检索相关信息来辅助大型语言模型回答问题，已经被证明能显著提高回答的准确性，减少模型产生的幻觉，尤其是在知识密集型任务中。朴素RAG（Naive RAG） 的核心是一个混合框架，它集成了嵌入模型、检索模型和生成模型：嵌入模型将原始数据转化为向量表示并嵌入到数据库中以便进行相似度搜索；检索阶段的核心是使用编码模型(如BM25、DPR、ColBERT等)基于问题检索相关文档，用检索到的上下文在生成阶段作为条件生成文本，能够保持知识或信息的时效性或准确性，提高LLM的预测能力与回答准确性；在生成阶段，查询或提问与检索到的信息共同被集成到提示词中以获得语言模型的响应。RAG和微调具有互补性，RAG适用于需要回答具体查询或特定信息检索任务的场景，而微调强调基础模型中的现有知识，修改或定制模型输出，为模型提供复杂指令的场景。在某些场景下通过结合RAG和微调技术，可以从不同层面增强模型的能力以实现最佳的模型性能。以提高检索质量为重点，高级RAG技术（Advanced RAG）引入了检索预处理和检索后处理步骤，通过查询分析（Qeury Analyzation）、混合检索（Mixed Retrieval）、结果重排（Re-ranking）等方式进行优化，进一步增强了RAG框架的能力。模块化的RAG（Modular RAG）建立在高级RAG的基础上，采用多种策略和组件来构建应对具体任务的RAG管道，被证明能显著地提升检索性能与LLM回答质量，例如RAG-Fusion采用多查询策略，将用户查询扩展到不同的角度，利用并行向量搜索和智能重新排序来发现显性和变革性知识，从而解决了传统搜索的局限性。矫正式检索增强生成(Corrective Retrieval Augmented Generation, CRAG)通过评估查询检索文档的整体质量，针对检索到的文档有选择地关注关键信息并过滤掉其中的不相关信息来提高生成阶段的鲁棒性；自我反思检索增强生成(Self-Reflective Retrieval-augmented Generation，Self-RAG) 旨在通过根据需求进行检索和自我反思来提升LLMs的生成质量，引入了按需检索和细粒度自我反思，使其具有更广泛的适用性、更强的鲁棒性和可控性。在实际中，需要针对特定的场景灵活地搭建RAG系统以取得最佳效果。

提示工程（Prompt Engineering）：为了让专用LLM输出任务特定的期望回答，提示工程（也称In-Context Prompting）旨在不更新模型参数的前提下，在提供尽可能多的上下文信息的同时，通过提供零样本(zero-shot learning)或少样本学习（few-shot learning）提升大模型在用例上的性能。许多增强模型推理能力的研究都基于提示工程，思维链(Chain-of-Thought, CoT)让模型在生成答案之前展示其推理过程提高了模型在需要复杂推理的任务上的性能、自我一致性(Self-Consistency)在CoT的基础上上增加了自洽性提高了整体推理的准确率、和自我修正(Self-refine)在不需要额外训练的情况下提供迭代式的自我精炼从而产生更高质量的输出。通过组合使用这些方法能让LLM应用程序具有更高的灵活性和性能，提示工程被广泛应用在构建LLM应用的各个环节。

基于大模型的智能体（LLM-based Agent）：指模拟人类和外部环境的交互过程，利用大语言模型作为智能体的核心协调者，加入记忆、感知、工具使用和规划等模块，来构建具有更高级认知能力的智能体，使其能够在相关任务中进行学习、推理和反思，以期完成更加复杂的任务。ReACT（Reasoning and Acting in Language Models）框架通过让语言模型交叉地进行推理、观察和行动，给出了构建LLM智能体的范式；思维树（Tree of Thoughts, ToT）采用树状的思维方式使用LLM模拟人类的决策过程，通过在子节点中递归地进行任务分解、想法生成、状态评价和最优路径搜索流程以获取答案，显著地提升了模型解决复杂任务的表现；语言智能体树搜索（Language Agent Tree Search，LATS）采用蒙特卡罗树搜索算法在可能的推理和行动步骤的组合空间上进行搜索，将LLM作为决策中心、价值函数和优化器，首次统一了LLM规划、行动和推理步骤，增强了LLM进行自主推理和决策方面的多样性。

