**《宋史探微》项目介绍**

各位好，今天我将为大家介绍由我们小组开发设计的“宋史探微”， 这一宋代 RAG 系统项目。

**一、项目背景与目标**

本项目聚焦宋代历史领域，旨在打造一个高效、精准的知识问答系统，助力人们深入了解宋代历史文化。

**二、模型架构与核心模块**

**（一）模型架构**

项目采用了”离线结合在线”的架构，整合多种技术实现高效可信的RAG。我们将系统划分为多个模块，总体来说能够支持本地文档与网络数据的多源历史文本、原始文本清洗分块的预处理；基于 Bi-encoder 和生成式大模型进行定向微调、结合FAISS 向量检索与 BM25 算法实现了混合检索；在用户层面：具有意图识别、安全检测以及多轮对话的交互式系统；能够对答案来源进行冲突检测，让系统更加的可靠。

**（二）核心模块**

* **数据处理与知识库模块**：支持处理 PDF、DOCX 等多种本地文档，还能定向爬取宋代历史相关网络数据，并对原始文本进行清洗、过滤及分块处理。
* **微调数据集与模型优化模块**：搭建了自动化生成 “问题——答案” 对的流程来生成专用的历史数据集，并对Bi-encoder嵌入模型和生成大模型进行了微调。
* **向量与索引模块**：以 FAISS 为核心向量数据库，结合 BM25 算法，利用 bge-large-zh-v1.5 等模型实现文本向量化及结果重排序。
* **前后端交互与用户输入模块**：前端用 JavaScript、HTML5/CSS3 实现交互逻辑与页面构建，后端以 Flask 为框架，用 SQLite 存储对话历史，
* **输入模块：**构建了双层安全检测体系和意图识别模块。
* **检索模块：**通过智能改写与递归检索机制，进行多路信息获取，通过向量与关键词检索结果的结合，让历史知识查询更全面
* **生成模块：**除了输出生成结果外，还添加了冲突检测机制。

**三、优化分析与数据成果**

**（一）检索模型优化**

在 SOFT 数据集和 HARD 数据集上的测试显示，“bm25+bi-encoder (finetuned)+reranker” 方法召回率都高于其他模型，性能优异，尤其在 HARD 数据集这种高难度测试中，更凸显了模型的鲁棒性与泛化能力。

**（二）生成模型优化**

微调后的生成模型在 ROUGE 和 BLEU（质量指标） 等指标上有显著提升，文本生成质量大幅提高。

**四、实操演示与总结**

谢谢大家！