基于混合架构的大语言模型智能问答系统研究

Research on Intelligent Question Answering System for Large Language Model Based on Hybrid Architecture

陶晓英(中国联通上海分公司,上海 200082)
Tao Xiaoying(China Unicom Shanghai Branch,Shanghai 200082,China)

摘 要:
探讨了大语言模型在企业政企营销知识智能问答方向的研究与应用。在开发技术上,采用检索增强生成技术[1],结合大模型微调、意图识别和向量库实现问答 *F*1 值初步达到 78.21%,构建闭环知识图谱体系,将问答 *F*1 值提升到 92.36%。在工程上,通过 vLLM 大模型加速机制提升系统性能,采取数据安全技术保障系统安全性,通过 API 及微服务模块化架构提升系统适配性及扩展性,并将系统应用于行业生产中,有助于加速大模型在企业中的应用落地。

关键词:
大语言模型;检索增强生成;知识图谱;vLLM
doi:10.12045/j.issn.1007-3043.2024.05.009
文章编号:1007-3043(2024)05-0048-08
中图分类号:TP311
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
It explores the utilization of large language model technology to develop an efficient intelligent question-answering system for empowering enterprise internal government-enterprise marketing guidance scenarios. In terms of development technology, the study adopts retrieval-augmented generation(RAG)technology, combines it with large model fine-tuning, intention recognition, and vector libraries to achieve a preliminary *F*1 score of 72.6% for question answering. By constructing a closed-loop knowledge graph system, the *F*1 score for question answering is improved to 91%. Engineering-wise, the system performance is enhanced through the vLLM large model acceleration mechanism, data security technology is employed to ensure system safety, and the system's adaptability and scalability are improved through Application Programming Interface(API) and microservices modular architecture. The system has been applied in industrial production. It may help to accelerate the application of large models in enterprises.

Keywords:
Large language model; Retrieval-augmented generation; Knowledge graph; vLLM

引用格式:陶晓英. 基于混合架构的大语言模型智能问答系统研究[J]. 邮电设计技术,2024(5):48-55.

0 引言

随着人工智能(AI)技术[2]的飞速发展,以 ChatGPT 大语言模型[3]为代表的生成式人工智能技术(AIGC)已成为一种新的生产力,在提供客户服务、推动业务流程自动化以及优化用户体验等方面展现出巨大潜力[4]。智能问答是大语言模型技术应用的一个重要场景,生成式大语言模型不仅能回答问题,还能生成连贯、自然的语言回复,这一技术的应用大大增强了用户互动的自然性和流畅性。本研究选择智能问答作为大语言模型应用的落脚点,在企业内部政企产品营销指引场景中探索大语言模型的应用。

然而,尽管大语言模型在智能问答场景中展现了巨大潜力,但在实际应用中仍存在不少挑战和限制[5],主要包括:

a) 模型的幻觉问题(Hallucination)。大语言模型有时会生成看似合理但实际上不准确或无根据的信息,这种现象被称为“幻觉”。在模型生成答案的过程中,这种幻觉现象尤为显著,可能会对用户产生误导。

b) 模型的精确度问题。大语言模型虽然能够提供流畅的语言输出,但在处理特定和复杂问题时可能表现出精确性的不足。这可能是因为模型对特定领

收稿日期:2024-03-01

域知识的理解有限，或者训练数据中缺少足够的相关信息。

c）模型数据偏见的问题。训练大语言模型所用的数据往往来自互联网[6]，这导致模型可能无意中学习并复制了数据中存在的偏见。这种偏见可能表现为性别、种族或文化上的歧视，导致模型的输出具有偏颇。

针对以上问题，业界通常采用大模型辅以检索增强生成（Retrieval-Augmented Generation，RAG）外挂向量库来进行专业领域知识的召回。以阿里千问QWen-14B模型为例，该模型在通用领域上具有较好的效能，但在处理政企产品营销领域的530篇语料数据集时，仅使用RAG外挂向量库进行知识召回，其问答精准度*F*1值仅为78.21%，无法满足行业B端用户需求。这是因为向量库的检索召回使用的是向量空间内积算法，该算法只能检索出相似度较高的内容，然后再由大模型进行总结回复，其精准度并不理想。

因此，本研究从技术和工程2个层面提出创新方案，以实现智能问答系统的高准确率和高效率。在技术层面，首先定义了智能问答系统的评估算法，用来准确衡量系统的有效精准度。针对通用RAG技术难以满足行业应用中对高精确性的要求这一问题，在系统设计上进行了一系列优化。通过引入知识图谱[7]、大模型微调等措施，使系统在政企产品营销领域的数据集上的问答*F*1值从78.21%提升至92.36%，相比优化前提升了14.15个百分点。在工程层面，使用了大模型加速器，使平均首字符响应时间缩短了20%，生成速度提升了12%。同时，还采用了数据安全保障及微服务模块化架构，以提升系统适配性及扩展性。