* 1. 本文研究内容

本文旨在研究并实现一系列方法，以一种提高大模型完成环评报告分析领域完成相关任务的能力，并基于这些方法打造环评报告智能分析工作流，构建端到端环评报告智能分析解决方案，为相关领域的专业人士提供更高效、准确和可靠的环评报告分析工具。面向环评报告智能分析系统的实现效果而言，本论文的目标是：

1. 实现对环评报告内容和逻辑要素的理解、审查和归纳：针对输入的环评报告，系统应能够对其进行信息获取、要素提取、内容理解、完备性审查、总结归纳；
2. 实现对评审人员针对环评报告的提问的准确分析与响应：针对评审人员的提问，系统应能对其进行查询分析及优化、文档检索及后处理、答案生成及评估、对话记忆持久化；
3. 实现能够完成复杂环评任务的智能工作流：系统应能完成环评工作中的特定复杂任务，如风险与影响因子定位汇总、报告意见书一键生成等。

针对以上目标，本论文的工作涵盖如下：

1. 本研究以RAG作为LLM能力增强的核心方法，面向环评报告研究并实现了一个高性能RAG框架，验证该框架的有效性。该框架包括了完整的、从原始输入到最终输出的解决方案，鉴于环评报告文档多为PDF形式，我们将基于OCR技术实现一个文档解析器以尽可能地获取高质量的数据；在RAG的核心部分，即数据嵌入、文档检索和答案生成阶段，我们采用了多种优化方式来达到理想效果；对于设计的RAG框架，我们将构造评估数据，搭建自动化评估流程，对RAG系统生成答案的相关性、准确性、整体质量等相关指标进行验证。
2. 本研究将基于上述RAG实现与前沿的智能体框架，进一步提高LLM的复杂任务完成能力。通过向LLM加入感知、记忆、行动、工具等模块，面向复杂的环评任务实现基于LLM的智能体群，采用制定逻辑流程、打造工具、采用多智能体协作等方式提高系统完成目标领域任务的能力。
3. 基于上述工作，本研究将结合Langchain、FastAPI、sqlalchemy和Streamlit等开发框架，实现一个端到端的环评报告智能分析解决方案，完成文档智能领域实践范本的构建，具体贡献为可一键部署的环评报告智能分析软件，该软件能够对环评报告内容和逻辑要素的理解审查和归纳、可以对评审人员针对环评报告的提问的准确分析与响应、能够完成复杂的环评任务。
   1. 论文组织结构

第一章为绪论，介绍本文的研究背景和意义，然后说明了当前打造垂直领域专用大模型时需要解决的问题及对应的大语言模型能力增强方法的研究现状，最后阐述了本论文的研究目标与具体工作。

