HSE SMART ASSISTANT: INTEGRATED VOICE ASSISTANT FOR STUDENTS

Presentation by Anastasiia Prokhorova

Retrieval-Augmented Generation | 2024

HSE

Higher School of Economics

PROBLEM

There exists no unique 100% reliable metric to evaluate RAG models.

NATIONAL RESEARCH UNIVERSITY

GOAL AND TASKS

GOAL

assess the effectiveness of Retrieval-Augmented Generation (RAG) models and determine the choice of metrics to use for evaluation

TASKS

- study the literature on metrics for evaluating machine learning models
- implement and integrate metrics
- analyze obtained results
- determine effective performance metrics
- analyze ways of model improvement

EXPERIMENTS SETTINGS

NATIONAL RESEARCH UNIVERSITY

MODELS

MODEL 1	bert-base-multilingual-cased + Cosine similarity
MODEL 2	cointegrated/rubert-tiny2 + Cosine similarity
MODEL 3	cointegrated/rubert-tiny2 + Euclidean similarity

MODELS

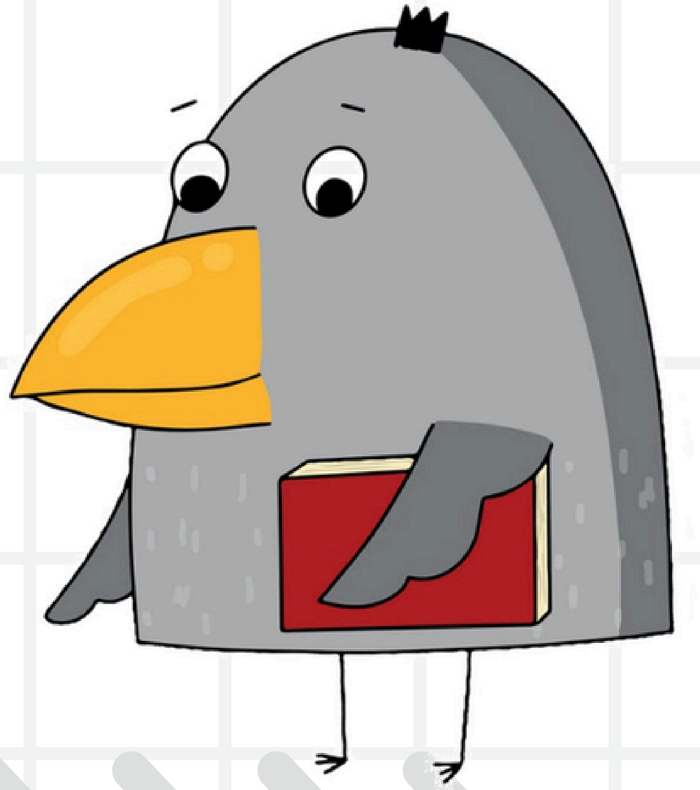
MODEL 4	Jaccard similarity
MODEL 5	Jaccard similarity + Lemmatization + Query Expansion + YandexGPT assistance
MODEL 6	cointegrated/rubert-tiny2 + Cosine similarity + Query Expansion + YandexGPT assistance

METHODOLOGY

DISTANCE METRICS	Cosine similarity, Euclidean distance, Jaccard similarity
RANKING METRICS	MAP@k, NDCG, MRR
COMPLEX METRICS	ROUGE-1, ROUGE-2, ROGE-L, BERTScore
LLM-BASED METRICS	ChatGPT-4o, YandexGPT-3

NATIONAL RESEARCH UNIVERSITY

DISTANCE METRICS



DISTANCE METRICS

COSINE SIMILARITY

$$Cosine\ similarity(A,B) = \frac{\mathrm{A} * \mathrm{B}}{||\mathrm{A}|| * ||\mathrm{B}||}$$

