



## RAG Time:

# Evaluate RAG with LLM Evals and Benchmarking



Mikyo King  
Arize AI  
Head of Open Source



Amber Roberts  
Arize AI  
ML Growth Lead

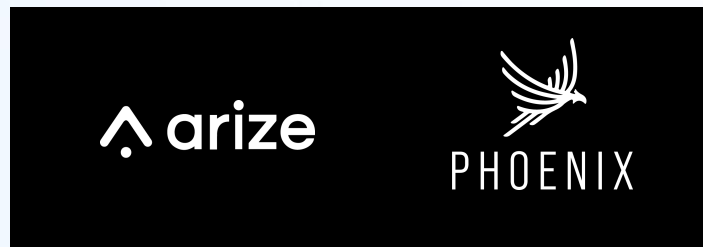
# Overview

## Presentation (10 minutes)

- What is Retrieval Augmented Generation (RAG)?
- What are Response Evals?
- What are Retrieval Evals?

## Colab Code Along (30 minutes)

- LLM Application Tracing Workflows (LLM Observability)
- Phoenix Traces and Spans
- Retrieval Evals and Metrics



Arize  
Community



Phoenix  
repo



Phoenix  
docs



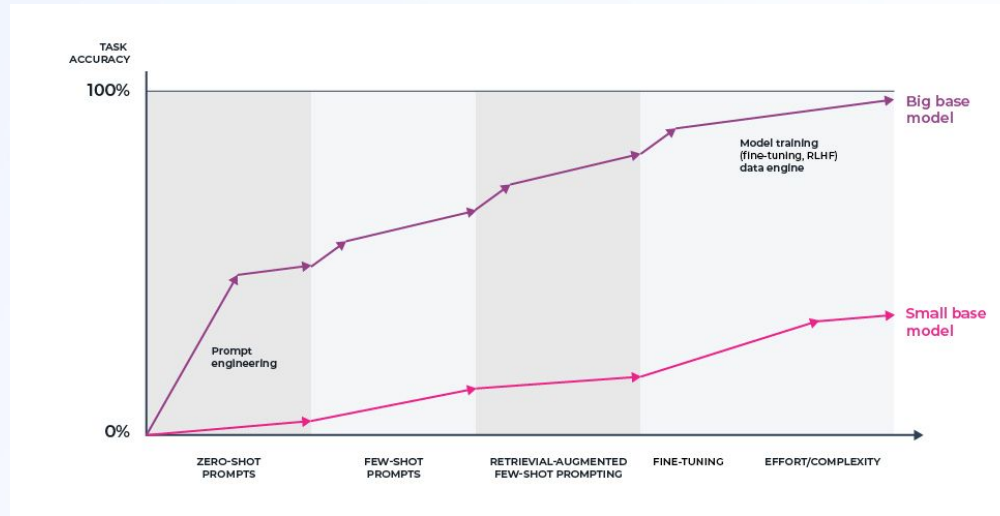
# Retrieval Augmented Generation (RAG): Pros and Cons

## Pros

- RAG can greatly improve the performance of your LLM application
- You can leverage LLM capabilities with proprietary data
- Advanced methods continue to come out for improved performance

## Cons

- Troubleshooting RAG workflows can become time consuming
- If system is not monitored, there can be several points of failure in the RAG system



Andrej Karpathy (<https://twitter.com/karpathy/status/1655994367033884672>)

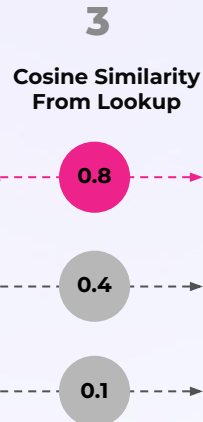
# How does RAG work?

