

基于 Dify 和 RAG 的大语言模型智能客服系统的设计与实现

摘要：随着大语言模型（LLM）与检索增强生成（RAG）技术的发展，智能客服系统逐渐向“LLM+知识增强”模式演进。本文基于 Dify 低代码平台，设计并实现了一套面向电商场景的智能客服系统。系统采用“问题分类+多知识库+多 LLM+人工兜底”的整体架构，通过 Question Classifier 实现用户意图识别，并结合 RAG 技术完成知识检索与增强生成，以提升回答准确性与业务可控性。同时，系统针对不同业务场景设计专业化 Prompt 与多模型协同机制。测试结果表明，该系统在问题分类、知识召回、多轮对话及异常处理等方面均具有较好的稳定性与实用性。

关键词：大语言模型；Dify 低代码平台；智能客服系统；检索增强生成；对话系统

0 引言

智能客服（基于 AI 的客户服务）正处于快速发展阶段。随着 AI 技术的迅速发展，智能客服系统成为网络零售商用来替代人工客服系统的选择^[1]。行业报告（如 Gartner、IBM、IDC）指出，85%的客服高管计划在 2025 年试点生成式 AI 解决方案^[2]，AI 客服在降低成本、提升效率方面的价值得到广泛认可。智能客服系统的发展历程是自然语言处理（NLP）技术演进史的缩影，其演进脉络可划分为四个阶段：人工坐席阶段、规则驱动阶段、机器学习阶段和大模型智能体阶段。每个阶段都反映了当时技术能力的边界与业务需求之间的匹配关系。早期研究多侧重于对话理解和对话管理的传统方法（如基于规则或浅层学习），存在意图识别准确率低下、多轮对话能力缺失、知识更新与维护成本高昂等多个核心问题。近十年智能客服研究集中在对话系统和知识问答技术的演进上。近年来涌现出以大语言模型（LLM）、检索增强生成（RAG）为核心的新技术路线。本人任职于一家拥有电商业务的连锁品牌企业。为有效降低运营成本并提升运营效率，企业内部立项智能客服系统项目。作为该项目的产品经理，我全面负责系统

的架构设计与交付落地工作。

1 智能客服系统概述

智能客服系统是一种面向行业的基于大规模知识处理的自动问答系统，涉及知识管理、自然语言理解、逻辑推理等技术，可为企业与海量用户之间的沟通提供一种有效的技术手段^[3]。智能客服系统的目标是将用户咨询的理解、知识匹配与应答生成等环节自动化，从而在降低人工成本的同时提升服务的响应速度与一致性。从技术定位上看，智能客服属于对话系统的产业应用，且以面向任务型对话为核心。在工程实现层面，传统任务型智能客服系统通常采用流水线（pipeline）架构，由自然语言理解、对话状态跟踪、对话策略与自然语言生成四个核心模块串联构成。这种架构便于分模块训练、评测与迭代，工程可控性强，但也存在误差逐级累积、对意图体系与标注语料依赖较重、跨业务域迁移成本高等落地痛点，尤其在意图边界模糊或长尾问题上表现受限。

近年来，大语言模型推动智能客服向“大模型+知识增强”的落地范式演进。受限于参数化知识难以及时更新、且存在事实性偏差，直接将大模型用于客服易产生与企业知识不符的“幻觉”回答。检索增强生成（Retrieval-Augmented Generation, RAG）通过在生成前从企业知识库检索相关文档并将其作为生成依据，将参数化记忆与非参数化外部知识相结合，从而提升回答的事实一致性与可更新性，已成为当前大模型客服落地的主流技术路线。在实际部署中，此类系统的工程重点通常包括：企业知识库（FAQ、产品文档、历史会话等）的结构化构建与切分、检索召回质量、生成内容的可控性与可溯源性、多渠道（网站、App、社交媒体、呼叫中心）统一接入，以及面向复杂或高风险问题的人工坐席兜底机制。综上，智能客服系统的研究与应用正由“单点算法优化”转向“端到端工程落地与知识协同”，而回答准确性、知识时效性与生成可控性仍是制约其规模化落地的关键问题。

