ВИШ
СПБПУ

РАЗРАБОТКА RAG-СИСТЕМЫ ДЛЯ КОНСУЛЬТИРОВАНИЯ ФИЗИКОВ-ТЕОРЕТИКОВ ПО КВАНТОВОЙ ЭЛЕКТРОДИНАМИКЕ

ВЫПОЛНИЛА:
ОРЛОВА В. С.
РУКОВОДИТЕЛЬ:
СЕМЕНДЯЕВ Р. Ю.

САНКТ-ПЕТЕРБУРГ
2026

АКТУАЛЬНОСТЬ

СУЩЕСТВУЮЩИЕ В ОТКРЫТОМ ДОСТУПЕ LLM ХОРОШО СПРАВЛЯЮТСЯ ТОЛЬКО С ОСНОВАМИ, ДАЛЬНЕЙШЕЕ ПОГРУЖЕНИЕ В ЭТУ ОБЛАСТЬ НЕДОСТАТОЧНОЕ ВВИДУ ОТСУТСТВИЯ ИНФОРМАЦИИ В ОТКРЫТОМ ДОСТУПЕ.

ЦЕЛЬ

РАЗРАБОТАТЬ RAG-СИСТЕМУ НА ОСНОВЕ НАУЧНЫХ СТАТЕЙ ПО КВАНТОВОЙ ЭЛЕКТРОДИНАМИКЕ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ ОТВЕТОВ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ПРИ КОНСУЛЬТИРОВАНИИ ФИЗИКОВ-ТЕОРЕТИКОВ.

ОБЛАСТЬ ИССЛЕДОВАНИЯ

УЛУЧШЕНИЕ РАБОТЫ БОЛЬШИХ ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ В СЛУЧАЕ ОТСУТСТВИЯ ПОДХОДЯЩЕГО КОНТЕКСТА ПРИ ОБУЧЕНИИ

ПРЕДМЕТ ИССЛЕДОВАНИЯ

RAG – ЭТО МЕТОД, УЛУЧШАЮЩИЙ ПРОЦЕСС ГЕНЕРАЦИИ ОТВЕТОВ МОДЕЛЬЮ ЗА СЧЕТ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ БАЗЫ ЗНАНИЙ.

