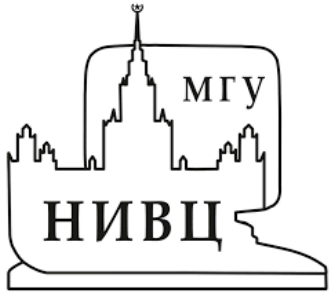


ИСП РАН

Подходы к интерпретируемости языковых моделей в автоматической обработке текстов

Лукашевич Н.В.

МГУ имени М.В. Ломоносова

Ведущий научный сотрудник НИВЦ МГУ

Зав. Кафедрой, профессор ВМК МГУ

Главный научный сотрудник ИСП РАН

Louk_nat@mail.ru

Риски применения нейросетевых моделей

- **Проблема постановки задачи** предполагает завышенные ожидания от применения методов машинного обучения.
- **Проблемы с данными –** качество данных не позволяет достичь приемлемого качества на целевой задаче.
- **Проблемы с методом –** данные хорошего качества, но проблема связана с применением метода машинного обучения
- **Необходимость объяснения -** данные и применяемый метод дают достаточно качественные результаты, но этого не достаточно для применения в реальной целевой задаче
- **Проблемы объяснения -** объяснения имеются, но их по каким-то причинам недостаточно для целевой задачи
- **Атаки на метод машинного обучения**

- Как оценить качество объяснения
- Важным является два компонента
  - **Понятность объяснения**
    - Например, объясняющее правило не должно быть слишком сложным
  - **Правильность объяснения (fidelity, faithfulness)**
    - Насколько объяснение соответствует реальному процессу в системе

Проверка достоверности фактов в LLM

- LLM стали рассматриваться как удобный инструмент для пользователя для поиска информации и ответов на вопросы
  - Например, для людей может быть легче получить совет от LLM относительно методов лечения, чем пойти к врачу
- Встраивание LLM в большие поисковые системы (типичный источник информации в течение многих лет) увеличивает риски получение недостоверной информации из Интернет
- Злоупотребление способностями LLM для написания злонамеренных текстов

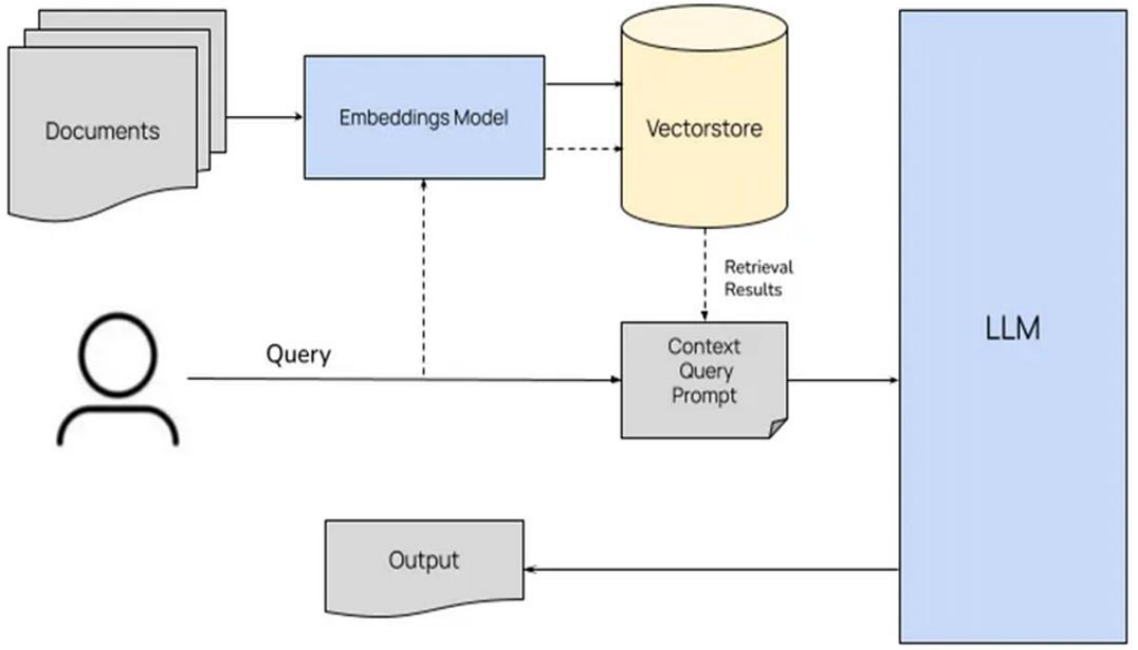
- Особенности LLM
  - Статическое знание, проблемы с обновлением
  - Связный, качественный текст, сообщающий неверную информацию
  - Способность порождения убедительных текстов
  - Способность порождения больших объемов текстов, содержащих не правильную информацию
  - Галлюцинации: LLM не знают, что они не знают