2 核心技术基础与平台特性

2.1 大语言模型

大语言模型是指基于海量文本训练，并且包含极多参数的语言模型。作为一种人工智能技术，它使用深度学习算法来处理自然语言的模型。它可以根据大量数据来识别、总结、翻译、预测和生成文本及其他内容。就是一种大语言模型的具体应用^[4]。研究表明，当模型与数据规模超过一定阈值时，其会“涌现”出上下文学习、指令遵循与逐步推理等小规模模型所不具备的能力，并在指令微调与人类反馈对齐后展现出接近通用的语言理解与生成能力。正是这些底层能力，使大语言模型在智能客服场景中具有显著的应用价值。一方面，其强大的语义理解与多轮对话能力，有助于提升对用户复杂、模糊问题的识别准确率，缓解传统规则式与检索式客服表达受限、难以应对长尾问题的不足；另一方面，依托少样本与指令学习能力，系统无需为每类业务单独标注大量语料即可较快适配新场景，从而在一定程度上降低开发与跨业务迁移成本。此外，将大语言模型与检索增强生成（Retrieval-Augmented Generation, RAG）相结合，可在生成前从企业知识库检索相关内容并作为回答依据，有效提升回答的时效性与事实一致性，在此基础上进一步支持情感安抚、意图澄清与高风险问题的人工兜底，从而整体改善智能客服的响应效率、服务质量与用户体验。综上，大语言模型已逐渐成为新一代智能客服系统的重要技术支撑。

2.2 检索增强生成

检索增强生成（RAG）是一种结合信息检索（无监督学习）和内容生成（有监督学习）的混合方法，旨在通过从外部信息源中检索相关信息来增强生成模型的能力，从而提高生成内容的相关性和质量^[5]。该方法将大模型的知识参数化与可随时更新的外部非参数化知识相结合，有效缓解了大模型知识滞后与“幻觉”问题。在智能客服场景中，RAG 的应用价值尤为突出：企业的产品文档、业务规则与历史会话等知识高度专有且更新频繁，难以全部固化于模型参数之中，而 RAG 可在不重新训练模型的前提下，将这些企业知识实时接入回答流程，使系统在保持自然流畅对话能力的同时，显著提升回答的准确性、时效性与可溯源性，并便于对答案来源进行核查与

管控。因此，“大模型+RAG”已成为当前智能客服系统落地的主流技术路线。

2.3 低代码平台 Dify

Dify 是一个开源的低代码大语言模型应用开发平台，其核心竞争力源于模块化、可扩展的技术架构设计^[6]。Dify 可用于构建 AI workflow：通过在可视化画布上编排 AI 模型、连接数据源、定义处理流程，直接将领域知识转化为可运行的软件。Dify 的核心设计理念是融合后端即服务（Backend as a Service, BaaS）与 LLM Ops，通过“可视化编排+API 优先”的方式降低 AI 应用的开发门槛，使开发者乃至非技术人员均可快速构建并持续运营生产级生成式 AI 应用。在层级结构上，Dify 采用模块化分层设计，自上而下大致可分为数据层、编排层与基础层：数据层负责数据集与提示词的处理及 RAG 知识检索增强，编排层以可视化编排工作站为核心协调各组件运行并保障应用质量，基础层则提供模型接入与存储等底层基础设施。在平台特性方面，Dify 内置可视化 workflow 编排、数百种主流模型的统一接入、高质量 RAG 引擎、Agent 框架、Prompt IDE 及可观测性监控等关键技术栈，并支持 Docker 一键私有化部署，可灵活集成至既有业务系统中。

3 智能客服系统架构设计

3.1 系统架构总体设计



图 1 智能客服系统架构图

本智能客服系统具备多个核心能力模块，共同构成完整的智能客服体系。其中，智能分类路由能力可以自动识别用户问题类型，并动态进入最合适的处理链路；知识检索增强能力通过向量检索与混合召回提升知识命

中率；多 LLM 请专业化处理能力则允许不同场景采用不同模型与 Prompt，从而提升回答专业性。系统还支持多轮对话记忆能力，可以在上下文中持续理解用户意图；人工兜底机制则能够在异常、低置信度或高风险场景下自动转人工处理，保证整体服务稳定性。

整个系统在设计上遵循“专业化、可扩展、可控性与稳定性”原则。首先，不同问题采用不同知识库与不同 LLM 进行处理，避免“大一统模型”导致的专业能力下降；其次，通过模块化 workflow 设计，使系统后续新增分类、知识库或模型时无需大规模重构。在风险控制方面，系统引入人工介入机制，对低置信度、高风险或无法处理的问题进行人工兜底，保证用户体验与业务安全。同时，系统强调“RAG 优先、模型生成辅助”的设计理念，尽量基于企业真实知识进行回答，降低模型幻觉风险。

整体架构既兼顾 AI 自动化能力，也保留人工可控性。

3.1.1 前端层（多渠道接入）

本智能客服系统的前端层是整个智能客服系统面向用户的统一接入入口，负责承载来自不同渠道的用户咨询请求，并将用户消息标准化后传递至后端 workflow 系统。该层通常支持 Web/H5、小程序、App、微信公众号以及第三方客服平台等多种接入方式，从而实现“多渠道统一客服能力”。本系统前端层保证了各渠道的交互体验一致，并承担了会话管理、用户身份识别、登录态传递、历史消息展示、会话保持等功能。

