Computer Science and Application 计算机科学与应用, 2025, 15(12), 161-172
Published Online December 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1512332

Hans汉斯

基于本地化大模型与RAG技术的银行业智能客服系统研究与应用

朱宏伟[1,2]，张　妍[1,3]，曹金发[2]

[1]河北省科技金融协同创新中心，河北　保定
[2]河北金融学院金融科技学院，河北　保定
[3]河北金融学院经济贸易学院，河北　保定

收稿日期：2025年11月3日；录用日期：2025年12月3日；发布日期：2025年12月10日

摘　要

为解决传统金融客服模式响应效率低、服务个性化不足及普惠客群覆盖有限等问题，特别是助力中小银行在普惠金融服务中突破技术投入有限、专业人才短缺等数字化转型困境，本研究设计并实现了一套全链路本地化部署的智能客服系统(AI Agent)。该系统采用Ollama部署大语言模型(LLM)作为核心推理引擎与向量嵌入模型，结合Xinference部署自动语音识别(ASR)、文本转语音(TTS)与重排序(Rerank)模型，并基于Dify平台实现可视化工作流编排、内容审查与应用程序接口(API)服务管理。同时，集成RAGFlow构建本地私有知识库，对金融文档进行向量化处理与检索，形成基于检索增强生成(RAG)的技术框架。结果表明，该系统在保障金融级安全合规的前提下，显著提升了客服响应效率，实现了7 × 24小时全天候运行，并通过智能化客户洞察与个性化服务，有效增强了中小银行在普惠金融领域的服务能力与竞争力。

关键词

智能客服，本地部署，检索增强生成(RAG)，大语言模型(LLM)，中小银行，私有知识库

Research and Application of an Intelligent Customer Service System for the Banking Industry Based on Localized Large Models and RAG Technology

Hongwei Zhu[1,2], Yan Zhang[1,3], Jinfa Cao[2]

[1]Hebei Center for Technology Finance and Collaborative Innovation, Baoding Hebei

文章引用: 朱宏伟, 张妍, 曹金发. 基于本地化大模型与 RAG 技术的银行业智能客服系统研究与应用[J]. 计算机科学与应用, 2025, 15(12): 161-172. DOI: 10.12677/csa.2025.1512332

[2]College of Financial Technology, Hebei Finance University, Baoding Hebei
[3]School of Economics and Trade, Hebei Finance University, Baoding Hebei

Received: November 3, 2025; accepted: December 3, 2025; published: December 10, 2025

Abstract

To address the issues of low response efficiency, insufficient personalized services, and limited coverage of inclusive customer groups in traditional financial customer service models, especially to help small and medium-sized banks overcome the digital transformation challenges such as limited technical investment and shortage of professional talents in inclusive financial services, this study designs and implements a fully localized and end-to-end deployed intelligent customer service system (AI Agent). The system adopts Ollama to deploy large language models (LLM) as the core inference engine and vector embedding model, combines Xinference to deploy automatic speech recognition (ASR), text-to-speech (TTS), and re-ranking (Rerank) models, and is based on the Dify platform to achieve visual workflow orchestration, content review, and API service management. At the same time, it integrates RAGFlow to build a local private knowledge base, vectorizes and retrieves financial documents, forming a technology framework based on retrieval-augmented generation (RAG). The results show that the system significantly improves the response efficiency of customer service while ensuring financial-level security and compliance, operates 24/7, and effectively enhances the service capabilities and competitiveness of small and medium-sized banks in the inclusive finance field through intelligent customer insights and personalized services.

Keywords

Intelligent Customer Service, On-Premises Deployment, Retrieval-Augmented Generation (RAG), Large Language Model (LLM), Small and Medium-Sized Banks, Private Knowledge Base

