

DOI: 10.13955/j.yzyj.2024.04.11.07

# 基于 LLM 和 RAG 的中邮网院智能客服系统研究

张丽静, 杜冬梅\*, 刘庆芳, 刘海云

(石家庄邮电职业技术学院, 河北 石家庄 050021)

**摘 要:** 中邮网院客服为百万邮政员工提供培训、学习、考试等全方位服务支持的同时, 存在着在学习高峰期回复不及时、全天候接待能力不足等问题。为解决这些问题, 文章设计并实现了基于大语言模型、LangChain 框架、pgVector 向量数据库以及表示学习等技术的智能客服系统原型, 旨在利用大模型理解用户复杂的自然语言输入, 随时随地以更高效、准确的方式回答用户所遇到的问题, 在降低中邮网院客服人工成本的同时, 提升用户体验感和满意度。经测试验证了基于检索增强生成和大语言模型实现智能客服的可行性, 为中邮网院后续进行智能客服系统建设和选型提供了参考。

**关键词:** 大语言模型; RAG; LangChain; 句向量表示学习模型; 向量数据库

**中图分类号:** F61

**文献标识码:** A

中国邮政网络学院(以下简称“中邮网院”)是面向中国邮政百万员工落实大规模教育培训, 支撑“人才强邮”战略, 提高全员素质, 提升企业核心竞争力的平台。中邮网院客服为邮政员工提供全方位的服务支持, 解决系统使用过程中遇到的有关培训、学习、考试等问题。目前, 人工运维方式存在一定的局限性。例如, 有些复杂或者专业性强的问题可能会因客服自身知识范围的限制而无法得到解决, 在学习高峰期存在回复不及时、全天候接待能力不足等问题。随着人工智能技术的发展, 上述问题可通过构建智能客服系统得以解决。

大语言模型(Large Language Model, LLM)通过深度学习技术, 能够理解和生成自然语言, 不仅能够捕捉语言的细微差别, 还能在复杂的对话中保持上下文的连贯性。但是大语言模型存在知识幻

想、数据实时性差、不具备特定领域知识等问题, 导致大模型不能直接应用于客服对话场景。

检索增强生成(Retrieval Augmented Generation, RAG)是一种结合了 LLM 和信息检索系统的技术。RAG 使大语言模型能够利用特定领域知识库提高文本生成的准确性和丰富度。当响应用户查询时, RAG 首先利用检索系统从特定领域知识库中检索相关内容, 然后将检索到的内容与原始查询一同输入到 LLM, 从而使大语言模型能够获取最新的信息, 解决模型的知识幻想、特定领域知识缺乏等问题, 并产生可靠的输出。

本研究旨在综合大语言模型、文本嵌入、向量数据库等相关技术, 搭建基于 LLM 和 RAG 的智能客服系统, 通过文本嵌入技术将中邮网院客服常见问题存放到本地向量数据库, 用于 RAG 进行

**基金项目:** 邮政应用技术协同创新中心资助项目(项目名称: 基于开源的大语言模型的企业知识服务研究; 项目编号: YB2023002)。

**作者简介:** 张丽静(1995—), 女, 河北邢台人, 硕士, 主要从事人工智能技术研究; 杜冬梅(1969—), 女, 河北辛集人, 通信作者, 硕士, 高级工程师, 主要从事软件工程和数据库管理研究; 刘庆芳(1981—), 男, 河北井陉人, 博士, 高级工程师, 主要从事知识工程与远程教育研究; 刘海云(1971—), 男, 北京人, 硕士, 正高级工程师, 主要从事信息技术、网络教育研究。

**收稿日期:** 2024-04-19

**本刊网址:** yzyj.sjzpc.edu.cn

知识检索，赋予大模型特定领域知识，利用大模型理解用户复杂的自然语言输入，以便随时随地以更高效、准确的方式回答用户所遇到的问题，在降低中邮网院客服人工成本的同时，提升了用户体验感和满意度。