ЗАДАЧИ

1. ПОДГОТОВКА ДАННЫХ
2. РЕАЛИЗАЦИЯ RAG-СИСТЕМ
3. АНАЛИЗ RAG-СИСТЕМ
4. РЕАЛИЗАЦИЯ ВЕБ-ИНТЕРФЕЙСА

RAG-СИСТЕМЫ

001

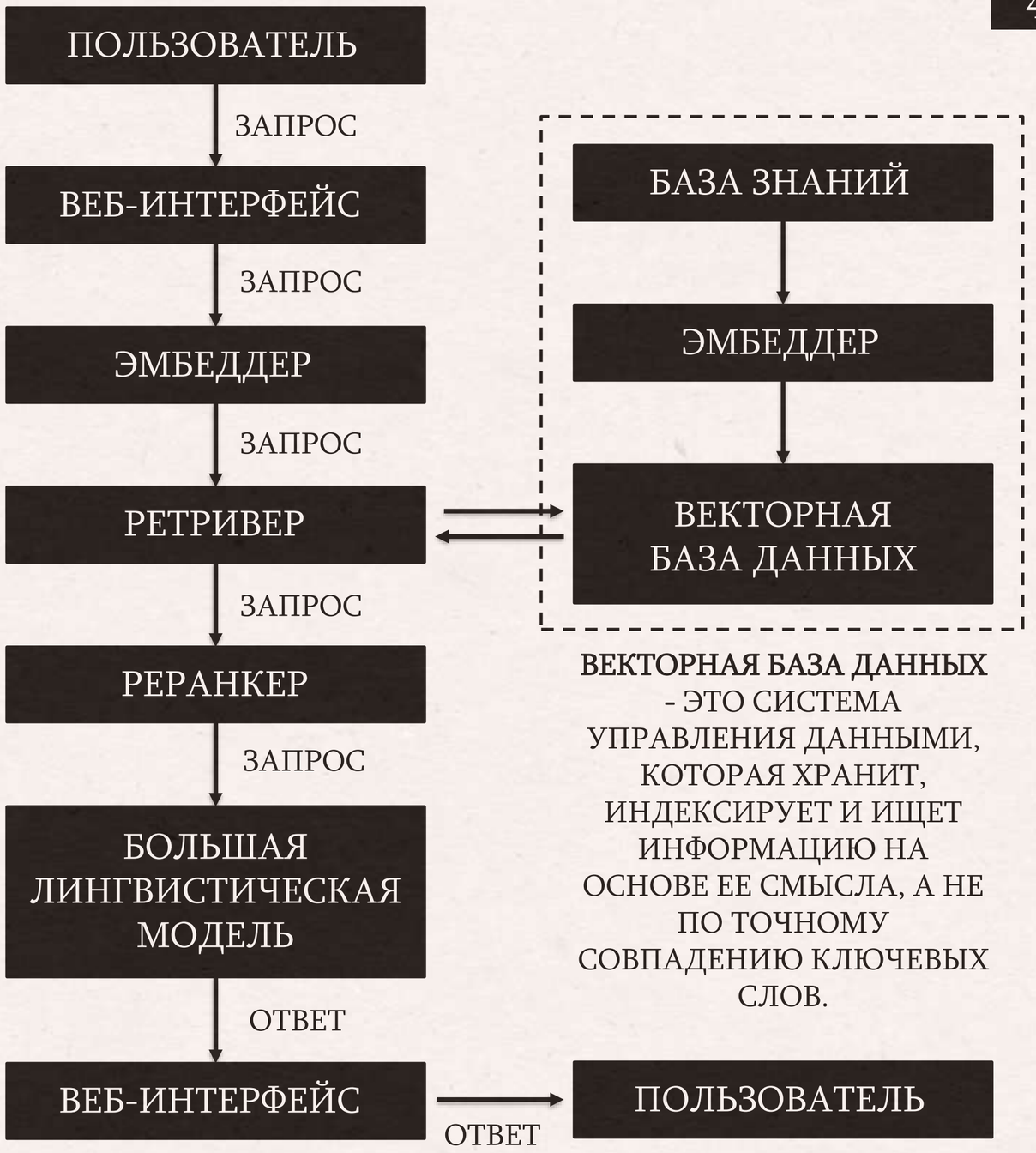
СХЕМА РАБОТЫ RAG-СИСТЕМЫ

ЭМБЕДДЕР – ЭТО НЕЙРОННАЯ СЕТЬ, КОТОРАЯ ПРЕВРАЩАЕТ ЧАСТИ ТЕКСТА (ДОКУМЕНТЫ, СТАТЬИ, ЧАНКИ) И ПОИСКОВЫЕ ЗАПРОСЫ В ЧИСЛОВЫЕ ВЕКТОРЫ (ЭМБЕДДИНГИ).

РЕТРИВЕР– СИСТЕМА, КОТОРАЯ ОТВЕЧАЕТ ЗА ПОИСК И ИЗВЛЕЧЕНИЕ НАИБОЛЕЕ ПОДХОДЯЩЕЙ К ЗАПРОСУ ИНФОРМАЦИИ С ПОМОЩЬЮ ФУНКЦИЙ СХОЖЕСТИ.

РЕРАНКЕР – ЭТО ЭМБЕДДЕР, КОТОРЫЙ ИСПОЛЬЗУЕТСЯ ДЛЯ ТОГО, ЧТОБЫ ПОСТАВИТЬ НА ПЕРВЫЕ МЕСТА БОЛЕЕ РЕЛЕВАНТНЫЕ ДАННЫЕ.

БОЛЬШАЯ ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ – ЭТО ГЛУБОКАЯ НЕЙРОННАЯ СЕТЬ, В КОТОРОЙ ПРИМЕНЯЕТСЯ АРХИТЕКТУРА ТРАНСФОРМЕР С МЕХАНИЗМОМ САМОВНИМАНИЯ.



ВЕКТОРНАЯ БАЗА ДАННЫХ - ЭТО СИСТЕМА УПРАВЛЕНИЯ ДАННЫМИ, КОТОРАЯ ХРАНИТ, ИНДЕКСИРУЕТ И ИЩЕТ ИНФОРМАЦИЮ НА ОСНОВЕ ЕЕ СМЫСЛА, А НЕ ПО ТОЧНОМУ СОВПАДЕНИЮ КЛЮЧЕВЫХ СЛОВ.