1. 引言

近年来，国家[1]持续推动做好科技金融、绿色金融、普惠金融、养老金融、数字金融“五篇大文章”，《“十四五”数字经济发展规划》《金融科技发展规划(2022~2025 年)》等政策(如图 1)明确提出要强化金融服务的包容性、可及性，同时推动金融机构数字化转型。中小银行作为普惠金融服务的“主力军”[2]，需重点覆盖县域、农村等[3]普惠客群，但受自身资源禀赋制约，其客服系统多依赖传统人工或规则式智能客服，在响应效率、服务精准度上存在明显差距。同时，随着金融消费场景持续丰富与用户金融认知不断提升，客户对金融客服的需求正朝着多元化、个性化、高效化方向深度升级。传统“千人一面”的标准化客服模式已难以匹配差异化需求，而大型银行虽已率先布局生成式 AI 客服，比如，中国工商银行所推出的“工小智”智能客服平台，利用先进的算法准确分析用户的问题并给予恰当的回答，而且整合了个性化推荐功能，向用户推送有关的金融产品信息，中国建设银行所打造的“小微”智能服务平台，凭借语音识别技术达成自然语言交流[4]，从而极大提升了自动应答的速度。但其解决方案在中小银行场景下面临成本高、适配性差等挑战。

Figure 1. Overview of national policies in recent years
图 1. 近年国家政策概况

1.1. 研究现状

当前大模型智能客服的研究与实践主要呈现三种技术路线，云端通用大模型 API 调用、领域微调大模型，检索增强生成架构。

第一类云端通用大模型 API 调用方案[5]，该方案具备良好的通用语义理解能力，但在实践中存在用户隐私数据泄露风险、模型幻觉率高、API 调用成本高等问题，难以在金融场景中采用。

第二类领域微调大模型[3]，通过注入金融领域的专业知识，以提高对金融领域的适应性，但是在算力、标注数据质量以及时间成本上提出了更高的要求，这对中小银行构成较高的门槛。

第三类检索增强生成架构，通过 RAG 技术将 LLM 与知识库相结合[6]，有效弥补了模型内部知识的滞后性与不准确性，降低大模型幻觉，使之“言之有据”。但是，现有的企业级 RAG 技术多依赖于商业软件或云服务，全链路本地化、低成本的开源集成方案仍待探索。

与此同时，AI Agent 技术的发展，掀起了新的浪潮，Agent 整合各种资源处理信息，能够处理更多 LLM 所不能处理的复杂场景，进一步拓展了大模型的应用边界，使其能够通过任务规划、工具调用与环境交互，完成更为复杂的业务流程，超越传统问答机器人的功能局限，为实现真正的智能化客户服务提供了新的可能。

1.2. 本方案创新点简述

针对上述问题，本文提出基于“本地大模型 + RAG 检索增强”混合架构的 AI Agent 金融客服解决方案，通过构建多模态交互、内容检测、私有化知识库等核心功能，实现金融客服服务的精准高效触达，为中小银行提供了一条成本可控、安全合规、易于运维的数字化转型路径。

本方案在降低代码量、数据自主可控、知识更新速度以及成本上具有显著优势。基于 Dify 平台，智能客服 Agent 支持可视化编排，大大降低了代码的需求量，同时针对复杂场景可以有更高的编排自主权，而且支持多 Agent 通过密钥区分，不同 Agent 实现不同功能，可开发符合实际需求的 Agent，更具个性化。Ollama、Xinference 部署的大模型、Dify 平台、RAGFlow 知识库整个系统都存储在本地，数据完全自主可控，有利于对数据进行二次开发(客户画像、推荐服务)，与通用大模型相比，本方案的回答言之有据，无 API 调用成本，且用户隐私泄露风险大大降低，与微调大模型相比[7]，本方案有更具个性化的知识库，不局限于金融术语，更包含本行的特色业务、特色政策，与本行紧密相连。

