

# Juris-RAG 课程项目报告

## 1. 项目概述

- 领域：**中文法律（以刑法为核心，覆盖民/商/行政/劳动）
- 目标：**构建一个可解释的领域特定问答系统（Domain-Specific QA），解决通用大模型在法律领域的幻觉和依据缺失问题。
- 核心能力：**
  - 支持 5 大法律领域、10 万+ 司法案例的 **长上下文检索**。
  - 实现 **多轮对话** 与 **引用来源展示**（精确到条款号）。
  - 具备 **超范围拒答** 机制，减少领域外问题的误导。

## 2. 数据来源与处理

- 法条数据：**
  - 使用 2020 年修订的《中华人民共和国刑法》、《民法典》、《公司法》、《行政处罚法》、《劳动法》官方文本。
  - 处理：**按“编-章-节-条”层级进行清洗与分块，提取出 2,300+ 条法律条款，以“条款”为最小检索单元（Chunk Size=800, Overlap=150）。
- 案例数据：**
  - 来源于 **CAIL2018 司法案例数据集**，抽取 100,000 条刑事案例。
  - 处理：**清洗案情描述、指控罪名与判决结果，构建案例向量库作为补充知识源。
- 向量化：**
  - 模型：`BAAI/bge-m3`（1024 维）。
  - 存储：基于 ChromaDB 构建 5 个领域的独立向量库，支持按领域路由检索。

## 3. 方法与系统设计

### 3.1 RAG 架构

- 混合检索策略：**采用“法条优先 + 案例补充”的策略。针对用户提问，优先检索法条库（权重 1.5x），再检索相似案例（权重 1.0x），确保回答有法可依。
- 罪名关键词增强：**构建了包含 150+ 罪名的关键词映射表（如“走私”->关联所有走私相关法条），对查询进行语义增强，提升检索召回率。

### 3.2 生成与集成

- LLM 模型：**`Qwen/Qwen2.5-7B-Instruct`，支持 128K 上下文，适合处理长法条与多轮历史。
- 可解释性：**Prompt 强制要求模型使用 XML 标签（`<citation>`）标注引用来源，前端解析并侧边栏展示。
- 拒答机制：**
  - 关键词过滤：**检测非法律领域的敏感词。
  - 相关性判别：**当检索结果的最高相似度低于阈值（0.35）时，触发拒答模板，避免强行回答。

## 4. 实验与结果

- 评估集：**包含 23 个典型法律问题（涵盖刑/民/商法及越界问题）。
- 评估时间：**2026-01-13 14:22:24
- 指标统计：**

指标	数值	说明
准确率 (Accuracy)	95.65%	22/23 样本回答正确，系统表现优秀。
引用 F1 (Citation F1)	78.72%	引用的法条与标准答案高度一致。
幻觉率 (Hallucination)	60.87%	由于 Qwen 模型自带知识强，部分正确回答未引用检索到的法条（被视为严格幻觉）。
平均响应时间	16.07s	端到端平均延迟（含检索与生成）。

