**基于检索增强的多智能体协作的招投标文档一键生成系统**

**项目描述：实现一套基于多智能体协作与检索增强生成(RAG)技术的招投标文档智能生成系统。**旨在解决传统招投标文档编制效率低，人工成本高等问题。

**个人职责：**

1任务规划：根据任务描述，结合预定义的输出Schema字段，生成结构化的多级递归目录大纲。

2.检索引擎：根据生成的章节标题，动态构建伪指令信息，从相关文件库中检索关键信息，指导模型生成符合上下文的高质量内容。

* 高级检索：采用混合检索引擎架构，结合字面召回和向量召回技术，通过重排模块对召回结果进行优化排序，最终合成高质量检索响应。
* 模块化检索：设计自定义检索器，结合大语言模型（LLM）对检索节点进行相关性评估与预回答生成(Hyde方法)，提升检索效率与结果质量。

3.记忆管理：

* 短期记忆：实时记录每个章节的用户指令及其关联的特定段落信息，并标识消息来源。
* 长期记忆：将生成的文本段落（索引节点）持久化存储到文档管理器中，支持后续任务的检索。

4.人类交互:

设计并实现实时反馈机制，支持用户对生成的大纲目录和文本段落进行交互式修改，系统根据反馈动态更新内容，确保生成结果符合用户需求。

**基于智能体软件操作流程自动化功能助手**

**项目描述：**采用智能体（Agent）技术，构建AI驱动的数字员工系统，实现软件操作的全流程自动化。该系统通过模拟人类操作行为，自动化处理软件中的复杂任务，显著提升效率并降低错误率。

**个人职责：**

1数据收集：收集目标软件的操作流程、步骤说明及功能API文档等相关信息，构建API调用拓扑关系图。

2.模型训练：采用Llama-Factory训练大语言模型，重点关注工具API调用和参数解析准确性。

3.构建Agent：规划模块(react,rewoo)、动作模块、反思模块，记忆模块，协同配合完成操作步骤。

4.功能测试和优化：测试智能体的理解能力、函数调用准确性等，收集用户反馈并进行优化；

**FAQ客服智能问答系统；**

**项目描述：**实现了一套基于Chat对话引擎的FAQ客服智能问答系统，优化客户体验提升服务效率。

**个人职责：**

1. 问题改写：基于历史对话消息，对用户当前问题进行语义改写，生成完整的问题表达。
2. 意图识别：利用LLM对用户输入进行路由，选择对话策略（闲聊、检索、反问）等。
3. 检索模块：基于QQ模式的多路召回与重排机制，从FAQ库中高效检索相关信息。

**大模型LLM微调； 项目负责人 项目描述：**本项目旨在利用开源大模型，通过学习特定领域的PDF文档，构建一个具备领域知识理解的私有化模型。**个人职责：**

1. 数据清洗：对解析后的文本块进行格式标准化和清洗处理，包括去除重复内容、规则过滤和有效性验证。

2. 数据合成：Self-Instruct方法通过模型自主生成与给定文本段落相关的问答对，构建指令数据集。

3. 模型微调：采用LoRA等参数高效微调方法，对开源LLM模型进行领域知识注入。

您好，我是，硕士毕业于东北大学控制理论与控制工程专业（双一流A+学科），目前在北京广联达担任高级算法工程师。我的核心方向是大模型应用与多智能体系统开发，拥有从算法设计到工程落地的全流程经验。

过去四年，我先后在浙江大华、中国平安和广联达负责多个AI项目。**在任职期间**，我作为项目负责人，主导了“招投标文档一键生成系统”的开发。通过结合多智能体协作与检索增强生成（RAG）技术，最终将文档生成效率提升70%以上。该项目代码已贡献至MetaGPT开源社区，并获得了原作者吴承霖的高度认可。

**其次,**我负责构建了“AI数字员工系统”，利用智能体技术实现软件操作全流程自动化。通过训练大模型精准解析API调用关系，并设计模块化Agent架构，系统上线后用户操作错误率降低40%。

**最后，**我主导了“FAQ客服智能问答系统”的研发，通过多路召回与LLM意图路由技术，将客户满意度提升了30%。

**技术深度方面**，我擅长大模型微调与领域知识注入，曾采用LoRA方法对开源模型进行高效训练，并利用Self-Instruct构建指令数据集。同时，我在社区贡献与技术分享上较为活跃，是MetaGPT开源贡献者，并在CSDN、GitHub等平台分享技术文章，累计粉丝量超2000+。