2. 系统架构设计

整体架构遵循模块化、可扩展和易于维护的设计原则，确保系统能够灵活适应中小银行的特定业务场景和资源约束。系统的总体架构如图 2 所示。

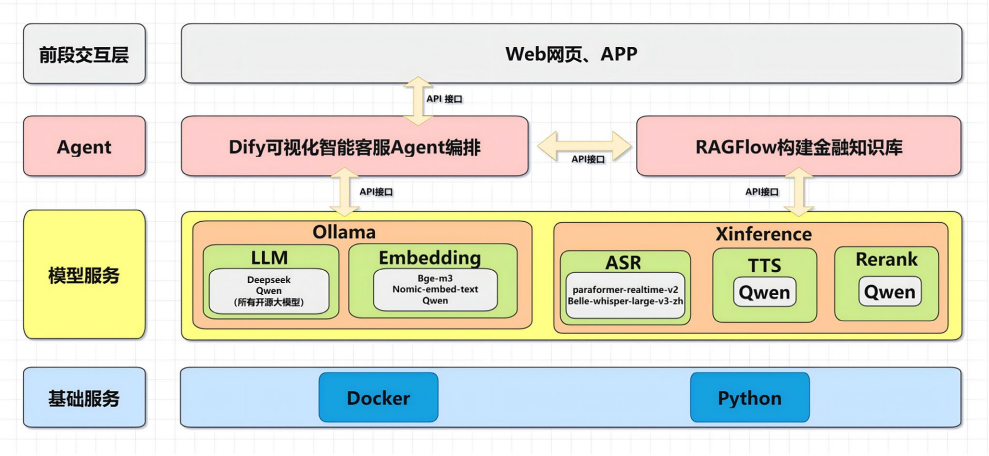

Figure 2. System architecture diagram
图 2. 系统架构图

系统采用四层分层架构，从下至上依次为：基础服务层、模型服务层、智能体(Agent)层和应用交互层。这种分层设计实现了各组件间的解耦，允许独立升级和扩展。

基础服务层作为系统的运行基石，采用容器化技术统一部署和管理核心组件，确保环境一致性和运维效率。模型服务层集成多种开源模型，为系统提供包括语言理解、语音交互、语义检索在内的核心 AI 能力。智能体(Agent)层，这是系统的“大脑”，负责理解用户意图、规划任务步骤、调用工具并生成回答，实现了从被动应答到主动服务的转变。应用交互层，为最终用户提供多样化的交互入口，确保服务能够无缝嵌入银行现有的各类渠道。

通过各层之间的协同工作，系统能够处理从简单的业务咨询到复杂的多轮对话和业务办理等多种客服场景。

2.1. 基础服务

基础服务层聚焦于底层基础设施和核心平台的部署，其核心目标是保证系统的高可移植性、易维护性和资源高效利用。

系统使用 Docker 容器化部署 Dify 和 RAGFlow。Dify 作为核心的 LLMOps (大语言模型运维)平台，提供了智能客服 Agent 的应用编排、工作流管理和对话服务等核心功能。而且其微服务架构确保了系统各组件(如 API 服务、Web 服务、工作节点)可以独立扩展和维护；RAGFlow 作为专业的 RAG (检索增强生成)引擎，同样通过 Docker 进行部署。它负责处理银行内部的各类非结构化文档(如产品说明书、监管政策、操作手册等)，实现从文档解析、智能分块到向量化索引的全流程自动化。其强大的多模态文档解析能力(支持 PDF、Word、Excel、Markdown 等诸多文件类型)和动态分块策略，为后续的精准知识检索奠定了坚实基础。

系统使用 Python 部署 Xinference，它是一个高性能且支持多类型模型的平台。在本架构中，它主要用于部署 ASR (语音识别)模型、TTS (文本转语音)模型以及用于提升检索质量的 Rerank (重排序)模型。需要说明的是 Xinference 也支持 LLM 和 Embedding 模型，但由于 Xinference 在多模型同时运行时对显

存要求较高，为节约成本，我们选择将 LLM 和 Embedding 模型用于 Ollama 部署和管理。

2.2. 模型服务

模型服务层为整个系统提供所需的人工智能大模型。当今开源大模型数量不断增多，最大开源平台 Hugging Face 单 LLM 大模型开源数量超 3000 多个，在模型上提供了更多的选择。考虑到中小银行对数据安全、成本和可控性的严格要求，我们全面采用开源模型，并通过 Ollama 和 Xinference 两个工具进行本地化部署与管理。

Ollama 直接部署于服务器中，它提供了极为简化的模型下载和运行环境，提供了多种机制和工具，帮助用户在本地实现推理效率与逻辑表现的优化。在本架构中，Ollama 主要负责部署和运行 LLM 模型和 Embedding 模型。LLM (如 DeepSeek、Qwen 等)是驱动智能对话的核心，而 Embedding 模型则用于将文本转换为向量，是 RAG 实现语义检索的关键。

Xinference 提供 ASR、TTS、Rerank 模型的部署和运行。ASR 和 TTS 模型使得 Agent 支持语音交互功能，使客户可以通过语音方式与客服系统进行交流，提升服务的便利性。Rerank 模型主要负责在 RAG 流程中，初步检索出大量相关文档片段后，对这些结果进行更精细的排序，筛选出最相关的几个片段，从而显著提升最终生成答案的准确性和相关性。