Меры по улучшению достоверности выдачи LLM

- Процедура Alignment в больших языковых моделях
  - обучение с подкреплением на основе обратной связи с человеком,
  - подбор промтов и фильтрация на основе машинного обучения
  - специальное обучение моделей для снижения галлюцинация
  - прямая оптимизация предпочтений с отдельными моделями вознаграждения за факты и выполнение инструкций.
- Но есть проблемы в небольших и открытых LLM
  - RAG (Retrieval augmented generation) – как важное средство для улучшения надежности и достоверности

RAG – порождение текста с использованием результатов информационного поиска

- Генерация с расширенным поиском (RAG) включает контекстную информацию из внешних источников в генерацию текста.
- RAG смягчает проблему LLM, создающих неточный контент, расширяя их возможности с внешними данными.
- Однако это требует эффективного поиска релевантных текстов и ох качественного отбора для подачи на вход LLM



RAG: Эксперименты на русских данных

- Небольшие модели
  - DeepSeek-LLM-7B-Chat 23
  - Meta-Llama-3.1-8B-Instruct
  - Mistral-7B-Instruct-v0.3
  - Qwen2.5-7B-Instruct
  - RuadaptQwen2.5-7B-Lite-Beta
  - YandexGPT-5-Lite-8B-instruct

- Русские вопросно-ответные датасеты:
  - XQuAD
  - TyDi QA
  - RuBQ
  - SberQuAD

  - Примеры вопросов:
    - Когда начался Карибский кризис?«
    - Где встречаются первые упоминания о строении человеческого тела?",

Пример вопроса в датасете


```
{
    "id": "2_sberquad",
    "question": "что встречается в протерозойских отложениях?",
    "answers": [
        "органические остатки"
    ],
    "normalized_answers": [
        "органический остаток"
    ],
    "context": [
        {
            "chunk": "В протерозойских отложениях органические остатки встречаются намного чаще, чем
                в архейских.",
            "is_relevant": true
        },
        {
            "chunk": "Они представлены известковыми выделениями сине-зеленых водорослей, ходами
                червей, остатками кишечнополостных.",
            "is_relevant": false
        },
        {
            "chunk": "Кроме известковых водорослей, к числу древнейших растительных остатков
                относятся скопления графито-углистого вещества, образовавшегося в результате
                разложения Corycium enigmaticum.",
            "is_relevant": false
        },
        {
            "chunk": "В кремнистых сланцах железорудной формации Канады найдены нитевидные
                водоросли, грибные нити и формы, близкие современным кокколитофоридам.",
            "is_relevant": false
        },
        {
            "chunk": "В железистых кварцитах Северной Америки и Сибири обнаружены железистые
                продукты жизнедеятельности бактерий.",
            "is_relevant": false
        }
    ],
    "metadata": {
        "tag": null,
        "is_answerable": true
    }
}
```