||A|| and ||B|| - magnitudes (Euclidean norms) of vectors A and B

EUCLIDEAN DISTANCE

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

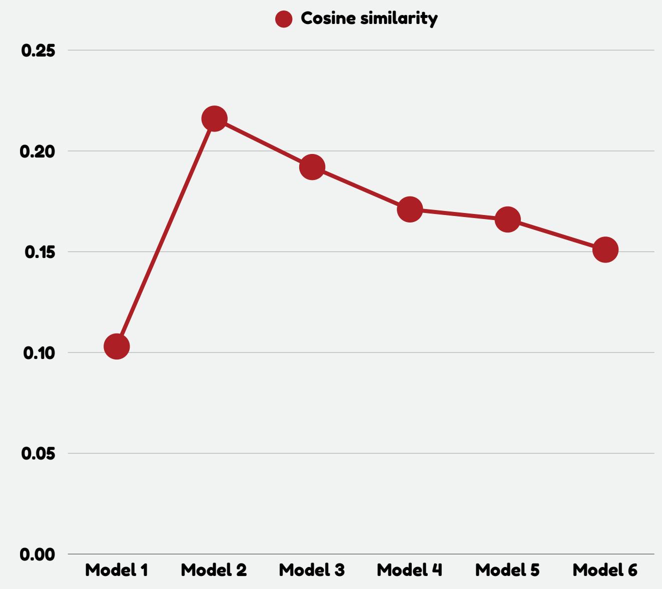
JACCARD SIMILARITY

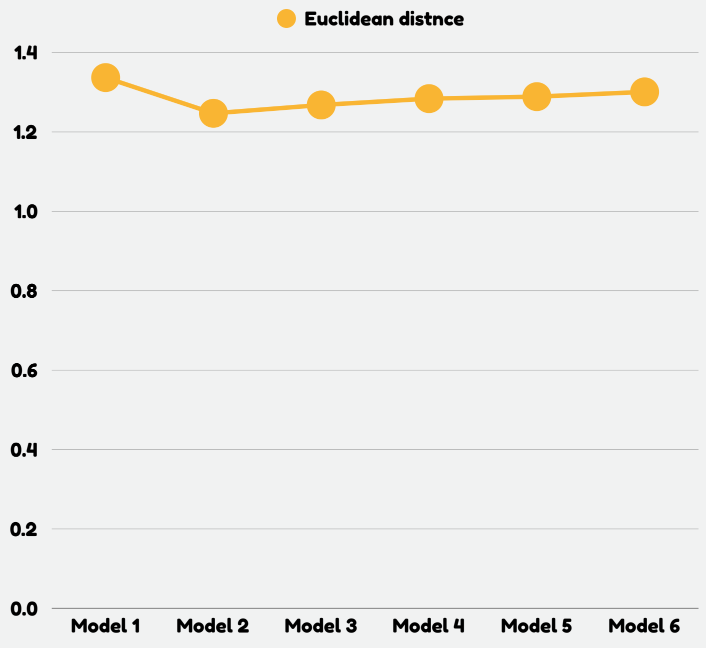
$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

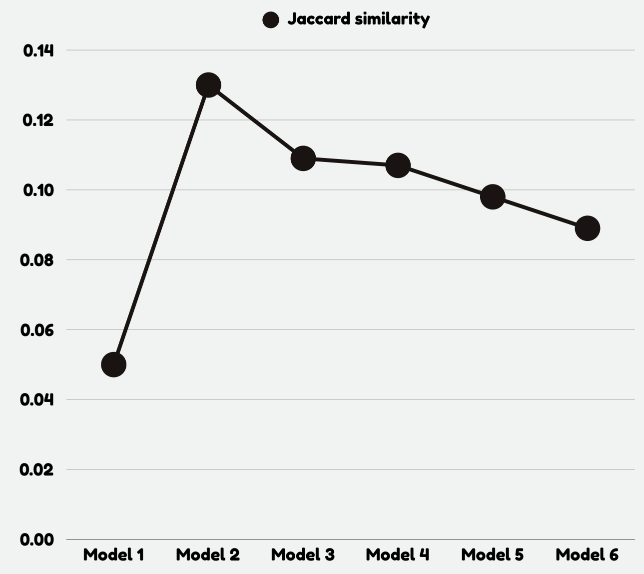
S1 and S2 - sets of words from answers 1 and 2

