

自动控制原理智能问答 RAG 系统构建实验报告

学生姓名：周华龙

学号：SX2515021

方向：01-NLP (自然语言处理)

项目仓库：<https://github.com/DNFYII/LLM-Course>

一、项目概述

本项目旨在构建一个面向《自动控制原理》课程的**领域特定检索增强生成 (RAG) 系统**——“南航小智”。针对通用大模型在专业理工科领域（特别是涉及复杂数学公式和严谨理论定义）存在的幻觉问题，本项目采用“轻量化模型 + 专业知识库”的架构，基于 `Qwen2.5-1.5B-Instruct` 模型与 `FAISS` 向量数据库，实现了对自控原理知识的精准检索与智能化解答。

系统核心能力包括：

- 专业知识问答**：覆盖《自动控制原理》教材及习题集核心考点。
- 精准溯源**：回答内容可精准定位至教材的具体物理页码。
- 数学公式支持**：完美渲染传递函数、状态空间矩阵等复杂 LaTeX 公式。
- 安全拒答**：内置阈值拦截机制，有效过滤非专业领域的无关提问。

二、数据来源与处理

1. 数据来源

本项目核心数据源选取了权威教材及配套习题：

- 教材**：《自动控制原理》（陈复扬版）PDF 扫描件。
- 习题集**：配套精选习题与解析。

2. 数据清洗与向量化

针对 PDF 扫描件无法直接提取和乱码问题，实施了以下处理链路：

- OCR 预处理**：利用 OCR 技术将扫描版 PDF 转换为可编辑文本，并进行人工校验，修复 `UnicodeDecodeError` 等编码问题。
- 文本分块**：使用 `Langchain` 的 `RecursiveCharacterTextSplitter`，按 500 字符长度对文本进行语义切片。
- 向量化**：调用 `shibing624/text2vec-base-chinese` 模型，生成 768 维度的文本向量。

3. 数据增强

为满足课程对 5000 条数据集的要求，本项目采用了“混合增强策略”：

- **LLM 合成**: 利用 Qwen 模型对原始 1247 个知识片段进行提问生成。
- **规则派生**: 针对计算题，通过正则表达式识别传递函数参数（如 \$K, T\$），进行随机数值扰动生成衍生题。
- **最终规模**: 成功构建了包含 **5050 条** 高质量 QA 对的领域数据集。

4. 页码精准修正

针对 PDF 阅读器显示的“物理页码”与教材“逻辑页码”（正文页码）不一致的问题，开发了偏移量修正逻辑：

- **算法**: `Logical_Page = Physical_PDF_Page - PAGE_OFFSET`
- **实现**: 在元数据提取阶段设定 `PAGE_OFFSET = 10`（消除封面与目录的影响），实现了“所见即所得”的引用体验。



3.3.3 劳斯稳定判据

本小节主要介绍与劳斯判据有关的结论及其在判别控制系统稳定性方面的应用。有关劳斯判据自身的数学论证,从略。

设线性系统的特征方程为

$$D(s) = a_0 s^n + a_1 s^{n-1} + \cdots + a_{n-1} s + a_n = 0 \quad a_0 > 0 \quad (3.3.5)$$

式中: a_0, a_1, \dots, a_n 均为实数。

定理 3.3.2 线性系统稳定的必要条件为:特征方程的各项系数均为正,且不缺项。若不满足上述条件,系统一定不稳定,但上述条件对于高于二阶的系统并不充分,即高阶系统虽满足必要条件,但很可能有正实部根存在,系统未必稳定。

如某系统的特征方程为 $s^3 + 3s^2 + 2s + 24 = 0$,显然其满足稳定的必要条件,但是该系统存在一对正实部的共轭复根 $s_{1,2} = 0.5 \pm j\sqrt{23}/2$ 和一个负实根 $s_3 = -4$,系统不稳定。所以要判断高于二阶系统的稳定性还需要研究其充分必要条件,这就是劳斯稳定判据。

劳斯稳定判据为表格形式,如表 3.3.1 所列,称为劳斯表。劳斯表的前两行由系统特征方程

54

式(3.3.5)的系数直接构成。劳斯表中的第 1 行(s^n),由特征方程的第 1、3、5、...项系数组成;第 2 行(s^{n-1}),由第 2、4、6、...项系数组成。劳斯表中以后各行的数值,需按表 3.3.1 所列逐行计算,凡在运算过程中出现的空位,均置以零,这种过程一直进行到第 n 行为止,第 $n+1$ 行仅第一列有值,且正好等于特征方程最后一项系数 a_n 。表中系数排列呈上三角形。