RAG: Результаты на русских датасетах

Датасет	Модель	Без контекста	Релевантный контекст	Полный контекст	Начиная с релевантного	Начиная с нерелевантного
XQuAD	DeepSeek	0.050	0.228	0.290	0.287	0.281
	Llama	0.099	0.650	0.674	0.276	0.369
	Mistral	0.086	0.705	0.670	0.597	0.569
	RuAdapt	0.097	0.639	0.628	0.624	0.611
	Qwen	0.099	0.556	0.541	0.569	0.609
	YandexGPT	**0.110**	**0.766**	**0.784**	**0.719**	**0.749**
TyDi	DeepSeek	0.087	0.530	0.336	0.318	0.312
	Llama	0.165	0.503	0.509	0.288	0.334
	Mistral	0.142	0.577	0.507	0.518	0.515
	RuAdapt	0.171	0.550	0.542	0.541	0.527
	Qwen	0.113	0.585	0.502	0.464	0.482
	YandexGPT	**0.228**	**0.644**	**0.655**	**0.611**	**0.630**
RuBQ	DeepSeek	0.178	0.493	0.314	0.256	0.191
	Llama	0.381	0.548	0.540	0.160	0.169
	Mistral	0.318	0.537	0.525	0.533	0.460
	RuAdapt	0.402	0.582	0.581	0.598	0.569
	Qwen	0.260	0.564	0.511	0.548	0.501
	YandexGPT	**0.557**	**0.597**	**0.638**	**0.640**	**0.658**
SberQuAD	DeepSeek	0.023	0.193	0.038	0.038	0.031
	Llama	0.046	0.278	0.320	0.276	0.424
	Mistral	0.049	0.424	0.374	0.437	0.418
	RuAdapt	0.056	0.351	0.360	0.363	0.356
	Qwen	0.039	0.447	0.449	0.448	0.449
	YandexGPT	**0.075**	**0.667**	**0.735**	**0.695**	**0.638**

Классические методы информационного поиска

- Классические лексические методы
  - Представление в виде мешка слов с весами
- Методы взвешивания
  - Tf.idf
  - BM25

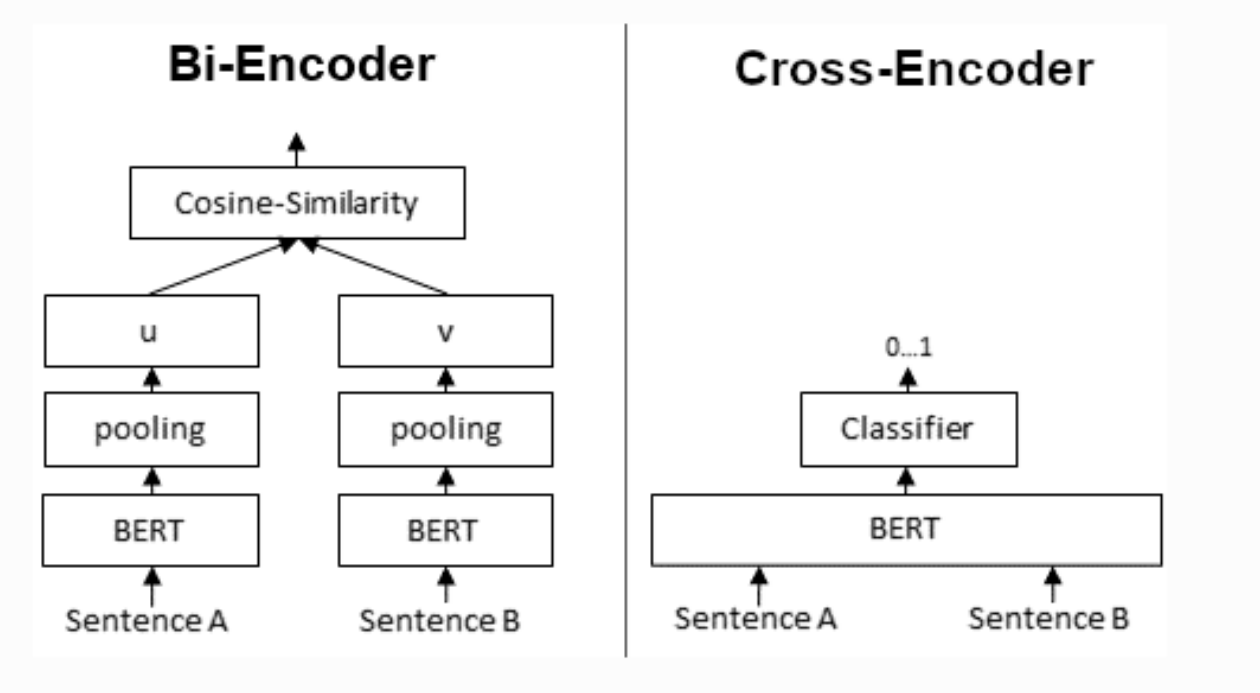
- Проблемы лексического несоответствия запроса и релевантного документа