我的优势在于既能深入技术细节，如设计检索算法与模型优化，也能从业务视角推动AI落地，解决实际效率与成本问题。未来希望能在贵公司的大模型应用场景中，持续输出创新解决方案。

以上是我的个人介绍，谢谢！

**问题1：你在“招投标文档一键生成系统”中提到了“Hyde方法”用于预回答生成。能否详细解释一下Hyde方法的原理，以及它在你的项目中是如何提升检索效率的？**

“Hyde方法的核心思想是通过大语言模型（LLM）生成一个假设性的回答，然后基于这个回答去检索相关文档。在我们的项目中，当用户输入一个章节标题时，系统会先让LLM生成一个假设性的段落内容，然后根据这个内容去检索相关的文档片段。这样做的好处是，生成的假设性回答能够更好地捕捉用户的意图，从而提升检索的精准度。

具体来说，Hyde方法在项目中帮助我们减少了无关文档的召回，尤其是在用户输入较为模糊的情况下。通过预生成的假设性回答，我们能够更准确地定位到与用户需求相关的文档片段。

**问题2：你在“AI数字员工系统”中使用了Llama-Factory训练大语言模型。能否分享一下你在模型训练过程中遇到过哪些挑战？比如数据不足或模型过拟合问题，你是如何解决的？**

“在训练Llama-Factory模型时，我们遇到的主要挑战是数据不足。由于目标软件的操作流程和API文档较为复杂，初始的训练数据量有限，导致模型在泛化能力上表现不佳。

为了解决数据不足的问题，我们采用了数据增强技术，通过生成合成数据来扩充训练集。具体来说，我们利用已有的API文档和操作步骤，生成了一些模拟的用户操作场景，并通过脚本自动生成对应的训练数据。

对于过拟合问题，我们引入了正则化技术，并在训练过程中使用了早停（early stopping）策略。

**问题3：你在FAQ系统中提到“基于QQ模式的多路召回与重排机制”。能否详细解释一下“QQ模式”是什么？它与其他召回模式（如BM25、语义召回）相比有什么优势？**

QQ模式的优势在于它能够兼顾召回的速度和精度。关键词召回速度快，适合处理大规模数据，而语义召回能够捕捉到更深层次的语义信息，适合处理复杂或模糊的用户问题。通过结合这两种召回方式，QQ模式能够在保证召回速度的同时，提升召回的精度。

**问题4：在“招投标文档一键生成系统”中，你提到“动态构建伪指令信息”来指导模型生成内容。能否举一个具体的例子，说明你是如何设计这些伪指令的？**

“在招投标文档生成系统中，伪指令的作用是帮助模型理解用户需求并生成符合上下文的内容。例如，当用户输入‘技术方案’作为章节标题时，系统会动态生成一条伪指令，比如‘请生成一份关于智能安防系统的技术方案，重点描述系统架构、功能模块和技术优势’。

**问题5：在“AI数字员工系统”中，你提到“反思模块”的设计。能否详细说明反思模块的作用是什么？它是如何帮助智能体优化操作的？**

“反思模块的核心作用是让智能体在执行任务后能够自我评估和优化。具体来说，当智能体完成一个操作步骤后，反思模块会分析执行结果，判断是否达到了预期目标。如果未达到目标，反思模块会生成一条反馈信息，指导智能体调整后续操作。

例如，在自动化生成工程报表的任务中，如果智能体在调用某个API时返回了错误，反思模块会分析错误原因（如参数错误或API调用顺序问题），并生成一条反馈信息，指导智能体重新调整参数或调用顺序。通过这种方式，反思模块能够帮助智能体不断优化操作，提升任务执行的准确性和效率。”

**问题6：你在“大模型LLM微调”项目中使用了LoRA方法。能否解释一下LoRA的原理，以及为什么选择LoRA而不是全量微调？**

“LoRA（Low-Rank Adaptation）是一种参数高效微调方法，其核心思想是通过低秩分解来减少微调参数量。具体来说，LoRA在原始模型的权重矩阵上添加了一个低秩矩阵，通过训练这个低秩矩阵来适应特定任务，而不需要修改原始模型的全部参数。