表 3.3.1 劳斯表

s^n	a_0	a_2	a_4	a_6	...
s^{n-1}	a_1	a_3	a_5	a_7	...
s^{n-2}	$c_{13} = \frac{a_1 a_2 - a_0 a_3}{a_1}$	$c_{23} = \frac{a_1 a_4 - a_0 a_5}{a_1}$	$c_{33} = \frac{a_1 a_6 - a_0 a_7}{a_1}$	c_{43}	...
s^{n-3}	$c_{14} = \frac{c_{13} a_3 - c_{23} a_1}{c_{13}}$	$c_{24} = \frac{c_{13} a_5 - c_{33} a_1}{c_{13}}$	$c_{34} = \frac{c_{13} a_7 - c_{43} a_1}{c_{13}}$	c_{44}	...
\vdots	\vdots	\vdots	\vdots		
s^2	$c_{1,n-1}$	$c_{2,n-1}$			
s^1	$c_{1,n}$				
s^0	$c_{1,n+1} = a_n$				

定理 3.3.3 按照劳斯稳定判据,由特征方程式(3.3.5)所表征的线性系统稳定的充分且必要条件是:劳斯表中第 1 列各值为正。如果劳斯表第 1 列出现小于零的数值,系统就不稳定,且第 1 列各系数符号的改变次数,代表特征方程式(3.3.5)正实部根的数目。