2.3. Agent

Agent 智能客服是我们整个系统的核心。Dify 提供了可视化的 Agent 编排界面(Workflow)，使得银行的技术人员可以通过添加、拖拽、连接节点并设置相关参数，来直观地设计复杂的客服流程，如通过条件分支节点判断用户是否输入“转接人工客服”或“转人工”之类的字眼来判断是否需要及时转接客服。其工作流引擎支持条件判断、循环、变量传递等逻辑，能够处理需要多步骤协作的复杂任务。

Agent 的核心能力体系主要体现在以下三个关键维度。

2.3.1. 工具调用能力(Function Calling)

Agent 能够基于对话上下文，自主调用内部系统或外部服务的工具与 API 接口(前提是 Workflow 中编排好使用该节点的触发条件)。例如，在响应用户查询账户余额的需求时，Agent 可主动调用 http 节点，访问核心业务系统中的查询接口(参考 5.3 示例图 9)，实现实时数据获取与反馈，显著提升了交互的功能性与实用性。

2.3.2. 知识与推理的深度融合

通过集成 RAGFlow 所提供的专业知识接口，Agent 成功将大语言模型的通用知识推理能力，与银行机构内部的私有知识体系(构建于 RAGFlow 知识库中)进行有机结合。该机制不仅赋予回答以常识合理性，更确保了信息内容的准确性与时效性，从而在根源上抑制了大模型常见的“幻觉”(hallucination)现象。例如，在回答关于某款理财产品收益率的问题时，其答案可直接来源于该产品最新的官方说明书，保障了信息输出的专业可信。

2.3.3. 记忆与状态管理机制

基于 Dify 对话系统所构建的上下文管理体系，Agent 能够有效记录并维护多轮对话中的状态信息。这一能力使其在复杂交互场景中保持逻辑连贯性，实现真正意义上的多轮次、有状态对话，从而更好地完成涉及多步骤、多条件查询的复合型任务。

2.4. 前端交互层

前端交互层作为系统与最终用户直接交互的界面层，承担着提供无缝、多渠道智能化服务体验的关

键职责。在本架构中，通过基于 Dify 平台提供的标准化 API 接口和 HTML 代码嵌入，实现了后台 Agent 核心能力与银行现有前端渠道体系的高效集成。

手机银行 App：在移动应用端嵌入智能客服模块，为客户提供全天候(7 × 24 小时)的业务咨询与业务办理服务，有效延伸传统银行业务的服务时间与空间边界。

微信小程序：依托银行已有的社交媒体生态，构建轻量化、便捷化的客户服务入口，实现用户高频使用场景下的即用即走式服务体验。

官方网站：在银行官方网页平台的客服中心集成智能客服能力，为传统网页端用户提供与传统渠道一致的服务体验。

本架构采用的 API 集成策略，既确保了系统核心能力的快速输出与灵活部署，又实现了前端表现层与后台业务逻辑的有效解耦。通过这一设计，银行机构无需针对不同渠道重复开发后台逻辑，显著提升了系统复用性与整体实施效率，为构建统一、连贯的全渠道智能服务体系奠定了坚实的技术基础。

3. 系统集成

3.1. 系统核心功能实现

3.1.1. 基于 RAG 的精准知识问答能力

为实现回答的准确性与权威性，引入知识库可以从根本上抑制大模型的“幻觉”现象[7]，大模型幻觉是指模型生成的内容看似合理且自信，但实际上是不真实、不准确或完全虚构的信息。这种现象在自然语言生成任务中非常常见，本系统基于 RAGFlow 构建了核心的精准知识问答能力。首先，我们将行内的产品说明书、业务规章、合规文件等非结构化文档注入 RAGFlow，利用其强大的解析与动态分块能力，构建起一个私有的、可实时更新的向量知识库。当用户发起查询时，该查询首先在知识库中进行语义向量检索，初筛出一批相关文档片段；随后，通过基于 Xinference 部署的重排序模型对结果进行精细化评分与排序，筛选出最相关的若干片段作为生成答案的可靠依据。最终，这些片段与用户问题共同构成提示词，送入 Ollama 管理的 LLM 中生成最终回答。此机制确保了系统的每一次回答均“言之有据”，显著提升了在专业金融咨询场景下的可信度。

