AI智能客服功能模块需求分析文档

1 需求分析

文华财经作为中国金融科技领域的领军企业，服务于数以百万计的专业投资者。当前，用户服务主要依赖以传统人工为主的客服体系，面临着服务时间受限、响应速度不均、重复性问题消耗大量人力等挑战。而近年来，以大语言模型（LLM）为代表的生成式AI技术取得了突破性进展。这为彻底变革传统的客户服务模式提供了历史性的机遇。文华财经自主研发的金融大模型，具备强大的自然语言理解、专业知识推理和文本生成能力，使其能够无限扩展服务容量、保证知识权威与一致性、自动化处理海量基础问题、赋能品牌科技形象。因此，构建一个智能、高效的AI客服系统，已不再是“可选项”，而是保持行业领先地位、提升核心竞争力的战略必然。

该AI智能客服系统的最终愿景是打造一个能够深度理解金融领域专业问题、具备多轮对话能力、并能无缝衔接人工服务的智能平台。该平台旨在显著提升客服效率，优化用户体验，并通过数据洞察反哺产品迭代，最终强化文华财经的品牌价值和市场竞争力。核心任务是开发AI客服系统的核心交互模块——一个功能完备的桌面端对话应用原型。此原型将专注于实现用户与AI之间最直接、最关键的交互流程，为后续构建完整的系统奠定坚实的技术和体验基础。需要实现的核心需求如下：

标准对话界面布局：构建一个经典的聊天界面。AI大模型的回答气泡及头像需居左显示。用户发送的消息气泡及头像需居右显示。界面底部必须是一个位置固定的用户文本输入区域及发送控件。

历史记录滚动与输入区固定：用户与AI的对话历史记录（无论长短）都应在中间区域显示，且内容超出可视范围时，用户可以通过滚动条平滑地滑动查看。在滚动过程中，底部的输入框必须保持其位置和大小不变，确保用户随时可以发起新的提问。

消息气泡动态自适应：宽度自适应： 当AI的回答内容较短（例如，不足一行）时，其消息气泡的宽度应由内容决定（fit-content），并整体容器靠左对齐。响应式布局： 当用户调整主窗口大小（如最大化、最小化或拖拽边缘）时，消息气泡（尤其是承载长文本的AI回答气泡）的宽度应能自适应变化，确保文本正确换行且布局美观。

AI回答流式匀速输出：为模拟自然对话、避免网络延迟带来的突兀感，模块必须调用文华财经大模型的流式API。AI的回答不能在生成完毕后瞬间显示在界面上，而应以一种平滑、匀速的“打字机”效