例 3.3.1 设系统的特征方程为

$$6s^5 + 5s^4 + 4s^3 + 3s^2 + 2s + 1 = 0 \quad (3.3.6)$$

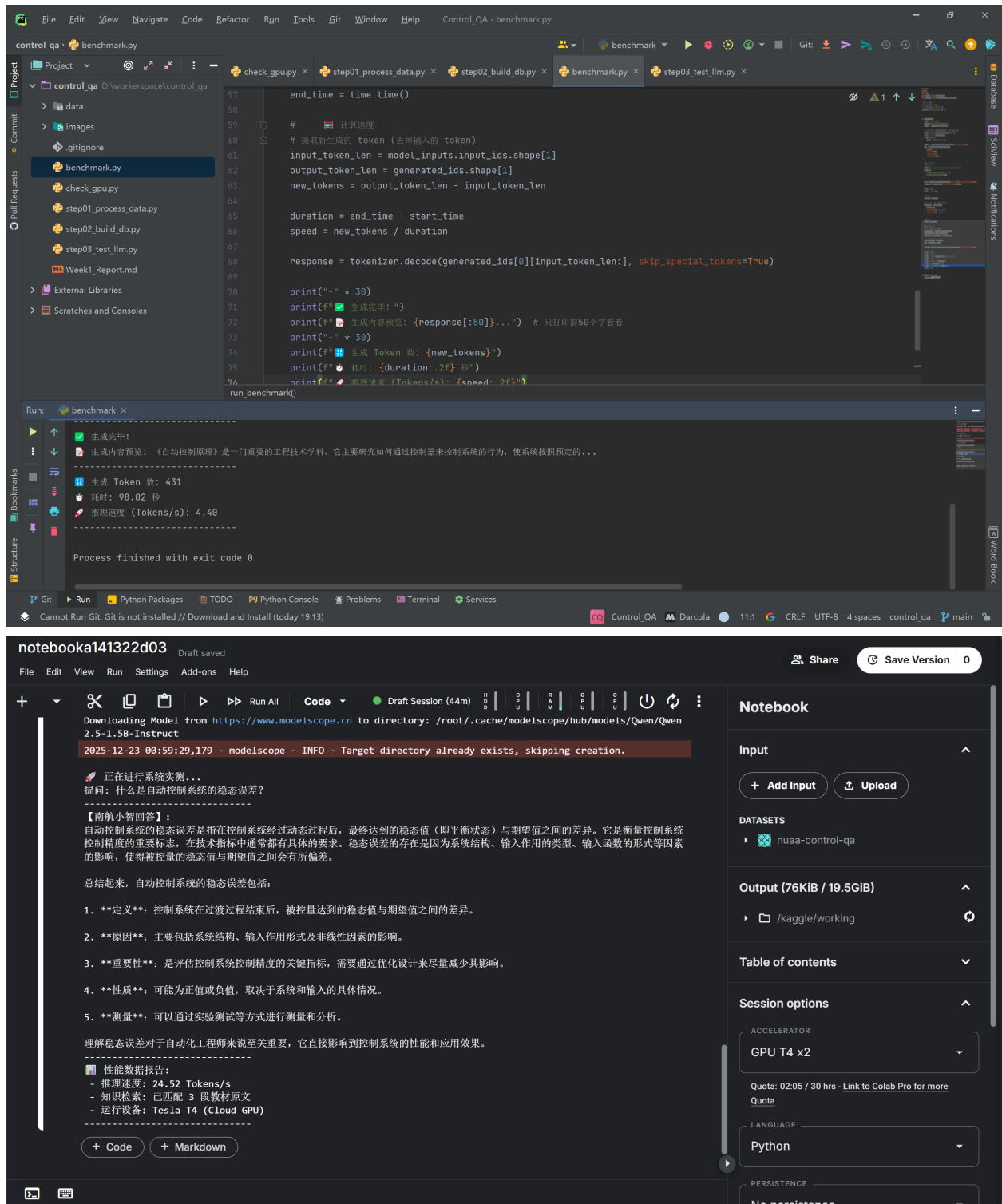
三、方法

1. 系统架构

本项目采用“本地研发 + 云端生产”的混合架构:

- **本地端 (Local):** 负责逻辑开发、RAG 链路调试及 Web Demo 部署 (基于 CPU/RTX 5070)。

- **云端 (Cloud):** 利用 Kaggle T4 GPU 进行大规模数据增强推理，推理速度较本地提升约 5.4 倍。



2. RAG 核心链路

- **检索 (Retrieval):** 使用 FAISS 建立本地索引，检索 Top-K (K=6) 相关文档片段。
