QA strategies to be adopted in ever evolving AI landscape

Shan Konduru

V3.0 - 22 May 2025

Agenda

Introduction

QA Challenges

Emerging Strategies

Tools and Technologies

Technical Demo

Q&A

Key Takeaways

Introduction

The AI revolution is here, marked by rapid advancements and significant financial backing.

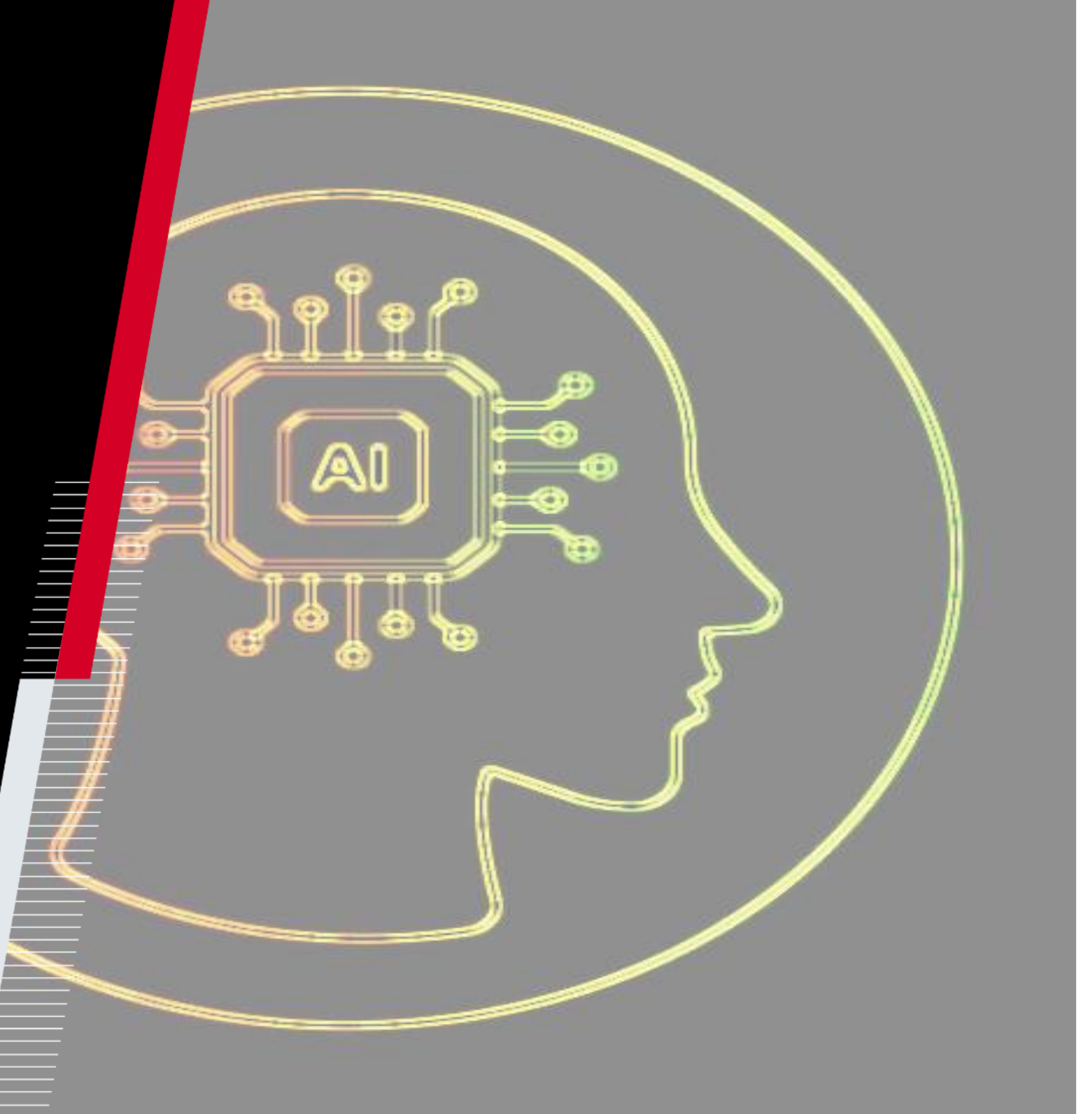
This session :

- Understand LLMs, AI Agents and Agentic AI
- Are our current Quality Assurance strategies fit for purpose in this new AI reality?
- We will explore the effectiveness of traditional approaches and outline essential QA strategies for navigating the evolving AI landscape.