果在气泡内逐字或逐词呈现，给用户带来流畅、不间断的视觉反馈。

2 整体方案

2.1 实现目标

项目的核心目标是利用文华财经提供的大模型能力，构建一个专业、高效、安全、用户友好的AI智能客服系统，服务于文华财经的用户群体（投资者、交易员、分析师等）。

2.1.1 核心功能目标

7x24小时即时响应：提供全天候、秒级的客户服务响应，解决常见问题，减轻人工客服压力。

精准解答金融/产品问题：深度理解用户关于文华财经软件功能、行情数据、技术指标、交易规则、金融基础知识、市场动态等领域的专业问题，并提供准确、清晰的解答。

高效任务处理：能够根据用户指令完成特定操作指引（如：如何设置某个指标、如何导出特定数据、如何查找某个功能等）。

个性化交互：根据用户历史交互、用户画像（如账户类型、使用习惯）提供更贴合的服务和推荐。

多轮对话能力：理解上下文，进行自然流畅的多轮对话，处理复杂咨询。

风险提示与合规性：在涉及投资建议、市场风险等内容时，自动嵌入合规风险提示语，确保交流内容符合金融监管要求。

2.1.2 用户体验目标

界面简洁直观：提供清晰易用的聊天界面，支持文字输入。

响应迅速流畅：对话过程无明显延迟，交互自然。

信息表达清晰专业：回答内容结构清晰，术语准确，避免歧义，必要时可引用数据或图表（如解释行情、指标时）。

无缝转接人工：当AI无法处理或用户明确要求时，能顺畅转接至人工客服，并传递对话上下文。

2.1.3 业务价值目标

显著提升客服效率：分流大量重复性、基础性咨询，降低人工客服成本。· 7x24小时即时响应： 提供全天候、秒级的客户服务响应，解决常见问题，减轻人工客服压力。

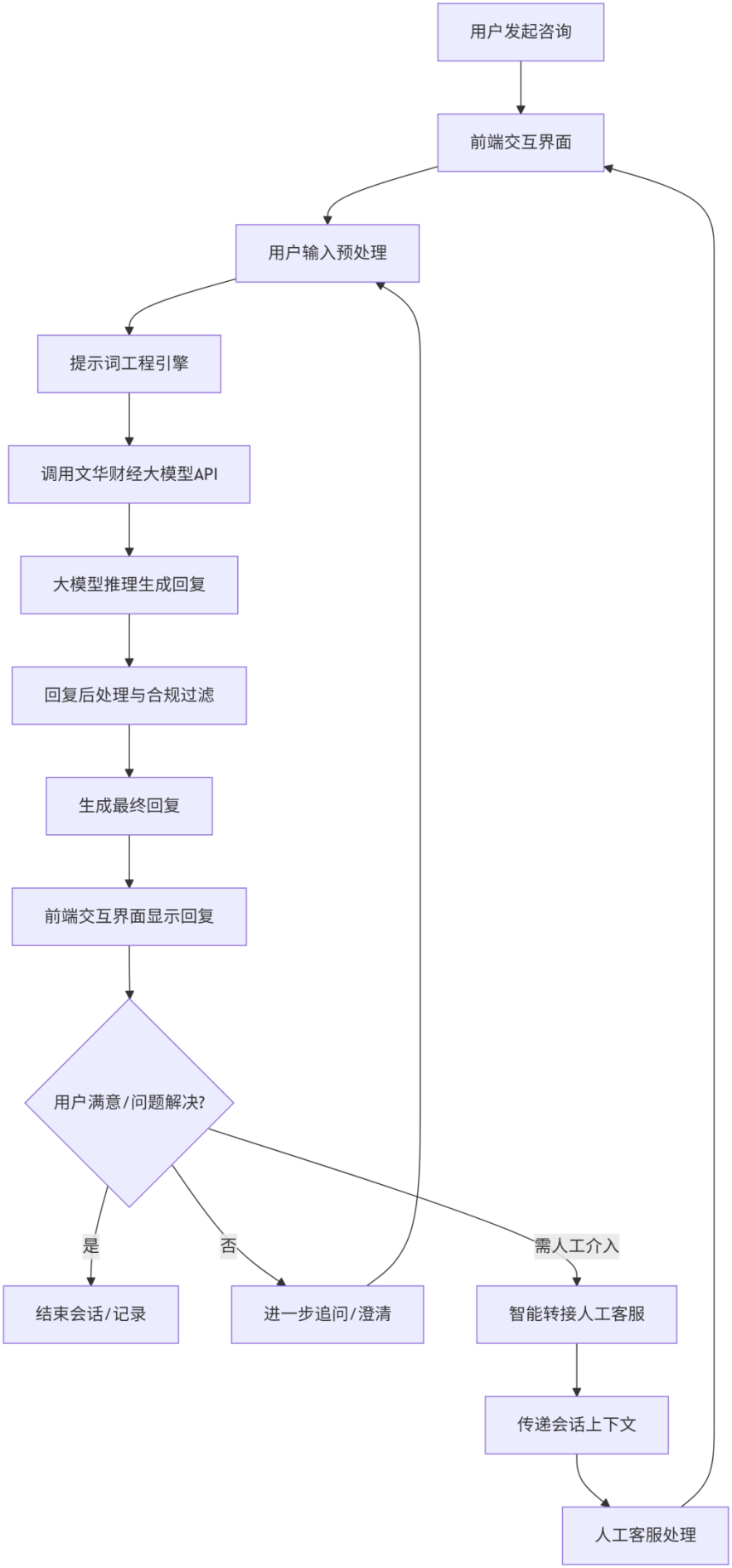
提升用户满意度：通过即时响应和专业解答，提升用户对文华财经产品和服务的整体满意度与忠诚度。

增强品牌专业形象：展现文华财经在金融科技与AI应用领域的领先地位。

收集用户反馈：通过对话分析，洞察用户痛点、需求热点，为产品优化和市场策略提供数据支持。

辅助用户决策（信息层面）：快速提供市场数据、功能解读等辅助信息（非投资建议），帮助用户更高效地进行分析和决策。

2.2 总体流程


用户发起咨询
前端交互界面
用户输入预处理
提示词工程引擎
调用文华财经大模型API
大模型推理生成回复
回复后处理与合规过滤
生成最终回复
前端交互界面显示回复
用户满意/问题解决?
是
否
需人工介入
结束会话/记录
进一步追问/澄清
智能转接人工客服
传递会话上下文
人工客服处理