3.1.2 网关层

网关层位于前端与核心 workflow 系统之间，主要负责请求统一接入、安全控制与流量治理，是整个系统的重要基础设施层。当用户请求进入系统后，API 网关会首先对请求进行身份校验、接口鉴权、频率限制与路由分发，避免恶意请求、接口滥用或系统雪崩等问题。例如，针对高频恶意调用、机器人刷接口等场景，可以通过限流熔断机制保护后端模型资源。同时，网关层还负责将不同渠道的请求统一转换为标准协议，屏蔽前后端之间的实现差异。

3.1.3 Dify workflow（核心引擎）

Dify workflow是整个智能客服系统的核心引擎，负责驱动用户问题从输入到最终回复的完整处理流程，也是整个 AI 客服系统实现“智能化”的关键部分。在整个架构中，Dify workflow相当于 AI 客服系统的大脑与调度中心。

3.1.4 模型层（LLM）

模型层主要负责自然语言理解、推理生成与对话回复，是智能客服系统中的 AI 能力核心。

3.1.5 数据层（知识库&数据存储）

数据层负责整个智能客服系统的知识沉淀与业务数据管理，是系统长期演进的重要基础。其中，知识库部分主要存储各类结构化或半结构化业务知识，例如：产品知识库、使用方法知识库、物流知识库、售后知识库等。这些知识来源于 FAQ、客服 SOP、商品文档、售后规则、运营资料等内容，并通过 Embedding 向量化后用于语义检索。

数据存储部分则负责保存：用户会话记录、用户画像信息、工单数据、操作日志、系统配置、反馈与标注数据。

3.2 Dify workflow编排

用户消息进入系统后，首先会通过 QuestionClassifier（问题分类器）进行意图识别，将用户问题自动划分为产品选购、使用方法、物流查询、售后退换、闲聊问候或其他问题等不同类别。不同分类会进入不同的知识检索与专业 LLM 处理链路，实现“按问题类型动态路由”。

在知识检索阶段，系统会从对应的领域知识库中进行向量检索与混合召回，例如产品问题进入产品知识库，物流问题进入物流知识库。随后，系统会将召回结果作为上下文传递给对应的专业 LLM 节点，由模型基于领域知识生成最终答案。

相比传统“一个模型回答所有问题”的方式，这种多分类+多知识库+多 LLM 的工作流结构能够显著提升回答准确率与稳定性。例如，售后场景

更强调流程规范与风险控制，而产品推荐场景更强调推荐能力与用户体验，不同 LLM 可以拥有不同 Prompt 与回答策略。

此外，Dify workflow 还支持变量管理、会话记忆、异常处理、人工介入、条件分支、工具调用以及后续 Agent 扩展能力，使整个系统具备较强的业务编排能力与可扩展性。

3.3 大模型选用

系统根据不同场景接入多个大语言模型，例如通义千问（Qwen）、DeepSeek、月之暗面、智谱 AI 等。不同模型在成本、速度、推理能力与中文理解能力上各有特点，因此系统采用了“多模型协同”的方式。

例如：

- 问题分类使用响应快、成本低的小模型；
- 产品推荐使用语言表达能力较强的模型；
- 售后场景使用更稳定、更保守的模型；
- 闲聊场景使用更自然、更具亲和力的模型。

模型层不仅仅是“生成回复”，更重要的是结合知识库进行 RAG（检索增强生成），避免模型幻觉，提高回答准确率。同时，不同 LLM 可以配置不同 Prompt、温度参数、记忆窗口与输出规则，以适配不同业务场景。

3.4 业务实时性保证机制设计

本系统采用 RAG 架构，将低等或中等业务变动频率的业务知识存储于知识库中，而非固化在大语言模型参数中。同时，本系统按业务模块拆分知识库，当 FAQ、客服 SOP、售后规则、运营资料发生变化时，无需每次全部重新上传大文档，只需通过单条新增、修改、删除进行增量更新，并重新生成向量索引即可，无需重新训练大模型。

对于订单状态、物流轨迹、库存价格、商品信息等实时性要求较高的数据，则通过 Dify workflow 调用内部业务系统 API 获取最新结果，再由 LLM 进行自然语言组织与回复。该方式既降低了模型训练与维护成本，也保证了系统知识更新的灵活性与业务响应的实时性。