- Q: Почему электрические **батареи** быстрее **разряжаются** на **холоде**?
- A: ***Батарейки*** *быстрее* ***садятся*** *на* ***морозе****, потому что*

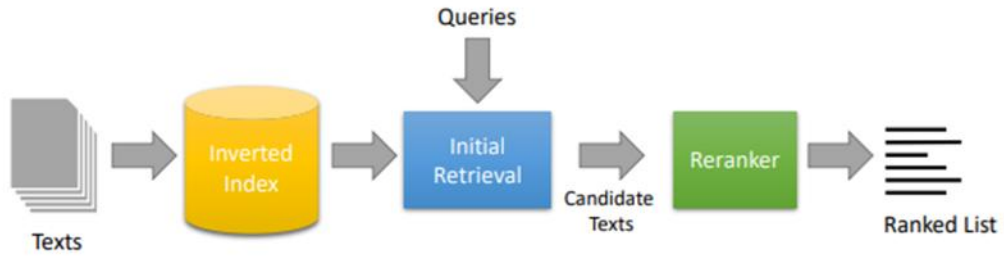
BM25= $$\sum_{t \in q} \log\left[\frac{N}{\mathrm{df}_t}\right] \cdot \frac{(k_1+1)\mathrm{tf}_{td}}{k_1((1-b)+b\times(L_d/L_{\mathrm{ave}}))+\mathrm{tf}_{td}}$$

Нейросетевые модели: Cross-encoders, biencoders and rerankers

- Основная идея: обучить плотные векторные представления запросов и текстовых фрагментов с помощью специальных кодировщиков, используя максимизацию различных метрик сходства
- А затем сравнить векторные представления запроса с документами
- Какое улучшение по сравнению с классическими методами?
  - Обучаются на одних данных, а применятся к другим данным



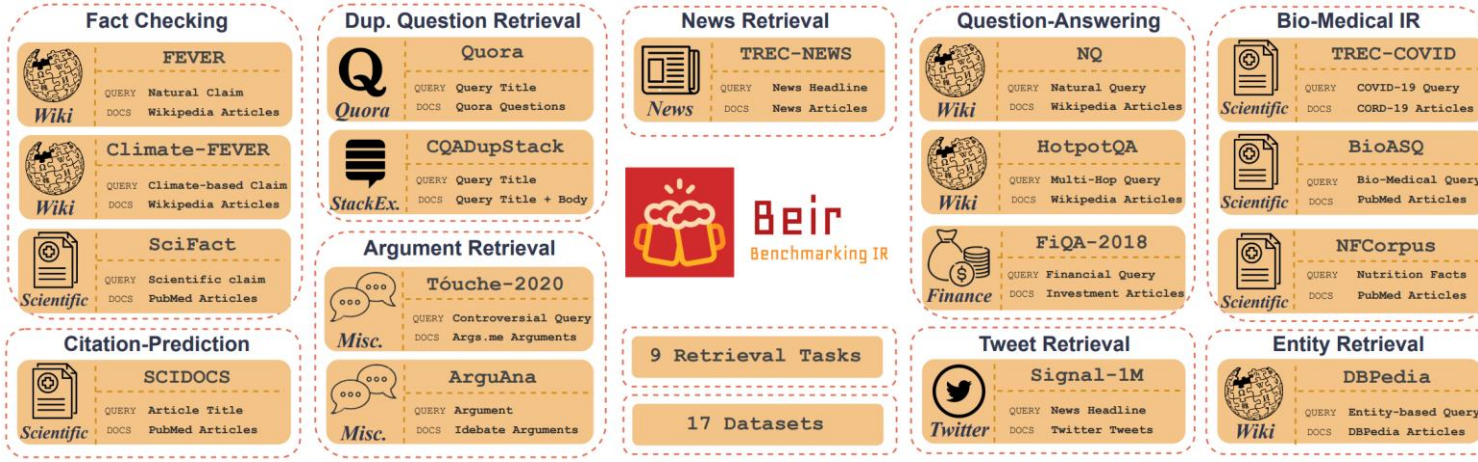
Reranker



Бенчмарк Beir (2021)

- Особенности обучения нейросетевых моделей
  - нужны большие объемы обучающих данных, есть подходы self-supervised,
  - Показываются высокие результаты на датасетах, на которых обучалась
  - Опубликованы обученные модели, но как изменится качество при переносе на другие данные?