## 1 相关技术研究

### 1.1 大语言模型

大语言模型，是一种由包含数百亿以上参数的深度神经网络构建的语言模型。2017年，Transformer模型提出并在机器翻译任务上取得巨大成功，为大语言模型研究奠定了基础，当前大语言模型基本都是基于Transformer结构。2018年，OpenAI和Google相继提出了基于Transformer的大规模预训练语言模型GPT和BERT。2019年，OpenAI发布15亿参数的GPT-2模型。2020年，Google发布110亿参数的T5模型，OpenAI发布拥有1750亿参数的GPT-3模型。2022年11月，ChatGPT的出现使大语言模型的研究达到了新的高峰；ChatGPT能够有效理解用户意图，并结合上下文情境给出合适的回答——不仅限于进行普通对话，还能处理复杂的任务，包括撰写文章和解答问题等。ChatGPT在众多任务中的表现超越了那些针对特定任务训练的监督式学习算法，凸显了大语言模型所蕴含的巨大潜力。2023年3月，OpenAI推出参数量高达1.8万亿的GPT-4模型，该模型拥有多模态理解能力，在众多基准考试中取得的分数超过了80%的应试者，展现出接近“通用人工智能（AGI）”的能力和水平。各大公司和研究机构也相继发布各类大模型，包括Google推出的Bard、智谱的ChatGLM、百度的文心一言、科大讯飞的星火、阿里云的通义千问等。大模型呈爆发式增长。

大语言模型强大的语言理解能力和泛化能力得益于Transformer架构中的自注意力机制，大语言模型能够处理长距离的依赖关系，这使得模型在理解复杂句子结构和上下文信息方面表现出色。同时，该类模型因其庞大的参数量和丰富的预训练数据，展现出强大的泛化能力，这意味着模型能够在多种语言任务和领域中迁移应用，而无需从头开始训练。

然而，大语言模型存在实时性差、缺乏领域

知识、可解释性差、知识幻想、安全性和隐私问题，只能生成和处理训练数据相关的问题，不能生成和获取实时的数据与信息。而且大语言模型的决策过程往往不透明，使得理解模型的工作原理和验证其输出的准确性变得困难。此外，大语言模型可能会无意中泄露训练数据中的敏感信息，或者被恶意利用来生成虚假信息。

### 1.2 LangChain 框架

随着大模型的爆发式增长，越来越多的开发者希望利用大模型厂商提供的API或者私有化部署开源大模型，以此来开发基于大语言模型的应用程序。虽然大语言模型的调用相对简单，但是要将其集成到应用程序中，仍然需要完成大量的定制开发工作，包括API集成、交互逻辑、数据存储等方面。自2022年起，多家机构和个人推出大量开源项目，旨在帮助人们快速创建基于大语言模型的端到端应用程序或流程。其中，LangChain框架作为一种利用大语言模型能力开发各种下游应用的开源框架，广受欢迎；其简化了开发大语言模型应用的难度，实现了数据感知和环境交互，能够使语言模型与其他数据源连接起来，并允许大语言模型与其环境进行交互。

LangChain包括六大核心组件：代理（Agents）、链（Chains）、索引（Indexes）、记忆存储（Memory）、模型（Models）、提示工程（Prompts），如图1所示。

Prompts用于动态生成模型输入，避免硬编码的需求。Indexes为检索索引，帮助从语言模型中提取相关信息。Chains组件链将多个组件组合在一起，解决特定的任务。Agents服务代理是语言模型与外部API交互的中介。

LangChain的设计使得大语言模型不仅能够处理文本，还能够更广泛的应用环境中进行操作和响应，大大扩展了其应用范围和有效性。通过LangChain组件，开发者可以构建出能够理解和执行复杂任务的智能系统。

