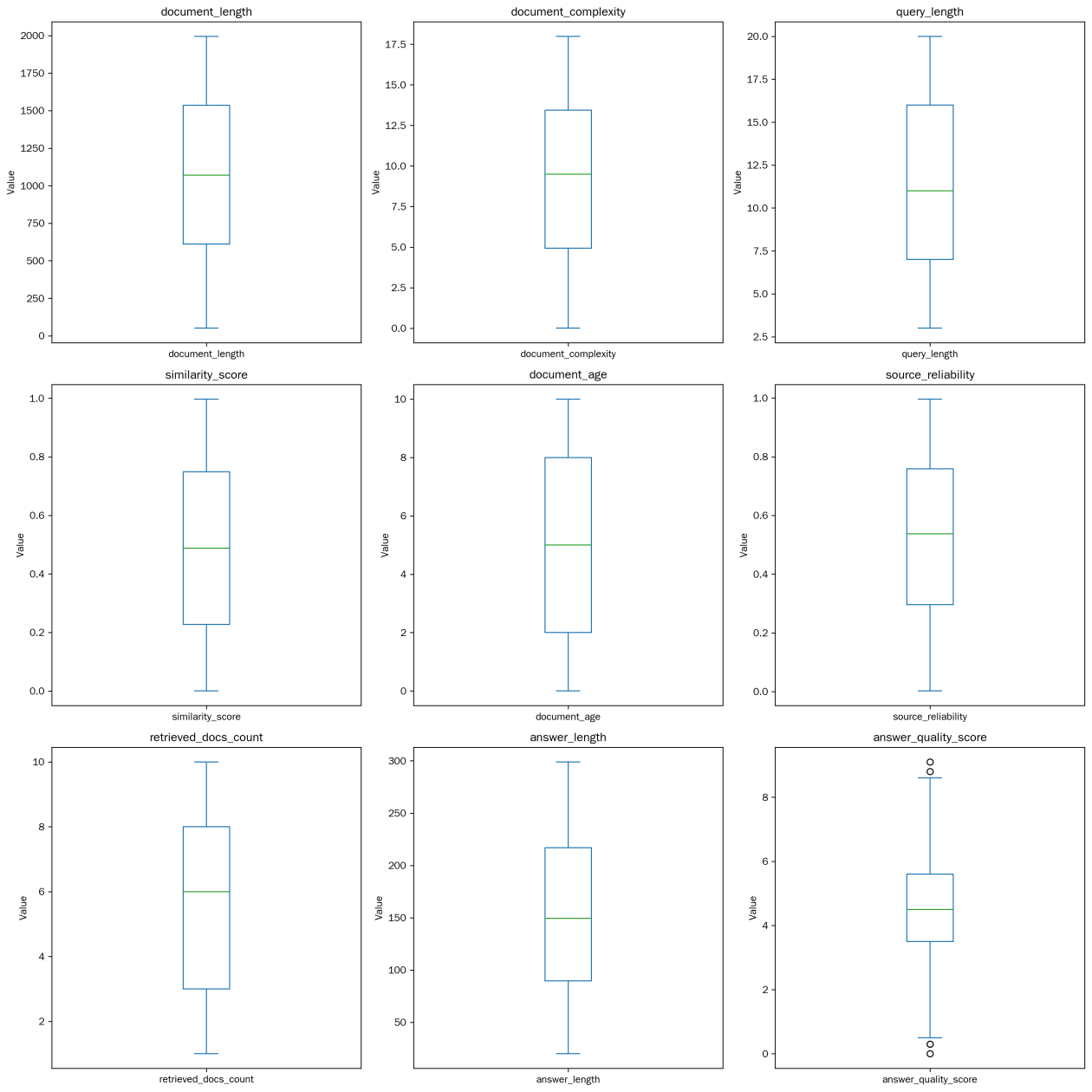
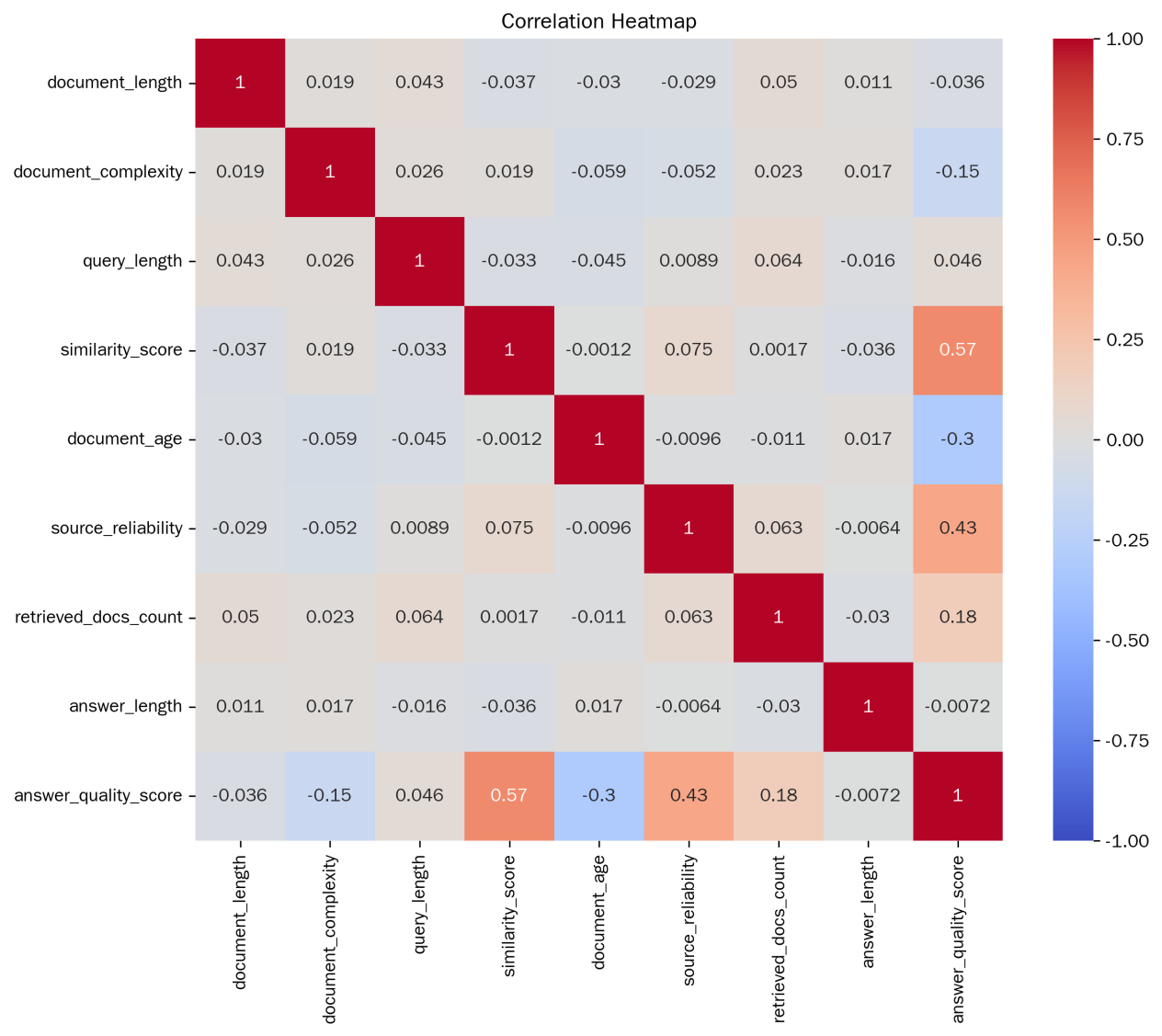
研究背景：随着大语言模型（LLMs）在知识密集型任务中的规模化应用，检索增强生成（RAG）技术因能有效解决LLMs上下文窗口有限、知识时效性滞后及生成内容“幻觉”等核心痛点，已成为提升LLM实用价值的核心技术路径；然而，当前RAG领域研究多聚焦于检索算法革新与生成阶段prompt工程改进，对RAG系统全链路关键特征与最终生成质量的量化关联分析存在明显缺口，且现有RAG相关数据集多为单一领域静态文档-查询-回答三元组，既未系统标注文档复杂度、来源可靠性、文档时效性等核心属性，也未提供基于业务逻辑的可量化生成质量标签，导致无法支撑特征重要性评估、性能预测模型训练等深度研究，进而制约RAG系统从经验化调优向数据驱动的智能化优化转型。在此背景下，本研究旨在为RAG系统性能优化、领域适配及相关预测模型训练提供支撑，推动RAG技术研究从定性分析向定量验证迈进。

