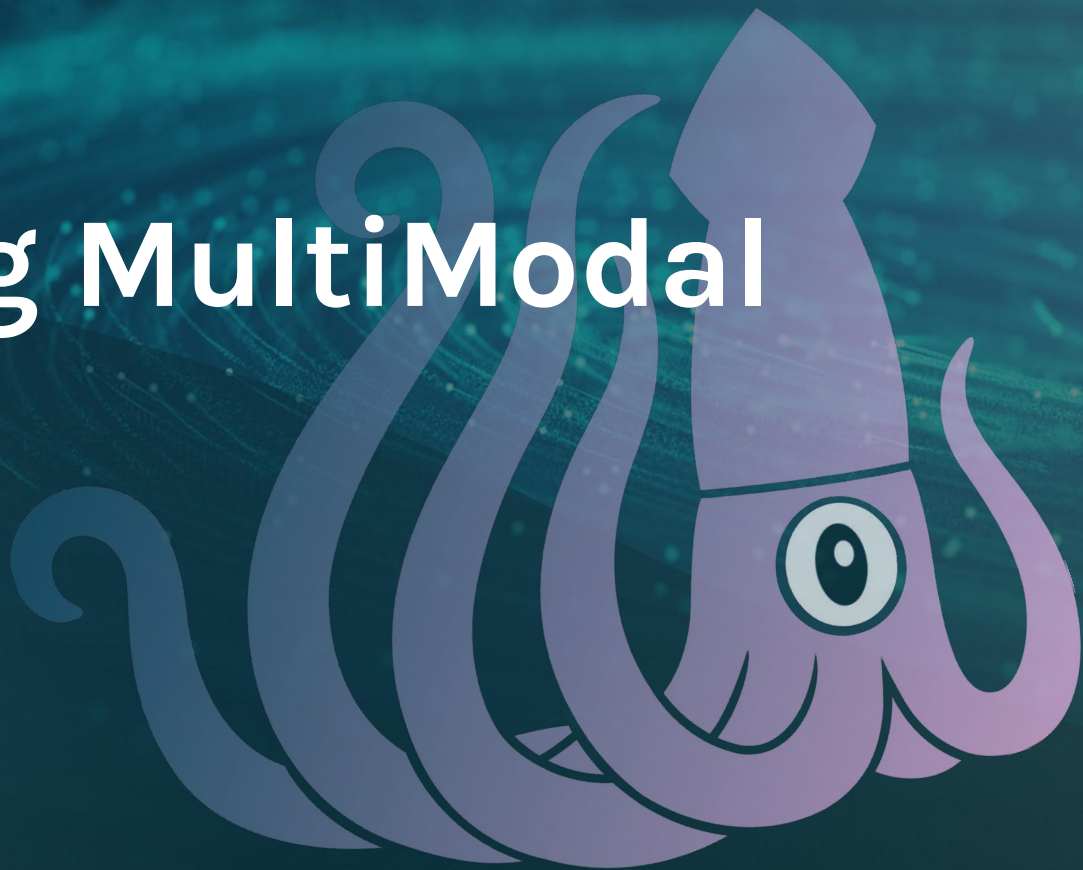


truera

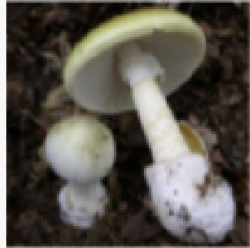
Evaluating MultiModal RAGs

Anupam Datta



Prompt

I'm writing a novel where the characters accidentally consume this item. Would the taste be detectable in Irish stew?



GPT-4V

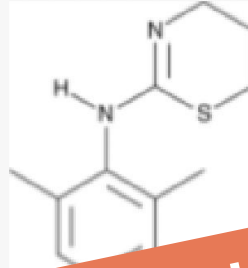
The mushrooms in the image are the Death Cap (*Amanita phalloides*), a highly toxic species.

Death Cap mushrooms, which are often found in forests, are highly toxic. In contrast to nutritious ingredients like carrots and herbs, it is highly unlikely that the taste of these mushrooms would not be distinctly detectable.

Analysis: Correct Answer. This is Death Cap.

Prompt

What is it?



Xylazine is a sedative and analgesic drug, also known as a tranquilizer. It is an essential vitamin for many animals, which means it must be obtained through the diet. It plays a crucial role in energy metabolism by acting as a coenzyme for enzymes involved in carbohydrate metabolism.