### 1.3 向量数据库

向量数据库是一种专门用于存储和检索向量数据的数据库系统。与传统的关系型数据库不同，向量数据库的核心功能是进行高效的相似度搜索，即通过计算一个向量与其他所有向量之间的距离来找到最相似的向量。这种数据库通常用于处理非结

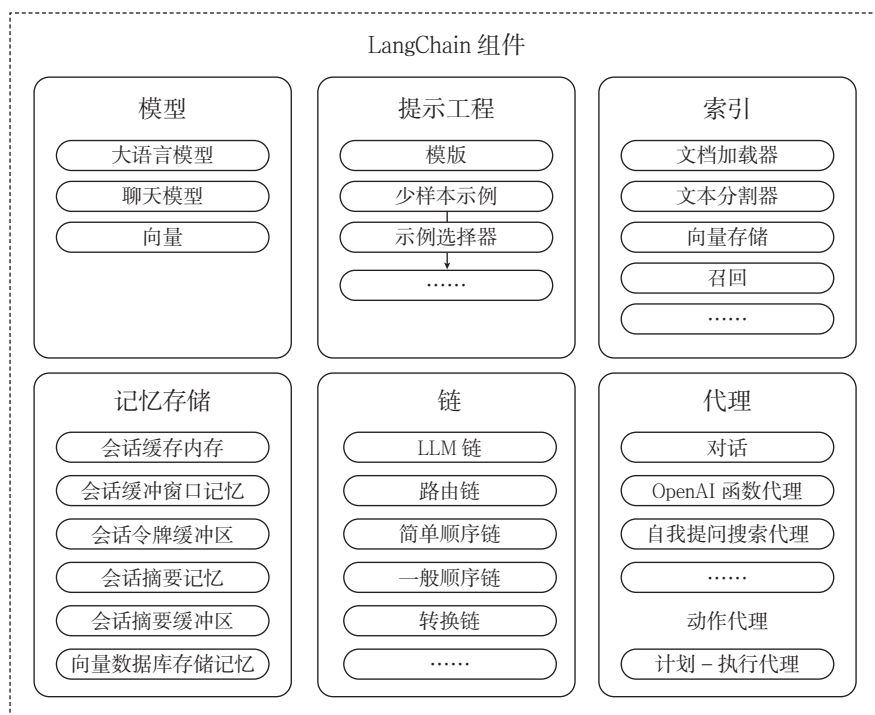


图1 LangChain 组件结构

构化数据，如文本、图像、音频和视频等，这些数据在经过向量化处理后，可以转换为高维空间中的点，从而便于进行快速的相似性匹配。

向量数据库的发展与人工智能技术的进展密切相关，随着深度学习和大数据处理技术的不断发展，向量数据库在处理复杂数据和提供智能服务方面的作用将越来越重要。

常见的开源向量数据库包括 pgVector、Milvus、Chroma 和 RedisVector 等。

Milvus 是基于原生向量设计的分布式向量数据库，可分布式计算、容灾备份、数据持久化等；但是系统复杂、资源消耗大，社区和支持比较少。

Chroma 是基于向量检索库 Faiss 构建的轻量级向量数据库，提供了更多功能和扩展性，但在使用时可能会遇到与 Faiss 相关的错误。

RedisVector 利用内存数据存储 Redis 实现高速的响应时间，但是受内存大小的限制，稳定性和成熟性较差，在处理大量复杂的向量操作方面，不如专业向量数据库高效。

pgVector 的优势在于其在 PostgreSQL 强大的开源关系型数据库基础上，通过插件形式实现，不需要进行重大的系统更改，同时扩展了 SQL 语法以处理向量操作。作为 PostgreSQL 的一部分，

pgVector 继承了 PostgreSQL 的可靠性、稳健性；同时 pgVector 可以处理大规模的向量数据，支持分布式架构，可以随着数据量的增长而扩展，具有良好的性能。