第二章为相关理论和技术，主要包括。

第三章为RAG框架及智能体的设计与实现，从

第四章为评估数据集的构建及实验验证，通过。

第五章为环评报告智能分析系统的构建，介绍了

第六章为总结与展望，总结了论文的主要工作和创新点，提出了存在的问题及可行的研究方向，为后续的研究提供参考。

在后面填加正文，然后删掉此句

论文书写过程中请使用“样式”来实现规范格式，如图 1.1所示。

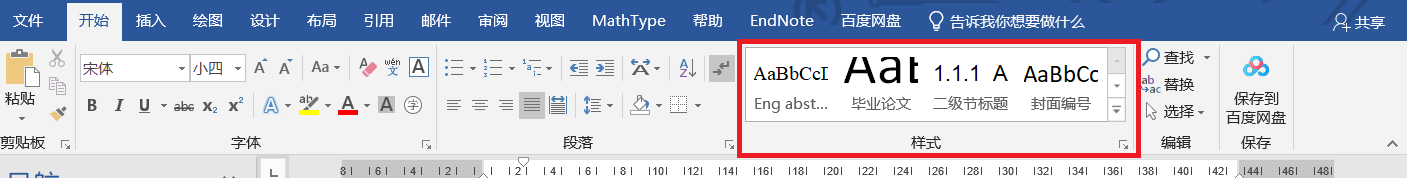


图 1.1 使用样式来实现格式规范化（不要动此图编号，可以改标题）

正确引用参考文献，是学术规范的体现，也表明作者了解研究现状[1]。

参考文献格式参照国家标准GB/T7714-2005, 标注的序号按正文中参考文献出现的先后顺序用阿拉伯数字连续编号，将序号置于方括号“[ ]”内。同一处引用多篇文献时，将各篇文献的序号在方括号中全部列出，各序号之间用“，”如[1， 2，…]的格式，标注方括号中的序号按增序排列[2-5]。同一文献在论文中引用多次的，只编一个序号。建议通过文献管理软件插入参考文献，在提交前全部转化为文本。可以在知网上下载正确的中文文献引用格式。

为使您的文档具有专业外观，Word 提供了页眉、页脚、封面和文本框设计，这些设计可互为补充。例如，您可以添加匹配的封面、页眉和提要栏。单击“插入”，然后从不同库中选择所需元素。

标注引用的时候不能随意用，按照学术规范，你看过的论文，使你产生了这样的想法要标注引用，你直接引原文，更要标注引用，否则就是学术不端，剽窃他人成果。

插入图片、表格时，一般每个图表单独成行，并且需要为图片和表格添加图题和表题，图题在图片下方，表题在表格上方，编号包含章节号。图片和图题可以直接复制下面图 1.2的“整个图和图编号”。

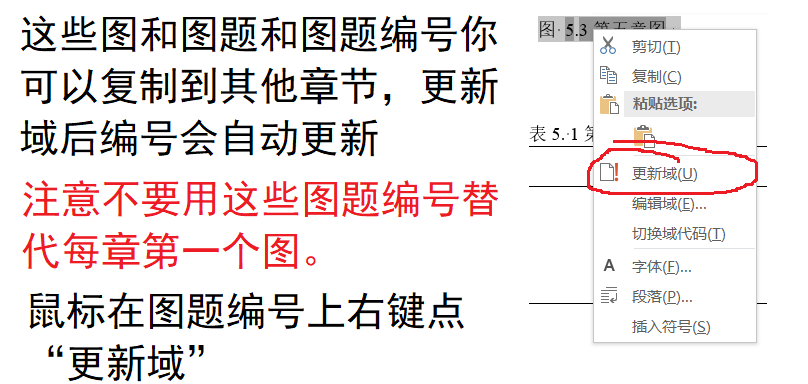


图 1.2第一章图的第二个 这是平凡的图和编号 请复制使用

在文中采用“交叉引用”方式选择相应的编号项目引用，可以通过“更新域”自动更新编号。图片交叉引用类型选择“图”这个标签，引用内容为“仅标签和编号”，如图 1.3所示。

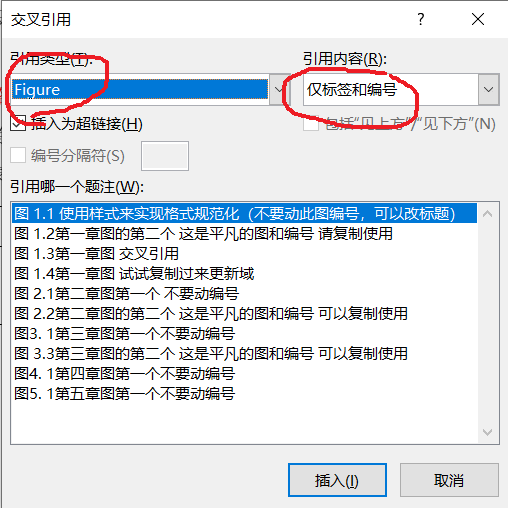


图 1.3第一章图 交叉引用选择“引用类型”和“引用内容”

