УДК 004.774:339.138-048.35

167. ОЦЕНКА КАЧЕСТВА RAG-СИСТЕМ (LLM + ПОИСК) В ПРИКЛАДНЫХ ЗАДАЧАХ

Костюкевич А.Ю., Васильева В.А., студенты гр.377901
Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Федосенко В.А. – Кандидат технических наук, доцент

Аннотация. В работе рассматривается задача оценки качества систем генерации ответа с поисковым дополнением в прикладной задаче вопросно-ответного взаимодействия по текстовому корпусу. Исследование направлено на сравнение двух подходов: генерации ответа без внешнего контекста и генерации ответа с предварительным извлечением релевантных документов. В практической части используется набор данных, содержащий вопросы, ответы и подтверждающие документы, что позволяет оценивать не только итоговый ответ, но и качество поиска. Эксперимент показывает, что подключение внешнего текстового контекста повышает точность и полноту ответа по сравнению с генерацией без поиска. Также установлено, что итоговое качество системы в значительной степени зависит от качества извлечения документов, а подача эталонного контекста позволяет определить верхнюю границу качества всей архитектуры. Полученные результаты подтверждают целесообразность применения систем генерации ответа с поисковым дополнением в задачах, где требуется опора на документы и фактическая обоснованность ответа.

Ключевые слова. Генерация ответа с поисковым дополнением, крупная языковая модель, извлечение документов, семантический поиск, лексический поиск, текстовый корпус, вопросно-ответная система, подтверждающий контекст, качество ответа, качество поиска, полнота ответа, точность ответа, векторное представление текста, многошаговое рассуждение, прикладная задача обработки текста.

Крупные языковые модели (далее – LLM) способны формировать связные ответы на естественном языке, однако в прикладных задачах этого недостаточно, поскольку без внешней опоры на документы они могут выдавать фактически неточные или слабо обоснованные утверждения. Одним из основных способов уменьшения этой проблемы является подход retrieval-augmented generation (далее – RAG), при котором система сначала извлекает релевантные текстовые фрагменты из внешнего корпуса, а затем использует их при генерации ответа. Такой подход применяется в вопросно-ответных системах, корпоративных помощниках, поисковых чат-ботах и других системах, где важна не только связность текста, но и его фактическая точность [1].

В данной работе проверяются две гипотезы. Первая гипотеза состоит в том, что RAG-система обеспечивает более высокое качество ответов, чем режим LLM-only, то есть генерация без внешнего контекста. Вторая гипотеза состоит в том, что итоговое качество RAG определяется не только самой языковой моделью, но и качеством этапа поиска, поэтому при подаче «идеального» контекста результат должен заметно улучшаться. Одновременно проверяются два тезиса: во-первых, RAG необходимо оценивать не только по качеству ответа, но и по качеству найденных документов; во-вторых, улучшение модуля поиска может давать существенный прирост качества даже без замены генератора [1], [2], [6].

Для корректного понимания дальнейших результатов необходимо кратко определить используемые термины. Под retrieval понимается этап поиска релевантных документов или фрагментов по пользовательскому запросу. Метод BM25 представляет собой классический лексический алгоритм ранжирования документов по совпадению слов запроса и текста документа; он особенно полезен там, где важны точные словесные совпадения [3]. Под dense retrieval понимается поиск не по словам, а по векторным представлениям текста, то есть по его смысловой близости. Для ускорения такого поиска обычно используется библиотека Facebook AI Similarity Search (далее – FAISS), предназначенная для быстрого поиска ближайших соседей в векторном пространстве [4]. Под groundedness в работе понимается степень опоры ответа на найденный контекст, то есть насколько сформулированный ответ подтверждается извлечёнными фрагментами. В современных работах по оценке RAG groundedness рассматривается как один из ключевых аспектов качества наряду с точностью ответа и качеством поиска [6].