- Бенчмарк Beir (2021): 18 датасетов, мера качества NDCG@10, оценка zero-shot, т.е. без обучения



Результат моделей на Beir (2021)

Model (→)	Lexical	Sparse			Dense				Late-Interaction	Re-ranking
Dataset (↓)	BM25	DeepCT	SPARTA	docT5query	DPR	ANCE	TAS-B	GenQ	ColBERT	BM25+CE
MS MARCO	0.228	0.296‡	0.351‡	0.338‡	0.177	0.388‡	0.408‡	0.408‡	0.401‡	**0.413**‡
TREC-COVID	0.656	0.406	0.538	0.713	0.332	0.654	0.481	0.619	0.677	**0.757**
BioASQ	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398	0.474	**0.523**
NFCorpus	0.325	0.283	0.301	0.328	0.189	0.237	0.319	0.319	0.305	**0.350**
NQ	0.329	0.188	0.398	0.399	0.474‡	0.446	0.463	0.358	0.524	**0.533**
HotpotQA	0.603	0.503	0.492	0.580	0.391	0.456	0.584	0.534	0.593	**0.707**
FiQA-2018	0.236	0.191	0.198	0.291	0.112	0.295	0.300	0.308	0.317	**0.347**
Signal-1M (RT)	0.330	0.269	0.252	0.307	0.155	0.249	0.289	0.281	0.274	**0.338**
TREC-NEWS	0.398	0.220	0.258	0.420	0.161	0.382	0.377	0.396	0.393	**0.431**
Robust04	0.408	0.287	0.276	0.437	0.252	0.392	0.427	0.362	0.391	**0.475**
ArguAna	0.315	0.309	0.279	0.349	0.175	0.415	0.429	**0.493**	0.233	0.311
Touché-2020	**0.367**	0.156	0.175	0.347	0.131	0.240	0.162	0.182	0.202	0.271
CQADupStack	0.299	0.268	0.257	0.325	0.153	0.296	0.314	0.347	0.350	**0.370**
Quora	0.789	0.691	0.630	0.802	0.248	0.852	0.835	0.830	**0.854**	0.825
DBPedia	0.313	0.177	0.314	0.331	0.263	0.281	0.384	0.328	0.392	**0.409**
SCIDOCS	0.158	0.124	0.126	0.162	0.077	0.122	0.149	0.143	0.145	**0.166**
FEVER	0.753	0.353	0.596	0.714	0.562	0.669	0.700	0.669	0.771	**0.819**
Climate-FEVER	0.213	0.066	0.082	0.201	0.148	0.198	0.228	0.175	0.184	**0.253**
SciFact	0.665	0.630	0.582	0.675	0.318	0.507	0.643	0.644	0.671	**0.688**
Avg. Performance vs. BM25		**- 27.9%**	**- 20.3%**	**+ 1.6%**	**- 47.7%**	**- 7.4%**	**- 2.8%**	**- 3.6%**	**+ 2.5%**	**+ 11%**

При переносе моделей на другие наборы данных качество обученных нейросетевых моделей значительно снижается

Русский бенчмарк rusBeir (2025)