表格采用三线表，顾名思义一般只有三条线，个别情况可以在表头增加一些横线或者第一列之后增加竖线（很少用），多数情况如表1.1所示。表格交叉引用类型选择“表”这个标签，引用内容为“仅标签和编号”。表格最好不要跨页。特别需要注意不要把书上或者网上截图的表格插入为图片。

以上文字都可以在读懂以后删除。

表1.1请使用三线表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 姓名 | 语文 | 数学 |
| 1 | 张三 | 100 | 100 |
| 2 | 李四 | 59 | 59 |

* + 1. 第二节标题

在此键入摘要正文

以上文字都可以在读懂以后删除。

尝试一下交叉引用，如图 1.4。



图 1.4第一章图 试试复制过来更新域

* + 1. 第二节标题

在此键入摘要正文

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 本文研究内容

1. 相关理论和技术
   1. 检索增强生成
      1. 检索增强生成概述及原理

检索增强生成（Retrieval Augmented Generation，RAG）泛指通过引用外部知识增强LLM能力的过程。最初的RAG旨在通过与预训练模型（PreTrained Models）结合额外知识来增强语言模型的能力，但随着ChatGPT的出现，LLM展现出了强大了上下文学习（In Context Learning，ICL）能力，但却在处理超出其训练数据或知识容量的任务时容易产生“幻觉”等现象，此时的RAG研究转为如何为LLM提供有益的信息以提高其在领域特定和知识密集场景下的回答和推理表现。

RAG的核心是一个混合框架，它集成了嵌入模型、检索模型和生成模型：嵌入模型将原始数据转化为向量表示并嵌入到数据库中以便进行相似度搜索；检索阶段的核心是使用编码模型(如BM25、DPR、ColBERT等)基于问题检索相关文档，用检索到的上下文在生成阶段作为条件生成文本，能够保持知识或信息的时效性或准确性，提高LLM的预测能力与回答准确性；在生成阶段，查询或提问与检索到的信息共同被集成到提示词中以获得语言模型的响应。

RAG的发展之路经历了朴素、高级、模块化的阶段，随着RAG研究的飞速发展，此时的RAG不再局限于在推理阶段对LLM进行增强，甚至开始与LLM微调技术相结合。RAG和微调具有互补性，RAG适用于需要回答具体查询或特定信息检索任务的场景，而微调强调基础模型中的现有知识，修改或定制模型输出，为模型提供复杂指令的场景。在某些场景下通过结合RAG和微调技术，可以从不同层面增强模型的能力以实现最佳的模型性能。

* + 1. 朴素RAG

朴素RAG（Naive RAG）遵循传统的“索引——检索——生成”三段式，索引（Index）阶段首先提取各种输入形式（如JPG、PDF、Excel、Word、HTML、Markdown等）的原始数据，将其转化为语言模型可读的文本格式，鉴于LLM的上下文限制，这些文本被不同的方法切分成为更小的块后，嵌入模型负责将文本块转换为向量表示以便其能够存储在向量数据库中，这对于检索阶段的相似性搜索（Similarity Search）或最大边际相关性(Maximum marginal relevance，MMR)搜索等基于向量的搜索算法至关重要。检索（Retrieval）阶段接收用户的查询，通过计算查询向量与向量库预料数据的相似性得分并挑选前K个块用作提示词中的上下文信息。生成（Generation）阶段的核心在于利用大模型的固有知识和提供的外部知识形成回答，在此步骤中可以集成记忆模块实现历史对话和持久化功能。

