

Методы тестирования порождающих моделей, использующих информационный поиск

Дарья Потапова¹, Наталья Лукашевич²

¹ МГУ им. М.В. Ломоносова, Москва, Россия
`darya.potapova03@yandex.ru`

² Научно-исследовательский вычислительный центр МГУ, Москва, Россия
`louk_nat@mail.ru`

Аннотация Применение порождающих моделей, использующих информационный поиск, или подхода Retrieval Augmented Generation (RAG), когда поисковый модуль извлекает релевантные документы по запросу пользователя, а порождающая модель формирует ответ, комбинируя найденный контекст со своими внутренними знаниями, позволяет преодолеть такие ограничения больших языковых моделей, как устаревание знаний и неспособность работать с динамично изменяемыми данными. Эффективность таких систем напрямую зависит от качества поиска, структуры контекста и формулировок инструкций, однако всесторонняя оценка такого подхода на текущий момент является сложной задачей не только из-за малого количества подходящих наборов данных и различия в требованиях к системам, но и из-за отсутствия стандартных метрик, которые позволяли бы измерять соответствие ответа предоставленному контексту. В работе проводилось сравнительное исследование шести языковых моделей (DeepSeek, Llama, Mistral, Qwen, RuAdapt, YandexGPT) в вопросно-ответной задаче с использованием подхода RAG. Эксперименты выполнены на четырех русскоязычных датасетах — XQuAD, TyDi QA, RuBQ и SberQuAD — приведённых к единому формату, подходящему под задачу. Были рассмотрены различные стратегии добавления и упорядочивания фрагментов в подаваемом модели контексте, а оценка моделей с помощью метрик Context Relevance, Utilisation, Completeness, Adherence и Exact Matching позволила выявить ограничения генеративных моделей при извлечении информации из релевантного контекста.

Ключевые слова: порождающая модель · информационный поиск
· бенчмарк · retrieval augmented generation

1 Введение

Современные большие языковые модели демонстрируют значительный прогресс в задачах обработки естественного языка. Однако, несмотря на все достижения, модели сталкиваются с рядом проблем: например, они подвержены галлюцинациям, т.е. формированию недостоверных ответов, и устареванию внутренних знаний.

Для решения этих проблем все большую популярность приобретают системы, основанные на подходе Retrieval Augmented Generation [1], или системы, использующие информационный поиск. Они сочетают извлечение релевантной информации с ее использованием в генеративной модели для формирования более точных и обоснованных ответов. Наиболее распространенным подходом является использование поисковых систем для получения внешнего контекста, что позволяет использовать большие объемы актуальной информации.

Тем не менее, подобные системы все еще недостаточно хорошо справляются с работой со сформированным контекстом, например, с выбором наиболее релевантной информации или объединением данных для ответа на сложные вопросы. Для их объективной оценки требуются специализированные наборы данных, или бенчмарки, которые позволяют измерить точность и надежность RAG в различных сценариях. Существует несколько исследований [2,3,4,5] на англоязычных наборах с различными требованиями и метриками, но системы, которые порождают в качестве ответа текст на русском языке, все еще нуждаются в дополнительном анализе.

В рамках работы были найдены и объединены в бенчмарк подходящие под формат задачи вопросно-ответные датасеты на русском языке. Полученный бенчмарк дает возможность проведения экспериментов технологий RAG для русскоязычных данных.

2 Обзор существующих подходов

Кратко опишем существующие англоязычные наборы данных и подходы к оценке RAG-систем.

Работа RGB [2] фокусируется на оценке систем по четырем критериям: устойчивость к зашумленным данным, умение отказываться от ответа, объединение информации из нескольких источников и устойчивость к фактическим ошибкам. Для оценки использовались новостные данные на английском и китайском языках и метрика Ассигасу, отвечающая за долю правильных ответов. Результаты показали, что подход улучшает ответы языковых моделей, однако рост шума снижает точность, а объединение данных и отказ от ответа остаются слабыми местами моделей.

Бенчмарк RAGTruth [3] ориентирован на обнаружение галлюцинаций в трех задачах, которые могут решаться с помощью подхода RAG: реферирования, вопросно-ответная и генерации текста. Данные были взяты из существующих датасетов, соответствующих каждой задаче, ответы к ним генерировались 6 языковыми моделями с последующей ручной проверкой. Для оценки использовались метрики точности (Precision), полноты (Recall) и F-мера. Проведенные эксперименты выявили низкую эффективность современных подходов, особенно в локализации ошибок внутри порожденного ответа.

Набор данных CRAG [4] заявлен как универсальный бенчмарк с требованиями к реализму систем, разнообразию решаемых задач, наглядности

результатов, надежности ответов и долговечности данных. Включает три уровня задач: реферирование на основе поиска, работа с графами знаний и End-to-End RAG. Предложено классифицировать ответы как точные/отсутствующие/некорректные с присваиванием соответствующих числовых меток, а показатель достоверности (Truthfulness) рассчитывать как среднее по всем ответам. Лучшие решения при исследовании достигали лишь 60.5% достоверности, что указывало на проблемы в работе моделей с зашумленными данными.

Авторы RAGBench [5] указали на отсутствие унифицированного подхода к оценке в исследованиях на бенчмарках RGB, RAGTruth и CRAG и предложили использовать для оценки фреймворк TRACe с метриками релевантности (Relevance), использования (Utilisation), полноты (Completeness) и подтвержденности (Adherence). В качестве данных использовались 12 датасетов для различных предметных областей. Эксперименты в очередной раз выявили сложность объективного измерения релевантности, что подчеркнуло необходимость стандартизированных инструментов тестирования.

3 Метрики для оценки систем RAG

Выбор метрик для оценки качества моделей, использующих информационный поиск, является нетривиальной задачей по следующим причинам:

1. сохраняются проблемы оценки качества порожденного текста;
2. возникает проблема оценки следования контексту при порождении ответа;
3. из-за отсутствия единых стандартов каждое новое исследование предлагает свой набор метрик, охватывающий только их конкретные требования — отсутствует универсальный подход к оценке, который учитывал бы все аспекты.