3.5 多场景客服问答的准确性保障设计

本系统从分类、知识、检索、生成、兜底、评估六个环节共同控制，保证不同场景下问答准确性，而不仅只依赖模型能力。

- (1) 首先，在入口层通过问题分类器对用户问题进行意图识别，将问题准确路由到产品选购、使用方法、物流查询、售后退换等不同分支，避免不同场景知识混用。对于边界模糊问题设置优先级，例如“涉及退款、破损、补发的问题优先归入售后退换”。
- (2) 其次，本系统按业务场景拆分知识库，保证产品、物流、售后、使用说明等知识相互隔离，减少召回污染。同时，知识库内容结构化维护，保持“一问一答一知识点”，并定期更新、去重和校验。
- (3) 第三，在检索阶段采用混合检索与 Rerank 机制，提高相关知识命中率。对于 FAQ 场景，控制合适的 Top K，避免召回过多无关内容。
- (4) 第四，不同场景使用独立 LLM Prompt。例如产品推荐强调对比与引导，售后退换强调凭证收集与不承诺结果，物流查询强调不得伪造物流状态。通过专业化 Prompt 约束模型输出。
- (5) 第五，对于知识库无命中、低置信度、模型异常、高风险售后等情况，设计必须转人工处理，避免模型强行回答。
- (6) 最后，通过对话日志持续评估分类准确率、知识召回率、回复准确率和转人工率，并根据错误案例反向优化分类规则、知识库和 Prompt，形成持续迭代闭环。

3.6 系统数据隐私与安全设计

本系统可从部署隔离、传输加密、权限控制、数据脱敏、日志审计、模型调用约束等方面保障数据隐私与安全。

- (1) 首先，系统采用私有化环境部署 Dify 平台，将知识库、会话记录、工单数据等核心数据存储在企业可控环境中，避免业务数据分散在不可控平台。数据库、Redis、向量库等组件仅允许内网访问，不对公网开放。
- (2) 其次，在数据传输层面，前端、网关与 Dify 服务之间统一采用 HTTPS 加密通信，防止用户咨询内容、订单号、联系方式等敏感信息在传输

过程中被截获。

- (3) 第三，在权限控制方面，系统应基于角色进行访问控制，不同角色仅能访问其职责范围内的数据。例如普通客服只能查看相关会话，管理员才能维护知识库与工作流配置。
- (4) 第四，在模型调用环节，对于外部 LLM API，应避免传递不必要的敏感信息。订单号、手机号、地址等信息可在调用前进行脱敏处理，实时订单、物流等数据优先通过企业内部 API 查询，LLM 仅负责组织回复，不直接持久化敏感业务数据。
- (5) 第五，系统需建立日志审计机制，对知识库修改、人工介入、接口调用、异常转人工等关键操作进行记录，便于后续追踪与问题排查。
- (6) 最后，对于高风险场景，如退款、投诉、隐私信息查询等，应设置人工审核与兜底机制，避免模型越权处理。通过上述措施，可以在保证智能客服效率的同时，降低数据泄露、越权访问和模型误用风险。

4 系统实现与测试



图 2 智能客服系统实现框架图

4.1 系统实现工具与环境

本系统采用本地化部署方式运行 Dify 平台及相关服务，模型推理通过外部大语言模型 API 完成。因此，部署环境主要承担工作流编排、知识库管理、向量检索、数据库存储及接口调用等任务。测试环境采用 4 核 CPU、8GB 内存、50GBSSD 的服务器配置；生产环境采用 8 核 CPU、32GB 内存、500GB SSD，并部署 Ubuntu22.04LTS、Docker 与 DockerCompose。系统运行依赖

PostgreSQL、Redis、向量数据库及 DifyWeb/API 服务。因采用外部模型 API，无需本地 GPU。

本系统当前采用外部大语言模型 API 的方式进行模型推理，而未采用本地私有化部署方案。其主要原因在于，基于 API 的模型服务能够显著降低 GPU 硬件成本与运维复杂度，同时具备较好的模型效果与稳定性，更适合中小规模电商客服场景的快速开发与部署需求。

此外，系统采用 RAG 架构，业务知识主要存储于外部知识库中，而非固化于模型参数，因此知识更新无需重新训练模型。后续若业务规模扩大或对数据安全要求进一步提升，可逐步考虑 Embedding 模型、Rerank 模型及部分业务模型的私有化部署，以增强系统自主可控能力。

4.2 系统实现-核心模块

4.2.1 知识库构建

通过 Dify 的“知识库”功能，创建智能客服专属知识库。知识库创建分为三个阶段：

(1) 选择数据源

主要添加了 FAQ、客服 SOP、售后规则、运营资料等作为数据源。

(2) 文本分段与清理

文本分段与清理是本过程的核心。考虑到知识库数据源为 FAQ/Q&A 型数据，每条数据：

- 本身已经是一个完整语义块（本来就是独立知识点）
- 不需要切割
- 不需要 overlap
- 不应该跨问答混合

这种场景下不应该让 Dify 自动乱切，否则会出现不同问题被拼进同一个 chunk 的问题，从而导致召回错乱、回答串味、相似问题互相污染等问题。故最终选择选择一个 FAQ = 一个 chunk 作为客服知识库黄金规则，相关配置如下表：