RESULTS

	Cosine similarity	Euclidean distance	Jaccard similarity
Model 1	0.103	1.337	0.050
Model 2	0.216	1.247	0.130
Model 3	0.192	1.268	0.109
Model 4	0.171	1.284	0.107
Model 5	0.166	1.289	0.098
Model 6	0.151	1.301	0.089







SUMMARY

- Model 2 shows the best distance results
- Such small results indicate that the models’ generated texts are significantly different from the reference texts
- Secondary metrics

NATIONAL RESEARCH UNIVERSITY

RANKING METRICS

RANKING METRICS

MAP@K

$$\frac{1}{Q}\sum_{q=1}^{Q}(\frac{1}{m_q}\sum_{j=1}^{k}(P(j) * rel(j)))$$

NDCG

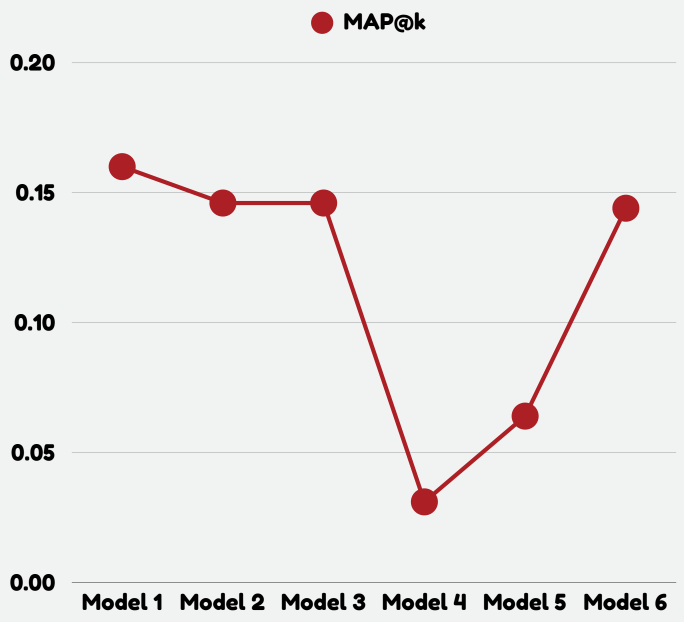
$$NDCG@k = \frac{DCG@k}{IDCG@k}$$
$$= \frac{\sum_{i=1}^{k}\frac{2^{rel_i}-1}{\log_2(i+1)}}{IDCG@k}$$