ПОДГОТОВКА ДАННЫХ

В КАЧЕСТВЕ ДАННЫХ ИСПОЛЬЗОВАЛИСЬ СТАТЬИ ПО КВАНТОВОЙ ЭЛЕКТРОДИНАМИКЕ.
СТАТЬИ ПРЕДОСТАВЛЕННЫ ПЕТЕРБУРГСКИМ ИНСТИТУТОМ ЯДЕРНОЙ ФИЗИКИ ИМ. Б. П. КОНСТАНТИНОВА (ПИЯФ).

ИНСТРУМЕНТЫ: OCR MISTRAL

О ЧЕМ СТАТЬИ?

СТАТЬИ ОПИСЫВАЮТ ИССЛЕДОВАНИЯ В ОБЛАСТИ ТЕОРЕТИЧЕСКОЙ АТОМНОЙ ФИЗИКИ И КВАНТОВОЙ ЭЛЕКТРОДИНАМИКИ (QED).

КЛЮЧЕВЫЕ ТЕМЫ:

1. РАСЧЕТЫ ОТДАЧИ ЯДРА,
2. ЭФФЕКТ БОРА-ВАЙСКОПФА,
3. G-ФАКТОР СВЯЗАННОГО ЭЛЕКТРОНА,
4. НАРУШЕНИЯ ЧЕТНОСТИ И ИЗОТОПИЧЕСКИХ СДВИГОВ ДЛЯ ПРОВЕРКИ QED В СИЛЬНЫХ ПОЛЯХ.

ВЕБ-ИНТЕРФЕЙС

АНАЛИЗ RAG-СИСТЕМ

002

АНАЛИЗ РАБОТЫ RAG-СИСТЕМЫ

CONTEXT PRECISION
ТОЧНОСТЬ
ПОИСКА

CONTEXT RECALL
ПОЛНОТА
ПОИСКА

МЕТРИКИ ПОИСКА

HALLUTINATION
ПРИСУТСВИЕ
ГАЛЛЮЦИНАЦИЙ

ANSWER RELEVANCE
КАЧЕСТВО
ОТВЕТА

МЕТРИКИ ГЕНЕРАЦИИ

ИНСТРУМЕНТЫ

RAGAS:
CONTEXT PRECISION
CONTEXT RECALL
FAITHFULNESS
ANSWER RELEVANCE

OPIK:
CONTEXT PRECISION
CONTEXT RECALL
HALLUTINATION
ANSWER RELEVANCE

АНАЛИЗ РАБОТЫ RAG-СИСТЕМЫ - ЭТАЛОН

ПРИМЕРЫ ФОРМИРОВАНИЯ ФИНАЛЬНОЙ ОЦЕНКИ ВОПРОСОВ ДЛЯ ЭТАЛОНА:

ВОПРОС:

КАК ВЫЧИСЛЯЮТСЯ ВЕДУЩИЕ ЧЛЕНЫ $\Delta E_C^{(1)}$ И $\Delta E_{TR(1)}^{(1)}$ ДЛЯ ВОДОРОДОПОДОБНЫХ АТОМОВ, И КАКОВА ИХ СУММА $\Delta E^{(1)}$?

ОТВЕТ:

$$\Delta E^{(1)} \equiv \Delta E_C^{(1)} \Delta E_{TR(1)}^{(1)} = \frac{M^2 - \varepsilon_A^2}{2M} = \frac{M^2 (\alpha Z)^2}{M 2 N^2}$$

ГДЕ ε_A — ЭНЕРГИЯ СОСТОЯНИЯ, А N — ЭФФЕКТИВНОЕ ГЛАВНОЕ КВАНТОВОЕ ЧИСЛО.

4 КРИТЕРИЙ СЛОЖНОСТИ ПОИСКА

ОБОСНОВАНИЕ:
В ТЕКСТЕ ОПИСАНЫ ФОРМУЛЫ ДЛЯ $\Delta E_C^{(1)}$ И $\Delta E_{TR(1)}^{(1)}$, НО СУММА $\Delta E^{(1)}$ НЕ УТОЧНЕНА.

3 КРИТЕРИЙ СЛОЖНОСТИ ГЕНЕРАЦИИ

ОБОСНОВАНИЕ:
ТРЕБУЕТСЯ ЗНАНИЕ МЕТОДОВ ВЫЧИСЛЕНИЯ ПЕРВЫХ ПОПРАВОК ДЛЯ ВОДОРОДОПОДОБНЫХ АТОМОВ.