配置项	配置值	配置说明
分段标识符	\n\n	适合 FAQ 型数据
分段最大长度	500	适合 FAQ 型数据

分段重叠长度 1 适合一个问题 = 一个 chunk，不需要 overlap

(3) 检索设置

选择“混合检索”，此方式能够平衡全文检索和向量检索（混合检索=向量检索+全文检索），兼顾知识库查询的全面性和准确性，相关配置如下表：

配置项	配置值	配置说明
索引方式	高质量	➢ 客服知识库最重要的是召回准确率，而不是省 token ➢ 经济模式可能匹配到用户的真实提问 ➢ 高质量模式：用 embedding 语义检索，可以理解语义接近的词句数据
Rerank	开启	重排序模型将根据候选文档列表与用户问题语义匹配度进行重新排序，从而改进语义排序的结果(embedding 负责粗召回，rerank 负责精排序)。
Top K	5	用于筛选与用户问题相似度最高的文本片段。系统同时会根据选用模型上下文窗口大小动态调整分段数量。
Score 阈值	0.4	对于 FAQ 问题，用户问法可能非常口语化，故相似度不一定特别高。若 score 值设置太高，可能直接召回失败。

4.2.2 workflow编排

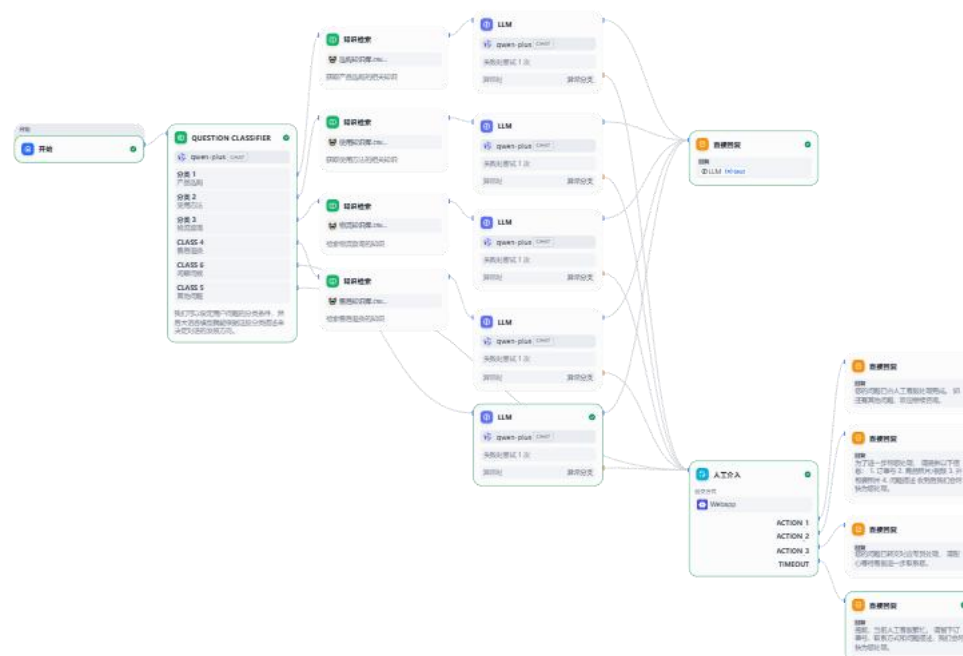


图 3 智能客服系统 workflow编排可视化画布（部分）

根据实际业务场景需要，本系统创建的是一个可以对结合知识库识别问题意图的聊天机器人（chatflow）。相关工作流编排核心节点如下：

(1) Question Classifier（问题分类器）

Question Classifier（问题分类器）节点本质上是在做 Intent Classification（意图识别），这个节点最重要的能力要求是：

能力	要求
中文理解	强
分类稳定性	强
成本	低
延迟	低
指令遵循	稳
FAQ 场景	强

故模型相关配置见下表：

配置项	配置值	配置说明
Model	qwen-plus	qwen-plus 非常适合做：中文客服、意图识别、分类路由、工作流编排、RAG 前置判断。
Temperature	0	低 temperature 可以保证分类更稳定。分类器不是创作节点，不需要发散。温度太高会导致同一个问题有时分到“售后”，有时分到“其他”。
聊天记忆	不开启	Question Classifier 是“单轮路由器”，而不是长期对话 Agent。它的任务只是针对当前问题进行分类。

根据实际业务场景，Question Classifier（问题分类器）节点共拆分多个分类：产品选购、使用方法、物流查询、售后退换、闲聊问候、其他问题等。另外，本系统向 Question Classifier（问题分类器）节点提供了清晰的 Prompt（指令），以帮助问题分类器更好的理解如何分类。

本系统采用 Router Rag 架构——由问题分类器识别用户问题类型，再路由到对应知识库（不同的分类连接不同的知识库）。

(2) 知识检索（多知识库 Router 架构）

由于客服 FAQ 的问题天然“高相似”，如果所有 FAQ 全混一起，会出现召回污染。故本系统采取多知识库 Router 架构，按业务分类拆分多个知识库，包括产品选购、使用方法、物流查询、售后退换等。知识检索的召回设置如下表：