Для оценки работы моделей были использованы метрики Context Relevance, Context Utilisation, Completeness и Adherence [5], а также метрика Exact Matching.

Определим и интерпретируем эти метрики. Здесь

- D — множество контекстных документов $\{d_1, \dots, d_n\}$;
- R_i — множество токенов в документе d_i , содержащих информацию, релевантную для ответа на запрос;
- U_i — множество токенов из документа d_i , реально использованных порождающей моделью для формирования ответа;
- $Len(x)$ — количество токенов.

Context Relevance. Определяется как доля релевантного контекста по отношению ко всему извлеченному к запросу. Низкая релевантность указывает на избыточность результатов поиска, что может привести к понижению точности ответа.

$$Context\ Relevance = \frac{\sum Len(R_i)}{\sum Len(d_i)} \quad (1)$$

Множество значений лежит на отрезке $[0, 1]$, где 0 — контекст не является релевантным к запросу, 1 — весь контекст является релевантным к запросу.

Context Utilisation. Определяется как доля извлеченного контекста, использованного моделью для порождения ответа. Низкая доля использования в сочетании с низкой релевантностью говорит о том, что поисковый модуль использует жадный алгоритм, а из низкой доли использования при высокой релевантности следует, что генеративная модель не способна эффективно использовать предоставленный контекст.

$$\text{Context Utilisation} = \frac{\sum \text{Len}(U_i)}{\sum \text{Len}(d_i)} \quad (2)$$

Множество значений лежит на отрезке $[0, 1]$, где 0 — извлеченный контекст не был использован для порождения ответа, 1 — весь контекст был использован для порождения ответа.

Completeness. Определяется как доля извлеченного релевантного к запросу контекста, использованного моделью для порождения ответа. Близость к 0 говорит о том, что использованный для порождения ответа контекст являлся нерелевантным, т.е. даже при высокой доле использования и высокой доле релевантности метрика Completeness может иметь низкое значение. Значение 1 достигается, когда весь релевантный контекст был полностью использован для порождения ответа.

$$\text{Completeness} = \frac{\sum \text{Len}(R_i \cap U_i)}{\sum \text{Len}(R_i)} \quad (3)$$

Множество значений лежит на отрезке $[0, 1]$.

Adherence. Является индикатором следования моделью строгой инструкции о том, что для генерации ответа нужно использовать только предоставленный контекст. 1 может быть получена при отсутствии парафраза в порожденном ответе; иными словами, модель должна отвечать только теми словами, которые использовались в извлеченных документах. Принимает значения из множества $\{0, 1\}$.

$$\text{Adherence} = [\forall a_i \in A \implies a_i \in D] \quad (4)$$

Exact Matching. Является индикатором наличия ожидаемого ответа в порожденном моделью ответе. При этом не учитывается семантика порожденного ответа, т.е. ожидаемый ответ может быть упомянут в порожденном тексте, но не являться, по мнению модели, реальным ответом на вопрос. Метрика является показательной только при наличии инструкции, что модель должна ответить кратко и не должна распространять свой ответ, включая в него дополнительную информацию, и при условии, что модель следует этой инструкции. Принимает значения из множества $\{0, 1\}$.

4 Данные и модели

4.1 Общее описание датасетов

В работе использовались следующие вопросно-ответные датасеты, приведенные к единому формату: русскоязычные части датасетов XQuAD (1190 вопросов) [6] и TyDi QA (6490 вопросов) [7], датасет RuBQ (2910 вопросов) [8] и датасет SberQuAD (50364 вопроса) [9].

Преимущество использования именно вопросно-ответных датасетов заключается в том, что на каждый вопрос из такого датасета можно дать точный ответ, который легко можно проверить на корректность. Основное требование к датасетам — наличие трех компонентов: вопроса, ответа и документов, которые можно использовать в качестве контекста, подаваемого генеративной модели для формирования ответа на заданный вопрос.

4.2 Преобразование данных

Предварительно все документы были очищены от специальных символов кодировки и диакритических знаков, документ контекста был разбит на фрагменты-предложения. Все полученные данные были преобразованы к единому формату:

- id — уникальный номер вопроса,
- вопрос,
- ответ или набор ответов — берутся из исходных датасетов,
- ответы, приведенные к своей нормальной форме,
- контекст к вопросу,
- метаданные.

Контекст представляет собой список предложений, каждое из которых имеет метку `true` или `false` — соответственно, является ли предложение релевантным для вопроса или нет.

Метаданные содержат информацию о возможности ответить на вопрос по заданному контексту и о типе вопроса (при наличии такового в исходных датасетах). Наличие подобной дополнительной информации позволяет анализировать, с какими типами вопросов система справляется лучше или хуже и насколько хорошо модель работает с вопросами, ответы на которые не содержатся в контекстных документах.

Для всех датасетов было расширено множество некоторых ответов: в тех случаях, когда ответ содержал дату в формате «ЧЧ месяц ГГГГ», во множество ответов был добавлен ответ с датой, преобразованной к виду «ЧЧ.ММ.ГГГГ».

Дополнительно для датасетов TyDi QA и XQuAD множество ответов расширялось с помощью языковой модели Llama-3.1-8B-Instruct. По заданному промпту модель должна была породить близкие по смыслу к исходному ответу последовательности. Полученные таким образом дополнительные ответы были проверены вручную. Так, для датасета TyDi QA было добавлено

4334 синонима, из них 1468 «удачных» (33%), для XQuAD — 2207 синонимов, из которых «удачных» — 541 (25%).

Например,

- «умением представить товар» → {«умением представить товар», «умением презентовать товар»};
- «не получить тюремного срока» → {«не получить тюремного срока», «избежать тюремного заключения», «избежать тюремного срока»};
- «Польской объединенной рабочей партии» → {«Польской объединенной рабочей партии», «Польской объединенной рабочей партии (ПОПР)», «ПОПР»} — аббревиатура встречается в подаваемом контексте и может считаться правильным ответом.