流程详细解释：

① **用户发起咨询**：用户通过Web、App或小程序等前端界面输入问题（文字或语音转文字）。

② **前端交互界面**：接收用户输入，进行基础格式化（如去除首尾空格、特殊字符过滤）。

③ **用户输入预处理**：使用小型模型或规则初步判断用户意图（如：功能咨询、数据查询、报错求助、投诉建议等），辅助后续提示词构建。随后识别用户问题中的关键实体（如：股票代码、指标名称、软件功能模块名、日期等）。最后初步过滤明显违规或敏感词汇。

④ **提示词工程引擎（核心）**：根据预处理结果（意图、实体）、用户画像（如有）、当前对话上下文、系统预设角色和规则，动态构建发送给大模型的提示词(Prompt)。提示词包含：系统角色设定（“你是一个专业的文华财经客服专家...”）、任务指令、用户问题、上下文信息、知识库引用提示、输出格式要求、合规性要求等。

⑤ **调用文华财经大模型API**：将精心构建的提示词发送给文华财经提供的大模型服务API。

⑥ **大模型推理生成回复**：文华财经大模型基于接收到的提示词进行推理计算，生成初步的回复文本。

⑦ **回复后处理与合规过滤**：使用规则引擎和关键词库对回复进行强制合规检查，确保包含必要的风险提示（如“市场有风险，投资需谨慎”、“以上信息仅供参考，不构成投资建议”），过滤任何可能涉及荐股、承诺收益等违规内容。对于涉及具体数据、功能操作步骤等关键信息，尝试与知识库或实时数据源进行二次校验。优化回复的排版、标点，确保可读性。如果支持图表，在此环节生成或引用。