数据集介绍

该数据集是用于RAG（检索增强生成）系统性能评估的一个数据集，共包含728条样本，涵盖特征变量、文本信息与预测变量三个模块，可用于RAG模型训练、性能优化及特征重要性分析。特征变量层面，包括RAG系统关键影响因素：文档维度、查询维度、检索维度及回答维度，包含技术、医学、法律等7个领域；文本信息层面，包含文档标题和查询文本；预测变量为回答质量分数（0-10分）。下面是各个变量的箱线图分析：



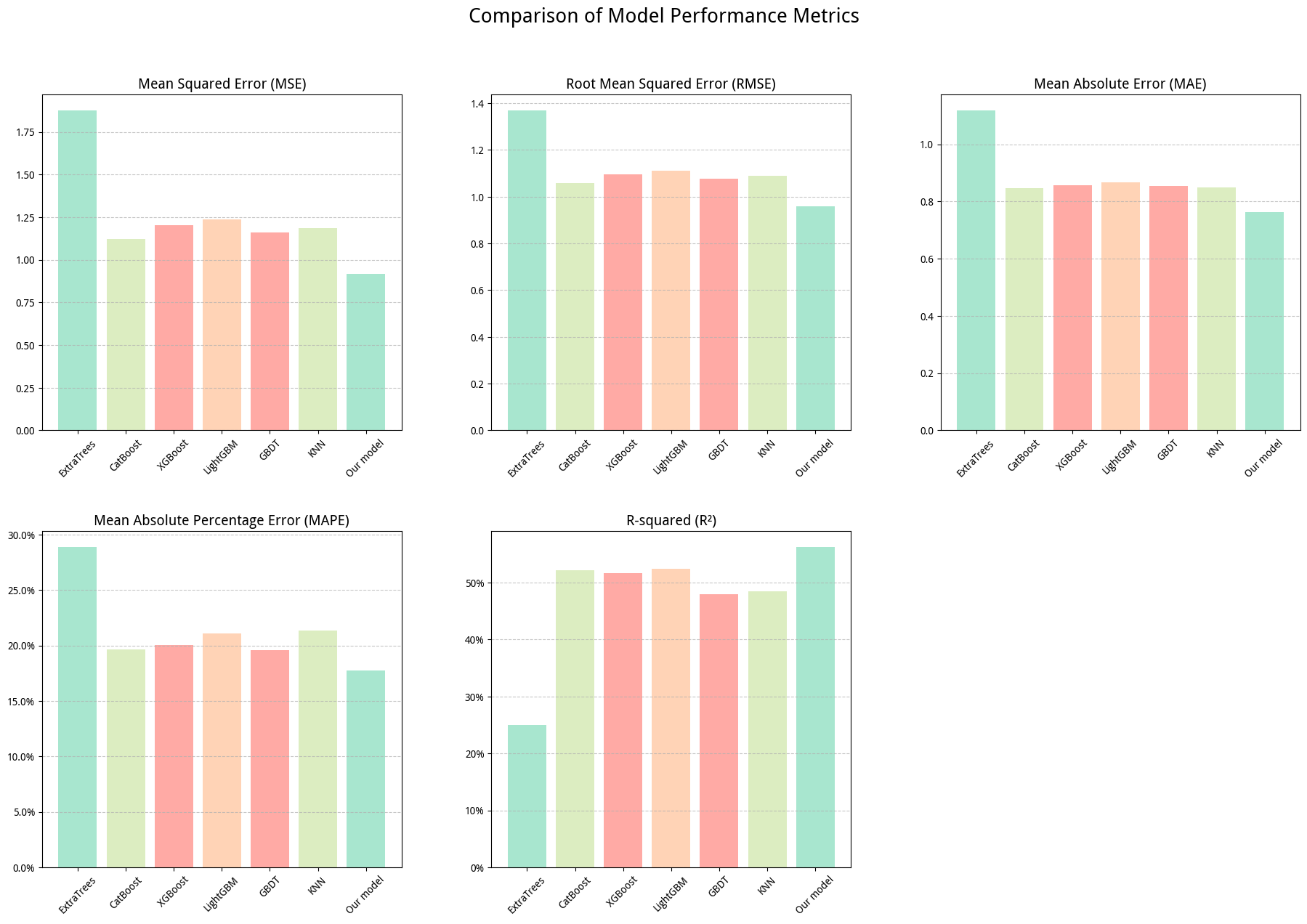
下文是相关性热力图，反应了回答质量分数与其他各个变量之间的相关性。



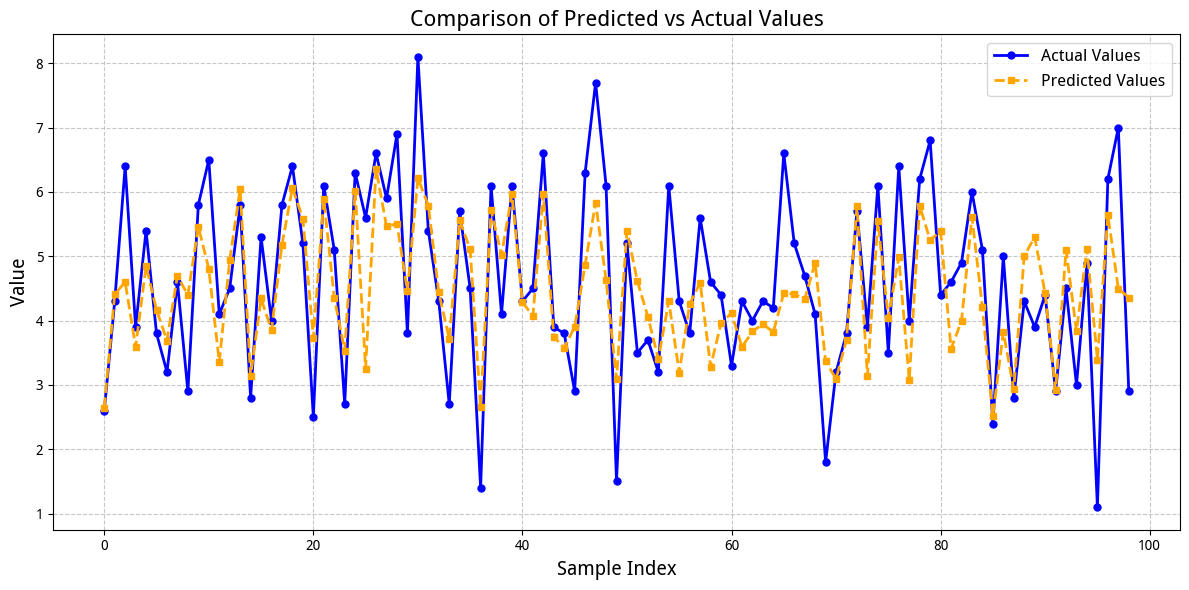
对比实验分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | MAPE | R² |
| ExtraTrees | 1.877 | 1.37 | 1.119 | 28.858 | 0.25 |
| CatBoost | 1.124 | 1.06 | 0.846 | 19.664 | 0.521 |
| XGBoost | 1.203 | 1.097 | 0.858 | 20.009 | 0.516 |
| LightGBM | 1.237 | 1.112 | 0.868 | 21.099 | 0.524 |
| GBDT | 1.162 | 1.078 | 0.855 | 19.598 | 0.48 |
| KNN | 1.184 | 1.088 | 0.849 | 21.372 | 0.485 |
| Our model | 0.919 | 0.959 | 0.762 | 17.731 | 0.562 |