Кроме этого, эти датасеты проверялись на корректность ответа относительно вопроса. Для этого к каждому вопросу из этих наборов данных с помощью языковых моделей были сгенерированы ответы с указанием использовать для формирования ответа данный контекст, после чего полученные ответы сравнивались с исходными, и в случае несовпадения изначальные ответы исправлялись. В качестве языковых моделей использовались Llama-3.1-8B-Instruct и Mistral-7B-Instruct-v0.3.

Полученные преобразованные данные опубликованы на платформе HuggingFace.co¹.

4.3 Модели

Для сравнения возможностей были выбраны компактные модели размером 7-8 млрд параметров (в скобках — максимальное число токенов на входе):

- **Международные**
 - DeepSeek-LLM-7B-Chat [10] (4 096 токенов)
 - Meta-Llama-3.1-8B-Instruct [11] (131 072 токена)
 - Mistral-7B-Instruct-v0.3 [12] (32 768 токенов)
 - Qwen2.5-7B-Instruct [14,15] (32 768 токенов)
- **Русскоязычные**
 - RuadaptQwen2.5-7B-Lite-Beta [13] (32 768 токенов)
 - YandexGPT-5-Lite-8B-instruct [16,17] (32 768 токенов)

Для всех моделей установлена температура 0.7 и применена 8-битная квантизация.

Выбор небольших моделей мотивирован желанием оценить, насколько эффективно может работать рассматриваемый подход даже при ограниченных вычислительных ресурсах. Дальнейший анализ и подтверждение факта, что добавление релевантного контекста дает заметный прирост качества даже для маленьких моделей, во-первых, позволяет обосновать использование RAG-подхода в реальных условиях с ограниченным доступом

¹ <https://huggingface.co/collections/bearberry/rusrag-benchmark-6830da5b51ec4cbbfe0c0f63>

к вычислительным мощностям, во-вторых, создает основу для дальнейшего масштабирования: с увеличением вычислительных ресурсов более мощные модели смогут использовать те же принципы, обеспечивая еще более высокое качество генерации.

5 Эксперименты

5.1 Общее описание

Целью экспериментов, в первую очередь, является нахождение наиболее успешного подхода с генеративной частью системы: так как выбранные данные уже содержат в себе весь необходимый для генерации ответа контекст, можно считать, что работа проводится с идеальной поисковой системой. Таким образом, проводится следующий набор экспериментов:

1. Замер результатов моделей при LLM Only подходе, т.е. решение вопросно-ответной задачи путем порождения ответа на основании внутренних знаний модели о мире. В данной серии экспериментов рассчитывается только метрика Exact Matching;
2. Замер результатов моделей при подаче ей в качестве контекста только релевантных, т.е. содержащих ответ на вопрос, фрагментов. В данной серии экспериментов рассчитываются все введенные ранее метрики;
3. Замер результатов моделей при подаче ей в качестве контекста всех относящихся к очередному вопросу фрагментов. В данной серии экспериментов рассчитываются все введенные ранее метрики.

В рамках тех экспериментов, где в качестве контекста подаются все имеющиеся фрагменты, дополнительно проводятся исследования влияния положения релевантных фрагментов внутри текста:

- Релевантные фрагменты находятся в своем естественном положении внутри текста, последовательность повествования не нарушается. Для более общей реализации, когда необходимые N фрагментов выбираются поисковым модулем, при таком подходе выбранные части не будут отсортированы в порядке возрастания/убывания показателя близости к запросу, а будут сохранять тот же порядок, в каком они встречались в документе;
- Релевантные фрагменты принудительно подаются первыми в списке всех фрагментов, что равносильно упорядочиванию по убыванию значения близости после выбора N фрагментов из исходных документов;
- Релевантные фрагменты принудительно подаются в конце списка всех фрагментов, что равносильно упорядочиванию по возрастанию значения близости после выбора N фрагментов из исходных документов.

Из датасета SberQuAD для проведения экспериментов взята выборка размером в 10 000 вопросов.

С результатами для метрики Context Relevance, на значение которой не влияет порожденный ответ, можно ознакомиться в Таблице 1. Среднее значение этой метрики для датасетов XQuAD, TyDi и SberQuAD совпадает

по моделям, но различается по типам экспериментов в зависимости от того, подавался ли только релевантный контекст или полный в любом из его форматов. Низкие значения доли релевантного контекста указывают, что для значительной части вопросов в датасете не существует корректных ответов в исходных данных.

Существенное различие наблюдается только для датасета RuBQ: фрагменты, из которых состоит контекст, длиннее, чем в других рассматриваемых датасетах, соответственно, полный контекст получается объемным и не может быть целиком передан модели DeepSeek, у которой контекстное окно равно 4096 токенам, поэтому он обрезался до требуемого количества токенов. Здесь можно наблюдать влияние положения релевантного контекста на метрику:

- Естественный порядок предложений: попадание релевантных фрагментов напрямую зависит от того, в какой части исходного документа они изначально расположены;
- Все релевантные фрагменты в начале контекста: в контекст попадает меньше нерелевантных документов, возможно, не попадает ни один;
- Все релевантные фрагменты в конце сформированного контекста: в контекст может не попасть ни один релевантный фрагмент, если текущий вопрос датасета содержит много длинных нерелевантных частей.

Таблица 1. Context Relevance по датасетам и моделям

Датасет	Модель	Релевантный контекст	Полный контекст	Начиная с релевантного	Начиная с нерелевантного
XQuAD	Все модели	0.904	0.239	0.239	0.239
TyDi	Все модели	0.922	0.396	0.396	0.396
RuBQ	DeepSeek	0.565	0.054	0.066	0.032
	Остальные модели	0.565	0.054	0.054	0.054
SberQuAD	Все модели	0.919	0.247	0.247	0.247

5.2 Анализ подходов с добавлением контекста

В Таблице 2 приведены результаты для метрики Exact Matching, которая оценивает долю порождений, совпадающих с эталонными ответами. Значения показывают, что все модели без контекста практически не могут найти правильный ответ на заданные вопросы.