3.1.2. 多模态交互

本系统通过部署多种专用模型并借助 API 接入 Dify 平台，构建了完整的智能多模态交互体系[8]。该系统不仅实现了文本、语音和图像三种主流交互方式的深度融合，还针对金融行业的特殊需求进行了专项优化，显著提升了金融服务的智能化水平和用户体验。

在文本交互层面，系统采用 NLU 语义解析技术结合 RAGFlow 向量知识库，构建了精准的语义理解与知识检索能力。当用户发起文本咨询时，系统能够在毫秒级内完成意图识别、实体抽取和知识匹配，并从结构化知识库中检索出最相关的专业内容，确保回复的准确性和时效性。

在语音交互层面，基于 Xinference 平台部署的 Whisper 语音识别模型实现了高质量的语音到文本转换，支持包括方言在内的多种语音输入。同时，系统集成的 TTS 技术能够将文本回复自然流畅地转换为语音输出，形成完整的语音交互闭环。这一设计特别适合移动金融场景，用户在进行复杂业务咨询时无需手动输入，大大提升了交互便利性。

在图像交互层面，系统通过 Ollama 视觉模型实现了多类金融文档的智能识别与理解。用户可直接上传理财产品说明书、身份证件、银行单据等图像资料，系统能够自动提取关键字段信息，并结合知识库进行深度分析，为用户提供准确的解答和建议。

这种全方位的多模态交互架构不仅突破了传统金融客服单一交互模式的局限，还通过多种感知渠道的协同工作，构建了更加自然、高效和人性化的智能服务体系。由此可以看出，该设计能够适应不同用

户群体的使用习惯，特别是在移动金融、远程业务办理等场景中展现出显著优势，为金融机构的数字化转型提供了有力的技术支撑。

多模态交互设计 App 端样例如图 3 所示，Web 端设计如图 4 和图 5 所示。

Figure 3. A multimodal example of the HarmonyOS App
图 3. 鸿蒙 App 端多模态示例

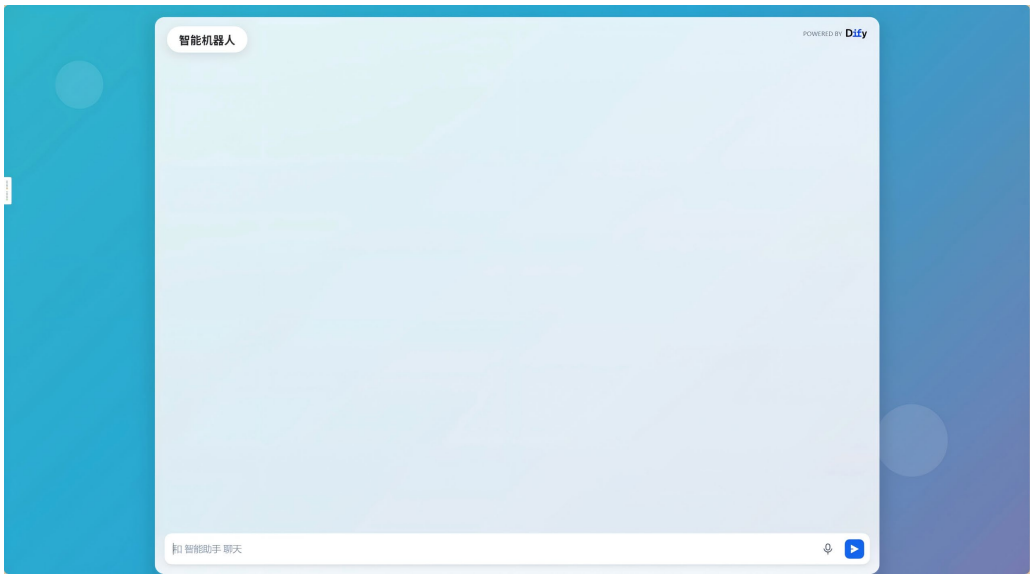

Figure 4. HTML example 1-occupying the entire page
图 4. HTML 示例 1-独占整个页面

Figure 5. HTML example 2-floating at the lower right corner
图 5. HTML 示例 2-右下角悬浮

3.2. 核心工作流与数据流通

为直观展示系统内部的数据流转与组件协同，我们选取一个典型的金融客服场景——“用户咨询高收益、低风险理财产品”进行逐步推演。该场景综合了语义理解、知识检索、工具调用等多方面能力。具体数据流程参考图 6，对话演示见 5.3 图 9。