配置项	配置值	配置说明
模式	Rerank	由于本系统已按业务拆分知识库，则不再做复杂权重混合，选择简单的 Rerank 模型。
Top K	4	Top K 太低会导致真正答案没被选中，而 Top K 太高则会导致串知识、回答混乱、token 增加等问题。
Score 阈值	关闭	目前已拆分知识库，不需要太严格等 score。

(3) LLM 回复

考虑到不同分类的回答规则明显不同（比如售后要严格按流程收集订单号/图片/破损证明；物流要调用查单接口/API；产品选购要做推荐话术、对比推荐），故本系统拆成多个 LLM 节点，每个 LLM 写对应角色 Prompt。每个 LLM 节点系统配置信息如下：

配置项	LLM 配置值				
	产品推荐	使用方法	物流查询	订单查询	闲聊
记忆	开启，记忆窗口为 5				
Temperature	0.3	0.2	0	0	0.7

从灵活性上看：产品推荐>使用方法>物流查询=订单查询。

本系统实现存在多个 LLM 分工协作：问题分类器 → 判断类别 → 路由到某一个 LLM。正常情况下，一次用户请求只会进入一个业务 LLM，例如产品推荐 LLM、物流 LLM、售后 LLM 或闲聊 LLM，因此冲突风险相对可控。但确实可能存在 LLM 协同冲突，主要包括：

冲突类型	示例
分类冲突	“XX 坏了能退吗” 既像产品问题又像售后问题
知识冲突	产品知识库和售后知识库规则不一致
Prompt 冲突	一个 LLM 鼓励推荐，另一个 LLM 强调风险保守
记忆冲突	闲聊上下文影响业务回复
输出冲突	不同 LLM 对同一问题给出不同处理建议

- 本系统采取以下处理方式有效避免多 LLM 协同冲突问题：
- ① 明确路由优先级
 - ② 统一全局客服原则
 - ③ 知识库分层管理
 - ④ 关闭跨业务记忆污染
- 例如：售后退换 > 物流查询 > 产品选购 > 使用方法 > 闲聊问候 > 其他问题。涉及退款、补发、破损、质量问题时，优先进入售后。
- 所有 LLM Prompt 都加共同约束：
- 不得编造政策、价格、库存、物流状态；
 - 涉及高风险问题必须转人工；
 - 以知识库内容为准。
- 通用政策、售后规则、物流规则应保持唯一来源，避免多个知识库中出现相互矛盾的规则。

闲聊 LLM、售后 LLM、物流 LLM 各自独立记忆，避免闲聊上下文影响售后判断。

⑤ 设置人工兜底

当分类置信度低、知识库无命中、用户问题跨多个业务域或模型回答不确定时，转人工处理。

本系统属于 Router + Multi-KB + Multi-LLM 架构，保证了“单次请求只由一个主 LLM 生成最终回复”，并通过路由优先级、统一 Prompt 规范和人工兜底控制边界，来有效避免多个 LLM 协同冲突。

(4) 异常兜底机制

本系统设置 Human Input / 人工输入节点作为异常兜底机制，用于让流程暂停，等待人工处理、审核或决策。针对以下场景， workflow 将流转至人工节点：

- 问题分类器：分类为【其他问题】
- 任意 LLM 异常

本系统已对接企微和内部工单系统，实现当流程流转至人工节点时，系统将给人工客服看的上下文信息会自动流转至企业内部相关客服角色并自动创建工单。本系统的 Human-in-the-loop 框架很好地帮助人工客服了解 AI 客服前面处理和回复了说明，以便完成顺滑的交接。

4.3 测试用例设计与测试方法

在智能客服系统的测试阶段，本文主要从功能正确性、分类准确性、知识召回效果、回复质量以及系统稳定性等多个维度开展测试。测试用例设计围绕产品选购、使用方法、物流查询、售后退换、闲聊问候以及其他问题六类典型场景展开，覆盖正常输入、模糊输入、多轮对话、异常输入等多种情况，以验证系统在真实业务环境下的适用性与鲁棒性。

在问题分类测试中，通过构建不同类别的用户问题样本，验证 Question Classifier 是否能够准确识别用户意图，并正确路由至对应知识检索与 LLM 节点。知识检索测试重点验证向量检索与混合检索机制的召回效果，包括知识命中率、相关性以及错误召回情况。LLM 回复测试则重点评估回答的准确性、完整性、逻辑性以及业务规范性，例如售后场景是

否按照预设流程收集订单号、图片等必要信息。

此外，系统还进行了异常处理与人工兜底测试，包括知识库无命中、模型异常、模糊问题以及越权问题等场景，验证系统能否正确转入人工介入流程。性能测试方面，主要关注系统响应时间、多轮对话稳定性以及高并发情况下的运行表现。