## 2 User query

*Do you support international calling?*

Query embedding

<1, 2, 3, 4>



1 Knowledge base of articles

Document Chunk Embedding	Document Chunk ID	4 Document Chunk
<1, 1, 2, 4>	1	<p><b>What countries support International Calling</b></p> <ul style="list-style-type: none"> <li>The International Calling feature is available to all countries when it is enabled. See <a href="#">How to Enable International Calling</a> for more information.</li> <li>Some countries may not be available when International Calling is enabled. This means that RingCentral has restricted International Calling to those countries. Contact Support if you need to reach any of the restricted countries. See <a href="#">How to Open a RingCentral Tech Support Case</a> for more information.</li> </ul>
<100, 309, 4, 7>	2	<p><b>Configuring Outbound Call Prefix</b></p> <ol style="list-style-type: none"> <li>Contact Support to enable Outbound Call Prefix in your account.</li> <li>Sign in to the RingCentral admin portal.</li> <li>Navigate to <a href="#">Admin Portal &gt; More &gt; Account Settings &gt; Outbound Call Prefix</a>.</li> <li>Toggle to enable the <a href="#">outgoing call prefix</a>.</li> <li>Enter the single-digit that users will dial before dialing an external number.</li> <li>Click <a href="#">Validate &amp; Save</a>.</li> </ol>
<59, 71, 73, 95>	3	<p><b>Calling within the same area code.</b></p> <ul style="list-style-type: none"> <li>The user is assigned a phone number 1-768-222-3333.</li> <li>A user would like to call 1-768-333-3333.</li> <li>The user dials the outbound call prefix, number 9 333 3333, and the system inserts the Default Area Code automatically. Default Area Code is available for Canada, Mexico, Australia, and China phone numbers. If the feature is not available for your country, dial-in-country numbers using the area code.</li> <li>The number can also be dialed as area code and number 768-222-3333.</li> </ul>

## 5 Prompt

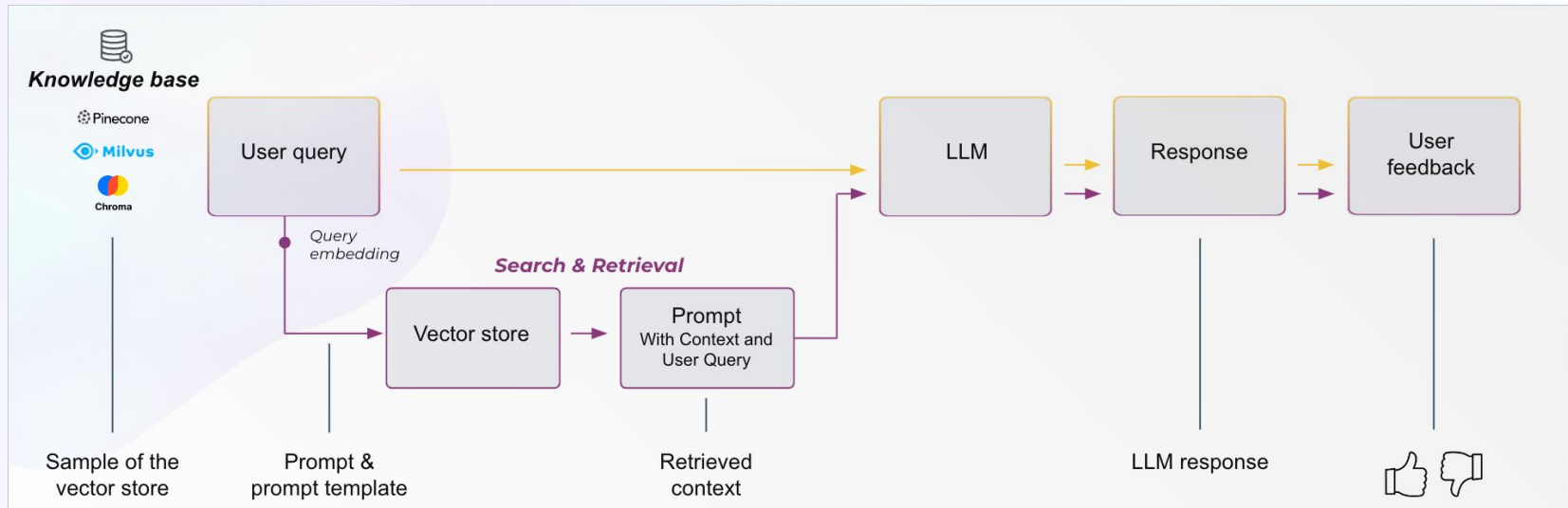
User is asking "Do you support international calling?"

Here's relevant content. Can you answer?