1 数据说明

1.1 数据使用

本研究共收集使用政企产品领域的语料530篇，涵盖Word、PDF和PPT等格式，其中一级产品大类11类，包括IDC、大数据、固化语音等；二级产品小类79类，包括IDC增值、IDC基础、固话模拟线、行业短信等；涉及产品252个，token量2 203 278个。

1.2 知识向量化

在本研究中，采用了通用RAG的外挂知识向量库的方式，因此需要对原始数据进行半自动化分片处理，程序自动标注原始文档，随后由人工手动检查并修订标注内容，并使用中文文本嵌入模型m3e-base（Moka Massive Mixed Embedding）对分片内容进行向量化，存储到Milvus的向量库中。

1.3 知识抽取

在本研究中，使用了知识图谱进行领域知识抽取，采用Agent框架[8]对原始文档语料进行自动化信息抽取，包括命名实体识别和事件抽取。其中，实体名称包括产品名称和产品对应的属性名称；事件抽取的schema是由属性名称和属性值组成的结构体。Agent框架集成了文档文段、属性名称替换和文本向量化存储等自定义工具，并设计了规划者与操作者2个角色。规划者对输入的query进行思考和解析，通过对环境的感知来规划相应的操作路径，并根据反馈意见进行路径纠正；操作者根据规划者提供的路径按步骤执行，并将执行结果自动反馈给规划者；实现了用户输入—路径规划—抽取操作—结果反馈—路径纠正的自动化信息抽取流程，从而将原始数据抽取为知识体系。

2 技术方法

在本研究中，通过评估算法定义智能问答系统的*F*1值，以此来有效评估智能问答系统的精准性，同时使用提示工程技术、增强检索生成技术和模型微调技术将智能问答系统的*F*1值提高到92.36%，达到B端用户的可用性要求。因此，本节将详细介绍评估算法机制，及使用提示工程技术、增强检索生成技术和模型微调技术，提升智能问答系统*F*1值的策略机制。

2.1 智能问答系统的评估算法

在智能问答系统设计前期[9]，首先对系统的精准性进行了定义，用以保障系统的正向迭代和调整。在评估算法上，采用了人工标注准确率+ROUGE-L算法自动获得召回率，综合准确率和召回率得到*F*1值，以此作为智能问答系统的精准性评估算法。

2.1.1 人工标注准确率

按类别进行分类定义：完全正确、基本正确、部分正确、完全不正确。

准确率：将所有预测为正确（即完全正确、基本正确和部分正确）的回答的权重之和除以系统给出的所有回答的数量。准确率（Precision）公式如下：

$$\text{Precision} = \frac{\sum_{i=1}^{n} w_i}{n} \tag{1}$$

其中，n表示预测的回答总数，w_i表示第i个回答的权重，权重取值：{[1，完全正确]，[0.75，基本正确

分],[0.5,部分正确],[完全错误,0]}。

2.1.2 ROUGE-L算法自动获得召回率

ROUGE-L算法通过计算系统生成的回答内容与参考答案之间的最长公共子序列(LCS)的长度自动计算召回率,如果生成文本中的大量内容与参考答案相匹配,则该文本的召回率相对较高。

ROUGE-L的召回率和准确率公式如下:

$$R_{lcs}=\frac{LCS(X,Y)}{m} \tag{2}$$

$$P_{lcs}=\frac{LCS(X,Y)}{n} \tag{3}$$

其中,m表示参考答案的长度;n表示系统生成答案的长度;LCS(m,n)表示系统生成答案与参考答案之间的最长公共子序列的长度;R_{lcs}表示ROUGE-L的召回率,即生成答案中与参考答案匹配的内容占参考答案总长度的比例;P_{lcs}表示ROUGE-L的准确率,即生成答案中与参考答案匹配的内容占生成答案总长度的比例;召回率(Recall)是ROUGE-L算法Recall和Precision的调和平均值:

$$\text{Recall}=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}} \tag{4}$$

其中,β是计算调和平均$F\beta$值的参数,用于调节准确率和召回率的权重。当β大于1时,更重视召回率;小于1时,更重视准确率;等于1时,两者权重相等,即$F\beta$值等于$F1$值。

由式(1)和(4)得出$F1$评估值计算公式:

$$F1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

2.2 智能问答系统$F1$值提升策略

使用开源通义千问Qwen-14B作为基座模型,结合提示工程技术、增强式检索生成技术和大模型微调技术,实现了智能问答系统$F1$值的提升。智能问答系统$F1$值提升策略流程如图1所示。