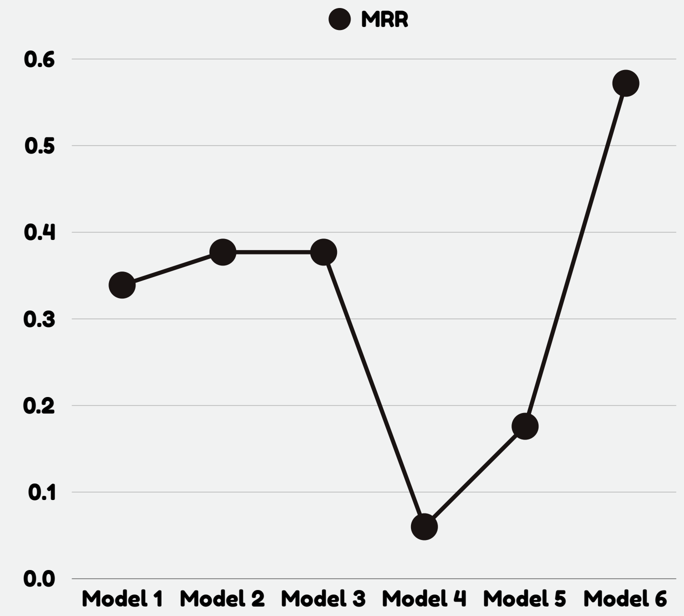
MRR

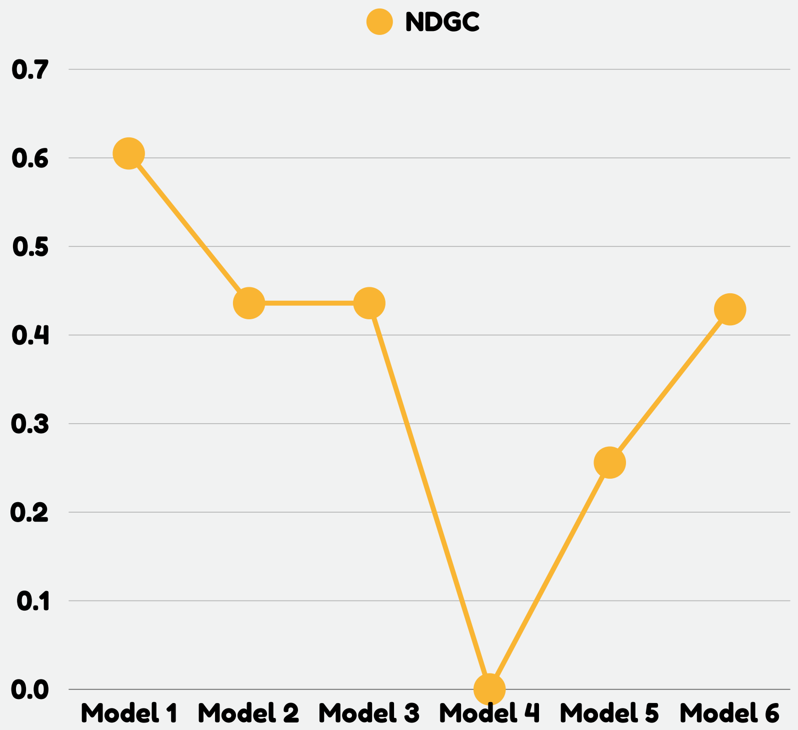
$$MRR = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{rank_q}$$

RESULTS

	MAP@k	NDCG	MRR
Model 1	0.160	0.605	0.339
Model 2	0.146	0.436	0.377
Model 3	0.146	0.436	0.377
Model 4	0.031	0.000	0.060
Model 5	0.064	0.256	0.176
Model 6	0.144	0.429	0.572







SUMMARY

- Model 6 shows better results
- High Reliability, especially MRR and NDCG
- Primary metrics