⑧ **生成最终回复**：将通过后处理的回复内容准备好。

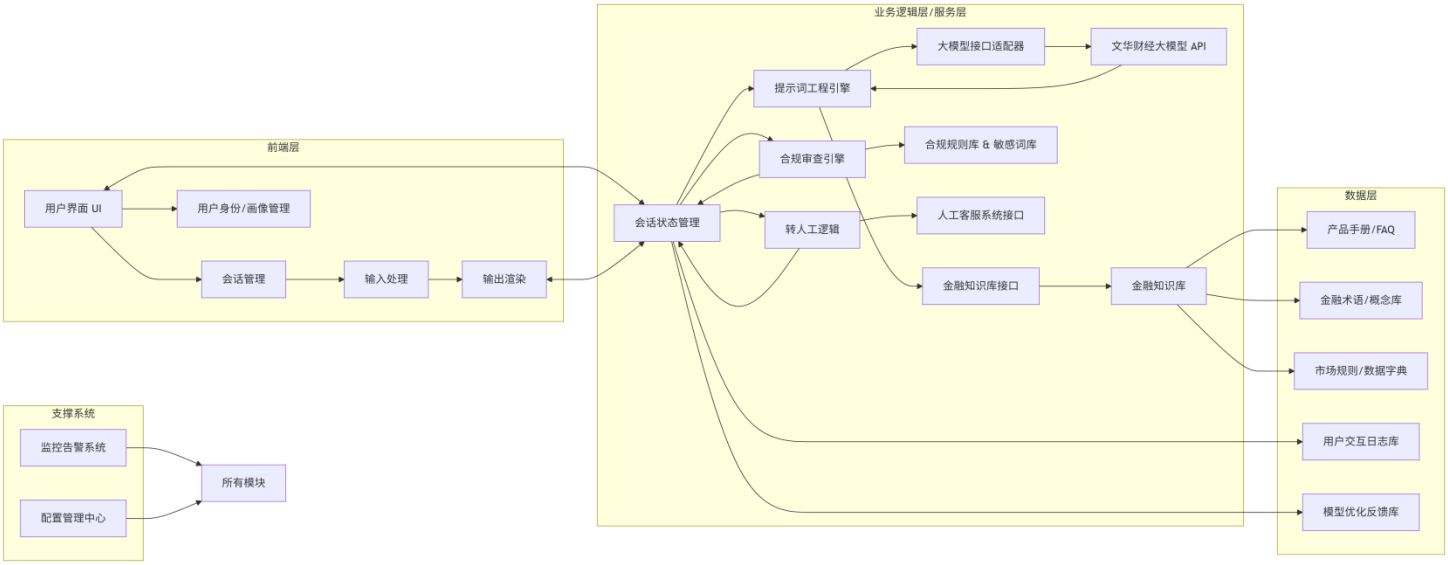
⑨ **前端交互界面显示回复**：将最终回复呈现给用户。

⑩ **用户反馈判断**：用户满意/问题解决则结束当前会话，记录日志用于分析优化。若用户追问/问题未解决则回到步骤3，将新输入和上下文一起处理。若用户明确要求或AI判断需人工介入则触发转人工流程。

⑪ **智能转接人工客服**：用户系统自动将当前对话记录（用户问题、AI回复历史）整理并传递给人工客服坐席系统，然后通知用户已转接人工。

⑫ **人工客服处理**：人工客服查看上下文，继续为用户服务，处理结果也可能反馈回系统用于模型优化。

2.3 产品架构与模块设计



模块详细解释：

① **前端层：**

用户界面 (UI)：提供用户与AI客服交互的窗口。通常是聊天窗口形式，支持文字输入/发送、显示对话历史（用户问 & AI答）、显示加载状态、转人工按钮、满意度评价按钮等。设计需简洁、专业、符合文华财经品牌风格。

会话管理：管理单个会话的生命周期（开始、进行中、结束），维护会话的唯一标识。处理新消息的接收和分发。

输入处理：接收用户原始输入（文本或语音转文本），进行基础预处理（如编码处理、基础清理）。

输出渲染：接收后端返回的最终回复内容，将其渲染展示在UI上。可能需要处理富文本、链接、或简单的图表展示（如图片或内嵌小图表）。

用户身份/画像管理：尝试获取用户基本信息（如账户类型 - 实盘/模拟盘、常用功能等），用于个性化服务和提示词构建。

② **业务逻辑层/服务层 (核心)：**

会话状态管理： 核心中枢：维护当前会话的完整上下文（多轮对话历史），协调调用其他模块（提示词引擎、合规引擎、转人工逻辑）。负责将用户输入和上下文传递给提示词引擎，接收大模型返回的回复，交给合规引擎处理，最终将合规后的回复返回前端。

提示词工程引擎 (Prompt Engineering Engine)：接收用户当前输入、完整对话上下文、用户画像信息。根据预定义的策略和模板，动态构建发送给大模型的提示词(Prompt)。策略包括：根据意图选择不同模板、嵌入上下文摘要、注入用户画像信息、引用知识库关键条目、强制加入系统指令和合规要求、设定输出格式。这一过程会使用精心设计的大量Prompt模板并持续优化。

大模型接口适配器：封装与文华财经大模型API的通信细节（如认证、参数构造、请求发送、响应解析、错误处理、重试机制）。将提示词引擎构建好的Prompt发送给API，并取回大模型生成的原始回复。

提示词引擎或后处理模块可能需要查询知识库来补充信息或校验事实。

合规审查引擎：接收大模型返回的原始回复，应用合规规则库（如强制在特定类型回答前/后添加风险提示语、禁止出现特定类型词汇如“稳赚”、“包赔”等）；应用敏感词库进行过滤和替换，并

进行逻辑检查（如是否遗漏了必要的风险提示），最后确保所有输出内容符合金融监管规定和文华财经内部合规政策。

转人工逻辑：根据预设规则（如用户多次表示不满、用户明确要求、AI连续多次无法理解、检测到高危词汇如“投诉”、“销户”）或AI自身置信度低时，触发转人工流程。负责整理会话上下文并传递给人工客服系统。

③ **数据层：**

金融知识库：比如产品手册/FAQ，即文华财经软件各功能模块的详细说明、常见问题解答；比如金融术语/概念库，即股票、期货、期权、技术指标、财务指标等专业术语的解释；比如市场规则/数据字典，即交易所交易规则、合约规格、数据字段含义解释。

合规规则库 & 敏感词库：存储需要强制添加的风险提示模板、禁止出现的词汇列表、需要替换的敏感词等。需要法务和合规团队深度参与制定和更新。

用户交互日志库：记录所有用户与AI的对话历史（脱敏处理）、时间戳、会话ID、用户ID（可选）、AI回复的原始版本和最终版本、转人工记录、用户满意度反馈等。用于分析、优化模型和提示词、审计。