Для всех датасетов при добавлении дополнительного контекста наблюдается значительный прирост метрики: максимальное увеличение при использовании только релевантного контекста относительно отсутствия контекста происходит для датасета XQuAD и модели YandexGPT, как и для случая использования полного контекста относительно отсутствия контекста. При этом ни в одном случае точность не превышает даже 0.8: модели все еще

недостаточно хорошо работают с поданным контекстом и, хотя ожидаемый ответ уже содержится в нем, модели не всегда способны извлечь его из документа.

Таблица 2. Сравнение Exact Matching по всем датасетам и моделям без упорядочивания

Датасет	Модель	Без контекста	Релевантный контекст	Полный контекст
XQuAD	DeepSeek	0.050	0.228	0.290
	Llama	0.099	0.650	0.674
	Mistral	0.086	0.705	0.670
	RuAdapt	0.097	0.639	0.628
	Qwen	0.099	0.556	0.541
	YandexGPT	0.110	0.766	0.784
TyDi	DeepSeek	0.087	0.530	0.336
	Llama	0.165	0.503	0.509
	Mistral	0.142	0.577	0.507
	RuAdapt	0.171	0.550	0.542
	Qwen	0.113	0.585	0.502
	YandexGPT	0.228	0.644	0.655
RuBQ	DeepSeek	0.178	0.493	0.314
	Llama	0.381	0.548	0.540
	Mistral	0.318	0.537	0.525
	RuAdapt	0.402	0.582	0.581
	Qwen	0.260	0.564	0.511
	YandexGPT	0.557	0.597	0.638
SberQuAD	DeepSeek	0.023	0.193	0.038
	Llama	0.046	0.278	0.320
	Mistral	0.049	0.424	0.374
	RuAdapt	0.056	0.351	0.360
	Qwen	0.039	0.447	0.449
	YandexGPT	0.075	0.667	0.735

При этом не прослеживается общая для всех моделей зависимость изменения результатов при добавлении полного контекста относительно релевантного. Для модели Mistral, например, использование только релевантного контекста всегда лучше, чем полного, но для Llama, в основном, отбрасывание нерелевантного уменьшает метрику, как и для YandexGPT, который стабильно демонстрирует лучший результат для полного контекста.

Для оценки строгости метрики Exact Matching мы дополнительно проанализировали случаи, когда ответ не сопоставился по этой метрике и при этом наблюдалась высокая косинусная близость ответа к эталонам (от 0.98). Таких примеров оказалось около 0.1% от всех предсказаний; ручная проверка показала, что 76% из них фактически корректны. Наиболее частая причина — перестановка слов при сохранении смысла. Таким образом, доля ложноотрицательных ответов, возникающих из-за вариаций формулировки внутри порождения, невелика и не влияет на сравнительные выводы.

Анализ изменения метрик из Таблицы 3 позволяет сделать следующие выводы:

- Метрика Completeness нелинейно зависит от подхода: в половине случаев модели больше используют для порождения именно релевантный контекст, когда им подавались все документы;

Таблица 3. Результаты по всем датасетам с группировкой по моделям без упорядочивания. Символом * помечены лучшие результаты с поправкой на то, что длина контекста была меньше, чем для остальных моделей

Датасет	Модель	Context Utilisation		Completeness		Adherence	
		Рел.	Полный	Рел.	Полный	Рел.	Полный
XQuAD	DeepSeek	0.100	0.056	0.100	0.153	0.153	0.165
	Llama	0.158	0.056	0.158	0.179	0.668	0.707
	Mistral	0.319	0.142	0.319	0.356	0.453	0.407
	RuAdapt	0.110	0.031	0.110	0.108	0.744	0.790
	Qwen	0.103	0.076	0.103	0.239	0.666	0.305
	YandexGPT	0.153	0.053	0.153	0.181	0.828	0.875
TyDi	DeepSeek	0.426	0.088	0.426	0.174	0.099	0.213
	Llama	0.207	0.069	0.207	0.154	0.404	0.682
	Mistral	0.326	0.133	0.326	0.261	0.403	0.500
	RuAdapt	0.126	0.053	0.126	0.126	0.687	0.739
	Qwen	0.335	0.050	0.335	0.118	0.220	0.675
	YandexGPT	0.173	0.078	0.173	0.182	0.716	0.771
RuBQ	DeepSeek	0.170*	0.022*	0.170*	0.087*	0.104	0.035
	Llama	0.058	0.009	0.058	0.066	0.406	0.603
	Mistral	0.098	0.011	0.098	0.073	0.411	0.622
	RuAdapt	0.032	0.003	0.032	0.031	0.630	0.795
	Qwen	0.095	0.003	0.095	0.029	0.268	0.730
	YandexGPT	0.046	0.006	0.046	0.020	0.661	0.683
SberQuAD	DeepSeek	0.109	0.009	0.109	0.026	0.168	0.037
	Llama	0.077	0.029	0.077	0.098	0.478	0.573
	Mistral	0.157	0.063	0.157	0.185	0.794	0.680
	RuAdapt	0.117	0.035	0.117	0.117	0.748	0.809
	Qwen	0.140	0.044	0.140	0.145	0.770	0.830
	YandexGPT	0.176	0.061	0.176	0.197	0.845	0.902

- Изменение метрики Completeness относительно Context Utilisation показывает, что модели в основном опираются именно на релевантный контекст и минимально захватывают нерелевантные фрагменты, хотя и сам релевантный контекст используют мало. Значимой корреляции между этой метрикой, в т.ч. и ее изменением, и метриками Exact Matching и Adherence выявлено не было;
- Метрика Adherence показывает, что модель DeepSeek хуже всего справляется со следованием инструкции о использовании для генерации только добавленного контекста. YandexGPT и RuAdapt демонстрируют стабильно высокие результаты с небольшим изменением между датасетами, и, с учетом высоких показателей Exact Matching ранее, можно предположить, что эти модели хорошо справляются с зашумленными данными. Между этой метрикой и метрикой Exact Matching самая большая корреляция: строгое следование инструкции об использовании только поданного контекста должно снижать возможность парафраза ответа, а значит, ведет к повышению точности.