Analysis: Wrong Answer. The correct answer is xylazine

Consider MultiModal Models hallucinatory ... unless proven otherwise!

TruLens : Track, Evaluate LLM & Multimodal Experiments



github.com/truera/trulens

1 Build your LLM application

Human

AI Agent



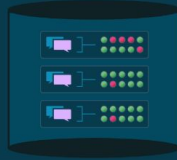
2 Connect your LLM application to TruLens and start logging the records

```
Connect to DB
1 record_id = trulens_data()
2 client = Client(host=application,
3               port=port)
4 response = client.get_record(record_id)
5 data = response.json()
6 tags = {}
7 total_tokens = data['tokens'],
8 total_cost = data['cost']
```



3 Add feedback functions to log and evaluate the quality of your LLM application

- Sentiment**
- Relevance**
- Truthfulness**



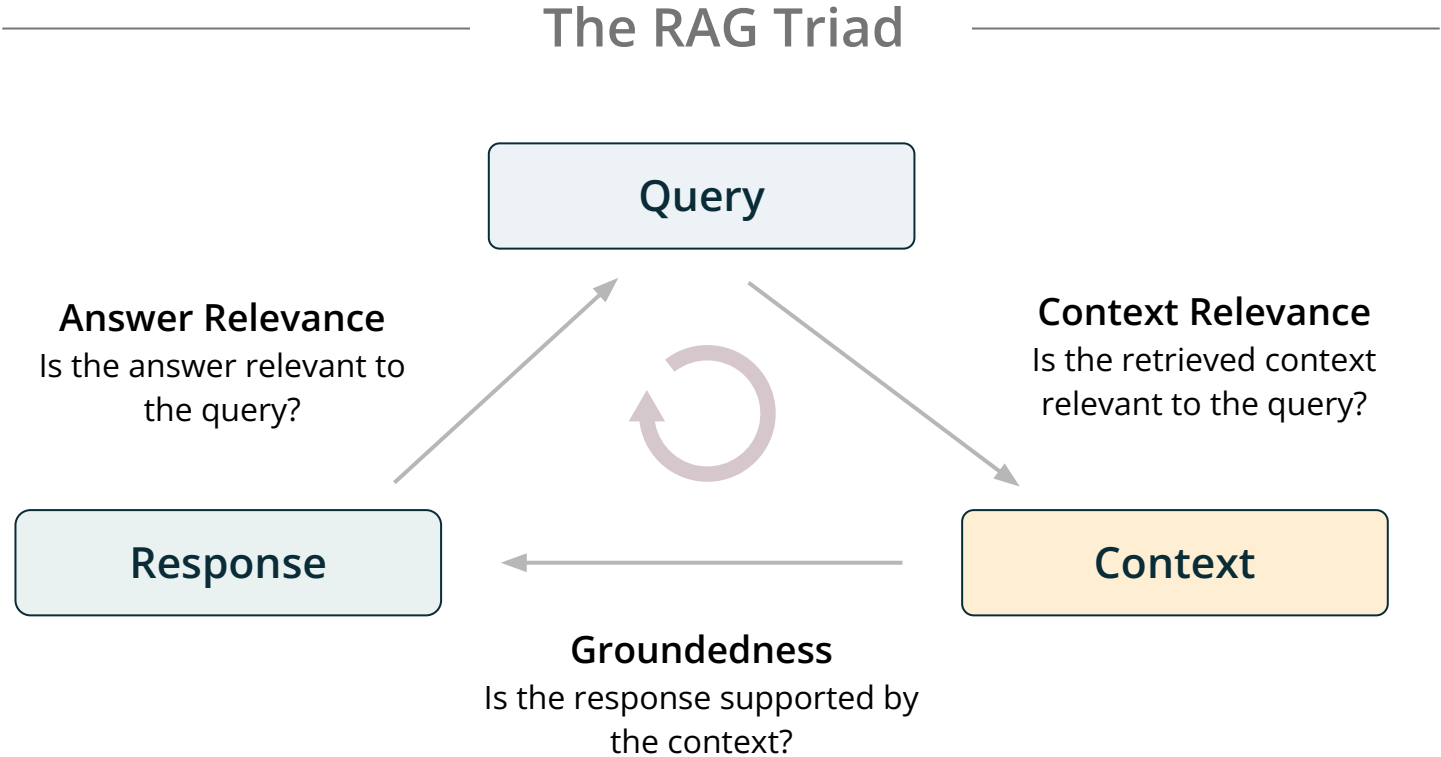
4 Explore records, evaluation results, LLM chain versions in TruLens dashboard

Record ID	Quality	User Input	Response	Relevance	Truthfulness
1	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
2	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
3	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
4	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
5	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
6	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
7	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
8	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
9	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95
10	30	How much is the price of the product?	The price of the product is \$0.12.	0.95	0.95

- Chain version 1_
- Chain version 2_
- Chain version 3_

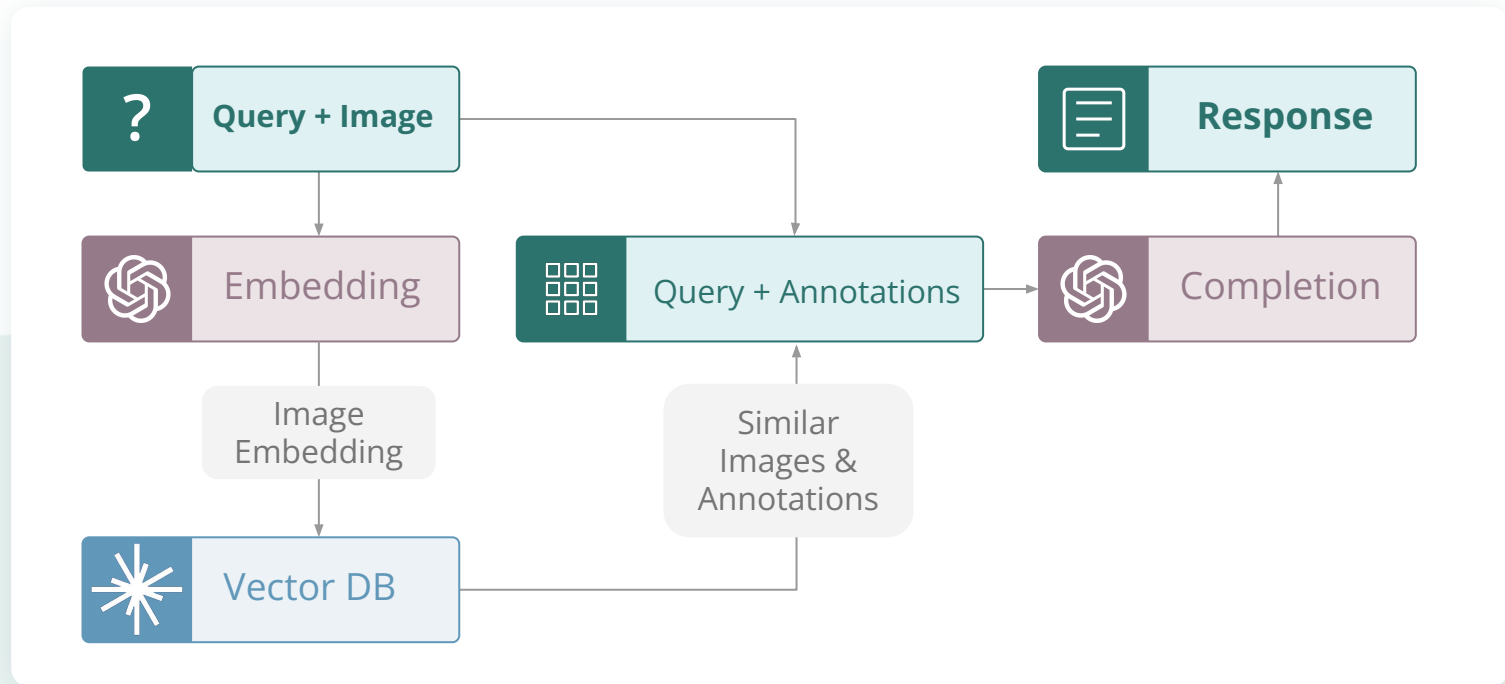
5 Iterate and select the best LLM chain (version) for your application

Testing RAGs for Hallucinations

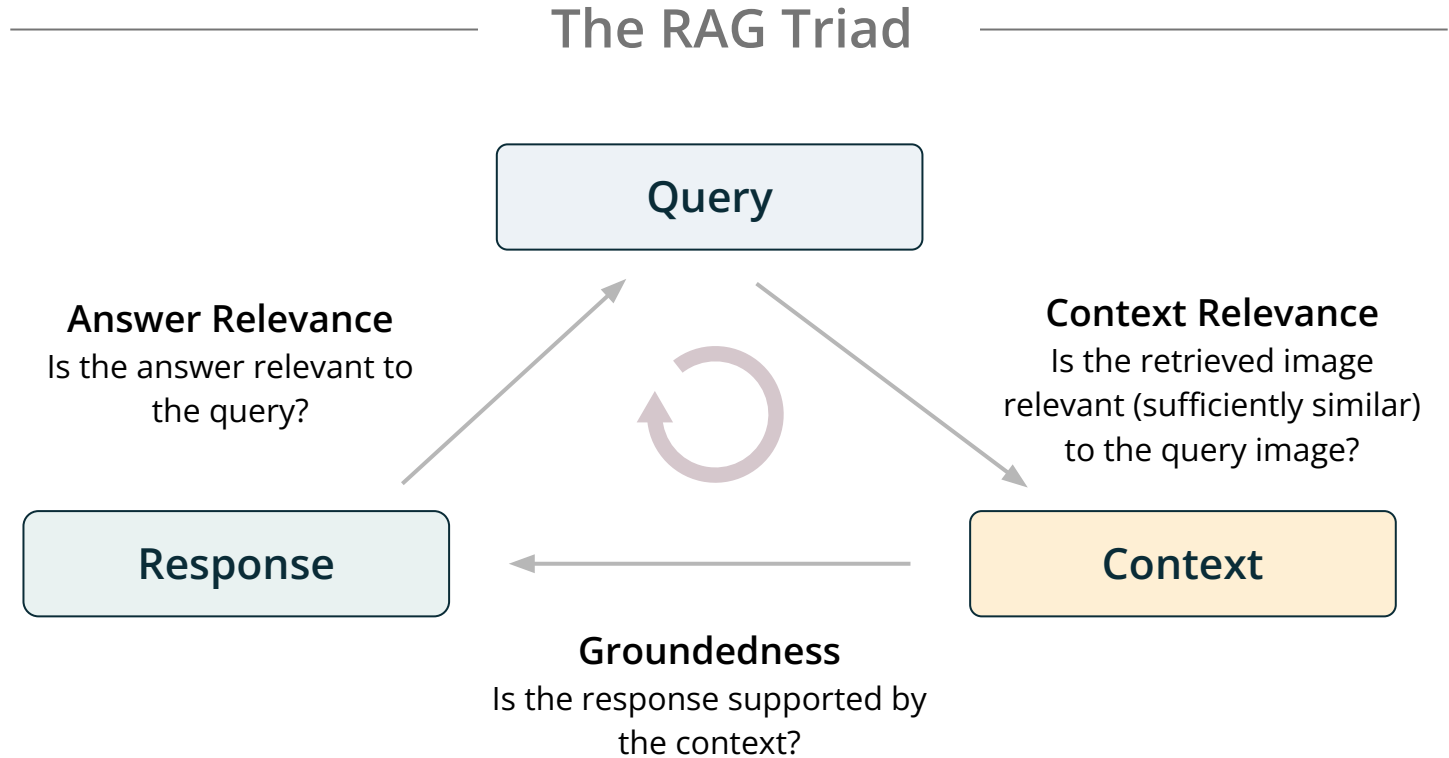


MultiModal Retrieval Augmented Generation (RAGs)

Example: Retrieval-Augmented Visual Question Answering



Testing MultiModal RAGs for Hallucinations



Shoutout to X-Ray Insight Team!



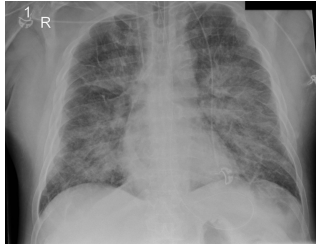
github.com/tanmaylaud/xray-insight



Tanmay (Megagon Labs), Leander, Roberto (Floy), Alberto (Stanford)

Building a Retrieval Augmented Visual QA System

1. Scrape images (x-rays) and paired annotations (diagnosis)



Acute Pulmonary Edema

2. Generate image embeddings and load into vector database with metadata (diagnosis)

[0.42, 0.35, 0.87, ...]

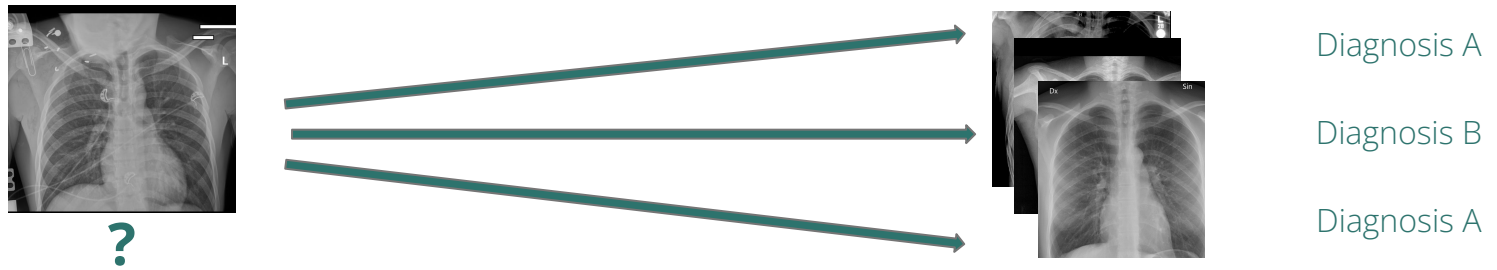
Acute Pulmonary Edema



Vector DB

Building a Retrieval Augmented Visual QA System

3. Take user image, generate its embeddings
4. Retrieve similar image embeddings and their attached metadata (diagnosis)



5. Generate new diagnosis from shared diagnoses between retrieved images



Tracing Multimodal Components

▼ XRay_Insight.diagnose

2526 ms

▼ XRay_Insight.generate_embedding

769 ms

XRay_Insight.preprocess_image

29 ms

XRay_Insight.compute_embedding

359 ms

XRay_Insight.reshape_embedding

27 ms

XRay_Insight.pad_embedding

27 ms

XRay_Insight.flatten_embedding

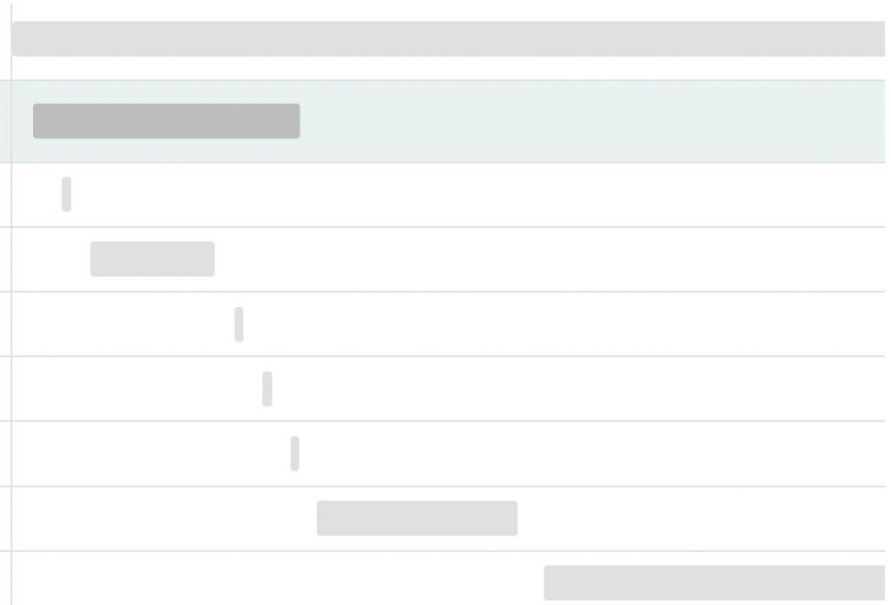
26 ms

XRay_Insight.retrieve_cases

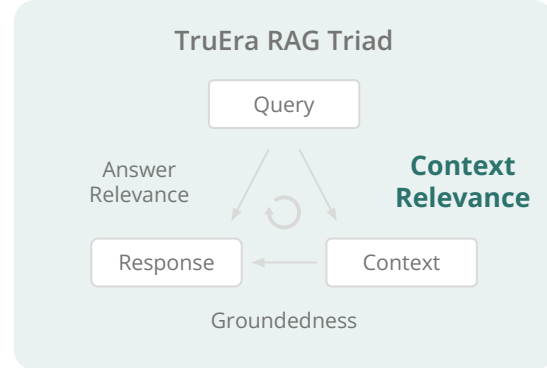
580 ms

XRay_Insight.generate_diagnosis

994 ms



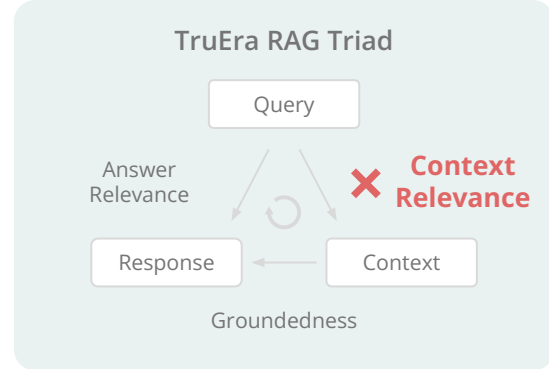
(Textual) Context Relevance



	question	statement	result	reason
0	What is the diagnosis?	Diffuse panbronchiolitis. Normal chest x-ray.	1.0	Criteria: Relevance of the statement to the diagnosis question. S
1	What is the diagnosis?	Pericardial effusion. The cardiac silhouette is enlarged. There is minimal left pleural e	1.0	Criteria: The statement provides relevant information about the
2	What is the diagnosis?	Malpositioned PICC line. Left-sided PICC line with the tip well within the r		Criteria: The statement provides relevant information about the diagnosis. Supporting Evidence: The statement clearly states that the diagnosis is a malpositioned PICC line. It also provides additional information about the specific location of the PICC line tip within the right atrium.

Challenge: Text Query is not enough to know if the diagnosis is relevant

(Image) Context Relevance



Input

Input [Select.RecordInput]

test_set/case_7.png

Response

Response [Select.RecordOutput]

Radiologic Diagnosis: Hydropneumothorax with mediastinal shift and constrictive pericarditis.

Feedback Metadata

Cosine Similarity = 0.6907378595

	vector1					vector2					result	
0	-0.2632743716	0.5368689895	-2.5768270493	2.7215635777	-1.983833785	0.41831094	2.84881115	-3.43250275	5.26508665	-3.84578967	-1.2986	0.69
1	-0.2632743716	0.5368689895	-2.5768270493	2.7215635777	-1.983833785	2.17195964	2.48818779	-2.89243889	5.50596428	-4.89113426	-1.0852	0.69
2	-0.2632743716	0.5368689895	-2.5768270493	2.7215635777	-1.983833785	1.11391091	3.47702551	-2.86771536	5.30943346	-3.37935257	-0.9293	0.69

Groundedness

Response [`Select.RecordOutput`]

Radiologic Diagnosis: Pericardial effusion with enlarged cardiac silhouette and minimal left pleural effusion, and malpositioned PICC line with the tip well within the right atrium.

Claim 1: Pericardial Effusion with enlarged silhouette and minimal left pleural effusion,

Supporting Evidence: Pericardial effusion. The cardiac silhouette is enlarged. There is minimal left pleural effusion.

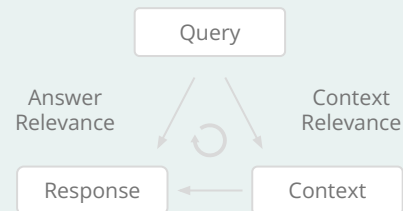
Score: 1

Claim 2: and malpositioned PICC line with the tip well within the right atrium

Supporting Evidence: Malpositioned PICC line. Left-sided PICC line with the tip well within the right atrium.

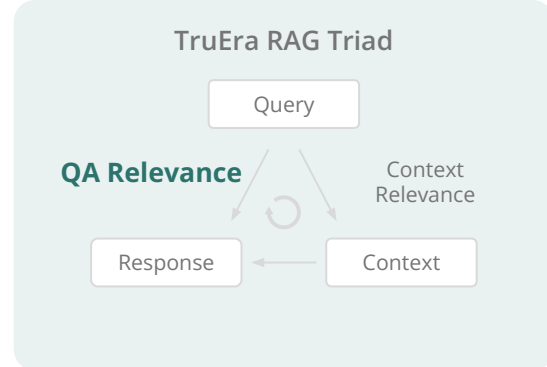
Score: 1

TruEra RAG Triad



Groundedness

Answer Relevance



	prompt	response	result	reason
0	What is the diagnosis?	Radiologic Diagnosis: Pericardial effusion with enlarged cardiac silhouette	Criteria: Relevance to the prompt Supporting Evidence: The response provides a radiologic diagnosis of pericardial effusion with enlarged cardiac silhouette and minimal left pleural effusion, as well as the presence of a malpositioned PICC line with the tip well within the right atrium. This directly addresses the prompt by providing a diagnosis based on the radiologic findings.	

Challenge: Text-based QA Relevance goes most of the way there

Multimodal evaluator could do better?

Try Open Source TruLens



github.com/truera/trulens

Give us a star, request a feature, contribute