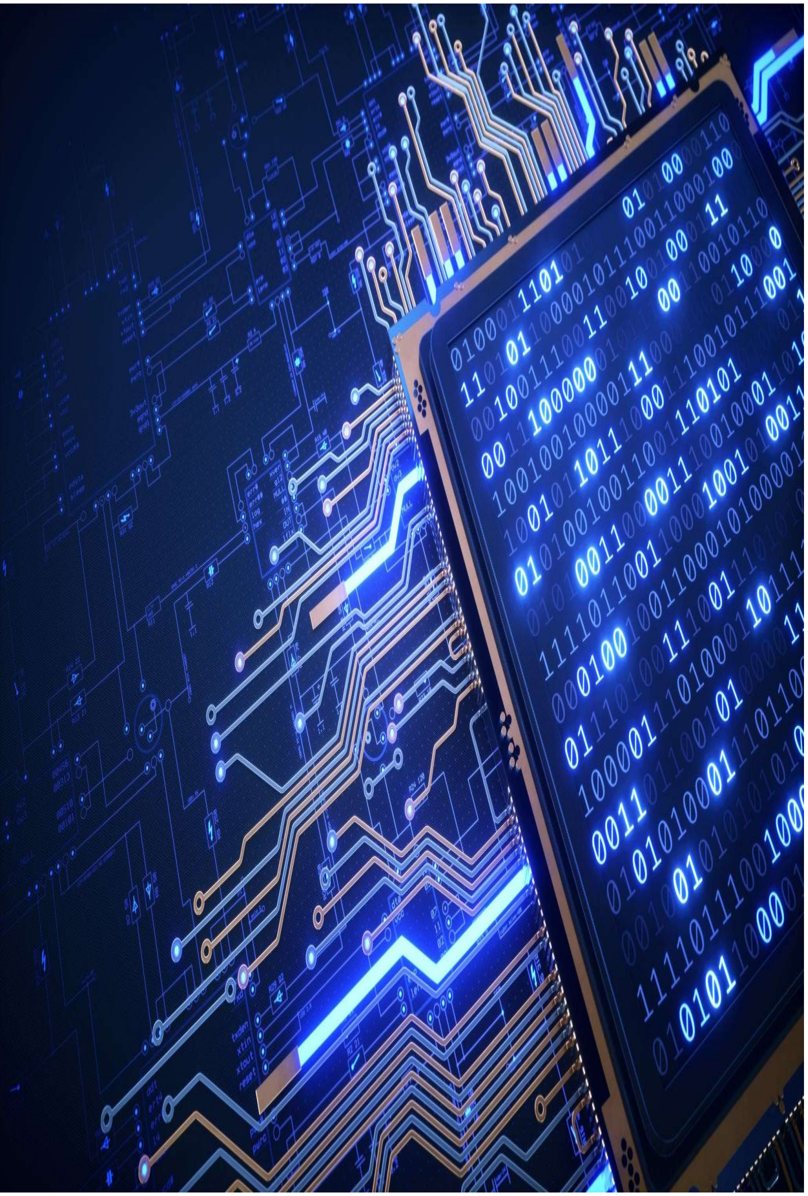


RAG2Rich

Barun Saha

Introduction

Retrieval-Augmented Generation (RAG) enables Large Language Models (LLMs) to generate responses based on private data

- Building an RAG system is easy, optimizing it is necessarily not

TruLens Eval enables systematic evaluation based on its RAG Triad metrics

- LLMs rate the answers and contexts
- So many parameters and score vectors, how to identify the optimal configuration?