Практическая часть исследования была выполнена в среде Kaggle Notebook на датасете HotpotQA, который предназначен для многошаговой задачи вопрос–ответ и содержит не только вопросы и эталонные ответы, но и supporting facts, то есть документы, на которые должен опираться правильный ответ. Именно поэтому HotpotQA удобен для сравнения обычной генерации и RAG-подхода: он позволяет оценивать как качество ответа, так и качество поиска релевантного контекста

[2]. В эксперименте использовалась подвыборка из 150 примеров, по которой был сформирован корпус из 1491 документа. В качестве генератора применялась instruction-tuned модель FLAN-T5-base, а в качестве плотностного энкодера – sentence-transformers/all-MiniLM-L6-v2, предназначенная для семантического поиска и сопоставления текстов [5], [7].

В эксперименте сравнивались четыре режима. Первый режим – LLM-only, при котором модель получала только вопрос без внешнего контекста. Второй режим – RAG-BM25, где в модель подавались три документа, найденные методом BM25. Третий режим – RAG-dense, где три документа извлекались с помощью dense retrieval и индекса FAISS. Четвёртый режим – Oracle-RAG, в котором в качестве контекста использовались эталонные supporting-документы из HotpotQA. Этот режим нужен не как реальный сценарий работы системы, а как верхняя граница качества: если ответ при идеальном контексте заметно лучше, значит основное ограничение находится именно на этапе поиска [2], [3], [4].

Для оценки результата использовались метрики exact match (далее – EM) и F1-мера (далее – F1), традиционно применяемые в HotpotQA. Кроме этого, для RAG-режимов оценивались метрики поиска Hit@3 и support-title recall, а также проверялось, содержится ли правильный ответ в найденном контексте. Такой набор показателей позволяет оценить систему сразу на двух уровнях: сначала выяснить, хорошо ли она нашла подтверждающие документы, а затем проверить, смогла ли она построить на их основе корректный ответ [2], [6].

```
Количество примеров: 150
Пример вопроса: Were Scott Derrickson and Ed Wood of the same nationality?
Пример ответа: yes
Поддерживающие документы: ['Scott Derrickson', 'Ed Wood']
```

Рисунок 1 – Пример записи датасета HotpotQA и supporting-документов

```
Размер корпуса: 1491
Title: Ed Wood (film)
Ed Wood is a 1994 American biographical period comedy-drama film directed and produced by Tim Burton, and star
Wood's life when he made his best-known films as well as his relationship with actor Bela Lugosi, played by Ma
rie, and Bill Murray are among the supporting cast.
```

Рисунок 2 – Размер сформированного корпуса: 1491 документа

Полученные результаты показали, что оба варианта RAG заметно превосходят режим LLM-only. У режима LLM-only значение EM составило 0,107, а F1 – 0,134. У режима RAG-BM25 значения выросли до 0,300 и 0,393 соответственно. У режима RAG-dense значения составили 0,227 и 0,343. Наилучший результат показал Oracle-RAG: EM = 0,553 и F1 = 0,680. Тем самым первая гипотеза подтверждается: использование внешнего контекста действительно улучшает качество ответа. Вторая гипотеза также подтверждается, поскольку разрыв между обычными RAG-режимами и Oracle-RAG показывает, что главный резерв улучшения лежит в повышении качества поиска.

Таблица 1 – Сравнение режимов LLM-only и RAG по основным метрикам

Система	**EM**	**F1**	**Semantic similarity**	**Hit@3**	**Support-title recall**	**Gold answer in context**	**Pred answer in context**
Oracle-RAG	0,553	0,680	0,785	1,000	1,000	0,940	0,880
RAG-BM25	0,300	0,393	0,536	0,913	0,623	0,580	0,740
RAG-dense	0,227	0,343	0,498	0,933	0,670	0,607	0,687
LLM-only	0,107	0,134	0,313	–	–	–	–

Примечание **–** Для режима LLM-only retrieval-метрики не вычисляются, поскольку внешний контекст в нём отсутствует.

