УДК 004.89

РАЗРАБОТКА ПЛАТФОРМЫ СРАВНИТЕЛЬНОЙ ОЦЕНКИ RAG-АРХИТЕКТУР ПО КАЧЕСТВУ И ЗАДЕРЖКАМ

Салимов Ш.К.[1]
Научный руководитель – канд. физ.-мат. наук, преподаватель Жуков Н.Н.[1]
[1]Университет ИТМО
e-mail: Shaliko9@yandex.ru

Работа выполнена в рамках темы НИР «Исследования и оптимизация архитектур Retrieval-Augmented Generation в распределенных системах хранения и обработки данных в веб-среде».

Введение

В современных Retrieval-Augmented Generation (RAG) системах применяются различные подходы к проектированию retrieval (извлечения релевантных документов) и embedding компонентов, а также механизмов объединения источников. При этом сравнимость архитектур низкая, так как нет единого метода оценки качества и системных метрик: используются различные датасеты, протоколы экспериментов и наборы метрик. Кроме того, в современных исследованиях проводятся несопоставимые друг с другом оценки, которые специфичны для конкретной предметной области и выбранной архитектуры. В ряде зарубежных исследований утверждается, что необходимы комплексные методы оценки производительности RAG, которые отражают значимые аспекты работы компонентов [1]. В отечественных работах проводятся конкретные измерения семантического поиска, а также применяются основные метрики Recall@k, MRR@k и временные показали при работе со специализированными коллекциями текстов [4,5]. Оценивание проводится не только по качеству ответа Large Language Model (LLM), но и через метрики классификации и извлечения релевантных документов из векторного хранилища. В условиях активного развития и роста применения RAG сравнение подходов усложняется из-за отсутствия комплексной платформы для проведения тестирования и замеров метрик конкретных архитектур. Таким образом, актуальной задачей является создание воспроизводимого подхода к оценке RAG-архитектур, который будет включать в себя подробный и информативный протокол тестирования, сохранение конфигурации и сбор метрик качества и задержек.

Основная часть

В рамках работы разработана платформа для сравнительной оценки RAG-архитектур. Эксперимент задается декларативной конфигурацией и поддерживает гибкую интеграцию новых методов оценки и тестирования. Поддерживается работа с различными датасетами, векторными хранилищами, условиями запуска и временными рамками каждого из этапов работы RAG. Кроме того, предусмотрены сценарии retrieval с single (поиск в одном источнике) и fanout_merge (параллельный опрос нескольких источников с объединением результатов) стратегиями, а также асинхронное выполнение fanout (параллельный опрос нескольких источников). Для объединения результатов реализованы Reciprocal Rank Fusion (RRF) и дедупликация документов при работе retrieval. Сбор метрик включает Recall@k, MRR@k и процентильные оценки задержек работы системы, включая end-to-end задержки сетевых запросов. Поддержка кэширования на этапах эмбеддингов и retrieval позволяет проводить тестирование, применяя оптимизации наиболее близким к реальным системам. Кроме того, реализована фиксация результатов оценки и сохранение параметров и конфигурации каждого из экспериментов, что позволяет наглядно увидеть особенности работы. Архитектура платформы спроектирована таким образом, чтобы обеспечить масштабируемость и возможность доработки под новые механизмы и сценарии

тестирования. Таким образом, применение разработанной платформы позволяет выявить сильные и слабые стороны различных архитектур и подходов.

Заключение

В работе разработана платформа для тестирования и сравнительной оценки RAG-архитектур, которая позволяет конфигурировать эксперименты декларативным языком и определять параметры запусков. Поддержка различных стратегий retrieval и объединение результатов с дедупликацией позволяют проанализировать особенности систем в разрезе их качества и сетевых задержек в распределенных системах. Платформа обеспечивает оценку качества retrieval и системных характеристик, таких как процентильные значения задержек и контроль временных границ этапов работы RAG. Применение системы и результаты ее работы представлены на типовых конфигурациях, включающими в себя кэширование, многоповторные сценарии и устойчивость к задержкам.

Литература

1. Yang X. et al. Crag-comprehensive RAG benchmark //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 10470-10490.
2. Tang Y., Yang Y. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries //arXiv preprint arXiv:2401.15391. – 2024.
3. Pipitone N., Alami G. H. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain //arXiv preprint arXiv:2408.10343. – 2024
4. Chen Li, Xiaoyu Wang, Tongyu Zong, Houwei Cao, Yong Liu. Predictive Edge Caching through Deep Mining of Sequential Patterns in User Content Retrievals // Computer Networks. 2023. Vol. 233. C. 109866 https://doi.org/10.1016/j.comnet.2023.109866
5. Лосев, Н. Исследование зависимости качества семантического поиска в RAG от масштаба модели эмбеддингов: сравнительный анализ точности извлечения признаков / Н. Лосев, Э. Гарбаренко. — Текст: непосредственный // LECTIO IBI – 2025. — СПб., 2025. — С. 469-477.
6. Каширина, И. Разработка и оценка RAG-системы для анализа семантических связей / И. Каширина, И. Осипов, В. Яковлев. — Текст: непосредственный // Вестник воронежского государственного университета. Серия: системный анализ и информационные технологии. — 2025. — № 2. — С. 114-126.