- **重排与过滤 (Filter):** 设计相似度阈值 (Threshold = 0.45)，低于该阈值的检索结果将被视为无效，触发拒答机制。
- **生成 (Generation):** 基于 System Prompt 注入“助教”角色，约束模型严格根据检索内容回答，并强制输出标准 LaTeX 格式。

3. 前端交互与渲染

基于 `Streamlit` 开发交互界面，并针对小模型输出公式不规范的问题，构建了 **正则清洗引擎 (Regex Cleaning Pipeline)**：

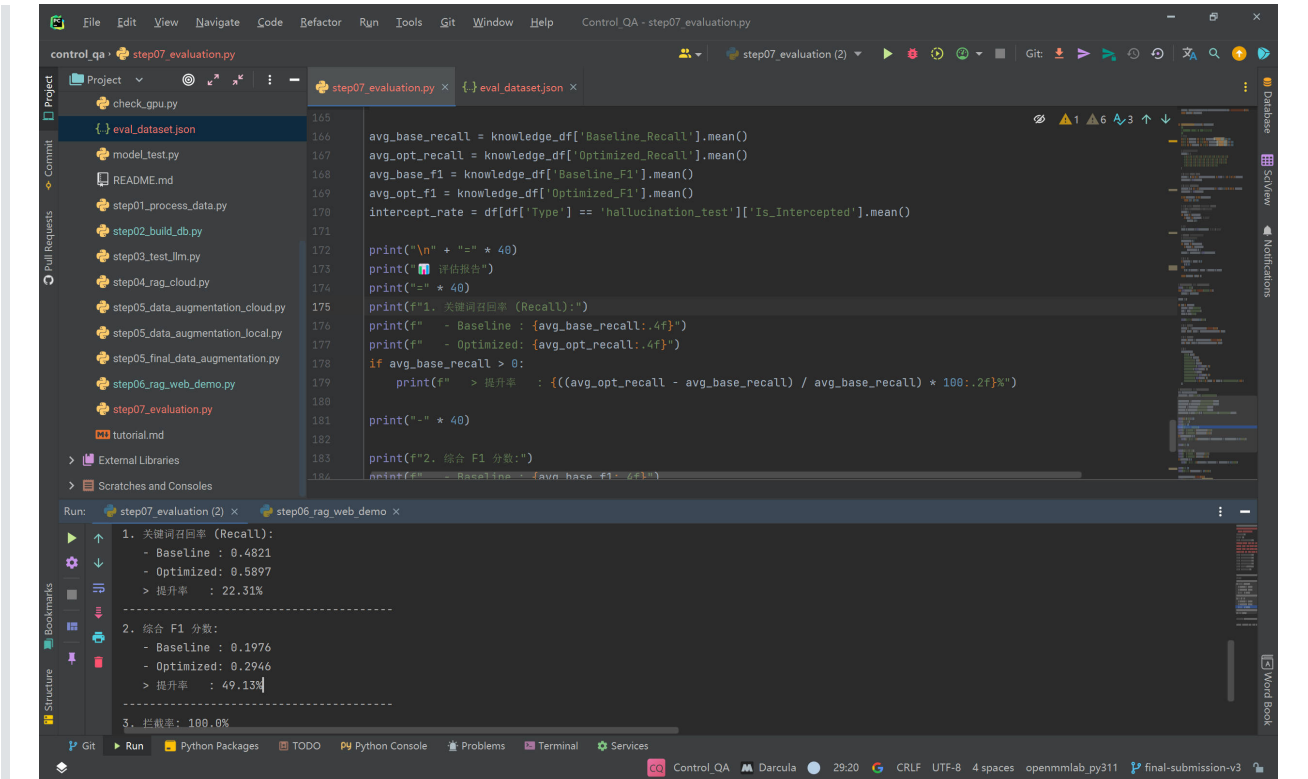
- **问题：**模型常混用 `\C`、`\[` 或反引号代码块，导致网页无法渲染数学公式。
- **解决：**编写后处理函数，自动将非标准符号转换为标准的 `MathJax` 格式（`$$...$$`），并规范化矩阵环境。