综上所述，本研究选择 pgVector 作为系统向量数据库。

## 2 系统设计

### 2.1 技术架构

智能客服系统基于 LangChain 框架结合 pgVector 作为相似性搜索引擎，搭建检索增强生成 (RAG) 应用，实现中邮网院客服的专业知识问答。技术架构分为服务层、应用层、数据库层、数据层及大模型层，如图 2 所示。

大模型层：对开源 ChatGLM3-6B、OpenAI、千问 AI、星火 AI、天工 AI、智谱 AI 等 LLM API 进行了调用封装，实现以统一的入口访问不同模型，支持模型切换。

数据层：包括源数据（中邮网院专业知识）和 Embedding API，源数据经过 Embedding 处理被向量数据库使用。

数据库层：搭建 pgVector 向量数据库，存放 Embedding 处理后的源数据。pgVector 支持欧几里

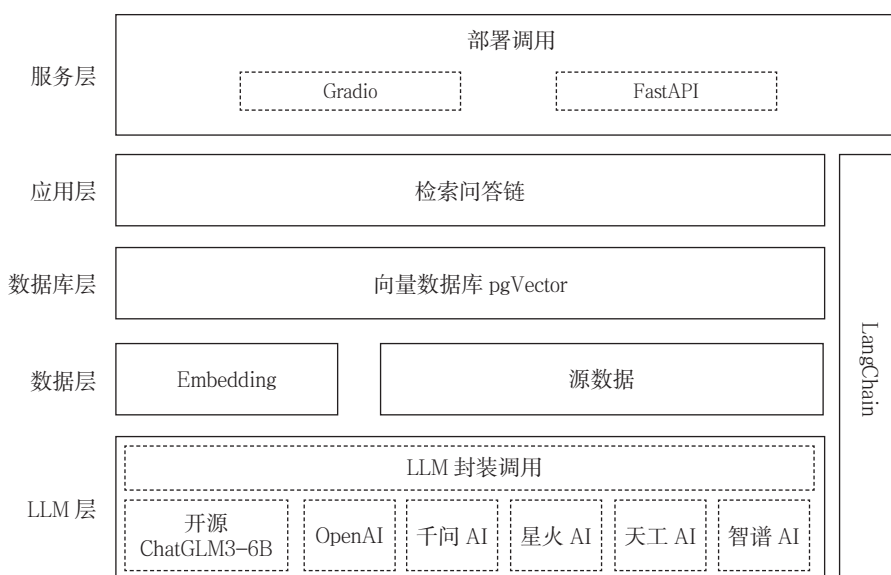


图 2 技术架构

德距离、余弦距离和曼哈顿距离等多种距离度量，可高效地存储、查询和处理向量数据；提供 k-最近邻（k-NN）搜索等高效索引选项，可保证搜索效率和准确性。

**应用层：**基于 LangChain 的检索问答链 RetrievalQA 作为系统的应用层。RetrievalQA 组合了检索器和问答链，检索器负责通过倒排序、向量空间模型等检索技术快速地从索引中找到包含查询关键词的文档，问答链通过规则和模型将用户的问题和检索到的文档信息，根据重要性组织成为有逻辑的答案。通过检索器和问答链的结合，RetrievalQA 能够高效地从大量文本数据中找到用户需要的信息，并将这些信息以易于理解的方式呈现给用户。

**服务层：**Gradio 与 FastAPI 结合实现交互式 Web 应用。Gradio 简单易用，可以快速地创建包含机器学习模型的交互式 Web 应用；FastAPI 是具有强大的类型提示和 JSON 支持的高性能 Web 框架。Gradio 创建用户界面，FastAPI 构建后端服务，二者结合使得用户能够以直观的方式与机器学习模型进行交互。

## 2.2 系统处理流程

智能客服系统主要功能是：通过 LangChain 框架将本地知识库内容与大语言模型的知识问答能力有效结合，对学员提出的问题进行个性化解答。系统处理流程如图 3 所示。