户主动提交的错误反馈。用于后续模型微调和提示词迭代。

④ **支撑系统：**

监控告警系统：监控整个AI客服系统的运行状态（API调用延迟、错误率、会话量、转人工率、合规拦截率等），设置阈值告警，确保服务可用性和及时发现问题。

· 配置管理中心：集中管理系统的可配置项，如Prompt模板、合规规则、敏感词库、转人工阈值、功能开关等，方便快速调整和上线。

3 开发设计

3.1 概要设计

3.1.1 组织结构

AI智能客服系统整体分为四层：**前端层、业务逻辑层/服务层、数据层、支撑系统。**

① **前端层：**负责用户交互、会话管理、输入输出处理。

② **业务逻辑层/服务层：**核心中枢，负责会话状态、提示词工程、大模型API对接、合规审查、转人工逻辑等。

③ **数据层：**存储知识库、规则库、日志、优化反馈等。

④ **支撑系统：**监控告警、配置管理。

3.1.2 功能模块

① **用户界面（UI）：**为用户提供与AI客服交互的窗口，支持消息输入、历史消息展示、满意度评价、转人工等操作，保证良好的用户体验。

② **会话管理**：负责维护每个用户会话的唯一标识、生命周期和上下文信息，确保多轮对话的连贯性和准确性。

③ **输入/输出处理：**对用户输入进行基础清理、格式化、分词、意图识别和实体提取，并对AI回复内容进行格式优化和渲染，提升交互质量。

④ **用户画像管理：**收集和管理用户的基本信息（如账户类型、常用功能、历史行为等），为个性化服务和提示词构建提供数据支持。

⑤ **提示词工程引擎：**根据用户输入、上下文、用户画像等动态生成适合大模型推理的Prompt，嵌入系统角色、合规要求、知识库引用等，提升AI回复的专业性和合规性。

⑥ **大模型接口适配器：**封装与文华财经大模型API的通信细节，包括认证、参数构造、请求发送、响应解析和异常处理，保障与大模型的高效稳定对接。

⑦ **金融知识库接口：**为系统提供访问内部金融知识库（如FAQ、产品文档、术语库、规则库等）的能力，支持AI回复的事实校验和信息补充。

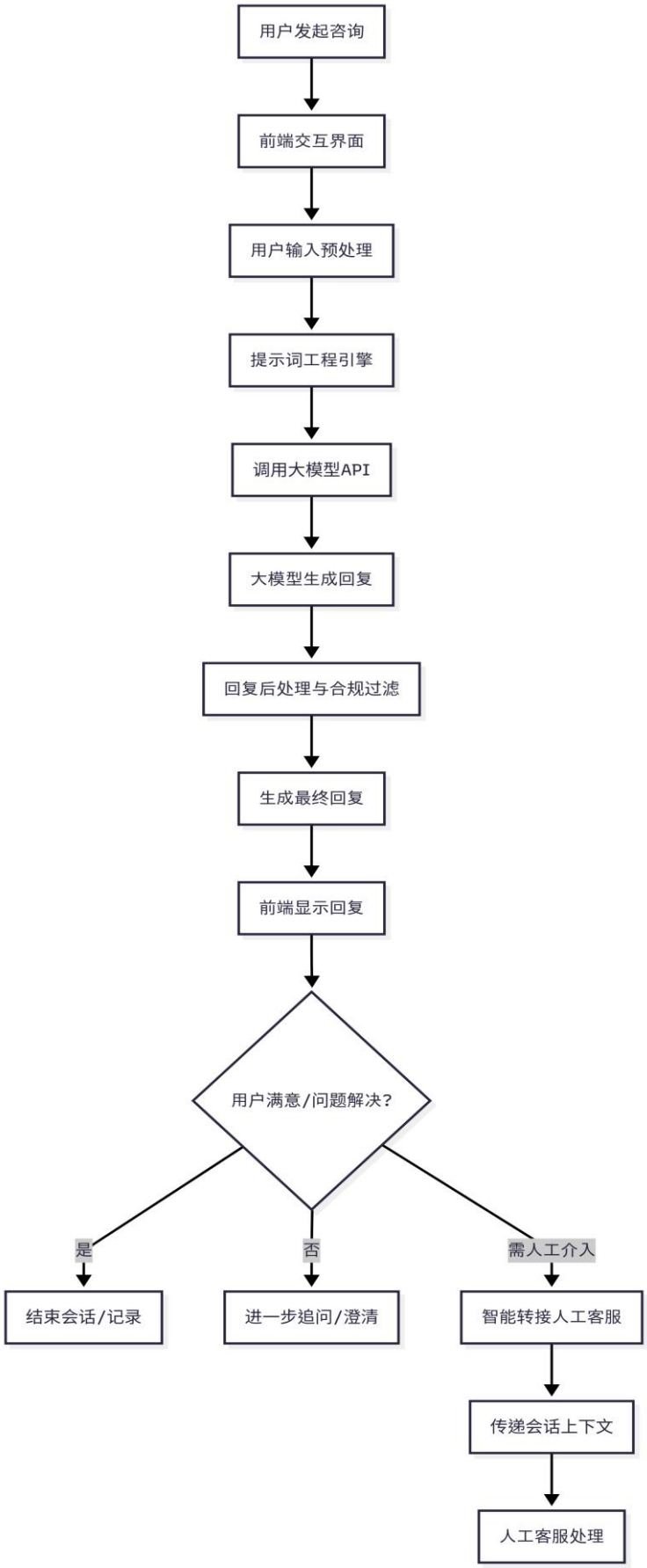
⑧ **合规审查引擎：**对大模型生成的回复进行合规性检查，应用规则库和敏感词库，强制添加风险提示，过滤违规内容，确保输出符合金融监管和公司政策。