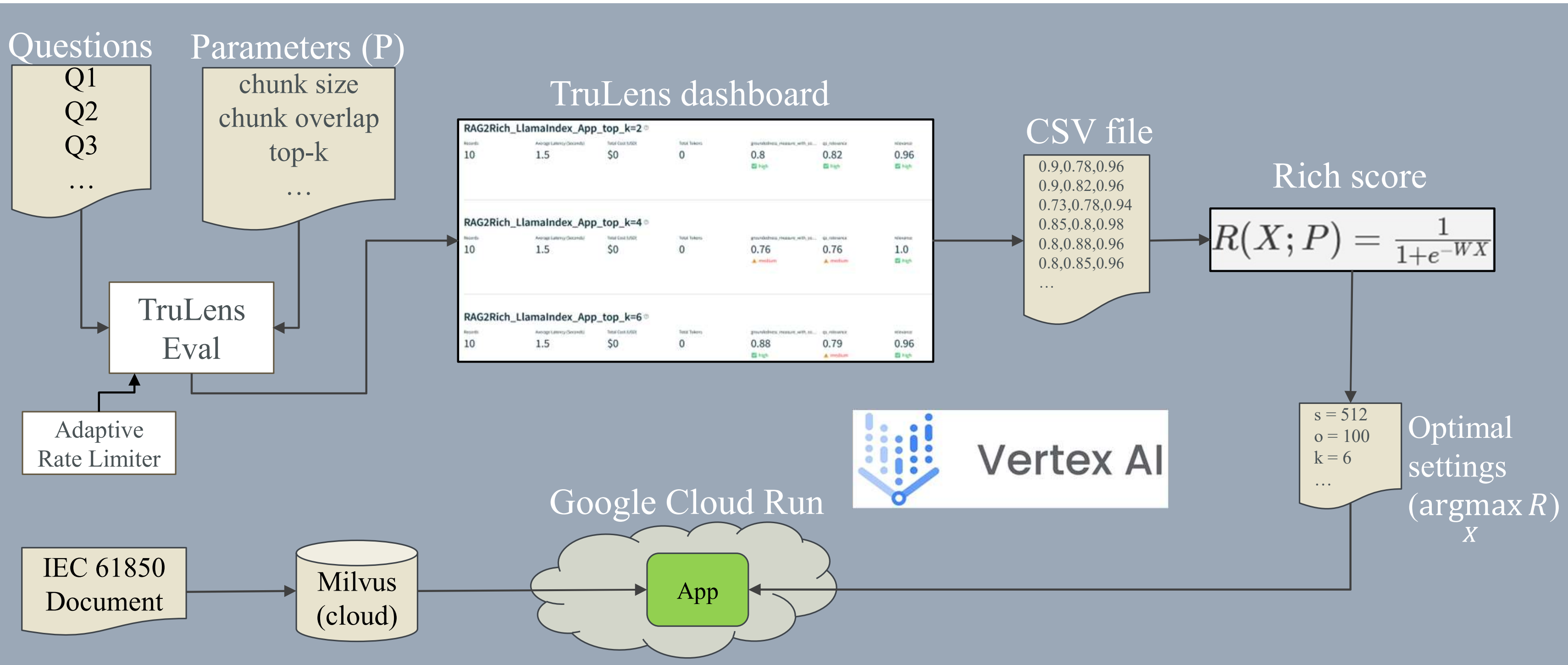
RAG2Rich solves this problem by:

- Designing the “Rich” metric to score the average “richness” of answers
- Using the RAG configuration corresponding to the optimal rich score

A scientific approach to optimize RAG-based systems

A premium subscription may offer parallel and distributed evaluations together with advanced optimizations

Use Case: Knowledge Base for Substation Automation Systems



$$R(X;P) = \frac{1}{1+e^{-WX}}$$

Optimal settings $(\arg\max_{X} R)$

RAG2Rich_LlamaIndex_App_top_k=2

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	groundedness_measure_with_co...	qs_relevance	relevance
10	1.5	$0	0	0.8 high	0.82 high	0.96 high

RAG2Rich_LlamaIndex_App_top_k=4

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	groundedness_measure_with_co...	qs_relevance	relevance
10	1.5	$0	0	0.76 medium	0.76 medium	1.0 high

RAG2Rich_LlamaIndex_App_top_k=6

Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	groundedness_measure_with_co...	qs_relevance	relevance
10	1.5	$0	0	0.88 high	0.79 medium	0.96 high

- X = [Groundedness, context relevance, and answer relevance]; measured using TruLens Eval
- The Rich score is computed using a weighted logistic function, which generates a scalar; W = [2, 3, 1]
- The optimal configuration is used with the RAG application, built with Chainlit, LlamaIndex, Vertex AI, and Cohere

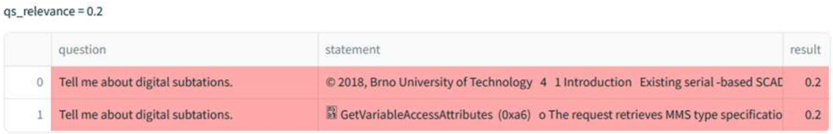
Manual Fine-tuning (1/3)

The third context results in a better response

c_size=512, c_overlap=75, top_k=**2**

Q: Tell me about digital subtations.