Из таблицы видно, что в рамках данного эксперимента режим RAG-BM25 оказался лучше RAG-dense по метрикам итогового ответа, хотя dense retrieval немного превосходит BM25 по одному из

показателей полноты поиска. Это означает, что в рассматриваемой постановке задачи точные лексические совпадения оказались полезнее чисто семантической близости. Такой результат важен практически: он показывает, что более «сложный» поиск не всегда автоматически означает более качественный ответ. При одинаковом генераторе качество RAG определяется тем, насколько подходящий контекст получает модель на вход.

	system	em	f1	semantic_similarity	hit_at_k	support_title_recall	gold_answer_in_context	pred_answer_in_context
3	rag_oracle	0.553333	0.680170	0.784782	1.000000	1.000000	0.940000	0.880000
1	rag_bm25	0.300000	0.392752	0.536481	0.913333	0.623333	0.580000	0.740000
2	rag_dense	0.226667	0.343273	0.498359	0.933333	0.670000	0.606667	0.686667
0	llm_only	0.106667	0.133833	0.312687	NaN	NaN	NaN	NaN

Рисунок 3 – Итоговая сводка метрик по всем режимам

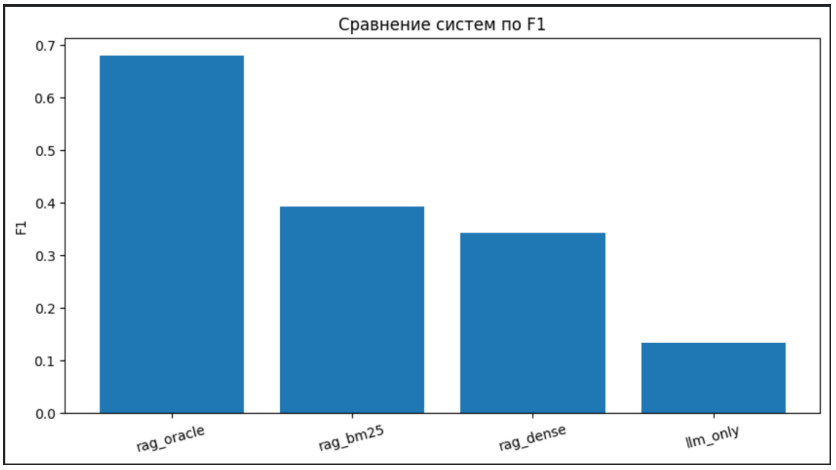

Рисунок 4 – Сравнение систем по метрике F1

Дополнительный анализ примеров ошибок также подтверждает различие между подходами. В режиме LLM-only модель часто отвечала «I don’t know» даже там, где ответ присутствовал в корпусе. В режимах RAG система чаще находила правильное направление ответа, однако при неидеальном retrieval могла извлекать нерелевантные или частично релевантные документы, из-за чего итоговая генерация всё равно оказывалась ошибочной. В режиме Oracle-RAG многие такие ошибки исчезали, поскольку модель получала корректный supporting-контекст. Следовательно, эксперимент показывает, что в прикладной системе вопрос–ответ недостаточно иметь только сильную языковую модель; необходимо отдельно проектировать и оценивать модуль поиска.

Худшие примеры для LLM-only

	question	gold_answer	prediction	retrieved_titles	em	f1	semantic_similarity
0	Were Scott Derrickson and Ed Wood of the same ...	yes	I don't know.	[]	0	0.0	0.188838
4	What government position was held by the woman...	Chief of Protocol	I don't know	[]	0	0.0	0.227374
8	What science fantasy young adult series, told ...	Animorphs	I don't know	[]	0	0.0	0.209875
20	2014 S/S is the debut album of a South Korean ...	YG Entertainment	I don't know	[]	0	0.0	0.168420
28	The arena where the Lewiston Maineiacs played ...	3,677 seated	I don't know	[]	0	0.0	0.003740