⑨ **转人工逻辑：**根据用户反馈、AI置信度、关键词等条件，智能判断是否需要转接人工客服，并整理会话上下文，保障服务连续性。

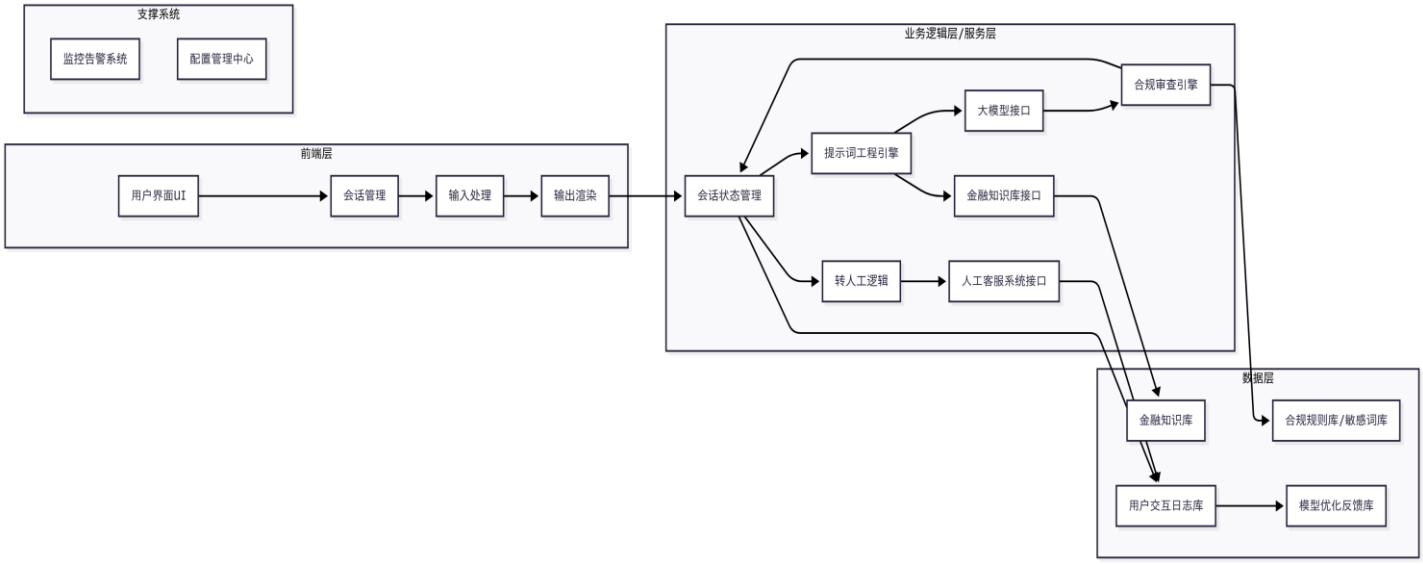
⑩ **日志与反馈收集：**记录所有用户与AI的交互日志、AI回复、人工介入、用户反馈等，为系统优化、模型迭代和合规审计提供数据支撑。