测试过程中，采用人工验证与实际对话测试相结合的方法，通过 Dify 对话日志与运行追踪功能，对 workflow 节点执行结果、知识召回内容及模型输出进行分析，从而持续优化分类规则、Prompt 与知识库结构，提高整体智能客服系统的准确率与稳定性。

4.4 测试结果与分析

4.4.1 本系统客服能力测试结果与分析

为验证基于 Dify 的大语言模型智能客服系统的实际效果，本文从问题分类、知识召回、LLM 回复质量、异常处理以及多轮对话稳定性等多个维度开展测试。测试结果表明，系统整体运行稳定，能够较好满足电商场景下的智能客服需求。

测试维度	测试内容	测试结果
问题分类	用户问题分类准确率	93%
知识检索	知识召回与匹配准确率	89%
LLM 回复	回复内容准确性与完整性	91%
异常处理	异常场景人工兜底成功率	100%
多轮对话	上下文记忆与稳定性	95%

通过对基于 Dify 的大语言模型智能客服系统进行多维度测试，结果表明系统整体运行稳定，能够较好满足电商场景下的智能客服需求。测试主要围绕问题分类、知识召回、LLM 回复质量、异常处理以及多轮对话稳定性等方面展开。

在问题分类测试中，系统基于 Question Classifier 对用户问题进行意图识别，并自动路由至对应业务流程。测试结果显示，系统能够较准确地区分产品选购、使用方法、物流查询、售后退换、闲聊问候等业务场景。在 200 条测试样本中，正确分类 186 条，整体分类准确率达到 93%。少量误判主要集中在同时包含多种业务语义的复杂问题中。

在知识检索测试中，系统采用向量检索与混合检索相结合的方式，并

通过 Rerank 模型进行结果重排序。测试结果表明，系统对标准 FAQ、物流规则以及售后流程等结构化知识具有较好的召回效果，知识召回命中率达到 89%。通过调整 Top K、知识分段与召回策略后，错误召回现象得到明显减少。

在 LLM 回复质量测试中，不同业务场景对应的专业 LLM 能够根据知识上下文生成较符合业务规范的回复。例如，在产品推荐场景中，系统能够根据用户描述的条件生成推荐建议；在售后场景中，系统能够按照预设流程引导用户提供订单号、商品照片等必要信息。综合人工评估结果，系统整体回复准确率达到 91%。

此外，在异常处理测试中，当知识库未命中、模型异常或问题超出系统处理范围时，系统能够自动进入人工介入流程，实现人工兜底，异常场景人工兜底成功率达到 100%。在多轮对话测试中，系统基于记忆窗口机制能够较好保持上下文连续性，多轮对话稳定性达到 95%，整体响应时间较短，能够满足中小规模电商客服场景的实际应用需求。

4.4.2 高并发性能测试结果

为验证系统在高并发场景下的运行稳定性，本研究采用模拟压力测试方式，对智能客服系统进行并发访问测试。测试环境为 8 核 vCPU、32GB 内存、500GB SSD 的生产部署配置，系统通过 Dify 工作流完成问题分类、知识检索、LLM 回复生成及异常兜底流程。测试工具模拟用户持续发送客服咨询请求，分别设置 50、100、200、300、500 个并发用户，统计系统平均响应时间、P95 响应时间、成功率及异常率等指标。

并发用户数	平均响应时间	P95 响应时间	请求成功率	异常率
50	1.8s	2.7s	99.8%	0.2%
100	2.4s	3.6s	99.3%	0.7%
200	3.7s	5.2s	98.6%	1.4%
300	5.1s	7.8s	96.9%	3.1%
500	8.6s	12.4s	92.5%	7.5%

测试结果表明，在 200 并发以内，系统平均响应时间保持在 4 秒以内，请求成功率达到 98%以上，能够满足中小规模电商客服场景的使用需求。当并发用户数提升至 300 以上时，系统响应时间明显增加，主要瓶颈集中在外部大语言模型 API 调用延迟、知识检索重排序耗时以及 workflow 节点排队