首先用户通过手机银行 App 发起咨询，“我想了解一下你们最近利率比较高的、风险相对而言比较低的理财产品”(假设为语音消息)。前端交互层捕获音频流后，调用部署于 Xinference 上的 ASR 模型(如 Whisper)，将其精准转换为文本 Query。

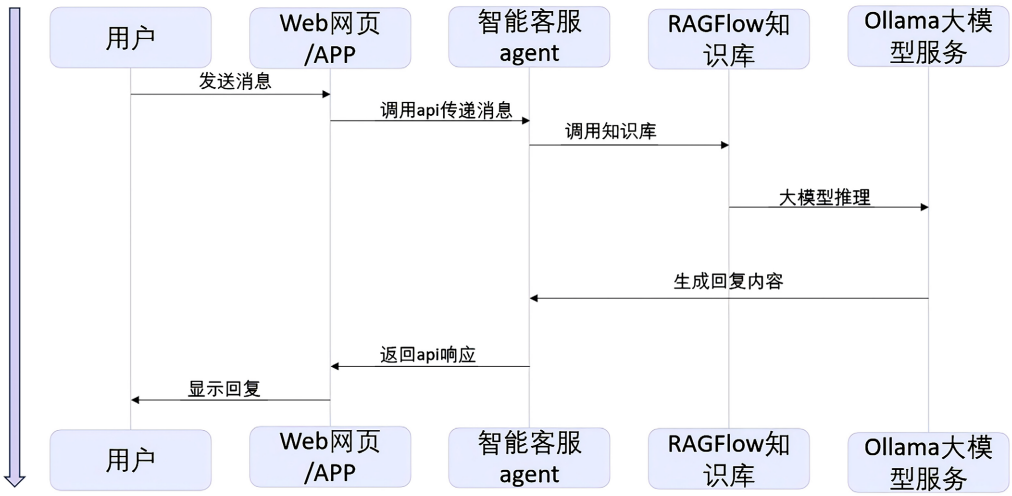

Figure 6. Example of data flow
图 6. 数据流动示例

然后，文本 Query 被发送至 Dify 平台的 Workflow。在 Workflow 中，LLM 节点(如 DeepSeek)对 Query 进行深度解析，准确识别出用户核心意图为“查询理财产品”，并成功提取出两个关键属性：“高收益

率”和“低风险”。传给下一个节点“知识检索”；知识检索节点基于解析出的意图和属性，自动调用RAGFlow 提供的检索接口，将“高收益、低风险理财产品”作为检索 Query。RAGFlow 从其已向量化的内部知识库(包含产品说明书、监管文件等)中，通过语义相似度匹配和 Rerank 模型的重排序，精准检索出最相关的几个知识片段(如“鑫安宝理财计划”和“稳盈增益”两款产品的风险等级、历史收益率说明)。

LLM 节点将 RAGFlow 返回的专业知识片段与用户问题相融合，生成一段准确、结构化且易于理解的回答。同时，Workflow 中的条件判断节点可被触发，执行 HTTP 请求，调用银行核心系统的实时产品接口，获取“鑫安宝”与“稳盈增益”的最新年化收益率，确保信息的时效性。

最终，生成的文本答案发送到前端，同时可通过部署在 Xinference 上的 TTS 模型转换为语音回复给用户，提升交互体验。

由此观之，本方案成功地将大模型的通用推理能力、RAG 的专业知识锚定能力以及外部系统的实时数据能力有机结合，实现了从“被动应答”到“主动感知、精准服务”的跨越。

4. 核心优势

4.1. 成本可控与资源优化

传统大型银行采用的先进生成式 AI 客服系统，通常依赖于大规模商业模型或需投入巨额资金进行自研，其建设与维护成本对于资源相对有限的中小银行而言难以承受。本方案通过全面采用并高效集成经过严格筛选的开源大模型(如 DeepSeek、Qwen 等)，结合 Ollama、Xinference 等轻量化本地部署与管理工具，显著降低了模型授权与推理的硬件成本。同时，基于 Dify 和 RAGFlow 等开源框架构建核心工作流与知识库管理能力，避免了昂贵的商业软件许可费用。这种技术选型策略，为中小银行提供了一条高性价比的数字化转型路径，使其能够在有限的 IT 预算内，获得接近大型银行水平的智能客服能力。