2.2.1 提示工程技术

在本研究中,使用提示工程技术设立了多种模版,主要采用了提示工程的思维链(Chain of Thought,CoT)的方式[10],在模版中先定义模型的角色和目标任务,再列举思维路径示例,让模型分解思维过程,并为每个步骤提供更多的推理能力,从而提高模型回答质量和逻辑推理能力,例如:

a)角色设定:你是一个政企产品营销顾问。

b)目标任务:根据历史对话内容,改写用户的问题,以补全相关的产品或属性信息。

c)思考步骤:判断用户的问题是否是政企产品的相关信息查询。如果不是政企产品相关信息查询,则不做处理,直接返回用户最后的问题。

(a)分析用户问题中是否包含了产品信息和属性信息。如果产品和属性信息都有,则不做处理,直接返回用户最后的问题。

(b)如果产品或属性信息不完整,则分析并理解历史对话内容,尝试改写用户最后的问题,补全产品或属性信息。

(c)比较改写后的问题与原始问题,检查改写是否改变了用户的意图,如果改写可能造成用户问题语义的改变,则放弃改写,直接返回用户的原始问题。仅输出原始问题或者改写后的问题。

d)成果示例:

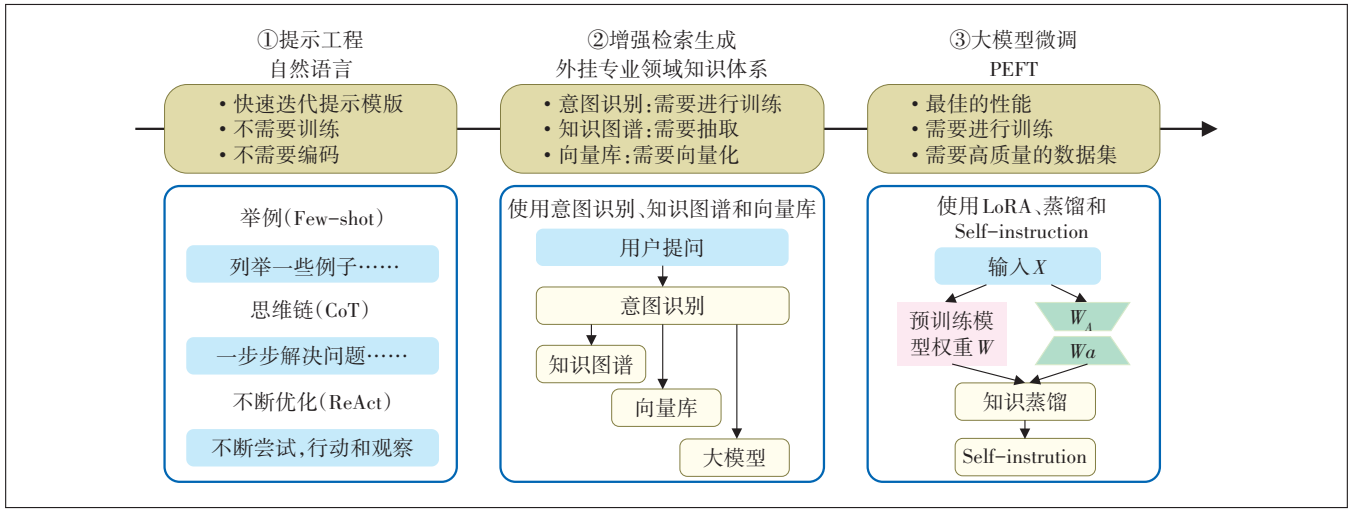

图1 智能问答系统$F1$值提升策略流程

（a）历史对话：{用户：5G烽火台的产品定位与目标客户。系统：XXXX。}用户问题：产品优势呢？输出：5G烽火台的产品优势是什么？

（b）历史对话：{用户：5G烽火台的产品定位与目标客户。系统：XXXX。}用户问题：中华人民共和国哪年成立的？输出：中华人民共和国哪年成立的？

（c）历史对话：{用户：行业短信的产品优势。系统：XXXX。用户：UID产品的已落地场景有哪些。系统：XXXX。}用户问题：目标客户呢？输出：UID产品的目标客户是什么？

2.2.2 增强检索生成技术

通用RAG通过外挂向量库，将专业领域知识进行向量化，并存储到向量库中。在进行检索时，通过向量内积将相似内容进行召回，再将召回内容传给大模型，由大模型进行总结并给出回答，这种方式的缺点是准确度不高，达不到B端用户的需求。智能问答系统采用的增强检索生成技术是在通用RAG的外挂向量库上集成了用户意图识别、知识图谱技术来提升准确率和召回率。