Source (↓)	Task (↓)	Dataset (↓)	Origin (↓)	Relevancy	Train	Dev	Test	Corpus	Avg. Word Lengths (D/Q)
BEIR	Bio-Medical IR	rus-NFCorpus	Translation	Binary	2,590	324	323	3,633	216.6 / 3.5
BEIR	Argument Retrieval	rus-ArguAna	Translation	Binary	—	—	1,406	8,674	147.8 / 173.8
BEIR	Fact Checking	rus-SciFact	Translation	Binary	809	—	300	5,183	185.8 / 11.2
BEIR	Citation-Prediction	rus-SCIDOCS	Translation	Binary	—	—	1000	25,657	153.1 / 9.8
BEIR	Bio-Medical IR	rus-TREC-COVID	Translation	3-level	—	—	50	171,332	138.9 / 8.5
BEIR	Question Answering (QA)	rus-FiQA	Translation	Binary	5,500	500	648	57,638	122.1 / 9.9
BEIR	Duplicate Question Retrieval	rus-Quora	Translation	Binary	—	5,000	10,000	522,931	9.8 / 7.9
BEIR	Duplicate Question Retrieval	rus-CQADupstack	Translation	Binary	—	—	13,145	457,199	117.6 / 7.6
BEIR	Argument Retrieval	rus-Touche	Translation	Binary	—	—	49	382,545	252.5 / 6.8
BEIR	Information-Retrieval	rus-MMARCO	Part of multilingual	Binary	502,939	6980	—	8,841,823	49.6 / 5.95
Open-Source Dataset	Information-Retrieval	rus-MIRACL	Part of multilingual	Binary	4,683	1,252	—	9,543,918	43 / 6.2
Open-Source Dataset	Question Answering (QA)	rus-XQuAD	Part of multilingual	Binary	—	1,190	—	240	112.9 / 8.6
Open-Source Dataset	Question Answering (QA)	rus-XQuAD-sentences	Part of multilingual	Binary	—	1,190	—	1212	22.4 / 8.6
Open-Source Dataset	Question Answering (QA)	rus-Tydi QA	Part of multilingual	Binary	—	1,162	—	89,154	69.4 / 6.5
Open-Source Dataset	Information-Retrieval	SberQuAD-retrieval	Originally Russian	Binary	45,328	5,036	23,936	17,474	100.4 / 8.7
Open-Source Dataset	Information-Retrieval	ruSciBench-retrieval	Originally Russian	Binary	—	345	—	200,532	89.9 / 9.2
Open-Source Dataset	Question Answering (QA)	ru-facts	Originally Russian	Binary	2,241	753	—	6,236	28.1 / 23.9
RU-MTEB	Information-Retrieval	RuBQ	Originally Russian	Binary	—	—	1,692	56,826	62.07 / 6.4
RU-MTEB	Information-Retrieval	Ria-News	Originally Russian	Binary	—	—	10,000	704,344	155.2 / 8.8
rusBEIR	Information-Retrieval	wikifacts-articles	Originally Russian	3-level	—	540	—	1,324	2,535.9 / 11.4
rusBEIR	Fact Checking	wikifacts-para	Originally Russian	3-level	—	540	—	15,317	219.2 / 11.4
rusBEIR	Information-Retrieval	wikifacts-sents	Originally Russian	3-level	—	540	—	188,026	17.8 / 11.4
rusBEIR	Fact Checking	wikifacts-sliding_para2	Originally Russian	3-level	—	540	—	118,025	35.7 / 11.4
rusBEIR	Fact Checking	wikifacts-sliding_para3	Originally Russian	3-level	—	540	—	188,024	53.6 / 11.4
rusBEIR	Fact Checking	wikifacts-sliding_para4	Originally Russian	3-level	—	540	—	188,023	71.4 / 11.4
rusBEIR	Fact Checking	wikifacts-sliding_para5	Originally Russian	3-level	—	540	—	188,022	89.3 / 11.4
rusBEIR	Fact Checking	wikifacts-sliding_para6	Originally Russian	3-level	—	540	—	188,021	107.1 / 11.4

Модели

Model	Based on	Parameters	Dim	Max input
Multilingual-E5-small	Multilingual-MiniLM	118M	384	512
Multilingual-E5-base	XLM-RoBERTa-base	278M	768	512
Multilingual-E5-large	XLM-RoBERTa-large	560M	1024	512
BGE-M3	BGE-M3	568M	1024	8192
USER-BGE-M3	BGE-M3	359M	1024	8192
LaBSE	LaBSE	471M	768	256
RoSBERTa	SBERT	404M	1024	512
FRIDA	FRED-T5	823M	1536	512
bge-reranker-v2-m3	BGE-M3	568M	1024	8192

Использование реранкеров – сначала любая модель, затем переупорядочивание первых 100 документов кросс-энкодером

Результаты на RusBeir (NDCG@10)