Худшие примеры для RAG-Dense

	question	gold_answer	prediction	retrieved_titles	em	f1	semantic_similarity
6	What government position was held by the woman...	Chief of Protocol	17-year-old	[Kiss and Tell (1945 film), A Kiss for Corliss...	0	0.0	0.071910
10	What science fantasy young adult series, told ...	Animorphs	Hamee.	[Animorphs, Victoria Hanley, The Hork-Bajir Ch...	0	0.0	0.090021
14	Are the Laleli Mosque and Esma Sultan Mansion ...	no	yes	[Esma Sultan Mansion, Laleli Mosque, Sultan Ah...	0	0.0	0.733495
22	2014 S/S is the debut album of a South Korean ...	YG Entertainment	WINNER.	[2014 S/S, BTS discography, Seventeen discogra...	0	0.0	0.209723
30	The arena where the Lewiston Maineiacs played ...	3,677 seated	4,000.	[Lewiston Maineiacs, Dwyer Arena, Androscoggin...	0	0.0	0.091691

Рисунок 5 – Примеры худших ответов для LLM-only и RAG-Dense

Таким образом, проведённое исследование подтверждает, что RAG-подход лучше обычной генерации без внешнего контекста в задачах, где ответ должен опираться на документы. Одновременно показано, что итоговый выигрыш RAG ограничивается качеством retrieval: при идеальном контексте модель отвечает значительно лучше, чем при обычном поиске. Практический смысл этого вывода состоит в том, что при разработке прикладных RAG-систем основной эффект часто достигается не заменой генератора, а улучшением поиска, фильтрации и компоновки контекста. Следовательно, для полноценной оценки таких систем необходимо одновременно анализировать качество ответа, качество найденных документов и степень опоры ответа на контекст [1], [2], [6].

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 Natural Language Processing Tasks [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2005.11401. – Дата доступа: 03.04.2026.

2. 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 Natural Language Processing Tasks [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2005.11401. – Дата доступа: 03.04.2026.

3. Robertson, S., Zaragoza, H. The Probabilistic Relevance Framework: Best Matching Twenty-Five and Beyond [Электронный ресурс]. – Режим доступа: https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf. – Дата доступа: 03.04.2026.

4. Johnson, J., Douze, M., Jégou, H. Billion-scale Similarity Search with Graphics Processing Units [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1702.08734. – Дата доступа: 03.04.2026.

5. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. Scaling Instruction-Finetuned Language Models [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2210.11416. – Дата доступа: 03.04.2026.

6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1910.10683. – Дата доступа: 03.04.2026.

7. Reimers, N., Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1908.10084. – Дата доступа: 03.04.2026.

8. Es, S., James, J., Espinosa Anke, L., Schockaert, S. Ragas: Automated Evaluation of Retrieval Augmented Generation [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2309.15217. – Дата доступа: 03.04.2026.

9. Es, S., James, J., Espinosa Anke, L., Schockaert, S. Ragas: Automated Evaluation of Retrieval Augmented Generation [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2309.15217. – Дата доступа: 03.04.2026.

10. Es, S., James, J., Espinosa Anke, L., Schockaert, S. Ragas: Automated Evaluation of Retrieval Augmented Generation [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2309.15217. – Дата доступа: 03.04.2026.

UDC 004.774:339.138-048.35

EVALUATION OF RAG SYSTEMS (LLM + SEARCH) IN APPLIED TASKS

Kostyukevich A.Y, Vasilyeva V.A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Fedosenko V.A. – PhD in Economics

Annotation. The paper considers the problem of evaluating the quality of answer generation systems with retrieval augmentation in an applied question answering task over a text corpus. The study is aimed at comparing two approaches: answer generation without external context and answer generation with preliminary retrieval of relevant documents. The practical part uses a dataset that contains questions, answers, and supporting documents, which makes it possible to evaluate not only the final answer but also the quality of retrieval. The experiment shows that the use of external textual context improves answer accuracy and completeness in comparison with generation without retrieval. It is also established that the overall quality of the system strongly depends on the quality of document retrieval, while the use of reference context makes it possible to determine the upper quality bound of the whole architecture. The obtained results confirm the usefulness of answer generation systems with retrieval augmentation in tasks that require document grounding and factual reliability.

Keywords. retrieval augmented generation, large language model, document retrieval, semantic search, lexical search, text corpus, question answering system, supporting context, answer quality, retrieval quality, answer completeness, answer accuracy, vector representation of text, multi-hop reasoning, applied text processing task.