УДК 004.8

Система интеллектуальной поддержки научных исследований с использованием больших языковых моделей

Харьковской Р.Р. (Университет ИТМО)
Научный руководитель - к.т.н. Федоров Д.А. (Университет ИТМО)

Введение. Современные научные исследования требуют автоматизации обработки и анализа больших объёмов информации. В связи с этим использование LLM в комбинации с RAG представляет перспективное направление для поддержки исследовательской деятельности. Настоящая работа посвящена разработке OpenSource-системы, которая помогает учёным на всех этапах подготовки научных статей: от уточнения темы исследования и формирования структуры до генерации обзоров литературы. Ключевое преимущество предлагаемого решения заключается в его способности адаптироваться к специфике исследовательских запросов и предоставлять точную, актуальную и структурированную информацию.

Актуальность темы обусловлена стремительным увеличением объёма научных публикаций, что затрудняет быстрый и точный доступ к актуальной информации [1]. Однако генеративные модели могут выдавать недостоверные либо устаревшие данные, поскольку обучены на статичных корпусах текстов. Для решения этой проблемы в систему интегрируется механизм RAG [2], который позволяет комбинировать возможности больших языковых моделей с актуальными данными из пользовательских источников. Это обеспечивает релевантность извлекаемой информации и облегчает анализ научных материалов. В результате исследователи смогут тратить меньше времени на поиск источников и обработку информации, сосредоточившись на аналитике и разработке новых научных идей.

Основная часть. Разрабатываемая система включает в себя:

- Веб-интерфейс для загрузки научных статей и ввода поискового запроса.
- Модуль поиска релевантной информации в загруженных документах с использованием RAG.
- LLM-агентов для структурирования информации, аннотирования статей и генерации кратких обзоров.
- Автоматизированную генерацию научного обзора на основе найденных данных.

Роль LLM-агентов в системе. LLM-агенты в данной системе представляют собой интеллектуальные модули, которые выполняют:

- Автоматический анализ и структурирование загруженных статей.
- Формулирование уточняющих вопросов для пользователя для повышения качества запроса.
- Генерацию аннотаций для загруженных статей.
- Составление научных обзоров на основе релевантных фрагментов текстов.

Оценка эффективности. Для оценки эффективности предложенного подхода проведён сравнительный анализ различных LLM (GPT-3, GPT-4, LLaMA, YandexGPT, GigaChat и DeepSeek), а также их комбинаций с RAG и использованием LLM-агентов. В качестве основных метрик выбраны [3]: ROUGE, BERTScore, Precision, Recall, F1-Score[3].

Заключение. Разработанная OpenSource-система на базе LLM, RAG и LLM-агентов призвана помогать исследователям в написании научных публикаций. Она повышает продуктивность за счёт автоматизации рутинных задач и обеспечивает более точный анализ данных из актуальных источников. Интеграция LLM-агентов позволяет повысить удобство

взаимодействия с системой и достичь более эффективного использования научной информации.

Список использованных источников:

1. Ziming, Luo LLM4SR: A Survey on Large Language Models for Scientific Research / Luo Ziming. — Текст : электронный // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2501.04306v1 (дата обращения: 13.02.2025).
2. Shailja, Gupta A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions / Gupta Shailja. — Текст : электронный // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2410.12837 (дата обращения: 13.02.2025).
3. Asli, Celikyilmaz Evaluation of Text Generation: A Survey / Celikyilmaz Asli. — Текст : электронный // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2006.14799 (дата обращения: 13.02.2025).

Харьковской Р.Р. (автор) _____

Федоров Д.А. (научный руководитель) _____