УДК 004.891

ПОДХОД К ОЦЕНКЕ КОММЕРЧЕСКИХ РУССКОЯЗЫЧНЫХ LLM ДЛЯ ИСПОЛЬЗОВАНИЯ В RAG АРХИТЕКТУРЕ НА ОСНОВАНИИ БЕНЧМАРКА CRAG

Иванов О.А. (ИТМО), **Титоренко А.А.** (ИТМО)
Научные руководители – кандидат технических наук, доцент Романов А.А. (ИТМО), **кандидат физико-математических наук, доцент Бойцев А.А.** (ИТМО)

Введение. В последние годы архитектура Retrieval-Augmented Generation (RAG), позволяющая сочетать преимущества языковых моделей с возможностью извлечения релевантной информации из внешних источников [1], стала одним из ключевых направлений в области обработки естественного языка. Разработка эффективных решений, использующих RAG-архитектуру, требует комплексной оценки ее составляющих. В англоязычных источниках существуют специализированные методы оценки - бенчмарки, такие как CRAG [2], UDA [3], RGB [4], предназначенные для оценки точности генерации ответов на основе внешних данных. Существует недостаток исследований по оценке производительности русскоязычных коммерческих моделей в контексте RAG. В данной работе представлен подход к анализу производительности некоторых популярных русскоязычных коммерческих больших языковых моделей в контексте RAG-архитектуры с использованием бенчмарка CRAG.

Основная часть. CRAG (Comprehensive RAG Benchmark) представляет собой комплексный бенчмарк для оценки систем генерации с использованием извлечения информации (Retrieval-Augmented Generation, RAG). Он включает в себя 4,409 вопросно-ответных пар, охватывающих пять тематических областей (финансы, спорт, музыка, кино, общие знания) и восемь типов вопросов (простые запросы о статических фактах, вопросы с условиями, сравнение сущностей, агрегацию данных, множественные запросы, многошаговые рассуждения, задачи с интенсивной постобработкой и вопросы с ложными предпосылками), что позволяет моделировать различные реальные сценарии пользовательского взаимодействия.

В рамках исследования проведена оценка производительности коммерческих моделей семейств YandexGPT [5], GigaChat [6] в контексте самостоятельных моделей и в рамках бейзлайн RAG-архитектуры и их сравнение с SOTA решениями на части набора данных CRAG. Исследуемые русскоязычные коммерческие большие языковые модели были интегрированы в процесс оценки CRAG с использованием публичных API.

Тестирование производилось на следующих задачах: суммаризация извлеченной информации, улучшение ответов за счет извлеченных знаний (графы знаний и веб-поиск), сквозной (end-to-end) RAG. В последнем случае система имела доступ к 50 веб-страницам и API, имитирующим реальное взаимодействие.

Производительность модели оценивалась с использованием метрик точности и достоверности (truthfulness, среднее значение оценок по всем вопросам) с классификацией ответов на четыре категории: идеальный ответ, приемлемый ответ, отсутствие ответа, некорректный ответ. За каждый правильный ответ начислялись баллы, за неправильный ответ баллы отнимались. Оценка и валидация ответов производилась в автоматическом режиме с помощью моделей GPT-4 и GPT-3.5 Turbo.

Выводы. В ходе исследования реализован подход к оценке и сравнению российских коммерческих больший языковых моделей семейства YandexGPT и GigaChat в контексте их применения в RAG-архитектуре с использованием части вопросов бенчмарка CRAG. Произведено сравнение производительности с SOTA решениями от OpenAI. Определены основные ограничения подхода и дальнейшие направления развития, включающие добавление и адаптирование вопросов из популярных русскоязычных бенчмарков и создание

собственных наборов данных, совместимых с CRAG.

Список использованных источников:

1. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint. – 2020. – URL: https://arxiv.org/abs/2005.11401.
2. Yang X., Sun K., Xin H., Sun Y., Bhalla N., Chen X., Choudhary S., Gui R., Jiang Z., Jiang Z., Kong L., Moran B., Wang J., Xu Y., Yan A., Yang C., Yuan E., Zha H., Tang N., Chen L., Scheffer N., Liu Y., Shah N., Wanga R., Kumar A., Yih W., Dong X.L. CRAG – Comprehensive RAG Benchmark // arXiv preprint. – 2024. – URL: https://arxiv.org/abs/2406.04744.
3. Li Y., Zhang J., Wu X., Liu C., Qian Z., Jin H. UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis // arXiv preprint. – 2024. – URL: https://arxiv.org/abs/2406.15187.
4. Chen J., Lin H., Han X., Sun L. Benchmarking Large Language Models in Retrieval-Augmented Generation // arXiv preprint. – 2023. – URL: https://arxiv.org/abs/2309.01431.
5. Yandex GPT // Яндекс. – URL: https://ya.ru/ai/gpt-4.
6. GigaChat // Сбер. – URL: https://giga.chat/.