⑪ **监控与配置管理：**实时监控系统运行状态（如API延迟、错误率、转人工率等），并集中管理系统配置项（如Prompt模板、规则库、功能开关等），保障系统高可用和灵活运维。

3.1.3 处理流程


用户发起咨询
前端交互界面
用户输入预处理
提示词工程引擎
调用大模型API
大模型生成回复
回复后处理与合规过滤
生成最终回复
前端显示回复
用户满意/问题解决?
是
否
需人工介入
结束会话/记录
进一步追问/澄清
智能转接人工客服
传递会话上下文
人工客服处理

3.1.4 模块关系



3.2 详细设计

3.2.1 系统设计

① **前端**：基于React/Vue等实现聊天窗口，支持消息输入、历史展示、满意度评价、转人工按钮。

② **后端**：采用Python（Flask/FastAPI）或Node.js，RESTful接口，负责会话管理、与大模型API通信、合规处理、日志记录。

③ **数据库**：MySQL/Redis存储用户会话、知识库、规则库、日志。

④ **监控**：Prometheus+Grafana监控API延迟、错误率、转人工率等。

3.2.2 子系统划分

① **前端子系统**：负责实现用户与AI客服的交互界面，包括聊天窗口、消息输入与展示、满意度评价、转人工按钮等。通过WebSocket或HTTP长轮询与后端保持通信，实时展示AI或人工客服的回复，提升用户体验。

② **会话与业务逻辑子系统**：作为系统的核心中枢，负责会话生命周期管理、上下文维护、输入预处理（如分词、意图识别、实体提取）、会话状态流转，以及协调各业务模块（如提示词引擎、合规引擎、转人工逻辑）的调用，确保多轮对话的连贯性和智能性。

③ **大模型与知识库子系统**：封装与文华财经大模型API的对接逻辑，负责Prompt构建、API请求与响应处理、异常重试等。同时提供金融知识库、FAQ、术语库等数据的查询接口，支持AI回复的事实校验和信息补充。

④ **合规与转人工子系统**：对AI生成的回复进行合规性审查，应用规则库和敏感词库，强制添加风险提示，过滤违规内容。根据用户反馈、AI置信度、关键词等条件，智能判断是否需要转接人工客服，并整理会话上下文，保障服务连续性和合规性。

⑤ **日志与监控子系统**：负责记录所有用户与AI/人工客服的交互日志、系统运行状态、API调用情况、用户反馈等。通过监控和告警机制，实时发现系统异常，支持系统优化、模型迭代和合规审计，保障系统高可用和安全运维。

3.2.3 功能模块详细设计

① **前端UI与会话管理：**本模块负责为用户提供友好、直观的交互界面，实现消息的输入、展示和会话管理。通过前端UI组件和会话管理机制，确保用户能够顺畅地与AI客服进行多轮对话，并支持满意度评价和人工转接等功能。

② **会话管理与输入处理：**本模块负责维护用户会话的唯一标识和上下文信息，对用户输入进行基础清理、分词、意图识别和实体提取。为后续的提示词构建和大模型推理提供结构化、标准化的输入数据，提升系统理解能力。

```
# 输入预处理伪代码
def preprocess_input(user_input):
    clean_text = clean_special_chars(user_input)
    intent = intent_recognition(clean_text)
    entities = entity_extraction(clean_text)
    return {'text': clean_text, 'intent': intent, 'entities': entities}
```

③ **提示词工程引擎：**本模块根据用户输入、上下文、用户画像等信息，动态生成适合大模型推理的Prompt。通过灵活的模板和策略，嵌入系统角色、合规要求和知识库引用，确保AI回复的专业性、针对性和合规性。

```
Prompt构建示例：
def build_prompt(user_info, context, intent, entities):
    prompt = f"你是文华财经AI客服专家。用户意图：{intent}，实体：{entities}。上下文：{context}。请专业、合规地回答。"
    return prompt
```

④ **大模型接口适配器：**本模块封装与文华财经大模型API的通信细节，负责Prompt的发送、响应的解析、异常处理和重试机制。保障系统与大模型服务的高效、稳定对接，是AI智能客服的核心推理能力来源。

```
API调用示例：
import requests
def call_llm_api(prompt):
    resp = requests.post('https://wenhua-llm/api', json={'prompt': prompt})
    return resp.json()['reply']
```