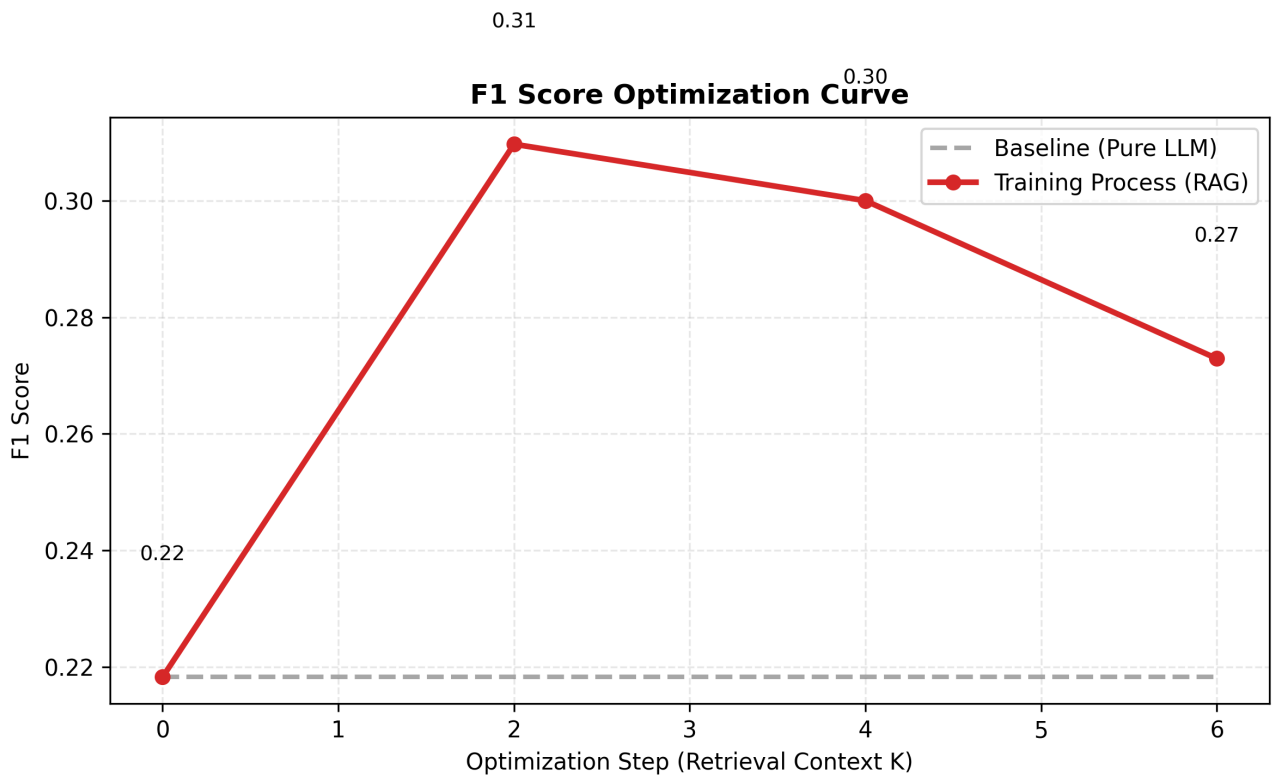
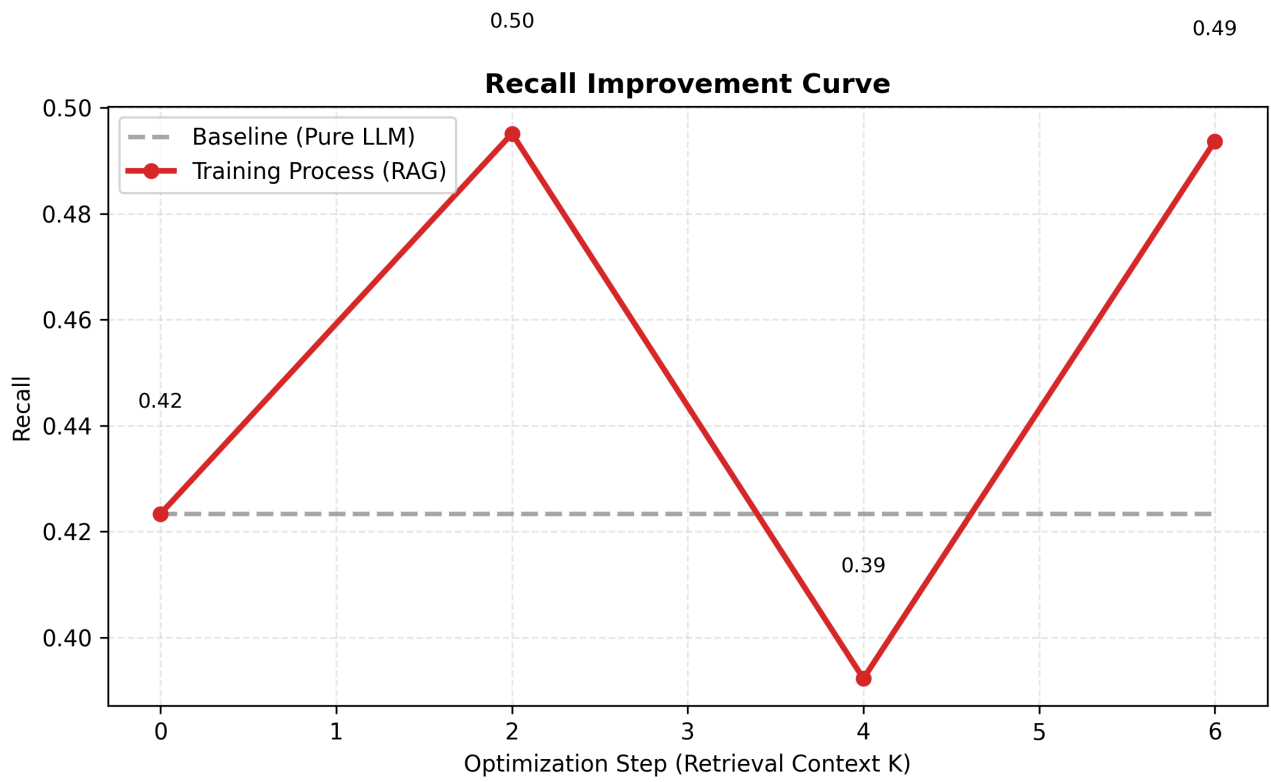
四、实验结果