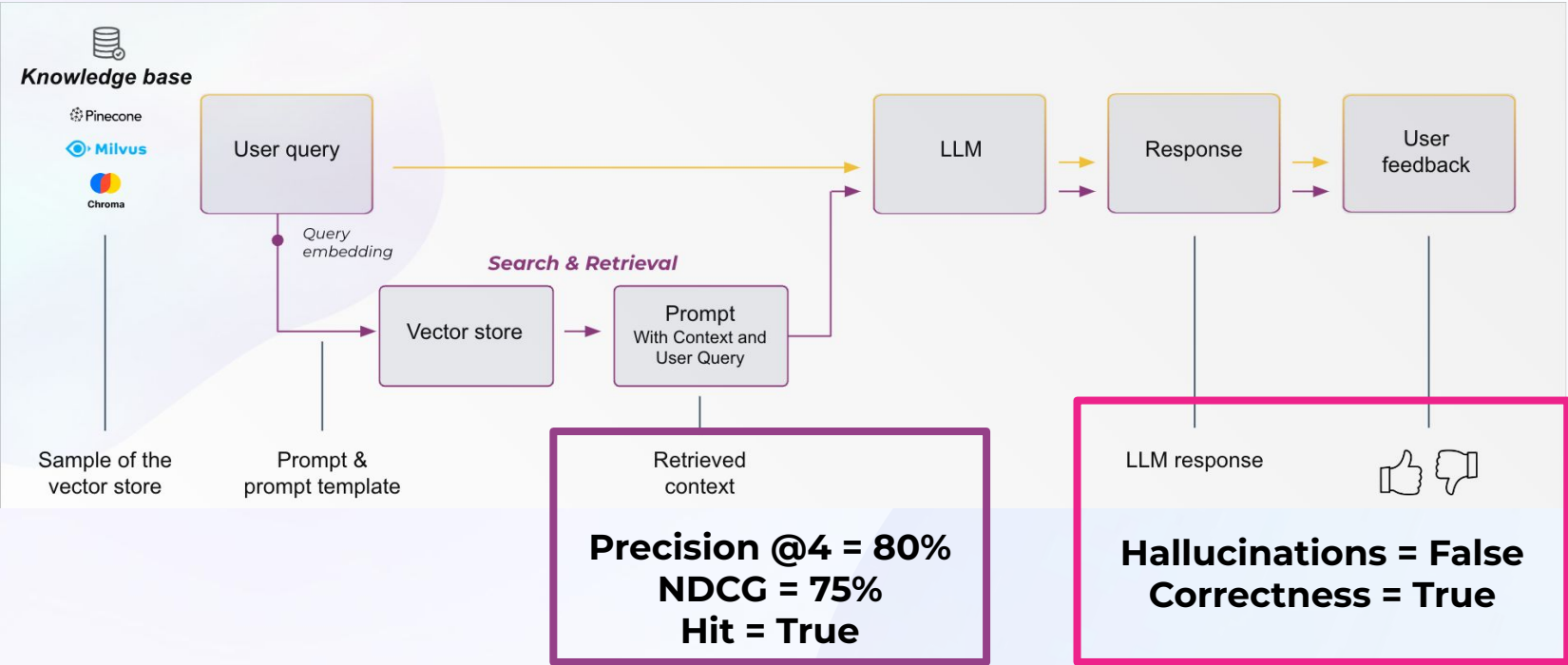
**What countries support International Calling**

- The International Calling feature is available to all countries when it is enabled. See [How to Enable International Calling](#) for more information.
- Some countries may not be available when International Calling is enabled. This means that RingCentral has restricted International Calling to those countries. Contact Support if you need to reach any of the restricted countries. See [How to Open a RingCentral Tech Support Case](#) for more information.

# Retrieval Evals vs Response Evals



# Retrieval Evals vs Response Evals

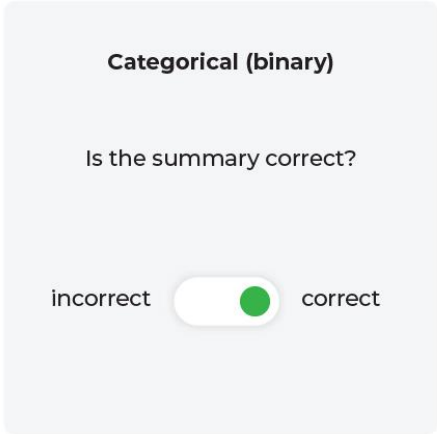


# Response Evaluations

***Measure the appropriateness of the response generated by the system when the context was provided.***

If LLMs do not have ground-truth labels evaluation can be done using the following response evaluation criteria:

- **QA Correctness** – whether a question was correctly answered by the system based on the retrieved data.
- **Hallucinations** – detect LLM hallucinations relative to retrieved context.
- **Toxicity** – identify if the AI response is racist, biased, or toxic.