朴素RAG是RAG技术的初级形式，从技术原理上看，它不可避免地存在以下缺陷：

1. 无效或冗余的检索：
2. 无法处理复杂查询：

高级RAG技术（Advanced RAG）引入了检索预处理和检索后处理步骤，通过查询分析（Qeury Analyzation）、混合检索（Mixed Retrieval）、结果重排（Re-ranking）等方式进行优化，进一步增强了RAG框架的能力。模块化的RAG（Modular RAG）建立在高级RAG的基础上，采用多种策略和组件来构建应对具体任务的RAG管道，被证明能显著地提升检索性能与LLM回答质量，例如RAG-Fusion采用多查询策略，将用户查询扩展到不同的角度，利用并行向量搜索和智能重新排序来发现显性和变革性知识，从而解决了传统搜索的局限性。矫正式检索增强生成(Corrective Retrieval Augmented Generation, CRAG)通过评估查询检索文档的整体质量，针对检索到的文档有选择地关注关键信息并过滤掉其中的不相关信息来提高生成阶段的鲁棒性；自我反思检索增强生成(Self-Reflective Retrieval-augmented Generation，Self-RAG) 旨在通过根据需求进行检索和自我反思来提升LLMs的生成质量，引入了按需检索和细粒度自我反思，使其具有更广泛的适用性、更强的鲁棒性和可控性。在实际中，需要针对特定的场景灵活地搭建RAG系统以取得最佳效果。

* 1. 第一节标题

在后面填加正文，然后删掉此句

本章的第一个图不要动编号，如图 2.1。



图 2.1第二章图第一个 不要动编号

* 1. 如何插入公式

文章中如果需要插入公式，公式需要居中，编号右对齐。实现手段有多种，本模板使用表格对齐，大家直接“拷贝-粘贴”表格及内部编号，使用Word自带工具编辑公式，交叉引用时注意选择“整项题注”，二次方程的求根公式如式(2.1)。注意不要去替换每一章第一个公式的编号。