4.2. 安全合规与数据隐私保障

金融行业对数据安全与合规性有着极其严格的要求。本方案坚持核心模型与数据本地化部署原则，所有用户交互数据、业务知识库及模型推理过程均运行在银行自身可控的私有化环境中，确保了敏感金融数据不出域，有效规避了因使用外部云服务可能引发的数据泄露与合规风险。此外，通过 RAG 检索增强生成技术，将模型的知识来源严格限定并锚定于经过审核的、最新的内部知识库(如产品手册、监管政策文件)，极大地抑制了大模型普遍存在的“幻觉”问题[7]，保证了信息输出的准确性与可靠性[2]，满足金融行业信息传播的严谨性要求。

4.3. 灵活适配与高效运维

中小银行的业务场景、客群特征与大型银行存在显著差异，直接套用为大型银行设计的解决方案往往面临“水土不服”的问题。本方案采用的模块化分层架构，以及基于 Dify 工作流的可视化 Agent 编排能力，赋予了系统高度的灵活性与可扩展性。银行技术人员无需深厚的算法背景，即可通过图形化界面快速配置和调整客服对话流程、业务规则以及工具调用逻辑，从而敏捷响应业务需求的变化。容器化(Docker)的部署方式进一步简化了系统的安装、升级与维护流程，提升了运维效率，降低了技术门槛。

5. 系统可行性分析与应用演示

5.1. 与传统方案的定性对比分析

通过表 1 对比可知，本方案在成本、安全性和知识准确性这三个中小银行最关切的维度上取得了最佳平衡，同时在灵活性和适配性上显著优于传统方案，完美契合了中小银行的资源禀赋与业务需求。

Table 1. Qualitative comparison of different intelligent customer service solutions
表 1. 不同智能客服方案的定性对比

对比维度	传统规则客服	通用大模型云服务	本文提出的混合架构方案
建设与维护成本	中等(需持续维护复杂的规则库)	高(按 Token 付费，长期成本不可控)	低(一次性硬件投入，采用开源模型，无持续授权费用)
数据安全性	高(数据本地化处理)	低(数据需上传至第三方云服务商)	高(全链路本地化部署，敏感数据不出域)
知识准确性	依赖规则覆盖度，僵化，更新滞后	存在“幻觉”风险，知识更新有延迟	高(通过 RAG 锚定权威、最新的内部知识源)
功能灵活性	低(无法处理规则范围外的问题)	高(通用语言理解和生成能力强)	高(兼具通用能力与专业的工具调用能力)
业务适配性	差(业务逻辑变更需修改代码)	一般(通过提示词工程调整，可控性弱)	强(通过 Dify Workflow 可快速配置和迭代业务逻辑)

5.2. 适用场景深度剖析

本系统凭借其模块化设计与高度灵活性，可广泛应用于金融服务多个核心环节：在高频业务咨询场景中，系统能够实现 7 × 24 小时即时响应费率、理财产品等常见问题，显著降低人工客服负荷与运营成本；在复杂业务办理场景下，可以智能引导用户完成贷款申请、信用卡办理等多步骤流程，有效提升业务转化率与用户体验；此外，作为内部知识赋能工具，该系统能够为客户经理及柜员提供精准的监管政策、产品细则与操作手册检索服务，大幅提升员工专业效率与业务合规水平。