A: Digital substations leverage communication protocols that utilize Ethernet and IP, such as the IEC 61850 standard. This international standard defines communication protocols for intelligent electronic devices at electrical substations. It enables modernized substation communication and takes advantage of Ethernet and IP capabilities.

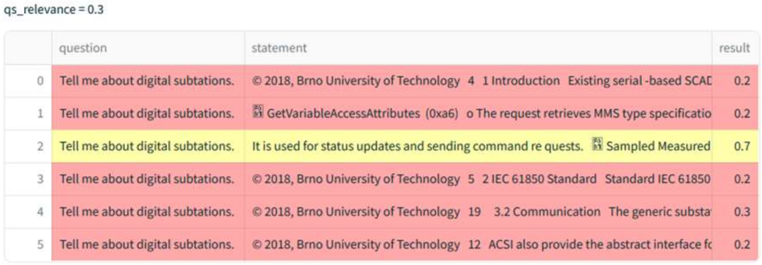


qs_relevance = 0.2

	question	statement	result
0	Tell me about digital subtations.	© 2018, Brno University of Technology 4 1 Introduction Existing serial -based SCAD	0.2
1	Tell me about digital subtations.	GetVariableAccessAttributes (0xa6) o The request retrieves MMS type specificatio	0.2

c_size=512, c_overlap=75, top_k=**6**

A: Digital substations utilize IEC 61850, an international standard defining communication protocols for intelligent electronic devices at electrical substations. This standard enables the integration of protection, control, measurement, and monitoring functions through a common protocol. It facilitates high-speed substation applications, station-wide interlocking, and other functions requiring intercommunication between intelligent electronic devices (IEDs). The well-described data modeling and specified communication services make IEC 61850 a key element in modern substation systems.



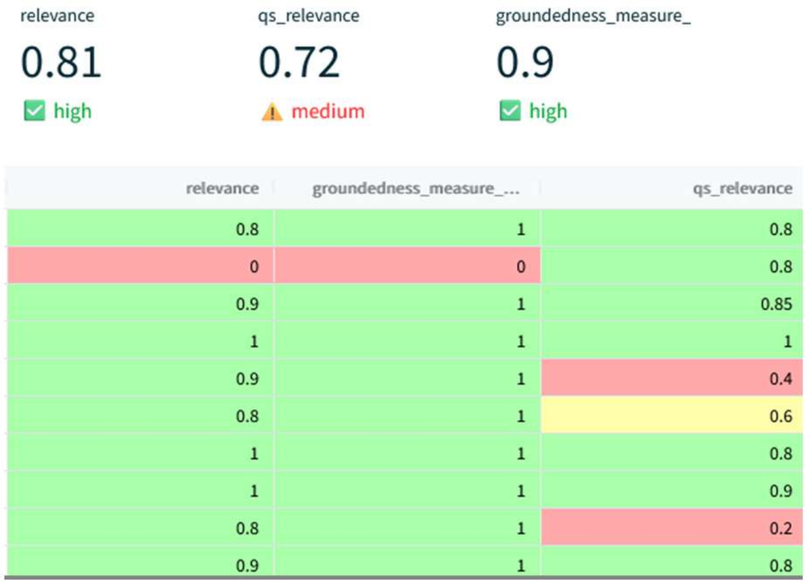
qs_relevance = 0.3

	question	statement	result
0	Tell me about digital subtations.	© 2018, Brno University of Technology 4 1 Introduction Existing serial -based SCAD	0.2
1	Tell me about digital subtations.	GetVariableAccessAttributes (0xa6) o The request retrieves MMS type specificatio	0.2
2	Tell me about digital subtations.	It is used for status updates and sending command re quests. Sampled Measured	0.7
3	Tell me about digital subtations.	© 2018, Brno University of Technology 5 2 IEC 61850 Standard Standard IEC 61850	0.2
4	Tell me about digital subtations.	© 2018, Brno University of Technology 19 3.2 Communication The generic substa	0.3
5	Tell me about digital subtations.	© 2018, Brno University of Technology 12 ACSI also provide the abstract interface f	0.2