NATIONAL RESEARCH
UNIVERSITY

COMPLEX METRICS

COMPLEX METRICS

ROUGE

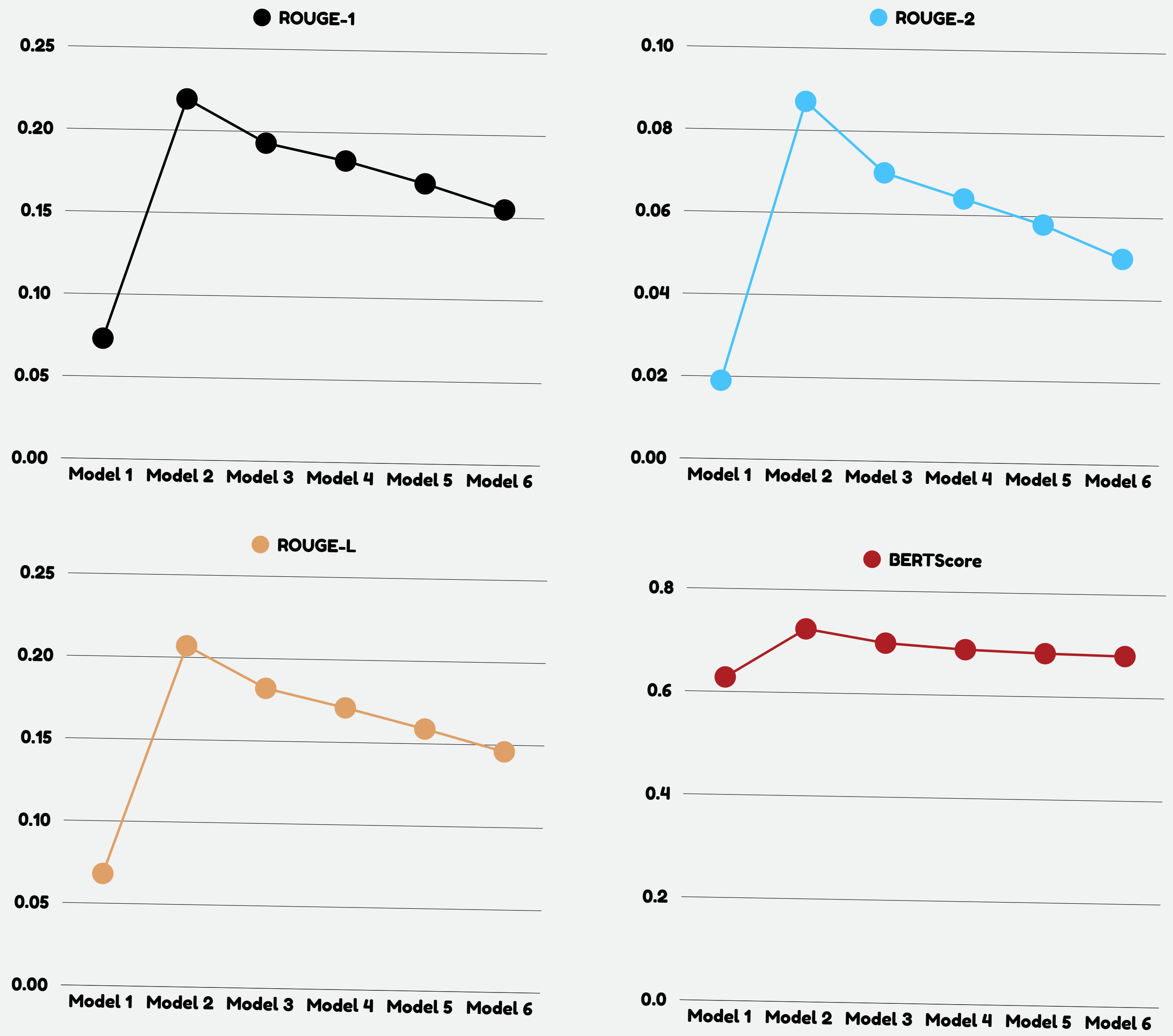
the number of matching n-grams between the model-generated text and a human-produced reference.

BERTSCORE

uses contextual embeddings from BERT to measure the semantic similarity between candidate and reference sentences

RESULTS

F1-score	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Model 1	**0.073**	**0.019**	**0.068**	**0.628**
Model 2	**0.219**	**0.087**	**0.207**	**0.724**
Model 3	**0.193**	**0.070**	**0.182**	**0.699**
Model 4	**0.183**	**0.064**	**0.171**	**0.689**
Model 5	**0.170**	**0.058**	**0.159**	**0.684**
Model 6	**0.155**	**0.050**	**0.146**	**0.681**



SUMMARY

- Model 2 shows the best results
- BERTScore captures the meaning and context of the words, not just their lexical appearance as ROUGE does
- Primary metric = BERTScore
- Secondary metric = ROUGE