2.2.2.1 用户意图识别

使用预训练好的BERT（Bidirectional Encoder Representations from Transformers）模型作为基础架构，利用其预训练的词向量和句子表征。通过收集到的意图识别数据集，包括产品属性查询、产品范围查询、系统咨询、名单制客户查询等10个意图的训练数据，对数据进行清洗，将其转化为BERT模型所需的输入格式，然后对不同意图的数据集进行实体标注，通过标注句法，额外融入词性、句法分析，进一步提升模型性能，使得实体提取和意图分类更加准确，再经过模型训练，构建出精准且具有鲁棒性的意图识别模型。

通过可视化的意图识别配置流程，对用户意图识别进行流程配置调整，以提升回答的$F1$值，用户意图识别配置流程如图2所示。

2.2.2.2 知识图谱

在数据准备阶段，对原始数据进行知识的抽取，将抽取的多元组构建成闭环的政企产品知识体系，使智能问答系统能够更精准地检索到专业领域知识，从而提升了回答的准确性和深度。

2.2.2.3 大模型微调

为了进一步提升增强检索生成技术在专业领域的应用效果，采用了LoRA（Low-Rank Adaptation）算法[11]对预训练语言模型进行微调，以融合专业领域的知识。LoRA通过在预训练模型的每个Transformer块中添加低秩分解矩阵，实现了参数高效的模型微调。

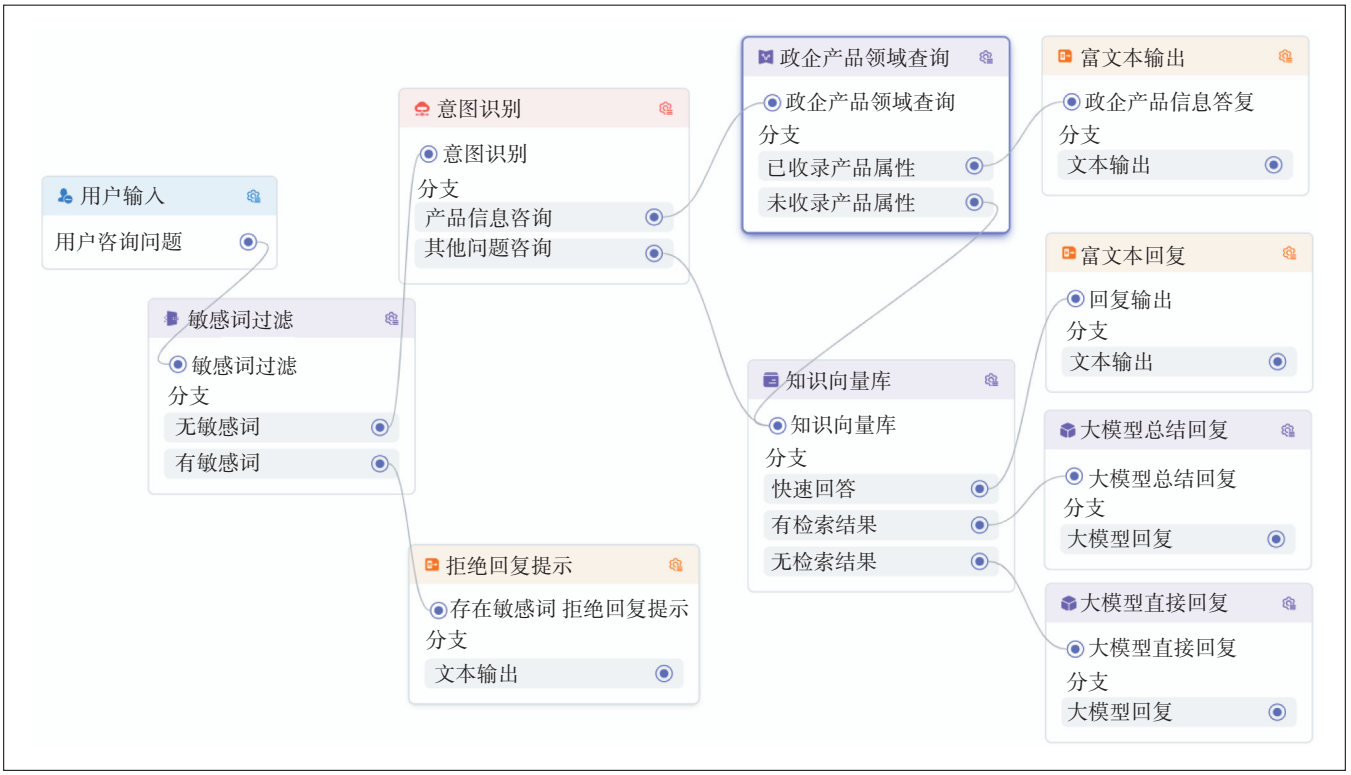

图2 用户意图识别配置流程

具体而言，对于模型中的每个权重矩阵 W，LoRA 定义了一个秩为 r 的矩阵 ΔW，其中 r 远小于 W 的维度。在微调过程中，使用了 1 500 条已标注的专业领域数据，只更新 ΔW 矩阵的参数，而保持预训练权重 W 不变。这样，就可以在不大幅增加模型参数量的情况下，对预训练模型进行高效的领域自适应。