**Categorical (binary)**

Is the summary correct?

incorrect  correct

# Retrieval Evaluation Metrics

*Assess the accuracy and relevance of the documents that were retrieved*

## nDCG

To measure the effectiveness of your top ranked documents.

Takes into account the position of relevant docs.

## Hit Rate

% of queries that have relevant context.

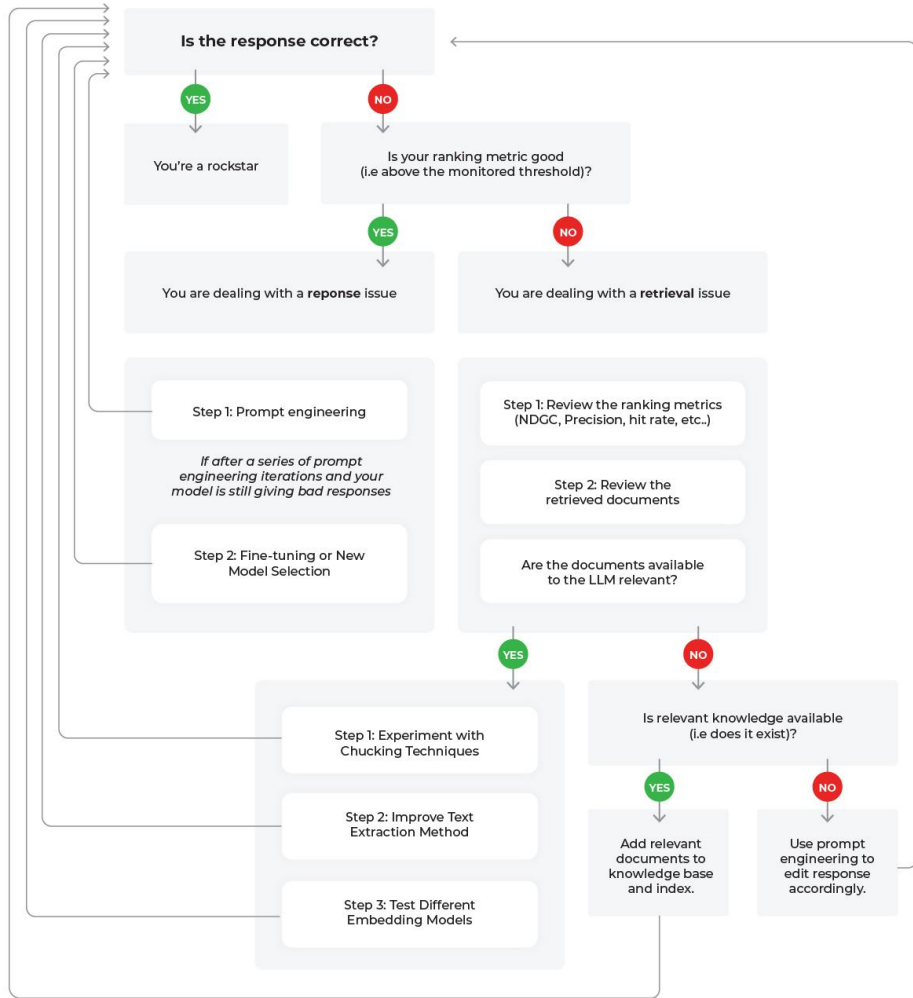
**Hit** is a binary metric (relevant document was or wasn't retrieved)

## Precision @K

Precision = % relevant documents, up to 'K' retrieved documents.

Precision@3 = 33%, if 1 out 3 docs is relevant.

# Example Workflows:





PHOENIX

---

**Now let's put these concepts into practice  
with a code-along session!**

# Thank you



[Sign up for free](#)



[Resources](#)



[Events](#)

# Resources

## Docs

Tutorials, examples and integrations

## Resource Hub

Collection of papers, case studies, tutorials, videos and interviews

## Community

Network with ML enthusiasts and stay up to date with AI observability advancements