⑤ **合规审查引擎：**本模块对大模型生成的回复进行合规性检查，应用规则库和敏感词库，强制添加风险提示，过滤违规内容。确保所有输出内容符合金融监管要求和公司内部合规政策，防范业务风险。

```
合规处理伪代码：
def compliance_check(reply):
    if contains_sensitive_words(reply):
        reply = filter_sensitive_words(reply)
    if not has_risk_warning(reply):
        reply += "\n【市场有风险，投资需谨慎】"
    return reply
```

⑥ **转人工逻辑：**本模块根据用户反馈、AI置信度、关键词等条件，智能判断是否需要转接人工客服。负责整理会话上下文并推送至人工客服系统，保障服务的连续性和用户问题的最终解决。

转人工触发伪代码：

```
def need_human_intervention(user_feedback, ai_confidence, user_input):
    if user_feedback == '不满意' or ai_confidence < 0.5 or '投诉' in user_input:
        return True
    return False
```

⑦ **日志与监控：**本模块负责记录所有用户与AI/人工客服的交互日志、系统运行状态、API调用情况和用户反馈。通过监控和告警机制，实时发现系统异常，支持系统优化、模型迭代和合规审计，保障系统高可用和安全运维。

3.3 开发周期

功能模块	负责人	预计时间
前段UI开发、会话管理与输入管理（前端层）	李金鹏、李子忆	7天
提示词工程引擎、大模型API对接（服务层）	陆奕舟	4天
合规查审引擎、转人工逻辑（服务层）	赵圣达	3天
日志与监控、系统联调与测试（数据层、支撑系统）	胡云璐	3天

4 非功能性需求

为确保原型产品的质量和可用性，除核心功能外，模块还需满足以下非功能性要求：

① **性能:**

启动响应：应用程序从启动到可进行交互的加载时间应控制在3秒以内。

UI流畅度：在输入文本、发送消息、滚动聊天记录等高频操作中，界面不应出现可感知的卡顿或延迟。

动产生影响。

② **可用性:**

直观易用：界面设计应遵循主流即时通讯软件的设计规范，确保用户无需学习即可上手。

操作便捷：支持通过键盘“Enter”键快速发送消息，支持在输入框内使用“Shift+Enter”进行换行。

清晰反馈：在进行API请求时，应有明确的加载状态提示（如AI头像旁出现动态图标）。

③ **可靠性与健壮性:**

异常处理：必须能够优雅地处理常见的异常情况，如：

网络中断/API连接失败：向用户弹出明确、友好的错误提示（例如：“网络连接异常，请检查后重试”），而非程序崩溃。

API返回错误：当API返回错误码时，应能捕获并向用户显示通用性提示（例如：“AI服务暂时不可用，请稍后重试”）。

④ **安全性:**

凭证管理：用于调用文华财经大模型API的认证凭证（如API Key），严禁硬编码在源代码中。必须通过外部配置文件、环境变量或安全的启动参数等方式进行加载，便于管理和保护。

5 后续跟踪

本次实习项目旨在构建一个坚实的核心原型。在此基础上，我们预见到该AI客服模块未来存在广阔的优化和迭代空间，这些方向可在后续工作中逐步实现。

① **功能增强：**

多轮对话记忆：引入会话上下文管理机制，使AI能够理解连续提问，进行有深度、有逻辑的对话。

富文本与多模态支持：升级UI渲染能力，支持AI回答中包含Markdown格式（如表格、代码块、列表）、可点击的超链接，甚至嵌入图片（如图表截图），使信息呈现更丰富、直观。

用户反馈机制：在每条AI回答后增加“赞/踩”或“满意/不满意”的按钮，收集用户对回答质量的直接反馈，为模型和提示词优化提供数据金矿。

② **智能化与个性化：**

意图识别与转人工：引入意图识别模块，当判断出AI无法解决复杂问题、用户情绪激动或用户明确要求时，触发“一键转接人工”流程，并将当前对话上下文无缝传递给人工坐席。

知识库增强（RAG）：对接文华财经内部的金融知识库（产品手册、交易规则、FAQ等），通过检索增强生成技术，确保AI回答的实时性、准确性和专业性，大幅减少“幻觉”现象。

③ **系统工程化：**

对话历史持久化：将用户的对话记录加密存储在本地，实现“历史消息云同步”，方便用户在不同设备上回顾。

端软件和移动App中。

监控与运维：建立完善的监控告警系统，对API调用成功率、响应延迟、用户满意度等关键指标进行实时监控，保障服务稳定运行。