Manual Fine-tuning (2/3)

Both groundedness and context relevance decreases when top_k = 3

c_size=512, c_overlap=75, top_k=**2**



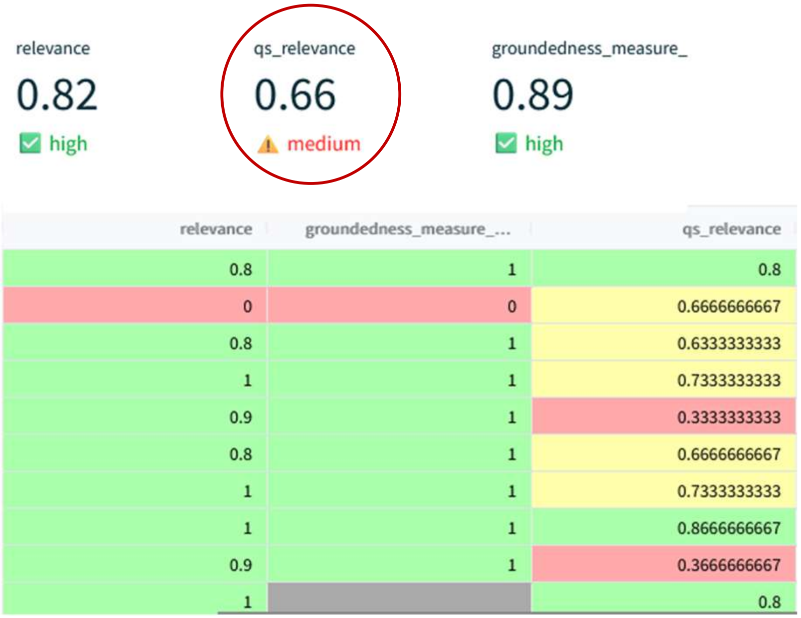
relevance
0.81
☑ high

qs_relevance
0.72
⚠ medium

groundedness_measure_
0.9
☑ high

relevance	groundedness_measure_...	qs_relevance
0.8	1	0.8
0	0	0.8
0.9	1	0.85
1	1	1
0.9	1	0.4
0.8	1	0.6
1	1	0.8
1	1	0.9
0.8	1	0.2
0.9	1	0.8

c_size=512, c_overlap=75, top_k=**3**



relevance
0.82
☑ high

qs_relevance
0.66
⚠ medium

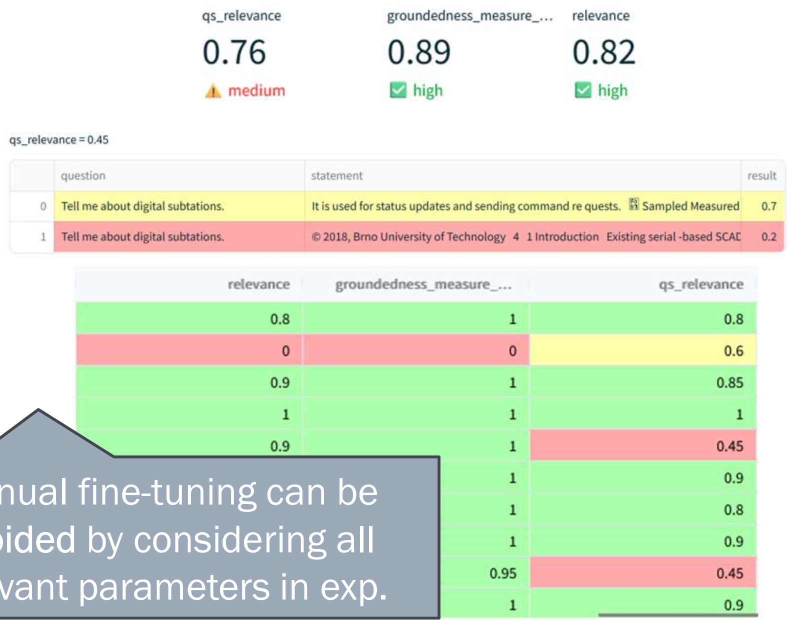
groundedness_measure_
0.89
☑ high

relevance	groundedness_measure_...	qs_relevance
0.8	1	0.8
0	0	0.6666666667
0.8	1	0.6333333333
1	1	0.7333333333
0.9	1	0.3333333333
0.8	1	0.6666666667
1	1	0.7333333333
1	1	0.8666666667
0.9	1	0.3666666667
1		0.8