5.3 Анализ влияния переранжирования

Рассмотрим результаты, полученные для экспериментов с переранжированием фрагментов документов перед формированием контекста и демонстрирующие различную чувствительность моделей к структуре контекста.

Для моделей DeepSeek, Llama и YandexGPT переранжирование контекста любым способом ухудшает метрику (Таблица 4), кроме использования

Таблица 4. Сравнение Exact Matching при анализе влияния упорядочивания контекста

Датасет	Модель	Полный контекст	Начиная с релевантного	Начиная с нерелевантного
XQuAD	DeepSeek	0.290	0.287	0.281
	Llama	0.674	0.276	0.369
	Mistral	0.670	0.597	0.569
	RuAdapt	0.628	0.624	0.611
	Qwen	0.541	0.569	0.609
	YandexGPT	0.784	0.719	0.749
TyDi	DeepSeek	0.336	0.318	0.312
	Llama	0.509	0.288	0.334
	Mistral	0.507	0.518	0.515
	RuAdapt	0.542	0.541	0.527
	Qwen	0.502	0.464	0.482
	YandexGPT	0.655	0.611	0.630
RuBQ	DeepSeek	0.314	0.256	0.191
	Llama	0.540	0.160	0.169
	Mistral	0.525	0.533	0.460
	RuAdapt	0.581	0.598	0.569
	Qwen	0.511	0.548	0.501
	YandexGPT	0.638	0.640	0.658
SberQuAD	DeepSeek	0.038	0.038	0.031
	Llama	0.320	0.276	0.424
	Mistral	0.374	0.437	0.418
	RuAdapt	0.360	0.363	0.356
	Qwen	0.449	0.448	0.449
	YandexGPT	0.735	0.695	0.638

Llama на данных из SberQuAD и YandexGPT на данных из RuBQ. Результаты RuAdapt меньше остальных зависят от способа подачи: здесь наблюдается минимальное среднее по всем датасетам отклонение.

Модель Mistral в основном показывает лучший результат, когда релевантный контекст идет в начале всего контекста; Qwen, наоборот, во всех случаях, кроме датасета RuBQ, при сравнении двух подходов к переранжированию порождает ответ более точно, когда релевантный контекст находится в конце.

Анализ результатов для Context Utilisation, Completeness и Adherence в Таблице 5 показывает, что в большинстве случаев переупорядочивание контекста снижает как долю фрагментов, которые модель использует при порождении, т.е. метрику Context Utilisation, так и Completeness. При этом при сравнении подходов переранжирования метрики показывают лучшие значения в основном для случая, когда релевантные фрагменты стоят в начале.

При рассмотрении результатов для Adherence выявляется, что Llama наиболее восприимчива к любому изменению естественного порядка документов, кроме случая с датасетом SberQuAD. Qwen и Mistral чаще всего улучшали свои показатели относительно исходного порядка, но конкретно при рассмотрении только этой метрики нельзя сказать, какой вариант для каждой модели лучше. Для RuAdapt применение разных подходов практически не влияет на значение метрик — модель нечувствительна к изменению подаваемой последовательности фрагментов.

Таблица 5. Результаты по метрикам при анализе влияния упорядочивания контекста. Здесь: FULL — эксперименты с полным контекстом без переранжирования, FIRST — релевантные документы в начале контекста, LAST — релевантные документы в конце. Символом * помечены лучшие результаты с поправкой на то, что длина контекста была меньше, чем для остальных моделей

Датасет	Модель	Context Utilisation			Completeness			Adherence		
		FULL	FIRST	LAST	FULL	FIRST	LAST	FULL	FIRST	LAST
XQuAD	DeepSeek	0.056	0.059	0.058	0.153	0.158	0.157	0.165	0.170	0.158
	Llama	0.056	0.016	0.021	0.179	0.056	0.078	0.707	0.356	0.485
	Mistral	0.142	0.048	0.050	0.356	0.151	0.159	0.407	0.728	0.689
	RuAdapt	0.031	0.031	0.034	0.108	0.108	0.114	0.790	0.801	0.784
	Qwen	0.076	0.031	0.034	0.239	0.106	0.115	0.305	0.693	0.758
	YandexGPT	0.053	0.048	0.042	0.181	0.156	0.147	0.875	0.842	0.918
TyDi	DeepSeek	0.088	0.035	0.033	0.174	0.064	0.062	0.213	0.059	0.063
	Llama	0.069	0.012	0.014	0.154	0.026	0.031	0.682	0.114	0.129
	Mistral	0.133	0.026	0.027	0.261	0.053	0.055	0.500	0.204	0.200
	RuAdapt	0.053	0.053	0.054	0.126	0.126	0.127	0.739	0.739	0.736
	Qwen	0.050	0.033	0.050	0.118	0.076	0.116	0.675	0.419	0.671
	YandexGPT	0.078	0.030	0.072	0.182	0.062	0.168	0.771	0.217	0.817
RuBQ	DeepSeek	0.022*	0.019*	0.018*	0.087*	0.077*	0.028	0.035	0.036	0.044
	Llama	0.009	0.002	0.002	0.066	0.012	0.012	0.603	0.263	0.274
	Mistral	0.011	0.004	0.004	0.073	0.034	0.027	0.622	0.765	0.705
	RuAdapt	0.003	0.003	0.004	0.031	0.033	0.031	0.795	0.802	0.797
	Qwen	0.003	0.003	0.003	0.029	0.030	0.029	0.730	0.746	0.729
	YandexGPT	0.006	0.005	0.005	0.020	0.047	0.039	0.683	0.800	0.842
SberQuAD	DeepSeek	0.009	0.009	0.008	0.026	0.025	0.024	0.037	0.036	0.034
	Llama	0.029	0.025	0.038	0.098	0.083	0.127	0.573	0.498	0.892
	Mistral	0.063	0.055	0.053	0.185	0.167	0.171	0.680	0.806	0.833
	RuAdapt	0.035	0.034	0.036	0.117	0.116	0.119	0.809	0.812	0.793
	Qwen	0.044	0.042	0.045	0.145	0.139	0.148	0.830	0.820	0.841
	YandexGPT	0.061	0.056	0.048	0.197	0.179	0.162	0.902	0.887	0.930