等环节。在 500 并发场景下，系统仍可保持基本可用，但异常率上升，用户体验有所下降。

考虑到 200 并发以及可以支持中小型电商平台客服场景需求，能够满足日均数万级咨询量场景、50 个在线客服坐席规模的稳定运行，目前宝兵法性能测试通过。

4.4.3 本系统与传统客服系统的对比测试结果与分析

本系统属于 LLM + RAG + Workflow + Human-in-the-loop 的新一代智能客服，而传统客服系统通常包括：

类型	核心技术
规则式客服	关键词/规则树
检索式客服	FAQ 检索

为了解本系统的改进价值，本研究构造实验数据集，从多个维度与传统客服系统进行对比实验。

(1) 实验数据集覆盖范围

实验数据集包含 300 条测试问题，覆盖用例场景包括：产品推荐、使用方法、物流查询、售后退换、模糊问题、长尾问题、多轮对话等。

(2) 实验指标

指标	指标说明
定量指标	回答准确率 ● 准确率用于衡量智能客服系统整体回答正确的比例，是评价系统效果和整体业务可用性最直观、最核心的指标之一 ● 其计算方式为：回答准确率 = 正确回答数量 / 总测试问题数量 ● 重点评估：回答是否符合用户问题、是否引用正确知识、是否存在明显错误、是否符合业务规则 ● 本次比对采用人工打分 5 分制： ➢ 完全正确：5 分 ➢ 基本正确：4 分 ➢ 部分正确：3 分 ➢ 错误较多：2 分 ➢ 完全错误：1 分
	用户满意度 ● 用户满意度用于衡量用户对客服回复效果与整体服务体验的主观评价。 ● 该指标能够综合反映：回答准确性、回复自然度、响应速度、用户体验、服务情绪 ● 其计算方式为：满意度 = 所有评分总和 / 总评价人数
定性	问题理解能力 ● 衡量系统对用户自然语言、模糊表达及真实意图的理解能力。

指标	多轮对话能力	衡量系统在连续对话中保持上下文记忆与语义连贯性的能力。
	长尾问题处理	衡量系统对低频、非标准化及未预设问题的处理能力。
	知识更新成本	衡量系统新增、修改知识内容时所需的人力与维护成本。
	回答自然度	衡量系统回复语言的流畅性、拟人化程度与交互体验。
	幻觉风险	衡量系统生成虚构、错误或与真实知识不一致内容的风险。
	可控性	衡量系统对回复内容、业务规则与风险边界的控制能力。
	开发成本	衡量系统从设计、开发到部署所需的技术与人力投入。
	扩展性	衡量系统后续新增业务、知识库或功能模块的适配能力。
	人工依赖	衡量系统对人工客服介入与人工处理流程的依赖程度。

(3) 实验结果

对比实验结果如下表：

指标	规则式客服系统	FAQ 检索式客服系统	本系统
回答准确率	62%	78%	91%
用户满意度	65%	79%	92%
问题理解能力	低	中	高
多轮对话能力	弱	弱	强
长尾问题处理	差	一般	强
知识更新成本	高	中	低
回答自然度	差	中	高
幻觉风险	无	低	中
可控性	强	强	中
开发成本	高	中	低
扩展性	差	一般	强
人工依赖	高	中	低

从对比实验结果看来：

- 本系统明显优于规则式和 FAQ 检索式的传统客服系统：本系统拥有更强的语义理解、长尾能力、多轮对话能力、可扩展性，且拥有更低的知识维护成本；
- 同时，本系统也存在一定的劣势：存在幻觉风险、Token 成本较高、响应时间较长、Prompt 依赖强、高风险场景需人工兜底等问题。

5 结语

本智能客服系统能够显著提升宠物用品电商平台的客服效率与用户体验。通过 AI 自动处理大量重复性问题，例如商品推荐、使用指导、物流查询与售后 FAQ，可以有效降低人工客服压力与运营成本。同时，多知识库与多 LLM 架构能够提高回答专业性与准确率，减少用户投诉与错误回复。

系统还能够实现 7×24 小时在线服务,提高用户响应速度与咨询转化率。对于企业而言,长期运行后,系统还会逐渐沉淀大量结构化业务知识与用户数据,形成企业专属 AI 知识资产。随着后续能力扩展,本系统还可以进一步演进为智能导购、智能运营助手、智能工单系统乃至企业级 AI Agent 平台。

参考文献

- [1] 张鹏,马俊,董绍增,等. 人工还是智能? AI 时代电商在线客服策略选择. 管理评论, 2023, 35 (11):166-178.
- [2] Gartner 官网. Gartner Survey Reveals 85% of Customer Service Leaders Will Explore or Pilot Customer-Facing Conversational GenAI in 2025. <https://www.gartner.com/en/newsroom/press-releases/2024-12-09-gartner-survey-reveals-85-percent-of-customer-service-leaders-will-explore-or-pilot-customer-facing-conversational-genai-in-2025#:~:text=Eighty,a%20survey%20by%20Gartner%2C%20Inc.> (2026-05-19).
- [3] 饶竹一,张云翔. 基于知识图谱的智能客服系统研究. 电力信息与通信技术, 2017, 15 (07):41-45.
- [4] 金星晔,左从江,方明月,等. 企业数字化转型的测度难题:基于大语言模型的新方法与新发现. 经济研究, 2024, 59 (03):34-53.
- [5] 刘雪颖,云静,李博,等. 基于大型语言模型的检索增强生成综述. 计算机工程与应用, 2025, 61 (13):1-25.
- [6] 陈碧萱. 基于 Dify 的大语言模型智能问答系统的设计与实现. 电脑知识与技术, 2026, 22 (03):45-47.