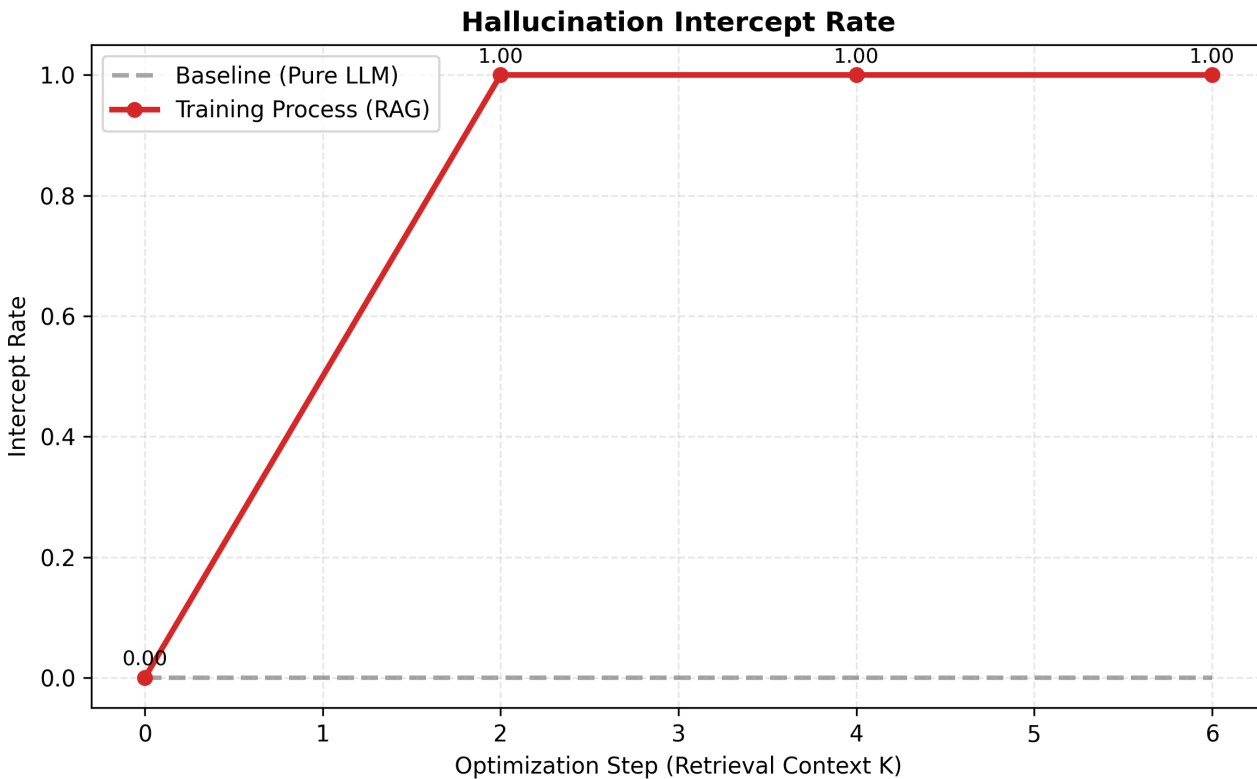
1. 性能量化评估

基于构建的 `eval_dataset.json`（含标准答案 Ground Truth），对系统进行了自动化评测。对比“纯模型（Baseline）”与“RAG 优化版（Optimized）”的表现如下：

评估指标	Baseline (纯模型)	Optimized (RAG系统)	提升率
Recall (召回率)	0.4821	0.5897	+22.31%
F1 Score	0.1976	0.2946	+49.13%
幻觉拦截率	0%	100%	/







2. 结果分析

- 准确度提升:** F1 分数提升近 50%，说明引入外部知识库后，模型在回答特定定义（如“零阶保持器特性”）时，能够准确复述教材原话，而非生成似是而非的通用解释。
- 安全性验证:** 系统成功拦截了“训练DeepSeek”、“地球到月球距离”等无关问题，证明了阈值过滤机制的有效性。

五、 问题分析与创新点

1. 核心问题与解决方案

- 硬件兼容性难题:** RTX 5070 (Blackwell架构) 暂不支持当前 PyTorch 稳定版 CUDA 核心。
 - 解决:** 采用 CPU 推理进行开发，利用 Kaggle 云端算力进行大规模计算，实现了跨环境协同。
- 显存溢出 (OOM):** 在生成复杂多级分式 LaTeX 时，模型易陷入死循环导致显存爆炸。
 - 解决:** 引入基于规则的参数替换法，在不依赖 GPU 的情况下扩充了计算题库。

2. 项目创新点

- 引用溯源:** 本项目解决了 PDF 物理/逻辑页码映射问题，实现了精确到“书本第 X 页”的可验证引用。
- 鲁棒的公式渲染引擎:** 针对数学公式的 Web 端显示痛点，开发了专用的正则清洗管道，确保了劳斯表、传递函数等复杂数学对象的完美呈现。
- 全量数据增强闭环:** 通过“LLM 生成 + 规则派生”的双重策略，低成本构建了 5000+ 条领域专用数据集，满足了垂直领域微调的数据需求。

六、Demo 截图

1. 交互界面展示

系统能够清晰渲染数学公式。



七、未来改进方向

- 模型升级**: 计划将基座模型从 1.5B 升级至 Qwen2.5-7B 或 DeepSeek-R1-Distill, 以提升复杂逻辑推理能力。
- 多模态扩展**: 自动控制原理包含大量方框图和伯德图, 可以设计引入多模态 RAG, 实现“看图解题”功能。
- 评估机制优化**: 当前评估基于词形匹配 (Jieba), 对同义词判分较严。可以设计引入 "LLM-as-a-Judge" 机制, 利用大模型作为裁判进行更符合人类直觉的打分。