在实践中，使用 1 500 条已标注的专业领域数据构建了微调数据集，并基于 LoRA 算法对预训练模型进行了微调。通过引入领域知识，显著提升了模型对领域实体、术语、概念等的理解和生成能力。微调后的模型生成的答案更加准确、专业。

除了 LoRA 微调，还采用知识蒸馏[12]和指令微调[13]等技术来进一步提升模型效能。知识蒸馏是一种将大型复杂模型（教师模型）的知识转移到小型简单模型（学生模型）的技术。通过最小化学生模型和教师模型输出分布之间的差异，学生模型可以在更小的参数规模下取得与教师模型相当的性能。在实验中，本研究使用 LoRA 微调后的大模型作为教师，将其知识蒸馏到一个更小的学生模型中，在保证效能的同时减小了部署所需的计算资源。

指令微调则是一种通过自然语言指令来引导语言模型执行特定任务的技术。不同于传统的微调方法使用任务特定的标注数据，指令微调只需要少量的自然语言指令样本，就可以让模型学会执行新的任务。我们利用指令微调来增强模型对领域任务的理解和执行能力。通过引入领域相关的指令，模型可以更好地理解用户意图，并根据指令生成符合要求的答案。

在整体大模型微调上，本研究通过开源 PEFT 项目中的 LoRA 算法进行微调，实现了参数高效的领域知识融合，并辅以知识蒸馏、指令微调等技术，全面提升了增强检索生成模型在专业领域的表现。这些方法有效地提高了模型对专业领域知识的理解和推理能力，使其能够生成更加准确、专业、符合用户需求的答案。

3 工程方法

3.1 性能优化

本研究在提升智能问答系统 $F1$ 值的同时，在工程上也对性能进行优化。采用了 vLLM 大模型[14]加速机制，通过优化计算资源的分配，减少了模型推理时间，使系统能够快速响应用户查询。同时，还对模型架构进行了精简和优化，采用了模型量化和知识蒸馏等技术，进一步减少模型的计算负担，提高了处理速度。

模型量化是一种通过降低模型权重和激活值的数值精度来压缩模型大小和加速推理过程的技术。在本研究中，采用了后训练静态量化（Post-Training Static Quantization，PTQ）方法，将训练好的模型权重从 FP32 精度量化为 INT8 精度，并通过最小化量化前后输出分布的差异来确定量化比例因子。通过模型量化，将模型大小压缩至原来的 1/4，推理速度提升了 2~3 倍，在保证模型性能的同时显著降低了部署所需的存储和计算资源。

通过实施上述优化策略，智能问答系统不仅提升了 $F1$ 值，而且显著提高了推理速度，降低了资源消耗，为实际应用部署提供了有力支持。

3.2 安全保障

本研究在开发和部署智能问答系统时，主要参照《中国联通人工智能隐私保护白皮书》[15]进行设计和研发，以确保数据安全和用户隐私得到充分保护。重点实施了以下隐私保护措施：一是应用隐私保护管控技术，对数据全生命周期进行安全管控；二是采用隐私保护数据加密技术，保障数据机密性；三是部署隐私保护攻击防御技术，以抵御针对人工智能系统的各类隐私攻击。通过这些措施，有效保障了系统数据安全和用户隐私。

3.3 伦理保障

在本研究中，智能问答系统采用了 Qwen-14B 大规模预训练语言模型。尽管大模型具有强大的语言理解和生成能力，但在其预训练阶段，可能会学习到数据中隐含的一些偏见，从而导致模型输出存在算法偏见问题[16]。根据中国信通院发布的《人工智能伦理风险分析报告》，算法偏见已成为大模型面临的主要伦理风险之一。比如微软小冰、Meta 的 Galactica 等对话大模型都曾因偏见问题被下线整改。