5.3. 应用演示示例

本文展示内容基于构建好的中国工商银行知识库，模拟的用户余额为 10,000 元，模型问答示例如图 7、图 8 所示。

Figure 7. Answer example 1
图 7. 问答示例 1

Figure 8. Answer example 2
图 8. 对答示例 2

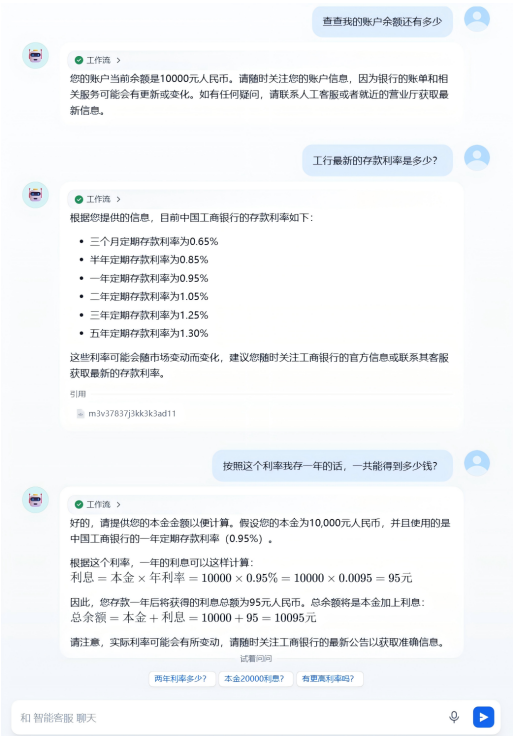

Figure 9. Context understanding and simulation interface call dialogue example
图 9. 上下文理解与模拟接口调用对话示例

6. 结论与未来展望

6.1. 结论

本研究针对中小银行在推进普惠金融进程中面临的客服响应效率不足、个性化服务能力欠缺及数字

化转型技术壁垒高等现实挑战，设计并实现了一套基于本地化大模型与检索增强生成技术的智能客服系统。该系统通过集成 Ollama、Xinference、Dify 与 RAGFlow 等开源组件，构建了一个全链路私有化部署的技术架构。该架构创新性地采用本地大模型作为核心推理引擎与 RAG 技术作为专业知识锚定的混合模式，在确保金融数据不出域、满足严格安全合规要求的前提下，有效提升了客服响应的准确性与智能化水平。系统具备多模态交互、工具调用及上下文状态管理等核心能力，能够胜任从常规业务咨询到复杂业务办理的多元化场景，显著降低了人工客服负荷，实现了 7 × 24 小时不间断的精准服务。实践表明，该方案成功在成本可控、安全合规与灵活适配之间取得了良好平衡，为资源禀赋有限的中小银行提供了一条可落地、可管控的智能化转型路径，对其提升普惠金融服务质效、增强市场竞争力具有重要的实践参考价值。

6.2. 未来展望

展望未来，本系统仍具有广阔的演进空间。鉴于绝大多数数据都存储在本地，我们可在严格遵循隐私合规的前提下，对脱敏后的用户行为与业务数据进行分析，构建用户画像，从而实现从“通用问答”到“个性化金融顾问”的跃迁，提供量身定制的产品推荐与财务建议。此外，通过引入专业化的任务型 Agent (如专注于反欺诈预警的“风控 Agent”和致力于客户关怀的“营销 Agent”)，形成多智能体协同网络。这些 Agent 通过共享状态与协作机制，能够打破传统客服系统的服务孤岛，形成业务合力，最终为中小银行客户构建一个更全面、精准与前瞻性的智慧金融服务生态。

参考文献

[1] 国务院办公厅. 国务院办公厅关于做好金融“五篇大文章”的指导意见(国办发[2025] 8 号) [EB/OL]. https://www.gov.cn/zhengce/content/202503/content_7010604.htm, 2025-03-05.

[2] 朱太辉, 张彧通. 农村中小银行数字化转型研究[J]. 金融监管研究, 2021(4): 36-58.

[3] 陆岷峰, 高伦. DeepSeek 赋能商业银行创新转型: 技术应用场景分析与未来发展路线[J]. 农村金融研究, 2025(2): 19-34.

[4] 毛茂郦. 人工智能技术重构会计领域的路径与挑战研究[J]. 商业观察, 2025, 11(19): 96-99.

[5] 刘其其, 宗亮, 韩慧宇, 等. 大语言模型在金融领域中的应用研究——以 GPT 系列大语言模型为例[J]. 北方金融, 2025(7): 14-20.

[6] 杜修平, 王崟羽. 检索增强生成赋能智能导学系统构建研究——基于本地大模型与私有知识库[J]. 中国电化教育, 2025(5): 117-127.

[7] 刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025, 36(3): 1152-1185.

[8] 戴国华, 武晓鸽, 詹文浩. DeepSeek 在终端本地部署的解决方案与发展研究[J]. 移动通信, 2025, 49(3): 100-106.