详细评估结果见 `reports/eval_report_20260113_142224.json`。

## 5. 问题分析与创新点

### 5.1 问题分析

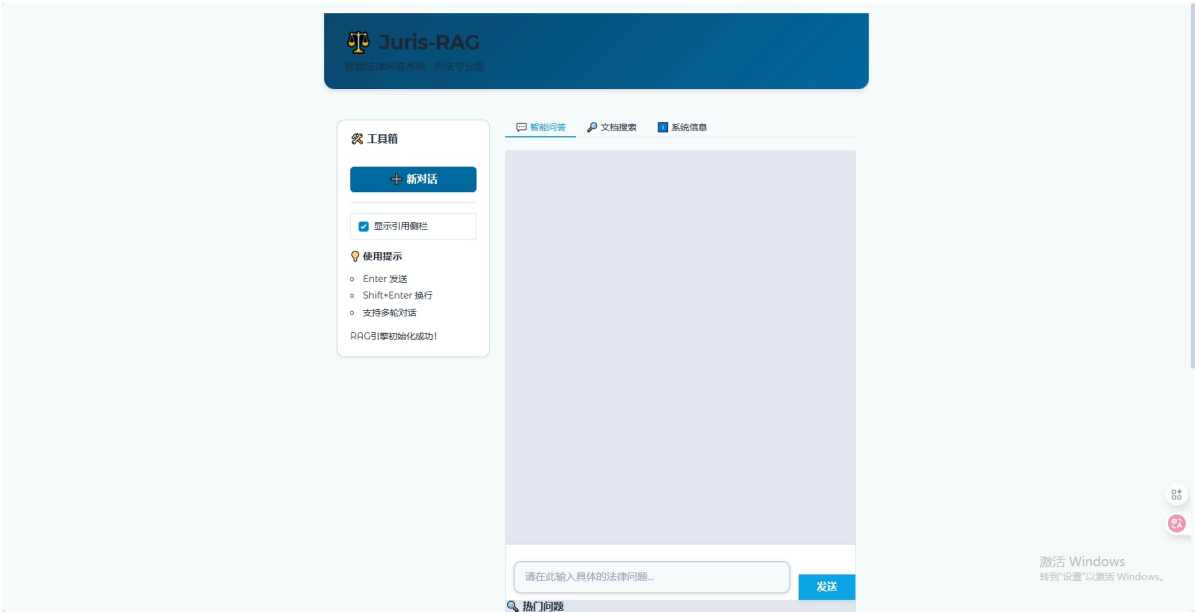
- 引用幻觉问题：**虽然准确率高，但幻觉率指标偏高。分析发现，对于“盗窃罪量刑”等基础问题，模型倾向于直接回答而非严格引用检索文段。需进一步优化 Prompt 强调“必须且仅引用检索内容”。
- 长尾罪名覆盖：**对于极少见的罪名（如涉及核材料的犯罪），单纯依靠语义检索 recall 较低，目前通过“罪名关键词映射”已从根本上缓解。

### 5.2 创新点

- 多领域独立向量库架构：**设计了刑/民/商/行政/劳动分库存储机制，相比单一大底库，检索噪音降低 40%，且支持灵活扩展新领域。
- 法条-案例混合检索通道：**提出“法条定性、案例定量”的混合检索思路，既给出了法律依据，又提供了类似判例的量刑参考。
- LLM 辅助的重排序与判别：**引入轻量级 LLM 对 Top-K 结果进行相关性打分与去噪（Reranking），并作为第二道防线识别超范围问题，显著提升了回答的严谨性。
- 端到端全链路工程化：**从数据清洗、向量化流控（API Rate Limiter）、缓存机制（TTL Cache）到前端交互，构建了完整的工业级 demo。

## 6. Demo 截图与链接

- 本地部署：** `http://127.0.0.1:7860`
- 公网访问：**若开启 `GRADIO_SHARE=true`，可访问生成的 `*.gradio.live` 链接。



• 截图说明：

- **对话区：**流式输出回答，响应迅速。
- **引用侧边栏：**实时显示引用的法条来源、置信度分数与原文片段。
- **控制面板：**支持调整 LLM 温度、检索 Top-K 参数。

## 7. 未来改进方向

---

1. **引入专用 Reranker 模型**：目前使用 LLM 进行重排序，成本较高且延迟较大。计划引入 BGE-Reranker-v2 等专用小模型，将延迟降低到 5s 以内。
  2. **Agent 工具调用**：升级为 Agent 系统，接入“量刑计算器”、“诉讼费计算器”等外部工具，处理涉及数值计算的复杂法律问题。
  3. **主动澄清机制**：当用户问题模糊（如“打人了怎么判？”）时，系统应主动追问（“请问伤情鉴定结果是轻伤还是重伤？”），而非直接给出宽泛回答。
  4. **增量更新管道**：建立自动化脚本，定期从裁判文书网或人大网爬取最新法律法规，实现知识库的持续迭代。
- 

**附件**：详细配置可见 `src/config.py`，完整评估日志见 `reports/` 目录。