为减轻智能问答系统的算法偏见风险，本研究采取了以下措施：一是在对大模型进行微调时，引入了无偏见的高质量数据，纠正模型原有的偏差；二是在应用部署阶段，针对敏感问题采用过滤机制，拦截有偏见或歧视性的不和谐内容；三是建立持续监管机制，通过人工抽检和人工反馈等方式，动态识别和修正系统中的算法偏见问题。通过在大模型基础上采取有针对性的改进措施，可以较为有效地预防和消除由算法偏见引起的伦理风险。

a）减轻算法偏见。智能问答系统通过使用专业领域数据，以及多元化和代表性的数据集来微调大模型，从而减少潜在的偏见。同时，采用算法和人工审核相结合的方法来识别和纠正偏见。

b）敏感词过滤。智能问答系统加强了敏感词过滤的功能，以确保不合理或不和谐词语的出现。

3.4 适配性与扩展性

本研究对智能问答系统也实现了适配性和扩展性。

a）适配性。系统提供了丰富的API接口和插件机制，允许快速集成到现有的IT架构中，如OA系统或其他业务应用，因此其强大的集成能力保障了系统的适配性。

b）扩展性。系统采用微服务模块化架构，其每个微服务模块负责特定功能，可以独立升级和扩展，保障了系统的便利扩展性。

4 整体架构

本研究最终构建出的政企产品营销指引智能问答系统技术架构分为6层，包括知识索引层、推理能力层、意图识别层、信息降噪层、人机交互层、触点层，智能问答系统技术架构如图3所示。

知识索引层：通过文档智能技术对文档进行标准化，对标准化后的文档进行数据增强生成QA与摘要，对文档进行基于语义的知识切片，对文档切片进行Embedding后存入Milvus向量库；通过对原始语料进行知识强化与挖掘，形成领域知识图谱。

推理能力层：基于大模型进行知识总结、信息对比、文本生成、数值计算以及智能推荐。

意图识别层：大模型+小模型的意图识别层，通过BERT模型训练意图识别对用户意图进行初步理解，以提高系统响应速度，通过LoRA微调大模型，开发意图识别模型，以提高对用户意图的理解能力。

信息降噪层：通过BART（Bidirectional and Auto-Regressive Transformers）模型对用户输入进行纠错，包含漏字纠偏、错别字纠偏、同音字纠偏，以提高用户体验。