Перестановка предложений внутри контекста не дала стабильного выигрыша ни в одной из рассмотренных метрик. Эффект зависит от конкретной модели и датасета. Естественный порядок следования предложений в документе можно считать хорошим, так как в среднем он не проигрывал явно одному из других порядков и в основном давал лучший результат.

5.4 Анализ отказов

Одним из требований, которое обычно устанавливается к RAG-системам, является требование об отказе пользователю в ответе в случаях, когда модель не уверена в том, что в добавленном контексте содержится ответ на вопрос. В рамках работы были проведены эксперименты, когда модели было предложено отказаться от ответа в случае ее неуверенности и сообщить об этом пользователю порождением текста «НЕТ ОТВЕТА». Ознакомиться с модифицированным промптом можно в Приложении 1.

На рис. 1 приводится статистика, посчитанная по метрике

$$Accuracy = \frac{\text{Количество (НЕТ ОТВЕТА) | Exact Matching} = [\text{True} | \text{False}]}{\text{Количество всех вопросов}}$$

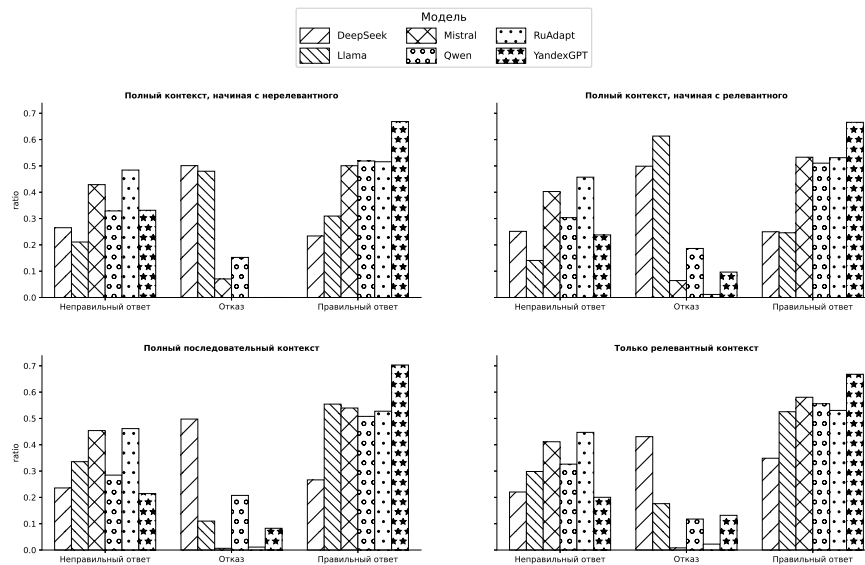


Рис. 1. Среднее на всех датасетах поведение моделей для разных подходов

Для каждого подхода приведен свой график с результатами моделей. На графиках с помощью столбчатой диаграммы модели сравниваются между собой по трем аспектам:

- Доля отказов;
- Доля верных ответов при порождении ответа, отличного от отказа;
- Доля неверных ответов при порождении ответа, отличного от отказа.

Обычно лучшим вариантом является та модель, которая демонстрирует наибольшее среднее количество верных ответов при наименьшем среднем количестве неверных. Излишняя осторожность модели и частые отказы при сохранении превышения среднего количества верных ответов над неверными можно считать хорошим показателем, так как в подобных системах галлюцинации моделей и порождение неверных ответов должны считаться более грубой ошибкой, чем отказы в ответах.

Среди выбранных моделей такой будет YandexGPT-5-Lite-8B-instruct (Таблица 6).

Модели Llama и DeepSeek чаще остальных порождали сообщение с отказом. С учетом того, что моделям все еще подавался контекст, содержащий ответ, это в очередной раз показывает, что модели недостаточно хорошо работают с ним: они предпочитали породить отказ, а не извлечь ответ из данных документов. Причины такого поведения моделей при работе с текстом документов, поданных вместе с инструкцией, требуют дополнительного тщательного исследования самого контекста.

Таблица 6. Статистика ответов моделей, %

Модель	Верные	Неверные	Отказы
YandexGPT	67.62	24.60	7.78
RuAdapt	52.62	46.25	1.13
Mistral	52.28	43.98	3.73
Qwen	51.66	33.27	15.07
Llama	39.96	26.30	33.75
DeepSeek	25.65	23.98	50.37

Отметим, что модели RuAdapt и Mistral похожи по поведению: они редко отказываются от ответа, и, хотя количество верных ответов превышает количество неверных, процент неверных достаточно большой — модели склонны порождать ответ даже при недостаточной уверенности, что снижает общую надежность системы по сравнению с более осторожными моделями. Вероятно, эти модели требуют более тщательного подбора промпта с применением других подходов к его составлению.