Model (→)	Lexical	Dense								Re-ranking			
Dataset (↓)	BM25	mE5-large	mE5-base	mE5-small	BGE-M3	USER-BGE-M3	RoSBERTa	LaBSE	FRIDA	BM25+BGE	E5+BGE	BGE+BGE	FRIDA+BGE
rus-NFCorpus	32.33	30.96	26.90	26.79	30.86	30.28	27.24	18.53	29.40	**34.83**	33.18	32.46	32.38
rus-ArguAna	41.49	49.06	39.40	39.59	50.75	46.52	49.38	25.52	41.90	52.91	**54.01**	53.87	52.24
rus-SciFact	65.60	63.49	63.46	60.46	62.42	58.25	53.90	29.07	63.43	70.40	**71.34**	69.64	70.09
rus-SCIDOCS	13.99	13.47	12.09	10.60	15.04	14.46	14.43	8.17	12.78	15.31	15.98	**16.21**	15.33
rus-TREC-COIVD	62.47	76.38	74.45	73.95	62.66	55.07	68.43	23.04	82.42	73.46	83.11	77.66	**85.97**
rus-FiQA	22.60	34.71	30.44	25.74	38.16	37.09	32.62	7.19	36.50	29.37	39.19	**40.21**	38.36
rus-Quora	61.34	80.18	78.62	74.99	80.28	79.87	68.73	72.39	74.54	71.01	**81.85**	81.59	80.38
rus-CQADupstack	24.49	32.08	28.43	27.31	32.57	31.40	27.85	21.36	27.69	28.86	33.78	**33.90**	31.44
rus-Touche	30.59	25.88	22.88	23.48	28.06	24.36	24.11	8.46	30.44	**32.72**	29.46	31.43	32.21
rus-MMARCO	15.25	34.04	30.27	29.07	29.51	27.92	20.16	9.06	33.55	24.12	**36.95**	34.52	36.33
rus-MIRACL	25.13	66.99	61.41	58.52	70.50	67.23	53.11	15.70	71.91	41.51	75.90	**76.44**	76.36
rus-XQuAD	96.19	97.33	95.84	95.66	95.97	95.63	93.90	69.77	93.49	98.85	**98.97**	**98.97**	98.24
rus-XQuAD-Sentences	82.36	88.84	86.37	85.41	86.91	85.42	83.20	75.33	86.89	89.93	**92.08**	91.69	91.60
rus-TyDi QA	35.80	59.41	55.91	55.23	58.34	57.86	52.06	28.05	59.84	50.12	**66.20**	65.78	65.84
SberQuad-retrieval	68.19	67.11	65.13	61.03	68.26	67.03	63.59	37.54	62.68	**70.34**	69.41	68.21	65.34
ruSciBench-retrieval	36.69	50.81	45.74	42.93	55.85	53.58	44.89	17.93	52.92	49.93	65.33	**69.05**	66.44
ru-facts	92.56	93.65	93.55	93.06	**93.91**	93.77	93.66	93.10	93.74	92.72	92.87	92.87	92.90
RuBQ	37.33	74.11	69.63	68.60	71.26	70.00	66.81	30.59	73.12	56.90	**77.03**	76.00	75.71
Ria-News	64.63	80.67	70.24	70.00	82.99	83.52	78.85	61.57	82.92	78.12	86.22	**86.85**	86.73
wikifacts-articles	84.28	66.09	63.04	67.86	74.50	79.41	74.13	45.17	75.47	**85.25**	83.06	83.91	83.07
wikifacts-para	61.31	50.15	49.51	34.71	54.55	57.53	50.66	14.78	55.17	**66.61**	59.95	63.76	63.85
wikifacts-sents	33.64	35.90	30.75	22.57	37.59	34.90	40.59	25.79	46.53	39.96	38.53	39.20	**49.41**
Avg.	48.69	55.78	52.41	48.94	56.04	54.94	51.72	31.40	56.63	56.35	61.31	61.31	**61.59**

Заключение

- Большие языковые модели имеют проблемы с достоверностью порождаемых текстов
- Одним их доступных решений повышения достоверности ответа LLM является подход RAG, в котором используются результаты информационного поиска
- Проблемы RAG
  - Обеспечение качества поиска
  - Разделение текстов на небольшие фрагменты (chunking)
  - Исследование влияния релевантного контекста на порождение (Hagstrom et al., 2024)