|  |  |  |
| --- | --- | --- |
|  |  | (2.1) |

下面的公式(2.2)是MathType输入的，字体更好看，感兴趣的同学自己装插件。

|  |  |  |
| --- | --- | --- |
|  |  | (2.2) |

下面继续输入正文。注意对公式内符号的说明也需要段首空两字符。

* 1. 在第二章插入表格和图片

你可以继续使用交叉引用来引用表格编号如表2.1，它可以通过更新域自动更新。

表2.1第二章第一个表题，请你不要动这个编号，替换表题内容即可

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

这里可以继续插入图片，如图 2.2。

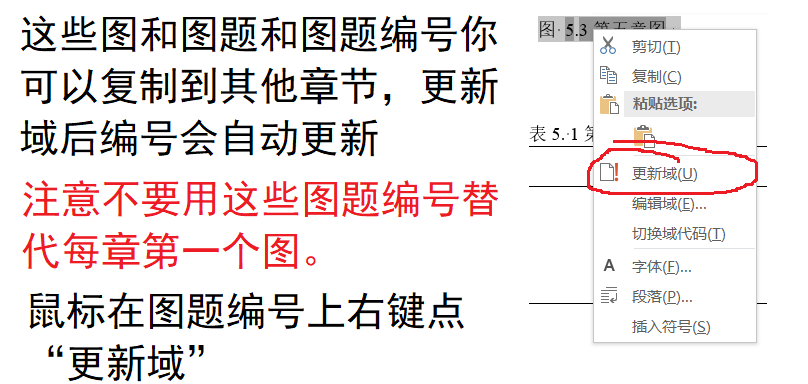


图 2.2第二章图的第二个 这是平凡的图和编号 可以复制使用

* + 1. 第二节标题

在后面填加正文，然后删掉此句

第二章的第二个表如表2.2，可以复制并替换到其他章节。

表2.2可以复制并替换到其他章节

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 本章小结

1. 章标题

在后面填加正文，然后删掉此句

* 1. 第一节标题

在后面填加正文，然后删掉此句



图 3.1第三章图第一个 不要动编号

* 1. 我要插入公式

本章第一个公式(3.1)的编号不要替换。

|  |  |  |
| --- | --- | --- |
|  |  | (3.1) |

可以复制这个公式(3.2)和编号到各个章节，但不能取代每章第一个公式的编号。

|  |  |  |
| --- | --- | --- |
|  |  | (3.2) |

后面继续输入正文。

如果表格卡在两页中间，不要跨页，调整文字内容或者少量留白。

表3.1本章第一个表题，请你不要动这个编号，替换表题内容即可

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

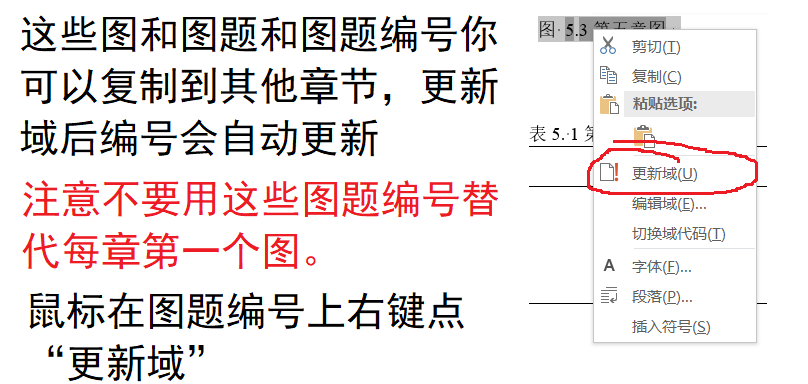


图 3.2第三章图的第二个 这是平凡的图和编号 可以复制使用

* + 1. 第二节标题

在后面填加正文，然后删掉此句

表3.2可以复制并替换到其他章节

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 本章小结

1. 章标题

在后面填加正文，然后删掉此句

* 1. 第一节标题

在后面填加正文，然后删掉此句



图 4.1第四章图第一个 不要动编号

表4.1第四章第一个表题，请你不要动这个编号，替换表题内容即可

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* 1. 第四章插入公式

本章第一个公式(4.1)的编号不要替换。

|  |  |  |
| --- | --- | --- |
|  |  | (4.1) |

后面可以输入正文。

* + 1. 第二节标题

在后面填加正文，然后删掉此句

表4.2可以复制并替换到其他章节

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 本章小结

1. 章标题

在后面填加正文，然后删掉此句

* 1. 第一节标题

在后面填加正文，然后删掉此句



图 5.1第五章图第一个 不要动编号

* + 1. 第二节标题

在后面填加正文，然后删掉此句

* 1. 我要插入公式

本章第一个公式(5.1)的编号不要替换。

|  |  |  |
| --- | --- | --- |
|  |  | (5.1) |

后面可以输入正文。

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 本章小结

1. 总结与展望

在后面填加正文，然后删掉此句

* 1. 第一节标题

在后面填加正文，然后删掉此句

* + 1. 第二节标题

在后面填加正文，然后删掉此句

* + - 1. 第三节标题不推荐使用

在后面填加正文，然后删掉此句

* 1. 展望

致 谢

参考文献

[1] Li R, Shi Y, Wang Z, et al. Densification behavior of gas and water atomized 316L stainless steel powder during selective laser melting[J]. Applied Surface Science, 2010, 256(13): 4350-4356.

[2] 于立娟, 孙鹏, 李广义, 等. 阻燃剂的研究进展[J]. 上海塑料, 2013, 28(3): 21- 23.

[3] 张健. 混合式步进电动机单步响应特性的研究[D]. 大连:大连交通大学, 2005.

[4] 楼鸿棣, 邹慧君. 高等机械原理[M]. 北京:高等教育出版社, 1990:30-34

[5] 姚建华. 一种刀具及其加工方法:, CN 1532304 A[P]. 2004.

附录A

在后面填加你的程序图或数据或代码，如果没有，把A删掉，只留一个附录，记得把这句话也要删掉。

附录B

在后面填加翻译，先中文，后英文，英文可以插入pdf，中文排版要美观整齐，然后删掉此句