6 Заключение

В рамках работы проведены эксперименты для оценки компактных (7-8 млрд параметров) генеративных моделей разных семейств (DeepSeek, Llama, Mistral, Qwen, RuAdapt и YandexGPT) на четырех адаптированных вопросно-ответных русскоязычных датасетах XQuAD, TyDi QA, RuBQ и SberQuAD. Получены оценки качества порождения с разными стратегиями добавления и упорядочивания контекста, и проведен анализ результатов, который показал, что при отсутствии контекста модели практически не способны дать правильный ответ, в то время как добавление релевантного контекста значительно улучшает точность. Сделаны выводы о проблемах моделей с извлечением ответа из поданного контекста, что не дает метрике Exact Matching достигать лучших результатов и влияет на количество порождаемых отказов в ответе, и о связи метрик Adherence и Exact Matching. Дополнительно проведены анализ соблюдения моделями требований, выдвигаемых для подобных систем, в частности, оценена способность моделей корректно отказываться от ответа при отсутствии информации в контексте, результаты чего еще раз подчеркнули, что даже при наличии ответа в контексте модели не всегда могут его корректно извлечь, что приводит к частым отказам.

Благодарности. Исследование выполнено за счет гранта Российского научного фонда № 25-21-00206².

² <https://rscf.ru/project/25-21-00206/>

Список литературы

1. Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey. CoRR (2023)
2. Chen J., Lin H., Han X., Sun L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 17754–17762 (2024)
3. Niu C., Wu Y., Zhu J. et al. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10862–10878 (2024)
4. Yang X., Sun K., Xin H. et al. CRAG-comprehensive RAG benchmark. In: Advances in Neural Information Processing Systems, pp. 10470–10490 (2024)
5. Friel R., Belyi M., Sanyal A. Ragbench: Explainable benchmark for retrieval-augmented generation systems. ArXiv preprint arXiv:2407.11005 (2024)
6. Artetxe M., Ruder S., Yogatama D. On the Cross-lingual Transferability of Monolingual Representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4623–4637 (2020)
7. Clark J. H., Choi E., Collins M. et al. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. In: Transactions of the Association for Computational Linguistics (2020)
8. Rybin I., Korablinov V., Efimov P., Braslavski P. RuBQ 2.0: An Innovated Russian Question Answering Dataset. In: Extended Semantic Web Conference (2021)
9. Efimov P., Chertok A., Boytsov L., Braslavski P. SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 3–15. Springer International Publishing (2020)
10. Bi X., Chen D., Chen G. et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. CoRR (2024)
11. Vavekanand R., Sam K. Llama 3.1: An in-depth analysis of the next-generation large language model (2024)
12. Jiang A. Q., Sablayrolles A., Mensch A. et al. Mistral 7B. ArXiv preprint arXiv:2310.06825 (2023)
13. Tikhomirov M., Chernyshev D.: Impact of tokenization on LLaMa Russian adaptation. In: Ivannikov Ispras Open Conference (ISPRAS), pp. 163–168. IEEE (2023)
14. Yang A., Yang B., Hui B. et al. Qwen2 Technical Report. ArXiv preprint arXiv:2407.10671 (2024)
15. Qwen2.5: A Party of Foundation Models, <https://qwenlm.github.io/blog/qwen2.5/>, дата обращения: 2025/05/12
16. Встречаем YandexGPT 5 — в Алисе, облаке и опенсорсе, <https://habr.com/ru/companies/yandex/articles/885218/>, дата обращения: 2025/05/03
17. Открываем instruct-версию YandexGPT 5 Lite, <https://habr.com/ru/companies/yandex/articles/895428/>, дата обращения: 2025/05/03
18. Tian F., Ganguly D., Macdonald C. Is Relevance Propagated from Retriever to Generator in RAG? In: European Conference on Information Retrieval, pp. 32–48. Springer (2025)
19. Zhang T., Kishore V., Wu F. et al. BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations.

Приложение 1

Ты — русскоязычный ИИ-ассистент, строго следующий инструкциям. Используй только предоставленный контекст для ответа на вопрос, даже если ты считаешь, что там есть ошибки. Строго соблюдай следующие правила:

1. Если ответ есть в контексте — дай краткий ответ (1-5 слов) без пояснений
2. Если ответа нет в контексте — ответь только 'НЕТ ОТВЕТА' без изменений
3. Никогда не повторяй вопрос и не добавляй посторонней информации

Примеры правильных ответов:

Пример 1:

Вопрос: Когда опубликовали статью Attention is All You Need?

Контекст: Статья Attention is All You Need вышла в 2017 году.

Ответ: В 2017 году.

Пример 2:

Вопрос: Кто автор 'Преступления и наказания'?

Контекст: Толстой обсуждал литературу с Достоевским в 1878 году.

Ответ: НЕТ ОТВЕТА

Пример 3:

Вопрос: Сколько голов забил игрок X?

Контекст: Нападающий X сделал хет-трик в последнем матче.

Ответ: Три гола.

Текущий вопрос: {}

Соответствующий контекст: {}

Твой ответ (только краткий ответ или НЕТ ОТВЕТА):

Список рецензий

Рецензия 1

3: strong accept

Статья посвящена тестированию генеративных моделей, использующих результаты информационного поиска при формировании ответа на вопрос пользователя. В статье приведены результаты экспериментов по оценке компактных (7-8 млрд параметров) генеративных моделей разных семейств (DeepSeek, Llama, Mistral, Qwen, RuAdapt и YandexGPT) на четырех адаптированных вопросно-ответных русскоязычных датасетах (XQuAD, TyDi QA, RuBQ и SberQuAD) с использованием специальных метрик оценки качества генеративных моделей Context Relevance, Context Utilisation, Completeness, Adherence и Exact Matching. Результаты экспериментов представлены в шести таблицах. Анализ полученных оценок качества порождения ответов моделями с разными стратегиями добавления и упорядочивания контекста показал, что при отсутствии контекста модели практически не способны дать правильный ответ, в то время как добавление релевантного контекста значительно улучшает точность ответов. Сделаны выводы о проблемах моделей с извлечением ответа из поданного контекста, что не дает метрике Exact Matching достигать лучших результатов и влияет на количество порождаемых отказов в ответе. Дополнительно, оценена способность моделей корректно отказываться от ответа при отсутствии необходимой информации в контексте. В целом статья написана в научном стиле, хорошо структурирована, результаты экспериментов хорошо оформлены, включает ссылки на релевантные публикации и содержит достаточно полный анализ результатов и необходимые выводы. Считаю, что данная статья может быть принята в качестве полной статьи после учета замечания рецензента.