我们选择LoRA而不是全量微调的原因主要有两点：

1. **计算资源节省**：全量微调需要更新整个模型的参数，计算成本非常高，而LoRA只需要更新少量的低秩参数，大大减少了计算资源的消耗。
2. **防止过拟合**：全量微调在小数据集上容易过拟合，而LoRA通过限制参数量，能够更好地保持模型的泛化能力。”

**问题7：在FAQ系统中，如果用户的问题非常模糊或包含歧义，系统如何确保能够给出准确的回答？你是否有设计特定的策略来处理这种场景？**

“在FAQ系统中，我们设计了多层次的策略来处理模糊或歧义问题。首先，系统会通过意图识别模块对用户问题进行语义分析，判断用户可能的意图。如果问题仍然模糊，系统会采用反问策略，向用户提出澄清问题，比如‘您是指XX功能吗？’或‘您需要了解XX的具体步骤吗？’。

此外，我们还设计了一个基于上下文的召回机制。如果用户之前有过相关问题的对话记录，系统会结合上下文信息来推断用户的真实意图，从而提高回答的准确性。”

**问题8：在“招投标文档一键生成系统”中，如果检索到的内容与用户需求不完全匹配，你会如何处理？是否有设计某种机制来动态调整生成内容？**

“在招投标文档生成系统中，如果检索到的内容与用户需求不完全匹配，我们会通过以下几个步骤来动态调整生成内容：

1. **反馈机制**：系统会实时提示用户检索结果的相关性，并允许用户对检索结果进行手动调整或重新检索。
2. **内容修正**：如果检索到的内容部分匹配，系统会利用大语言模型（LLM）对内容进行局部修正，确保生成的内容更符合用户需求。例如，如果检索到的技术方案缺少某个功能模块，系统会自动补充相关内容。
3. **多轮迭代**：系统支持多轮交互，用户可以对生成的内容进行多次修改，系统会根据用户的反馈动态调整生成策略，直到用户满意为止。”

**问题9：在“AI数字员工系统”中，如果智能体在执行任务时遇到未预见的错误（如API调用失败），系统会如何应对？你是否设计了某种容错机制？**

1. **错误检测**：系统会实时监控API调用的返回状态，如果检测到错误（如调用失败或返回异常），系统会立即触发错误处理流程。
2. **重试机制**：对于临时性错误（如网络波动），系统会自动进行重试，最多重试3次，以确保任务能够顺利完成。
3. **备用方案**：如果重试后仍然失败，系统会启动备用方案。例如，如果某个API调用失败，系统会尝试调用功能相似的备用API，或者提示用户手动干预。

**问题10：你在“大模型LLM微调”项目中提到“通过学习特定领域的PDF文档构建私有化模型”。能否分享一个具体的业务场景，说明这个模型是如何解决实际业务问题的？**

“在合同审查场景中，我们利用大模型LLM微调技术构建了一个私有化模型，专门用于处理法律文档。具体来说，模型通过学习大量的合同范本和法律条款，能够自动识别合同中的关键条款（如违约责任、付款方式等），并生成审查报告。

**问题16：你在简历中提到“技术分享博主，CSDN粉丝量2100+”。能否分享一个你最近写的技术文章，并说明它对你的职业成长有什么帮助？**

这篇文章对我的职业成长帮助很大。首先，通过写作，我进一步加深了对RAG技术的理解，并发现了许多在实际应用中的优化点。其次，文章发布后，我收到了很多读者的反馈和建议，这些反馈帮助我不断优化我的技术方案。最后，通过分享技术文章，我在社区中建立了良好的口碑，也为我的职业发展带来了更多机会。”

**问题17：你未来3-5年的职业规划是什么？你希望在大模型领域或智能体技术方向有哪些突破？**

“未来3-5年，我的职业规划是继续深耕大模型和智能体技术领域，并希望在以下几个方面取得突破：

1. **大模型应用**：我希望能够将大模型技术应用到更多实际业务场景中，特别是在金融、法律等垂直领域，构建具备领域知识理解的私有化模型。
2. **智能体协作**：我计划进一步研究多智能体协作技术，探索如何让多个智能体在复杂任务中高效协作，提升系统的整体性能。