КРИТЕРИИ ОЦЕНКИ ВОПРОСОВ

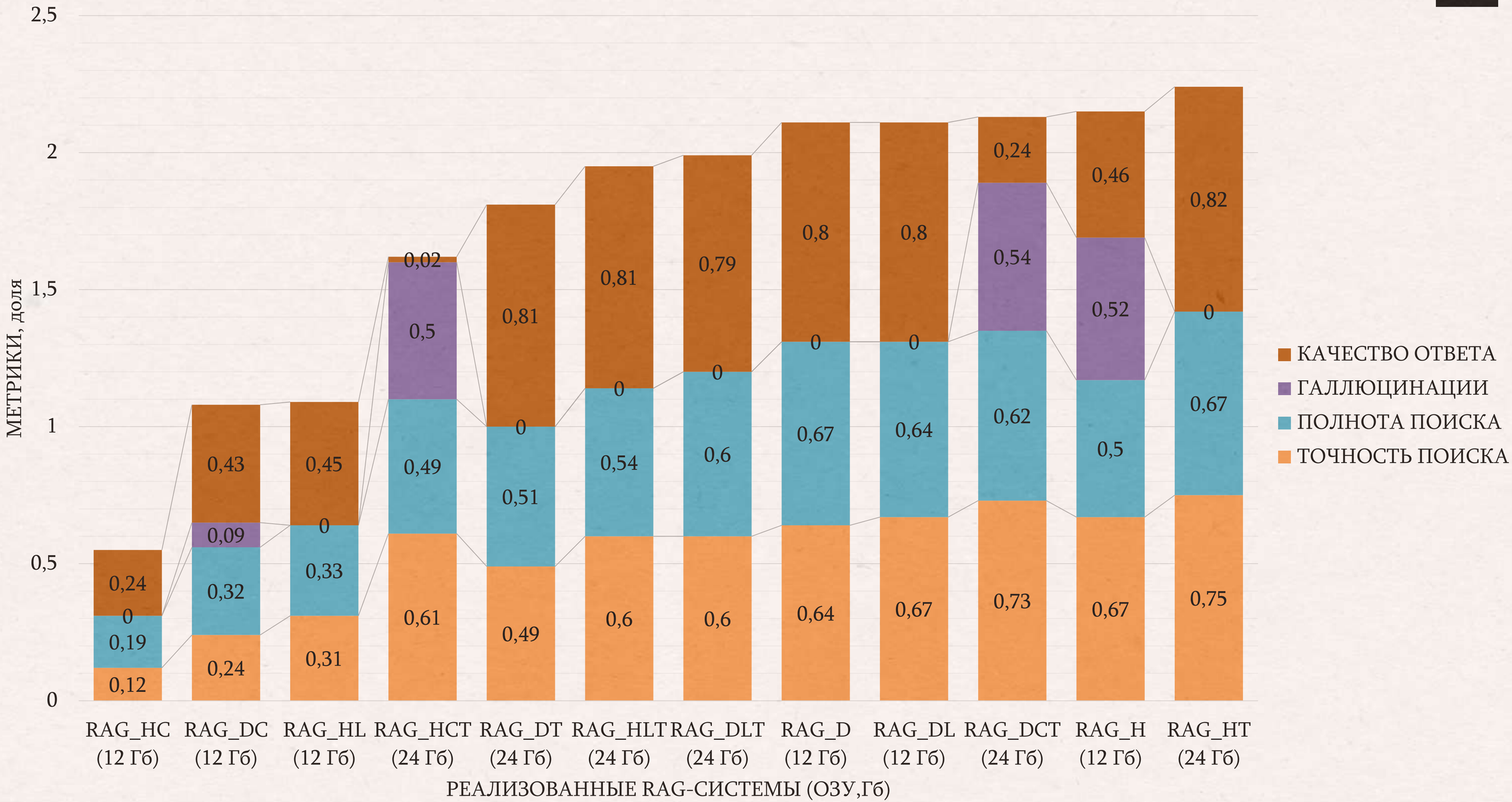
КРИТЕРИЙ 1: СЛОЖНОСТЬ ПОИСКА	
БАЛЛЫ	ОПИСАНИЕ
1	ПРЯМОЕ СОВПАДЕНИЕ
2	СИНОНИМЫ И ПАРАФРАЗ
4	МУЛЬТИДОКУМЕНТАЛЬНЫЙ ПОИСК
5	ПОИСК ПО СКРЫТОМУ КОНТЕКСТУ