首先，收集整理中邮网院客服历年的问题及

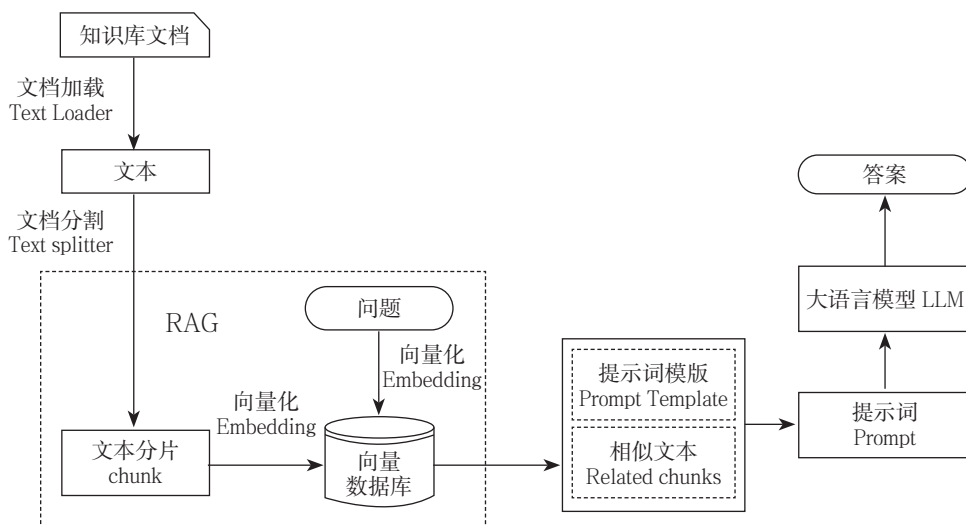


图 3 系统处理流程



相关答案,形成本地知识库。本地知识库文档通过 Text Loader 进行加载,加载过程中将文档内容转换为文本格式,以便后续处理。然后,通过 Text Splitter 按照一定规则(如段落、句子、词语等)将文档分割成较小的片段,帮助模型更好地理解 and 处理文档内容。分片后的文本通过 Text Embedding 模型,如 text2vec-base-chinese、m3e-base、bge-base-zh 等进行向量化处理,将文本转换为数值向量,使得模型更容易处理和理解文本。向量化后的文本片段存储到向量数据库 pgVector 中完成数据持久化,以便后续根据用户问题进行高效检索和匹配操作。用户提出问题后,将问题同样转换为向量形式,在向量数据库中进行相似度检索,找出与问题最接近的文本向量,形成最相似答案。最后,将找到的最相似答案作为上下文和问题一并添加到 Prompt 中,提交给大语言模型(LLM),大模型理解后组织语言输出问题答案。

### 3 实验过程与分析

实验采用 Intel64 核 CPU、128G 内存高性能计算能力的服务器,配置 NVIDIA A30 显卡, CUDA 版本为 12.0,操作系统为 Ubuntu 22.04 LTS,安装 PyTorch 深度学习框架。实验内容共包括三个部分,分别是:本地知识库构建、评测数据集构建、大语言模型选型对比。实验旨在为系统的效果提供实验证据。

#### 3.1 本地知识库构建

本地知识库基于中邮网院客服历年积累的问题及相关答案,人工整理涵盖中邮网院账号、登录、

培训、学习、考试等多方面的问题,构建形成 97 条问答对数据。部分数据内容如表 1 所示。

LangChain 支持 PDF、Word、Markdown、CSV 等多种格式的文档加载,分别对应不同的文档加载器:PyPDFLoader、Docx2txtLoader、UnstructuredMarkdownLoader、CSVLoader 等。经过对比分析不同文档格式及 LangChain 对应的文档加载器,本文最终使用 CSV 文件存储本地知识,使用 LangChain 提供的 CSVLoader 对文件进行加载,每一行作为一条知识。