Manual Fine-tuning (3/3)

(Cohere) Reranking improves the context relevance; marginal decrease in groundednes

c_size=512, c_overlap=75, top_k=**3**, top_n=2



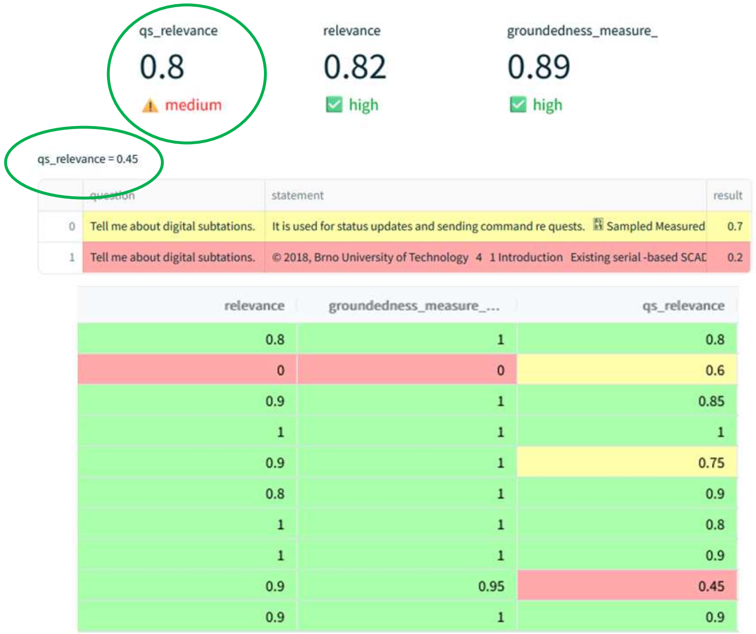
qs_relevance	groundedness_measure_...	relevance
0.76	0.89	0.82
⚠ medium	☑ high	☑ high

qs_relevance = 0.45

	question	statement	result
0	Tell me about digital subtations.	It is used for status updates and sending command re quests. Sampled Measured	0.7
1	Tell me about digital subtations.	© 2018, Brno University of Technology 4 1 Introduction Existing serial -based SCAD	0.2

relevance	groundedness_measure_...	qs_relevance
0.8	1	0.8
0	0	0.6
0.9	1	0.85
1	1	1
0.9	1	0.45
	1	0.9
	1	0.8
	1	0.9
	0.95	0.45
	1	0.9

c_size=512, c_overlap=75, top_k=**4**, top_n=2



qs_relevance	relevance	groundedness_measure_
0.8	0.82	0.89
⚠ medium	☑ high	☑ high

qs_relevance = 0.45

	question	statement	result
0	Tell me about digital subtations.	It is used for status updates and sending command re quests. Sampled Measured	0.7
1	Tell me about digital subtations.	© 2018, Brno University of Technology 4 1 Introduction Existing serial -based SCAD	0.2

relevance	groundedness_measure_...	qs_relevance
0.8	1	0.8
0	0	0.6
0.9	1	0.85
1	1	1
0.9	1	0.75
0.8	1	0.9
1	1	0.8
1	1	0.9
0.9	0.95	0.45
0.9	1	0.9

Manual fine-tuning can be avoided by considering all relevant parameters in exp.

Challenges & Future Directions

The “best-effort,” adaptive rate limiter addresses the LLM call rates to some extent by adding delay

- Use multiple LLMs from different service providers running in parallel?

Answer relevance and groundedness measures are zero when no result is found

- Affects the overall average

Complement the evaluations with a human-generated baseline

Scale up the number of documents

Support multimodal contents, such as images and tables

Demonstration

https://rag2rich-vc4smj6o3q-uc.a.run.app/

Thank you!