КРИТЕРИЙ 2: СЛОЖНОСТЬ ГЕНЕРАЦИИ	
БАЛЛЫ	ОПИСАНИЕ
1	ПРОСТОЕ ИЗВЛЕЧЕНИЕ ФАКТА
2	СЖАТИЕ И СУММАРИЗАЦИЯ
3	ПОНИМАНИЕ ЛОГИКИ И ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ
4	СРАВНЕНИЕ И АНАЛИЗ
5	СИНТЕЗ И ЛОГИЧЕСКИЙ ВЫВОД

АНАЛИЗ РАБОТЫ RAG-СИСТЕМ

СОСТАВЛЯЮЩИЕ RAG-СИСТЕМ						РЕАЛИЗОВАННЫЕ RAG-СИСТЕМЫ											
						1	2	3	4	5	6	7	8	9	10	11	12
№	РОЛЬ	ТИП		МОДЕЛИ	ОЗУ, ГБ	RAG_D	RAG_DT	RAG_H	RAG_HT	RAG_DC	RAG_DCT	RAG_DL	RAG_DLT	RAG_HC	RAG_HCT	RAG_HL	RAG_HLT
1	ПЕРЕВОД	T	LLM	QWEN3-8B	24		✓		✓		✓		✓		✓		✓
2	ЭМБЕДДЕР		SBERT	LABSE	3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	РЕТРИВЕР	D	DENSE	FAISS	0,05	✓	✓			✓	✓	✓	✓				
4		H	HYBRID	FAISS + BM25	0,08			✓	✓					✓	✓	✓	✓
5	РЕРАНКЕР	C	BERT	MS-MARCO-MINILM-L-12-V2	0,3					✓	✓			✓	✓		
6		L	LLM	QWEN3-RERANKER-4B	12							✓	✓			✓	✓
7	ГЕНЕРАТОР		LLM	QWEN3-4B	12	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
АНАЛИЗ РАБОТЫ RAG-СИСТЕМ																	
МЕТРИКИ ПОИСКА																	
1	CONTEXT PRECISION (ТОЧНОСТЬ ПОИСКА)					0,64	0,49	0,67	0,75	0,24	0,73	0,67	0,60	0,12	0,61	0,31	0,60
2	CONTEXT RECALL (ПОЛНОТА ПОИСКА)					0,67	0,51	0,50	0,67	0,32	0,62	0,64	0,60	0,19	0,49	0,33	0,54
МЕТРИКИ ГЕНЕРАЦИИ																	
3	HALLUCINATION (НАЛИЧИЕ ГАЛЛЮЦИНАЦИЙ)					0	0	0,52	0	0,09	0,54	0	0	0	0,50	0	0
4	ANSWER RELEVANCY (КАЧЕСТВО ОТВЕТА)					0,80	0,81	0,46	0,82	0,43	0,24	0,80	0,79	0,24	0,02	0,45	0,81

МЕТРИКИ RAG-СИСТЕМ



ВЫВОДЫ

003

ВЫВОДЫ

В РАБОТЕ БЫЛИ…

ПОДГОТОВЛЕНЫ И ПЕРЕВЕДЕНЫ В ТЕКСТОВЫЙ ФОРМАТ СТАТЬИ ПО КВАНТОВОЙ ЭЛЕКТРОДИНАМИКЕ.

РЕАЛИЗОВАНЫ 12 РАЗНЫХ RAG-СИСТЕМ.

РЕАЛИЗОВАН ВЕБ-ИНТЕРФЕЙС ДЛЯ УДОБНОГО ИСПОЛЬЗОВАНИЯ RAG-СИСТЕМЫ

ПРОАНАЛИЗИРОВАНА РАБОТА RAG-СИСТЕМ С ПОМОЩЬЮ ФРЕЙМВОРКОВ RAGAS И OPIK.

“

ATTENTION IS ALL YOU NEED*

— ASHISH VASWANI,
NOAM SHAZEER,
NIKI PARMAR,
JAKOB USZKOREIT,
LLION JONES,
AIDAN N. GOMEZ,
ŁUKASZ KAISER

*НАЗВАНИЕ СТАТЬИ ПО LLM, В КОТОРОЙ ОПИСЫВАЕТСЯ МЕХАНИЗМ САМОВНИМАНИЯ И ТРАНСФОРМЕРЫ