#### 3.2 评测数据集构建

智能客服系统的性能依赖于多个方面,包括向量表示学习模型选型、大语言模型选型等。为验证这两个方面对于系统性能的影响,构建包含 50 个问题的问题集用于主观评估,将问题按照相关程度划分为三种类型,分别为无关问题(Completely Unrelated Question, CUQ)、部分相关问题(Some Related Question, SRQ)、完全相关问题(Completely Related Question, CRQ)。其中,完全相关问题占比 40%,部分相关问题占比 50%,无关问题占比 10%。部分评测数据集示例如表 2 所示。

本实验邀请三位专业人员进行评分,将评测数据集中的每个问题输入到智能客服系统中,根据系统给出的答案,采用十分制的方式进行评分,最终每个模型的得分为每位专家对 50 个问题打分的平均分。

#### 3.3 大语言模型选型

本系统集成当下热门的开源、闭源大语言模型,包括 ChatGLM3-6B、OpenAI、智谱 AI、星

表 1 问答对数据示例

问题	答案
如何重置密码?	打开手机 App,绑定手机后,点击“忘记密码”即可重置密码;如已登录需要修改密码:邮政员工请点击“中邮网院”App—“我的”—“安全中心”—“重置密码”,进行修改;邮储银行员工请点击“邮银 E 学”App—“我的”—“齿轮”设置—“密码设置”,进行修改
修改手机号后,收不到验证码	手机收不到验证码主要因为:1. 获取验证码手机号是否欠费停机。2. 该手机号为运营商黑名单。3. 该手机将验证短信自动屏蔽。如果手机号为运营商或是短信网关黑名单的情况,需要多发几次验证码请求,在第四次就会收到验证码
如何查询证书	从网院首页点击“学员中心”—“我的档案”—“证书”,进行查询
手机扫码时提示无效的二维码	1. 请先刷新二维码。2. 请使用中邮网院客户端,清除缓存后重新登录。3. 使用“中邮网院”App 扫描二维码。4. 如果仍有问题请拨打 400-010-1890

表 2 评测数据集示例

问题	问题类型
如何重置密码？	CRQ
如何修改照片？	CRQ
在哪里查看证书？	CRQ
课程总是有一个学习进度 94%，到不了 100% 怎么办？	SRQ
直播一打开就是白屏怎么回事？	SRQ
理财经理新人培训班有一个章节三次考试都没通过怎么办？	SRQ
个人考勤情况在哪里查看？	CUQ
集团公司征文在哪里上传？	CUQ
店招检测入口在哪？怎么使用？	CUQ