交互收集层：支持语音和文本形式的人机交互能力。

触点层：通过触点层，把智能问答服务开放给其他应用系统调用。

5 实验验证

5.1 精准性实验

本研究使用原始数据通过商业生成式大模型和人工的方式生成上千条问答对，这些问答对分为有专

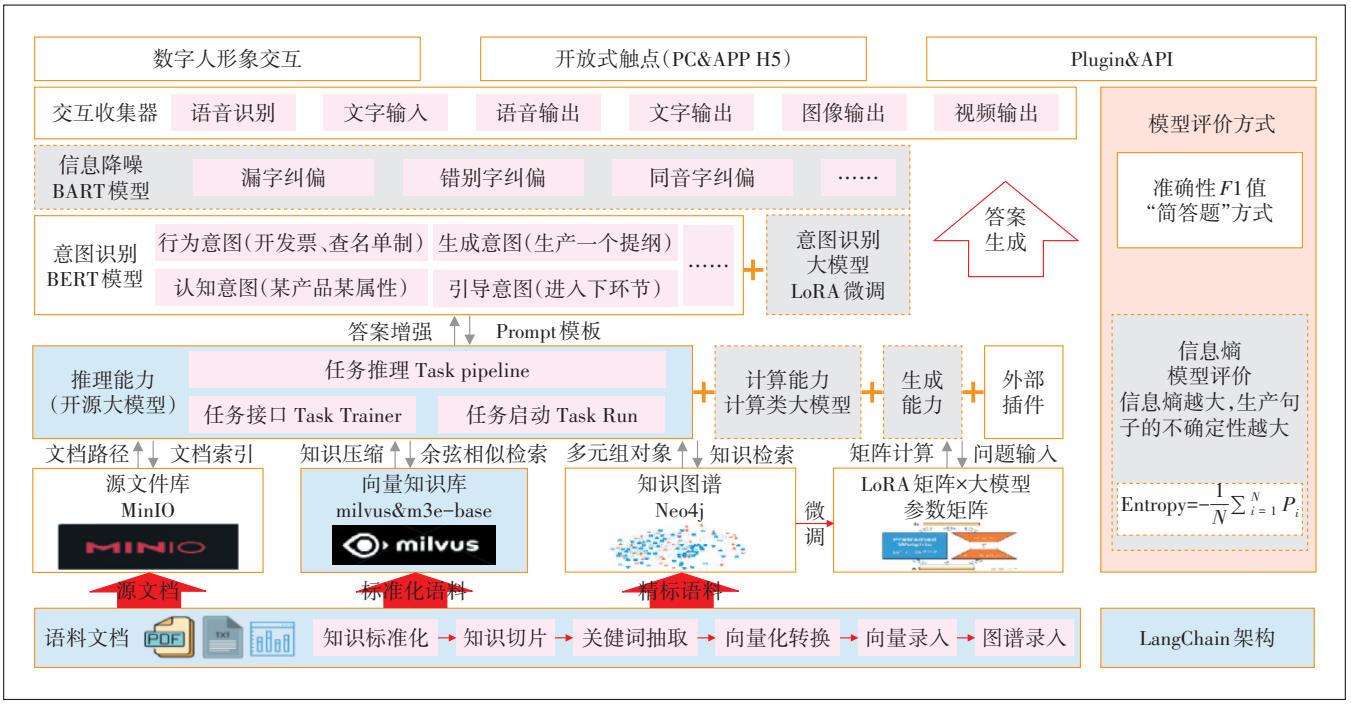

图3　智能问答系统技术架构

业知识固定的Q/A对、抽取为知识图谱的专业知识问答对和将专业知识向量化到向量库的问答对。使用这些问答对，在多台A800-80G GPU上，针对3种不同的技术策略进行了多次实验，以测试智能问答系统对不同问题类型的处理效果，并对*F*1值进行均值对比。

5.1.1 第1种技术策略实验

主要采用了大模型与向量库的技术组合，将政企的专业领域数据进行分片处理并向量化，存入向量库中。在问答检索阶段，系统通过向量库的内积算法进行相似度检索，以召回相关性的内容，再将召回的内容输入给大模型，通过prompt的提示模版，引导大模型对召回的相关内容进行总结答复。

在第1种技术策略的实验中，共使用1 200条问答对进行了10次实验，智能问答系统的*F*1平均值为57.61%。

5.1.2 第2种技术策略实验

通过引入意图识别技术，使系统能较为精准地识别用户意图，进行更为精准的检索，提升召回率和准确率。第2种技术策略的实验共使用1 200条问答对进行了10次实验，智能问答系统的*F*1平均值为78.21%。

5.1.3 第3种技术策略实验

通过引入知识图谱技术，将政企专业领域的知识进行实体、关系、属性和内容的抽取，在提高准确率的同时也构建出政企的知识体系。第3种技术策略的实验共使用1 200条问答进行了10次实验，智能问答系统的*F*1平均值提高到92.36%，为最优技术策略。

3种技术策略实验结果如表1所示。

表1 3种技术策略实验结果

实验	第1种技术策略				第2种技术策略				第3种技术策略			
	Q/A问答对/%	知识图谱问答对/%	向量库问答对/%	平均/%	Q/A问答对/%	知识图谱问答对/%	向量库问答对/%	平均/%	Q/A问答对/%	知识图谱问答对/%	向量库问答对/%	平均/%
1	97.10	99.50	80.30	92.30	93.50	61.20	79.30	78.00	97.10	99.50	80.30	92.30
2	96.00	99.80	81.20	92.33	95.20	62.00	81.60	79.60	96.00	99.80	81.20	92.33
3	96.90	99.70	79.00	91.87	92.80	64.40	80.50	79.23	96.90	99.70	79.00	91.87
4	95.80	99.60	82.10	92.50	94.30	60.50	78.90	77.90	95.80	99.60	82.10	92.50
5	97.50	99.90	80.80	92.73	91.60	63.70	82.20	79.17	97.50	99.90	80.80	92.73
6	96.60	99.40	79.90	91.97	90.90	62.80	80.10	77.93	96.60	99.40	79.90	91.97
7	95.50	99.30	81.60	92.13	93.10	61.90	79.80	78.27	95.50	99.30	81.60	92.13
8	97.30	99.80	80.50	92.53	92.40	63.30	81.30	79.00	97.30	99.80	80.50	92.53
9	96.20	99.50	81.90	92.53	94.70	60.80	79.50	78.33	96.20	99.50	81.90	92.53
10	97.70	99.70	79.50	92.30	91.30	64.10	80.90	78.77	97.70	99.70	79.50	92.30

综合实验结果如表2所示，系统方案采用的第3种技术策略为最佳技术策略，*F*1值达到92.36%，显著提升了智能问答系统的性能、精确度和用户体验。实验结果对比如图4所示。

5.2 性能实验

基于多台A800-80G GPU的环境，对不同问题类型进行智能问答系统的性能实验。在有线网络的环境下，使用1 000条问答对进行首字符响应的性能实验。实验结果显示，无加速情况下首字符的平均响应时间为3.5 s，加速后首字符的平均响应时间为2.8 s，平均首字符响应时间缩短了20%。不同问题类型进行智能问答系统的性能实验结果如表3所示。

表2 综合实验结果

技术	技术策略	*F*1均值/%
第1种	提示工程+向量库+大模型	57.61
第2钟	提示工程+意图识别+向量库+大模型	78.21
第3种	提示工程+意图识别+向量库+知识图谱+大模型	92.36

