智能对话系统技术规格文档  
（版本：Release 2.1 | 更新日期：2024-03-25）

# 一、系统架构概览

## 1. 会话管理子系统

**实现机制**

存储协议：采用原子化写入（Atomic Write）

文件结构：

json

{

"session\_id": "SID\_2024032501",

"messages": [

{"role": "user", "content": "..."},

{"role": "assistant", "content": "..."}

]

}

性能指标：

| **项目** | **指标值** |
| --- | --- |
| 最大会话数 | 500 |
| 单会话历史容量 | 10,000 tokens |
| 加载速度 | 200会话/秒 |

## 2. 双模型处理引擎

**架构设计**

**组件对比**

| **模块** | **主对话模型** | **摘要模型** |
| --- | --- | --- |
| 模型版本 | DeepSeekR1:70B | DeepSeekR1:1.5B |
| 最大上下文 | 4K tokens | 16K tokens |
| 响应延迟 | <800ms | <1.5s |
| 典型调用场景 | 实时对话,文件解析 | 知识检索，摘要生成 |

# 二、核心功能实现

## 1. 双模响应机制

**流式响应**

技术实现：SSE（Server-Sent Events）协议

分块策略：

python

for chunk in response\_stream:

if len(chunk) > 20: *# 智能分块*

yield chunk[:20] + "�"

else:

yield chunk

**同步响应**

质量保障：

完整性校验（CRC32）

敏感词过滤（预置 5,000+ 词库）

## 2. 上下文管理

**窗口策略**

动态调整算法： W = max(5, min(20, 0.2 \* T)) # W: 保留轮数，T: 总对话轮数

**记忆优化**

| **数据类型** | **保留策略** |
| --- | --- |
| 用户身份信息 | 永久存储 |
| 临时对话记录 | 30天自动清理 |
| 文件摘要 | 关联会话生命周期 |

## 3. 文件处理系统

**处理流程**

文件上传 → 2. 格式校验 → 3. 分块解析 → 4. 原子化存储

**性能基准**

| **文件类型** | **处理速度** | **内存占用** |
| --- | --- | --- |
| TXT (500KB) | 0.8s | 120MB |
| DOCX (1MB) | 2.1s | 250MB |
| Markdown (300KB) | 0.5s | 80MB |

# 三、RAG增强架构

## 1. 知识检索流程

**四阶段模型**

mermaid

graph TB

A[Query理解] --> B[向量检索]

B --> C[语义重排]

C --> D[证据增强]

D --> E[生成优化]

**检索性能**

| **数据集规模** | **召回率@3** | **响应延迟** |
| --- | --- | --- |
| 10,000文档 | 89.7% | 220ms |
| 100,000文档 | 82.3% | 480ms |

## 2. 混合检索策略

​**向量检索**：cosine相似度阈值 >0.78

​**关键词检索**：BM25算法（k1=1.2, b=0.75）

​**融合公式**：*Score*=0.6∗*Vsim*​+0.4∗*Kscore*​

# 四、扩展能力说明

## 1. 插件接口规范

**预处理接口**

python

def text\_preprocess(

raw\_text: str,

context: dict

) -> tuple[str, dict]:

"""输入文本预处理

参数：

raw\_text: 原始输入文本

context: 当前会话上下文

返回：

处理后的文本

更新后的上下文

"""

**后处理接口**

python

def response\_postprocess(

model\_output: str,

metadata: dict

) -> str:

"""输出结果后处理

参数：

model\_output: 模型原始输出

metadata: 包含检索结果等元数据

返回：

最终响应内容

"""

## 2. 扩展案例

​**法律文档插件**

条款自动编号

风险条款高亮（基于正则表达式）

​**医疗术语插件**

| **功能** | **实现方式** |
| --- | --- |
| 术语标准化 | SNOMED CT 映射表 |
| 剂量校验 | 药典数据库查询 |



A computer screen shot of a computer code

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated