



# LLM Retrieval Evaluations: Advanced Techniques



**Jerry Liu**  
LlamaIndex  
Co-Founder and CEO

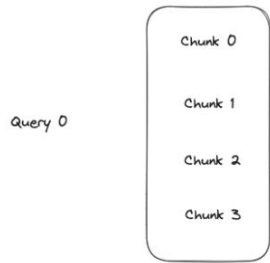


**Jason Lopatecki**  
Arize AI  
Co-Founder and CEO

# Retrieval Eval

## Retrieval Eval per Chunk

### Is Chunk Relevant?



"Chunk 0"

You are comparing a reference text to a question and trying to determine if the reference text contains information relevant to answering the question. Here is the data:

```
[BEGIN DATA]
*****
[Question]: {query}
*****
[Reference text]: {reference}
[END DATA]
```

Compare the Question above to the Reference text. You must determine whether the Reference text contains information that can answer the Question. Please focus on whether the very specific question can be answered by the information in the Reference text. Your response must be single word, either "relevant" or "irrelevant", and should not contain any text or characters aside from that word. "irrelevant" means that the reference text does not contain an answer to the Question. "relevant" means the reference text contains an answer to the Question.

### Relevance Eval

"Chunk 0"

1

"Chunk 1"

1

"Chunk 2"

1

"Chunk 3"

0

MRR

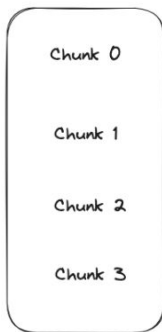
Precision @ k

# Q&A Eval

## Question & Answer Eval

Is Answer Correct?

Query 0



Answer

"Chunk 0"  
+  
"Chunk 1"  
+  
"Chunk 2"  
+  
"Chunk 3"

Concatenated String



You are given a question, an answer and reference text. You must determine whether the given answer correctly answers the question based on the reference text. Here is the data:

```
[BEGIN DATA]
*****
[Question]: {question}
*****
[Reference]: {context}
*****
[Answer]: {sampled_answer}
[END DATA]
```

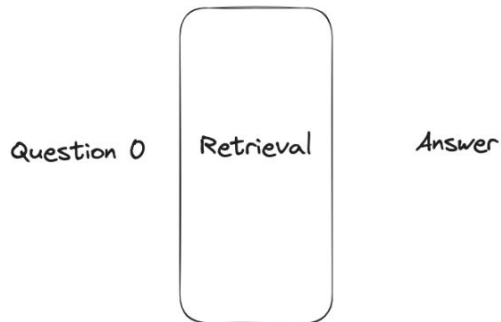
Your response must be a single word, either "correct" or "incorrect", and should not contain any text or characters aside from that word. "correct" means that the question is correctly and fully answered by the answer. "incorrect" means that the question is not correctly or only partially answered by the answer.

# Context versus Relevance

Query	Context Returned 1	Context Returned 2	Relevance 1	Relevance 2	Response	Q&A Eval
Do you need a prediction ID for the training set?	A `prediction ID` is an ID that indicates a unique prediction event. A prediction ID is <b>required</b> to connect predictions with delayed actuals (ground truth). Learn how to send delayed (latent) actuals here.	Ensure Training and Validation records must include <b>both</b> prediction and actual columns	Relevant ●	Relevant ●	Yes, a prediction ID is required for the training set. ●	Correct
How do I configure permissions for GBQ?	There are two ways to setup access permissions with Arize Configure An Individual Bucket Policy Give Arize permission to access individual buckets #configure-an-individual-bucket-policy Configure Multiple Buckets Via Role Based Permissions Assign Arize a role to access multiple buckets using external IDs #configure-multiple-buckets-via-role-based-permissions	1. <b>In Google Cloud console</b> : Navigate to the BigQuery SQL Workspace 2. Select the desired table or view, navigate to the <b>Details</b> tab and click "Edit Details". Under the <b>Labels</b> section, click "Add Labels". Add the following label: * Key as <b>arize-ingestion-key</b> _ _ _ ...	Irrelevant ●	Relevant ●	To configure permissions for Google BigQuery (GBQ), you can either configure an individual bucket policy or assign Arize a role to access multiple buckets using external IDs.	Correct
Can I copy a dashboard?	Templates are designed as starting points for dashboard and model analysis. Once a dashboard is created from a template, it can be edited and customized as desired.	To add a widget, simply: * Click the Edit Dashboard icon in the top right corner. * Select or drag and drop the widget onto an area of the dashboard.	Irrelevant ●	Irrelevant ●	Yes, you can copy a dashboard. To do so, click the Edit Dashboard icon in the top right corner and select the Copy Dashboard option.	Incorrect

# Human versus AI

Is the answer correct as compared against the Human Label?



Human Label

You are comparing a human ground truth answer from an expert to an answer from an AI model. Your goal is to determine if the AI answer correctly matches, in substance, the human answer.

```
[BEGIN DATA]
*****
[Question]: {question}
*****
[Human Ground Truth Answer]: {human_answer}
*****
[AI Answer]: {ai_answer}
*****
[END DATA]
```

Compare the AI answer to the human ground truth answer, if the AI correctly answers the question, then the AI answer is "correct". If the AI answer is longer but contains the main idea of the Human answer please answer "correct". If the AI answer diverges or does not contain the main idea of the human answer, please answer "incorrect".

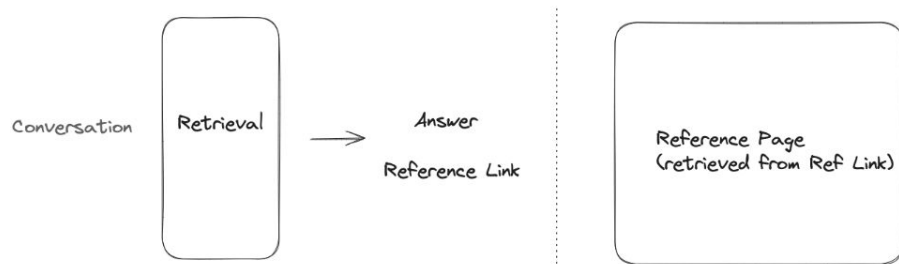
# Human vs AI: Validating Your Eval with Labels

Query	Human - Ground Truth	AI Response	Q&A Eval	Human vs AI
Do you need a prediction ID for the training set?	Yes, a prediction ID is required in the training set	Yes, a prediction ID is required for the training set.	Correct ●	Correct ●
How do I configure permissions for GBQ?	You can either use a bucket policy or use a Role that controls access to the external buckets. ...	To configure permissions for Google BigQuery (GBQ), you can either configure an individual bucket policy or assign Arize a role to access multiple buckets using external IDs.	Correct ●	Correct ●
Can I copy a dashboard?	To add a widget, simply: * Click the Edit Dashboard icon in the top right corner. * Select or drag and drop the widget onto an area of the dashboard.	No, we don't support copying dashboards.	Correct ●	Incorrect ●

Use Human vs AI to confirm your Q&A Eval is correct

# Reference Link Evals

Is the reference Link returned correct?  
Does the Reference Link contain content to Answer the Question?



You are given a conversation that contains questions by a CUSTOMER and you are trying to determine if the documentation page shared by the ASSISTANT correctly answers the CUSTOMER'S questions. We will give you the conversation between the customer and the ASSISTANT and the text of the documentation returned:

[CONVERSATION AND QUESTION]:

{conversation}

\*\*\*\*\*







[DOCUMENTATION URL TEXT]:


{document\_text}

[DOCUMENTATION URL TEXT]:

You should respond "correct" if the documentation text answers the question the CUSTOMER had in the conversation. If the documentation roughly answers the question even in a general way please answer "correct". If there are multiple questions and a single question is answered, please still answer "correct". If the text does not answer the question in the conversation, or doesn't contain information that would allow you to answer the specific question please answer "incorrect".

# Human vs AI: Validating Your Eval with Labels

Query	Ref Link	AI Response	Q&A Eval	Ref Link Eval
Do you need a prediction ID for the training set?	<a href="https://docs.arize.com/arize/sending-data-guides/model-schema-reference#5.-prediction-id">https://docs.arize.com/arize/sending-data-guides/model-schema-reference#5.-prediction-id</a>	Yes, a prediction ID is required for the training set.	Correct 	Correct 
How do I configure permissions for GBQ?	<a href="https://docs.arize.com/arize/sending-data-methods/google-bigquery/google-bigquery-faq">https://docs.arize.com/arize/sending-data-methods/google-bigquery/google-bigquery-faq</a>	To configure permissions for Google BigQuery (GBQ), you can either configure an individual bucket policy or assign Arize a role to access multiple buckets using external IDs.	Correct 	Correct 
Can I copy a dashboard?	<a href="https://docs.arize.com/arize/quickstart#extra-credit-create-a-dashboard">https://docs.arize.com/arize/quickstart#extra-credit-create-a-dashboard</a>	No, we don't support copying dashboards.	Correct 	Incorrect 

 Incorrect link

# Explanations

query	reference	Label	explanation
0 what triggered the civil war	<p>The American Civil War (ACW), also known as the War between the States or simply the Civil War (see naming ), was a civil war fought from 1861 to 1865 between the United States (the "Union" or the "North") and several Southern slave states that declared their secession and formed the Confederate States of America (the "Confederacy" or the "South"). The war had its origin in the issue of slavery , especially the extension of slavery into the western territories. Foreign powers did not intervene. After four years of bloody combat that left over 600,000 soldiers dead and destroyed much of the South's infrastructure, the Confederacy collapsed, slavery was abolished, and the difficult Reconstruction process of restoring national unity and guaranteeing rights to the freed slaves began. In the 1860 presidential election , Republicans, led by Abraham Lincoln , opposed expanding slavery into United States' territories . Lincoln won, but before his inauguration on March 4, 1861, seven cotton-based slave states formed the Confederacy. Outgoing Democratic President James Buchanan and the incoming Republicans rejected secession as illegal. Lincoln's inaugural address declared his administration would not initiate civil war. Eight remaining slave states continued to reject calls for secession. Confederate forces seized numerous federal forts within territory claimed by the Confederacy. A Peace Conference failed to find a compromise, and both sides prepared for war. The Confederates assumed that European countries were so dependent on " King Cotton " that they would intervene; none did and none recognized the new Confederate States of America. Hostilities began on April 12, 1861, when Confederate forces fired upon Fort Sumter , a key fort held by Union troops in South Carolina. Lincoln called for each state to provide troops to retake the fort; consequently, four more slave states joined the Confederacy, bringing their total to eleven. The Union soon controlled the border states and established a naval blockade that crippled the southern economy. The Eastern Theater was inconclusive in 1861–62. The autumn 1862 Confederate campaign into Maryland (a Union state) ended with Confederate retreat at the Battle of Antietam , dissuading British intervention. Lincoln issued the Emancipation Proclamation , which made ending slavery a war goal. To the west, by summer 1862 the Union destroyed the Confederate river navy, then much of their western armies, and the Union at Vicksburg split the Confederacy in two at the Mississippi River. In 1863, Robert E. Lee's Confederate incursion north ended at the Battle of Gettysburg . Western successes led to Ulysses S. Grant command of all Union armies in 1864. In the Western Theater, William T. Sherman drove east to capture Atlanta and marched to the sea , destroying Confederate infrastructure along the way. The Union marshaled the resources and manpower to attack the Confederacy from all directions, and could afford to fight battles of attrition through the Overland Campaign towards Richmond, the Confederate capital. The defending Confederate army failed, leading to Lee's surrender to Grant at Appomattox Court House on April 9, 1865. The American Civil War was one of the earliest true industrial wars . Railroads, the telegraph, steamships, and mass-produced weapons were employed extensively. The mobilization of civilian factories, mines, shipyards, banks, transportation and food supplies all foreshadowed World War I . It remains the deadliest war in American history , resulting in the deaths of an estimated 750,000 soldiers and an undetermined number of civilian casualties. Historian John Huddleston estimates the death toll at ten percent of all Northern males 20–45 years old, and 30 percent of all Southern white males aged 18–40.</p>	relevant	<p>The question asks about what triggered the Civil War. The reference text provides information about the origins of the Civil War, stating that it had its origin in the issue of slavery, especially the extension of slavery into the western territories. It also mentions the 1860 presidential election where Republicans, led by Abraham Lincoln, opposed expanding slavery into United States' territories. Lincoln won, but before his inauguration, seven cotton-based slave states formed the Confederacy. This information directly answers the question, making the reference text relevant.</p>
1 when did classification of races begin	<p>Race is a classification system used to categorize humans into large and distinct populations or groups by anatomical , cultural , ethnic , genetic , geographical , historical , linguistic , religious , or social affiliation. First used to denote national affiliations, the term began to be used to relate to physical traits in the 17th century. In the early 20th century the term was often used, in a taxonomic sense , to denote genetically differentiated human populations defined by phenotype . While biologists sometimes use the concept of race to make distinctions among fuzzy sets of traits, others in the scientific community suggest that the idea of race often is used in a naive or simplistic way, i.e. that among humans, race has no taxonomic significance: all living humans belong to the same species, Homo sapiens and subspecies, Homo sapiens sapiens. Social conceptions and groupings of races vary over time, involving folk taxonomies that define of individuals based on perceived traits. Scientists consider biological essentialism obsolete, and generally discourage racial explanations for collective differentiation in both physical and behavioral traits. Since the second half of the 20th century the associations of race with the ideologies and theories that grew out of the work of 19th-century anthropologists and physiologists has led to the use of the word race itself becoming problematic. Although still used in general contexts, it is now often replaced by other words which are less ambiguous and emotionally charged, such as populations , people (s), ethnic groups or communities depending on context.</p>	relevant	<p>The question asks about when the classification of races began. The reference text provides information that the term 'race' began to be used to relate to physical traits in the 17th century. This directly answers the question, making the reference text relevant.</p>
2 how much is a cord of wood	<p>A cord of wood The cord is a unit of measure of dry volume used in Canada and the United States to measure firewood and pulpwood . A cord is the amount of wood that, when "ranked and well stowed" (arranged so pieces are aligned, parallel, touching and compact), occupies a volume of . This corresponds to a well stacked woodpile high, long, and deep; or any other arrangement of linear measurements that yields the same volume. The name cord probably comes from the use of a cord or string to measure it.</p>	relevant	<p>The question asks for the amount of a cord of wood. The reference text provides a detailed explanation of what a cord of wood is, including its volume and how it is measured. Therefore, the reference text is relevant to the question.</p>

# Statistics

---

# Latest Phoenix Eval Benchmarks

Evals (F1 Scores)	GPT-4 Turbo	GPT-4	GPT 3.5 Turbo	Claude	Palm 2
<i>Retrieval Evals</i>	0.78	0.78	0.59	0.34	0.69
<i>Hallucination Evals</i>	0.82	0.81	0.75	0.87	0.61
<i>Q&amp;A Evals</i>	0.98	0.96	0.83	0.78	0.97
<i>Toxicity Evals</i>	0.83	0.91	0.87	0.54	Toxic content blocker not usable for Eval
<i>Summarization Evals</i>	0.76	0.83	0.18	0.67	0.63
<i>Code Generation Evals</i>	0.83	0.85	0.85	0.31	0.85
<i>Reference Link Evals</i>	0.79	0.89	0.58	0.58	0.80

# Do Functions and Explanations Affect Results?

without\_function\_calling the ordinary prompt completion  
 with\_function\_calling asks the LLM to put its answer in a JSON object that only accepts enum as input  
 with\_explanation asks the LLM to provide explanation alongside its answer in the same JSON object

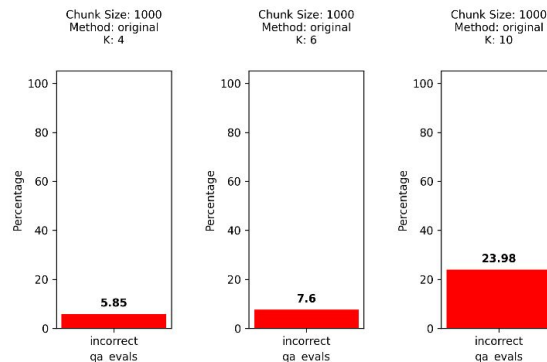
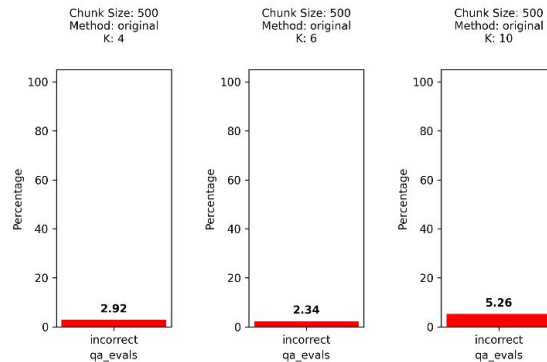
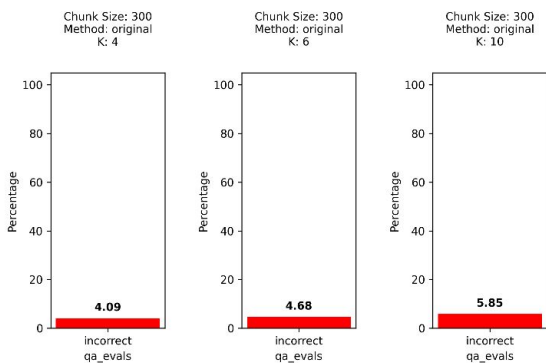
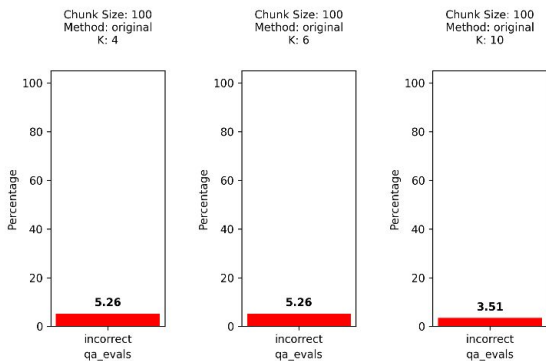
## Relevance

		N=100		N=61			N=39		
		Median Process Time (ms)	Accuracy	Irrelevant			Relevant		
				Precision	Recall	F1	Precision	Recall	F1
gpt-4	without_function_calling	922	77%	93%	67%	78%	64%	92%	76%
	with_explanation	7,810	76%	88%	70%	78%	65%	85%	73%
	with_function_calling	1,818	74%	87%	67%	76%	62%	85%	72%
gpt-3.5-turbo	without_function_calling	233	44%	100%	8%	15%	41%	100%	58%
	with_explanation	1,439	46%	100%	11%	21%	42%	100%	59%
	with_function_calling	516	46%	89%	13%	23%	42%	97%	58%
gpt-3.5-turbo-instruct	without_function_calling	120	39%	0%	0%	0%	39%	100%	56%

## Hallucination

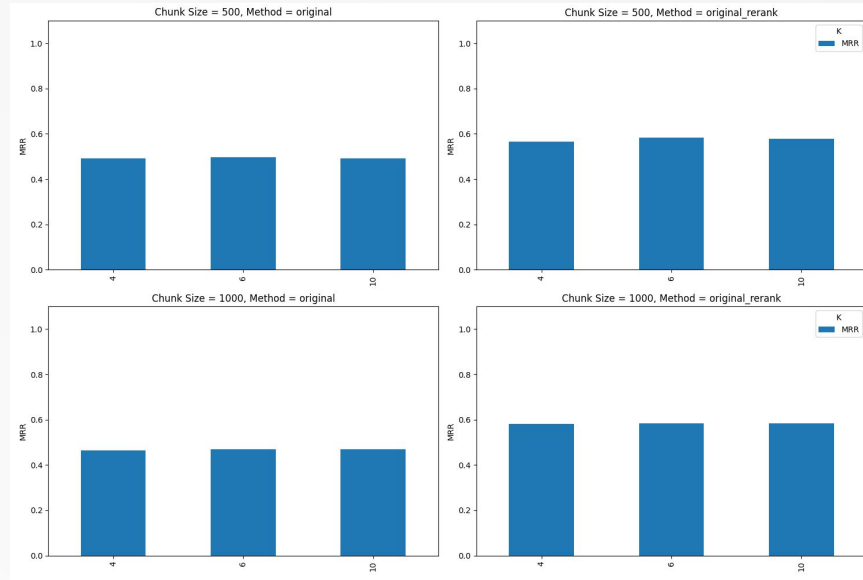
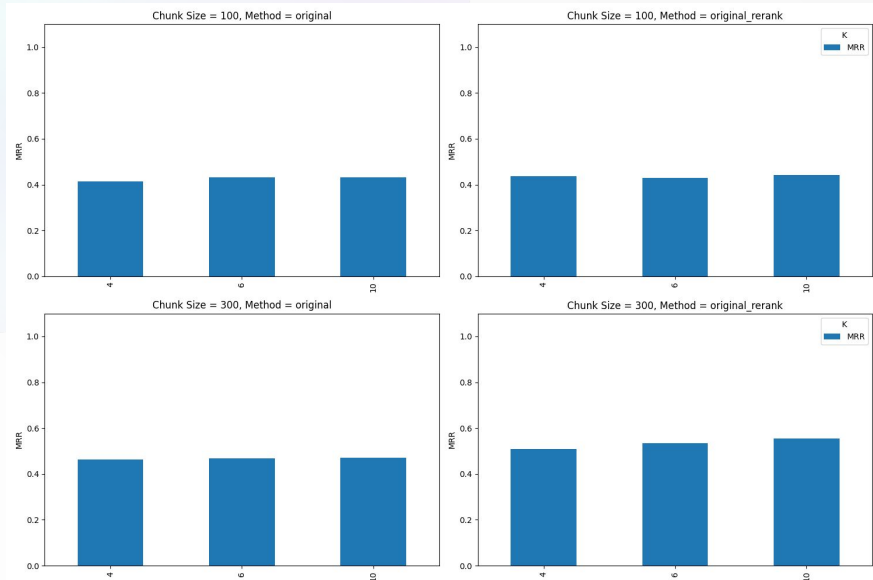
		N=100		N=48			N=52		
		Median Process Time (ms)	Accuracy	Hallucinated			Factual		
				Precision	Recall	F1	Precision	Recall	F1
gpt-4	without_function_calling	798	85%	92%	75%	83%	80%	94%	87%
	with_explanation	7,200	88%	95%	79%	86%	83%	96%	89%
	with_function_calling	1,658	88%	95%	79%	86%	83%	96%	89%
gpt-3.5-turbo	without_function_calling	175	79%	86%	67%	75%	75%	90%	82%
	with_explanation	1,491	76%	83%	62%	71%	72%	88%	79%
	with_function_calling	388	77%	84%	65%	73%	73%	88%	80%
gpt-3.5-turbo-instruct	without_function_calling	135	79%	91%	62%	74%	73%	94%	82%

# Question and Answer Eval - Sweep



# Results: Chunk Retrieval Eval

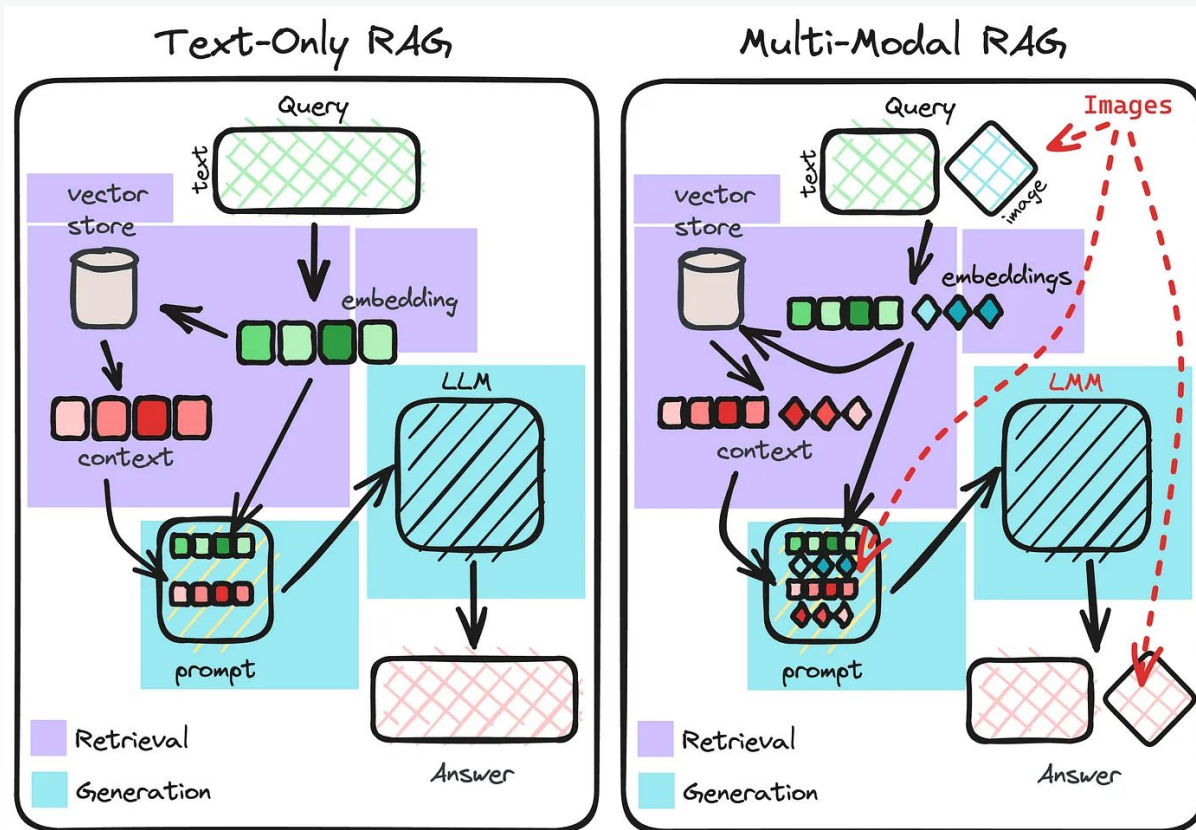
## Precision MRR



# Multi-modal Eval with LlamaIndex

---

# Overview



# Multi-modal RAG vs. Text RAG

Build Consideration	Text-Only RAG	Multi-Modal RAG
Index Data*	<p>Encode text-only data and store in an index (e.g., a Vector Index).</p> <p>Thus need an encoder for text data, such as text-embedding-ada-002.</p>	<p>Encode text data as well as image data and store them in a separate index (or namespace/collection).</p> <p>Thus need two encoders: 1 for text (e.g. text-embedding-ada-002) and another for image (e.g. clip).</p>
Generative Model	Choose an LLM (e.g., GPT-4, GPT-3.5, LLaMa-2, etc).	Choose an LMM (e.g., GPT-4V, LLaVA, etc).

Table 1: Build considerations for RAG systems and how they differ text-only versus multi-modal scenarios.

# Multi-modal RAG vs. Text RAG

Query Pipeline	Text-Only RAG	Multi-Modal RAG
User query	User submits a text-only query	User submits a query containing both image and text
Retrieved documents	Text query is encoded and used to retrieve relevant encoded text data.	Text and image query data are encoded and used to retrieve relevant text as well as image data.
Response generation	Retrieved text data is used as context to prompt the LLM generator to produce an answer to the query.	Retrieved text and image data are used as context to prompt the LMM generator to produce an answer to the query.

Table 2: The pipeline for querying a RAG and how they differ text-only versus multi-modal scenarios.

# Evaluating Multi-modal RAG (Retrieval)

Separate out image and text modalities

Compute retrieval score for each modality

	Hit Rate	Mean Reciprocal Rank
Text	0.95	0.88
Images	0.88	0.75

Table 3: Retrieval evaluation in multi-modal scenario.

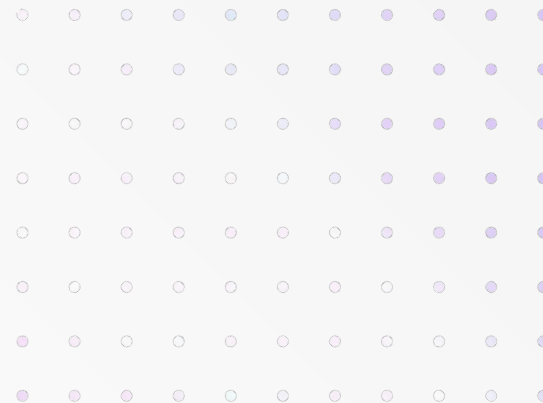
# Evaluating Multi-modal RAG (Generation)

Instead of using GPT-4 as a judge,  
**use multi-modal models (LMMs).**

Judge the quality of joint image  
and text context.

## Example metrics:

- Relevancy (multi-modal)
- Faithfulness (multi-modal)



# Resources

LlamaIndex + Arize Phoenix



Multi-modal Evals





# Thank you

Visit us at [arize.com](https://arize.com)