NATIONAL RESEARCH UNIVERSITY

LLM-BASED METRICS

LLM-BASED METRICS

CHATGPT-4O

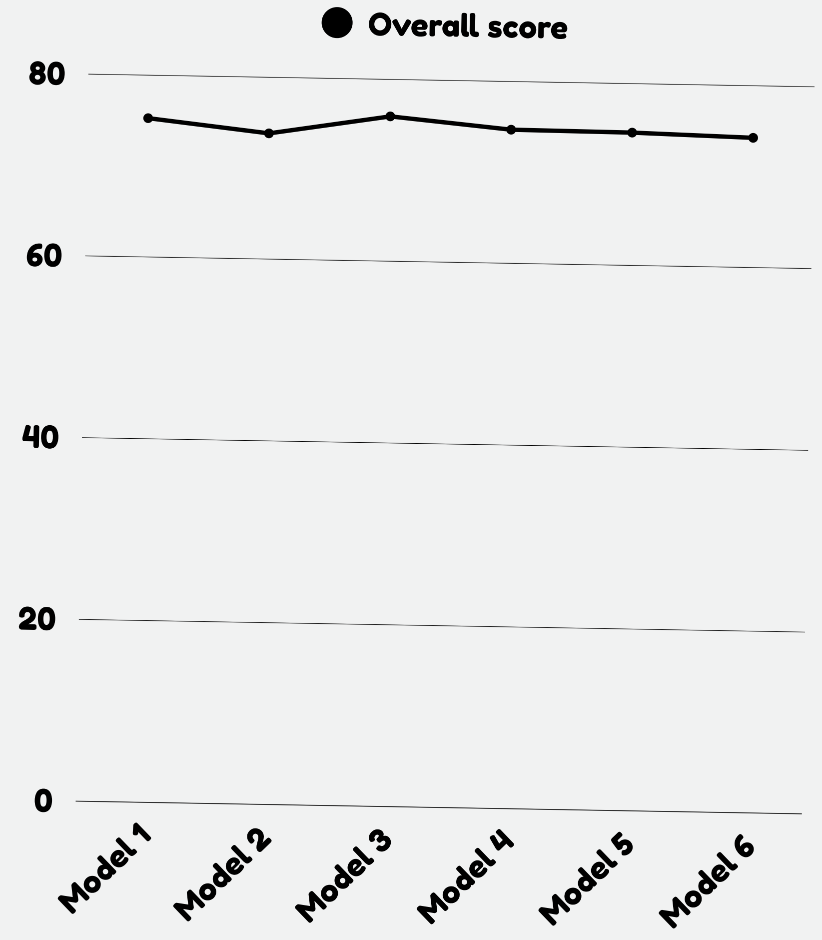
an enhanced version of GPT-4 language model, optimized for specific functionalities and improved performance

YANDEXGPT-3

based on the Transformer architecture and is trained on a huge set of data, including texts, images and other types of information.

RESULTS

ChatGPT-4o	Accuracy	Consistency	Logic	Completeness	Overall
Model 1	74.70	73.64	76.52	76.21	75.27
Model 2	75.16	74.04	72.31	73.81	73.83
Model 3	75.24	77.55	74.84	76.07	75.93
Model 4	74.71	74.66	73.77	75.66	74.70
Model 5	74.78	74.56	74.16	74.93	74.61
Model 6	72.95	75.42	74.67	74.04	74.27



SUMMARY

- Model 3 is thought to be the best
- High Reliability
- Primary metrics

COMPLETE VIEW

	BEST	SECOND BEST	THE WINNER
DISTANCE METRICS	Model 2	Model 3	**Model 2**
RANKING METRICS	Model 6	Model 2 & Model 3	
COMPLEX METRICS	Model 2	Model 3	
LLM-BASED METRICS	Model 3	Model 1	

WHY SUCH RESULTS?

FACTOR 1
Model Architecture

FACTOR 2
Data

FACTOR 3
Metrics Sensitivity

FACTOR 4
Query Expansion

FACTOR 5
Inherent Variability

CONCLUSION

MAP@K, MRR and NDCG are essential for evaluating RAG models as they directly measure the model's ability to retrieve and rank relevant documents accurately

+

BERTScore is also highly valuable for evaluating the quality of generated responses

=>

should be heavily weighted in decision-making processes for selecting RAG models

FURTHER WORK

- work on different prompts of the models and evaluate results
- combine several RAG models
- add re-ranking step in models
- test models on broader range of documents

REFERENCES

- Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 9459-9474.
- Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. – Cambridge university press, 2008.
- Voorhees E. M. et al. The TREC-8 Question Answering Track Evaluation //TREC. – 1999. – Т. 1999. – С. 82.
- Baron J. R., Lewis D. D., Oard D. W. TREC 2006 Legal Track Overview //TREC. – 2006.
- Voorhees E. M., Buckland L. Overview of the TREC 2003 Question Answering Track //TREC. – 2003. – Т. 2003. – С. 54-68.
- Papineni K. et al. Bleu: a method for automatic evaluation of machine translation //Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – С. 311-318.
- Lin C. Y. Rouge: A package for automatic evaluation of summaries //Text summarization branches out. – 2004. – С. 74-81.

NATIONAL RESEARCH UNIVERSITY

THANK YOU