实验结果分析：在各项模型评估指标中，Our model 的表现均优于 ExtraTrees、CatBoost、XGBoost、LightGBM、GBDT 和 KNN 这六种模型。具体来看，Our model 的 MSE 为 0.919，显著低于其他模型中最小的 CatBoost 的 1.124 以及最大的 ExtraTrees 的 1.877；RMSE 为 0.959，低于其他模型中最小的 CatBoost 的 1.06；MAE 为 0.762，小于其他模型中最小的 CatBoost 的 0.846；MAPE 为 17.731，低于其他模型中最小的 GBDT 的 19.598；R² 为 0.562，高于其他模型中最大的 LightGBM 的 0.524。整体而言，Our model 在误差类指标（MSE、RMSE、MAE、MAPE）上均实现了最小化，在拟合优度指标（R²）上实现了最大化，综合性能更优。



预测值与实际值的折线图



算法原理：该算法将群智能优化算法与BiLSTM-Attention模型结合实现回归任务。群智能优化算法通过初始化种群生成候选超参数组合，经迭代搜索优化L2正则化系数、初始学习率和隐藏层节点数，搜索过程中通过边界检查确保参数在合理范围，并借助食物存储机制保留更优解。BiLSTM-Attention模型以优化后的参数构建，通过序列输入层接收预处理数据，双向LSTM层捕捉时序特征，ReLU激活层引入非线性，自注意力层强化关键特征权重，dropout层抑制过拟合，全连接层和softmax层输出回归结果。群智能优化算法通过最小化1-准确率的适应度函数引导参数寻优，最终使模型在优化后的参数配置下实现高效回归。