火 AI、天工 AI、千问 AI。以上多个大语言模型在基于相同的本地知识库、Embedding 模型以及参数的情况下，对于用户问题的回答，以问题“如何重

置密码”为例，如表 3 所示。

由于不同大语言模型的训练语料、模型结构存在一定差异，因此在本任务上模型效果也存在一

表 3 大语言模型回答示例

大语言模型	问题：如何重置密码？
ChatGLM3-6B	您好，要重置中邮网院平台的密码，您可以按照以下步骤操作：首先，打开手机 App 并绑定您的手机号码。接着，点击“忘记密码”选项。选择相关的服务即可进行密码重置。请注意，具体操作路径可能会因平台版本更新而有所变化，建议您在操作过程中参考相应的用户手册或联系客服寻求帮助
OpenAI	打开中邮网院手机应用，在绑定手机后，点击“忘记密码”即可重置密码。如果已经登录需要修改密码，可以按照以下步骤进行操作： 1. 邮政员工：打开“中邮网院”应用，进入“我的”—“安全中心”—“重置密码”进行修改。 2. 邮储银行员工：打开“邮银 E 学”应用，进入“我的”—“齿轮”设置—“密码设置”进行修改。 如果密码输错 10 次被锁，请等待 1 小时后再尝试输入密码。如果在登录时输入完账号和密码点击登录没有反应，请检查网络是否正常，账号密码是否输入正确，并尝试清除浏览器缓存后重新登录。如果仍然遇到问题，请拨打 400-010-1890 客服热线反馈
智谱 AI	重置密码的方法如下： 1. 打开手机 App，绑定手机后，点击“忘记密码”即可重置密码。 2. 如已登录需要修改密码，邮政员工请点击“中邮网院”App—“我的”—“安全中心”—“重置密码”，进行修改。 3. 邮储银行员工请点击“邮银 E 学”App—“我的”—“齿轮”设置—“密码设置”，进行修改。 如果您在操作过程中遇到问题，请拨打 400-010-1890 客服热线反馈
千问 AI	打开手机 App，绑定手机后，点击“忘记密码”即可重置密码；如果是邮政员工，请进入“中邮网院”App 的“我的”—“安全中心”—“重置密码”；如果是邮储银行员工，请进入“邮银 E 学”App—“我的”—“齿轮”设置—“密码设置”进行修改
星火 AI	打开手机 App，绑定手机后，点击“忘记密码”即可重置密码；如已登录需要修改密码：邮政员工请点击“中邮网院”App—“我的”—“安全中心”—“重置密码”，进行修改；邮储银行员工请点击“邮银 E 学”App—“我的”—“齿轮”设置—“密码设置”，进行修改
天工 AI	打开“中邮网院”App，点击“忘记密码”按钮，按照提示操作即可重置密码

定的差异。基于评测数据集，三位专家分别进行打分，评测结果如表4所示。

表4 大语言模型评分结果

大语言模型	专家1	专家2	专家3	平均分
ChatGLM3-6B	6.897	6.483	6.671	6.684
OpenAI	8.517	8.241	8.382	8.380
智谱 AI	8.379	8.172	8.242	8.264
千问 AI	8.069	7.966	8.024	8.020
星火 AI	8.296	7.931	8.163	8.130
天工 AI	7.034	7.241	7.143	7.139

通过评分结果可以看出，OpenAI 回答质量最优，其次是智谱 AI、星火 AI、千问 AI、天工 AI、ChatGLM3-6B。由于 OpenAI 在中国大陆地区访问受到限制，本文最终选用智谱 AI+bge-base-zh+pgVector 来构建智能客服系统。

#### 4 结束语

本研究利用大语言模型、LangChain 框架、pgVector 向量数据库以及表示学习等技术，设计和实现了一个基于本地知识库的智能客服系统原型，进行了本地私有化部署，并通过实验对比了主流开源以及闭源的大语言模型、句向量表示学习模型，分析了各项超参数对最终结果的影响效果。研究验证了基于检索增强生成和大语言模型实现智能客服的可行性，可以为中邮网院后续进行智能客服系统建设和选型提供参考。

#### 参 考 文 献

- [1] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J/OL]. [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf), 2018-06-11
- [2] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019 (8)
- [3] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020 (33)
- [4] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training

of deep bidirectional transformers for language understanding [C]. //Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019

- [5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. //Advances in Neural Information Processing Systems, 2017

- [6] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. The Journal of Machine Learning Research, 2020 (1)

- [7] Zeng A, Liu X, Du Z, et al. GLM-130b: An open bilingual pre-trained model [C/OL]. //The Eleventh International Conference on Learning Representations (ICLR). <https://openreview.net/forum?id=-Aw0rrrPUF>, 2023-02-01

- [8] Du Z, Qian Y, Liu X, et al. Glm: General language model pretraining with autoregressive blank infilling [J]. arXiv preprint arXiv, 2021 (3)

- [9] Bai J, Bai S, Chu Y, et al. Qwen technical report [J]. arXiv preprint arXiv, 2023 (9)

- [10] Wei T, Zhao L, Zhang L, et al. Skywork: A more open bilingual foundation model [J]. arXiv preprint arXiv, 2023 (10)

- [11] Yang L, Yang H, Cheng W, et al. SkyMath: Technical Report [J]. arXiv preprint arXiv, 2023 (10)