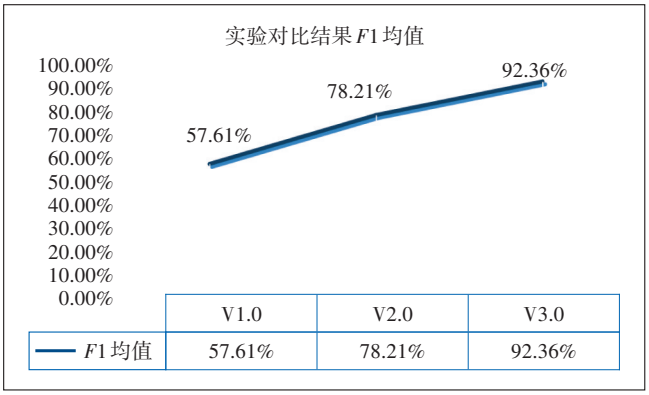

图4 实验结果对比

表3　不同问题类型进行智能问答系统的性能实验结果

实验	Q/A问答对/s	知识图谱问答对/s	向量库问答对/s	大模型/s	平均/s
无加速	1.4	1.6	4.1	3.4	3.5
使用加速	1.5	1.5	3.1	2.3	2.8
实验	Q/A问答对	知识图谱问答对	向量库问答对	大模型	平均

同时，也对大模型平均每秒生成的token数进行了检测与对比。实验结果显示，无加速的生成速度为20 tokens/s，加速后的生成速度为22.4 tokens/s，生成速度提升了12%。

6 结束语

政企产品营销智能问答系统以开源通义千问Qwen-14B作为大模型底座，综合运用RAG、BERT意图识别、大模型微调以及知识图谱构建等先进技术，精准识别用户意图，降低大模型幻觉、精准度低和偏见问题。在政企产品营销领域的530篇语料数据集上进行的实验中，通过引入知识图谱、大模型微调等优化措施，将智能问答系统的$F1$值从78.21%提升至92.36%，相比优化前提升了14.15个百分点，达到了行业应用的标准，符合一线使用要求。整体系统采用微服务模块化架构，保障了系统的便利扩展性。系统提供了丰富的API接口和插件机制，可以很好地适配各类应用系统和场景。此外，通过vLLM大模型加速机制和构建知识体系的闭环，保证性能稳定。

随着AI技术的不断进步和行业需求的不断变化，智能问答系统的发展将持续加速。未来，系统将更加强调深度学习、自适应学习和多模态理解能力。同时，随着多智能体技术的集成，系统将变得更加角色化、智能和灵活，能够在更多的领域和场景中发挥作用，为用户提供更加丰富、高效、个性化的服务。

参考文献：

[1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: 9459-9474.

[2] KASNECI E, SEßLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and Individual Differences, 2023, 103: 102274.

[3] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30: 681-694.

[4] 王苏静，渠继鹏，唐楠. 从ChatGPT到AIGC人工智能重塑千行百业[M]. 北京：电子工业出版社，2023.

[5] WU T Y, HE S Z, LIU J P, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122-1136.

[6] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2024: 27730-27744.

[7] 慈颖，秦留洋，韩惠婕. 基于航天装备数据的知识图谱体系研究[J]. 计算机测量与控制，2023，31(5)：249-254.

[8] XI Z H, CHEN W X, GUO X. The rise and potential of large language model based agents: a survey[EB/OL]. [2023-12-24]. https://arxiv.org/pdf/2309.07864.pdf.

[9] 黄巧. 基于知识的问答系统的设计与实现[D]. 济南：山东大学，2022.

[10] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2024: 24824-24837.

[11] HU E J, SHEN Y L, WALLIS P. LoRA: low-rank adaptation of large language models[EB/OL]. [2023-12-24]. https://arxiv.org/abs/2106.09685.

[12] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[EB/OL]. [2023-12-24]. https://arxiv.org/abs/2109.01652.

[13] HINTON G, VINYALS O, DEAN J, et al. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.

[14] KWON W, LI Z H, ZHUANG S Y, et al. Efficient memory management for large language model serving with pagedattention[C]//Proceedings of the 29th Symposium on Operating Systems Principles. New York: Association for Computing Machinery, 2023: 611-626.

[15] 中国联通研究院，中国联通网络安全研究院，下一代互联网宽带业务应用国家工程研究中心. 中国联通人工智能隐私保护白皮书[R/OL]. [2023-12-24]. http://221.179.172.81/images/20231129/92411701242535945.pdf.

[16] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots: can language models be too big?[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2021: 610-623.

作者简介：
陶晓英，高级工程师，硕士，主要从事电信运营商数字化转型与数据要素赋能等方面的研究工作。