Замечание.

В статье имеется описка – в предложении «С учетом того, что моделям все еще подавался контекст, содержащий ответ, это в очередной раз показывает, что модели достаточно недостаточно хорошо работают с ним ...» присутствует лишнее слово «достаточно».

Ответ: Описка была исправлена.

Рецензия 2

2: accept

Статья “Методы тестирования порождающих моделей, использующих информационный поиск” описывает подход к подбору данных и проведению тестирования различных моделей в режиме вопрос-ответ.

В статье достаточно подробно приводится принцип отбора данных. А также приводятся результаты тестирования по различным метрикам, позволяющие оценить применимость разных моделей для ответов на русском языке.

В качестве замечаний, отметим неполноту описания способа подсчёта метрик. Например, не ясно, считалась ли метрика Exact Matching автоматически или вручную. Если подсчёт был автоматическим, то как именно вычислялось соответствие. То же в отношении других метрик.

Также, не ясно, есть ли влияние предметной области текста на качество ответов моделей. Возможно, следовало бы разделить набор данных по предметным областям (хотя бы две разные) и сравнить полученные результаты, чтобы оценить гипотезу их различности.

И, очень полезным было бы оценить взаимное перекрытие ложноположительных и ложноотрицательных ответов на одни и те же вопросы различными моделями. Если перекрытие малое, то становится возможным оценивать степень галлюцинаций за счёт сравнения с другими моделями.

Ответ: Метрика Exact Matching считалась автоматически: производился поиск ожидаемого ответа, приведенного к своей нормальной форме, внутри порожденного ответа, так же приведенного к нормальной форме. Анализ влияния предметной области действительно не проводился, так как датасеты исходно не содержали информации о предметных областях каждого вопроса.

Если под ложноположительными и ложноотрицательными ответами авторы рецензии подразумевают соответственно случаи, когда модель порождает ответ при условии, что на вопрос его нельзя дать по подаваемому контексту, и когда модель, наоборот, порождает сообщение с отказом отвечать при возможности дать ответ по контексту, то считаем важным указать, что в случае наших исследований невозможно таким образом выделить ложноположительные случаи: считается, что на каждый вопрос из датасетов реально породить ответ по контексту. Случаи порождения отказов действительно стоит анализировать дополнительно: в пункте 5.4 указано, что некоторые из моделей плохо работают с контекстом, и требуется большое исследование вопроса, почему это может происходить, и выявление причин или, возможно, зависимостей между типами вопросов или предлагаемых в качестве контекста фрагментов и этими отказами.

Если же под ложноположительными и ложноотрицательными случаями понимаются ошибки в терминах метрики Exact Matching (ложноотрицательные — ответ верный, но не совпал по форме с эталоном; ложноположительные — семантически неверен), то по ложноотрицательным был проведен небольшой анализ (пункт 5.2), показывающий, что такие ошибки возникают нечасто и, вероятно, оценивание взаимного перекрытия не даст показательных результатов.

Рецензия 3

2: accept

Статья представляет собой значимый вклад в область оценки Retrieval-Augmented Generation (RAG)-систем, особенно для русскоязычных данных.

Авторы проводят систематическое исследование, сравнивая несколько компактных языковых моделей на адаптированных датасетах, и предлагают методику оценки, основанную на ряде метрик. Работа актуальна, учитывая растущую популярность RAG-подхода и отсутствие стандартизированных бенчмарков для русского языка.

В качестве замечаний отметим:

1. В работе рассматриваются только модели с 7–8B параметров. Было бы полезно включить более крупные модели для оценки влияния на результаты вклада именно БЯМ.
2. Хотя авторы отмечают, что модели часто отказываются отвечать даже при наличии правильного контекста, причины этого явления (например, влияние токенизации, качества промптов) не исследуются глубоко.

Статья представляет собой важный шаг в стандартизации оценки RAG-систем для русского языка. Авторы показали, что:

- добавление релевантного контекста значительно улучшает качество ответов;
- порядок контекста влияет на результаты;
- метрика Adherence тесно связана с Exact Matching.

Ответ:

1. Как было указано в статье, небольшие модели рассматривались для оценки эффективности подхода в условиях ограниченности ресурсов и для создания основы для дальнейшего масштабирования.
2. Исследование причин частых отказов при наличии правильных контекстов планируется в дальнейшем.

Рецензия 4

3: strong accept

Статья представляет методологически сильное исследование, которое вносит вклад в решение проблемы достоверности генерации при помощи retrieval-систем. Значимым вкладом исследования является то, что это первая комплексная оценка русскоязычных моделей в RAG-задачах. Безусловной заслугой является публикация унифицированного бенчмарка для RAG-систем на русском языке. Подход, при котором тестируются сравнительно небольшие языковые модели, кажется интересным и плодотворным, т.к. делает результат актуальным с практической точки зрения. Статья рекомендуется к печати с минимальными доработками. В частности, желательно снабдить формулами все применяемые метрики, а не только первые две. Кроме того, необходимо устранить лексическую ошибку на с. 13 "...модели достаточно недостаточно хорошо работают с ним...". В качестве потенциального улучшения исследования предлагается провести больше анализов ошибок, учитывающих типы вопросов. Также требует уточнения, как размечалась релевантность фрагментов - вручную или полуавтоматически?

Ответ: Лексическая ошибка на с. 13 была устранена. Формулы добавлены ко всем метрикам. Релевантность фрагментов размечалась следующим образом: исходно в TyDi QA, SberQuAD, RuBQ — вручную, в XQuAD — полуавтоматически. Это следует из статей, которыми сопровождалась данные датасеты.