CPKC

LLMs are Spectacular

- Unprecedented Language Understanding
- Impressive Text generation capabilities
- Ability to Learn and Adapt
- Scale and Complexity
- Impact they generate on Industries
- Scope for Future innovation and potential



AI Agents are Ingenious

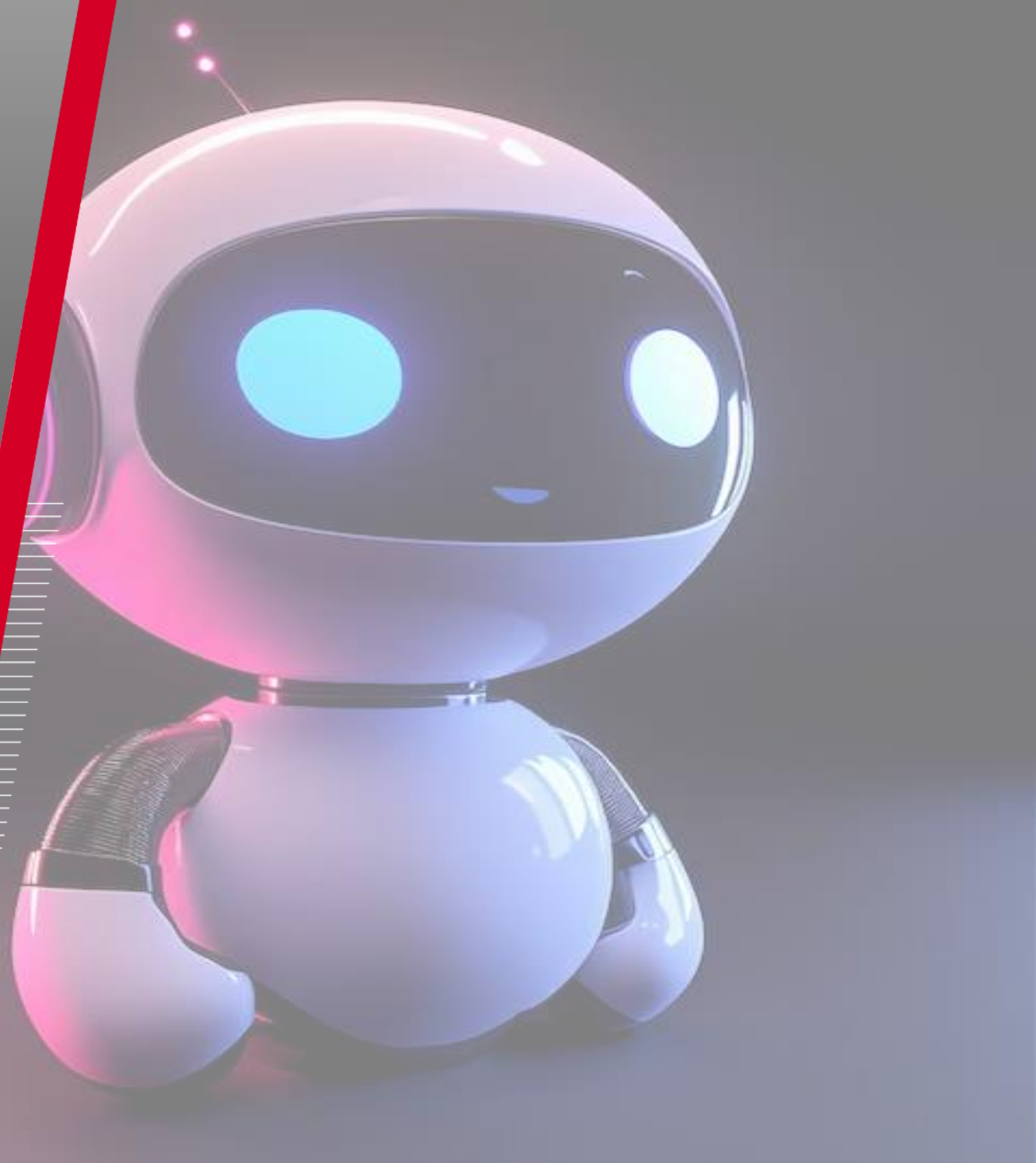
Goal Oriented behaviors

They Learn and Adapt

Take actions to reach Goal

Make autonomous decisions

Intelligent Interactions



Agentic AI is Unstoppable

Synergy of Autonomous Agents

Drive towards greater automation

Limitless application possibilities

Continued cycle of Innovation

Potential for Exponential Growth

Potential to address real life problems

CPKC

QA Challenges

How do we effectively assure the quality of AI in a world where outputs are often non-deterministic?

What methodologies can truly evaluate the intelligence and reasoning of these complex systems?

How can QA handle the immense range of input and output data?

1. Ensuring **bias and fairness**,
2. Testing **explainability** and **interpretability**
3. Testing **dynamic data** and **evolving AI** systems
4. Simulating **real-world** complexities
5. Defining relevant **success metrics**
6. Addressing critical **ethical considerations** and **alignment**

Emerging QA Strategies

Addressing the multifaceted challenges inherent in ensuring the quality of Artificial Intelligence necessitates a shift towards more adaptive and AI-aware QA methodologies.

1. Metric-Driven Evaluation
2. Comprehensive Test Case Design
3. Continuous Monitoring and Evaluation
4. Bias and Fairness Assessment
5. Explainability and Interpretability Evaluation
6. Human-in-the-Loop Evaluation
7. Robustness and Adversarial Testing

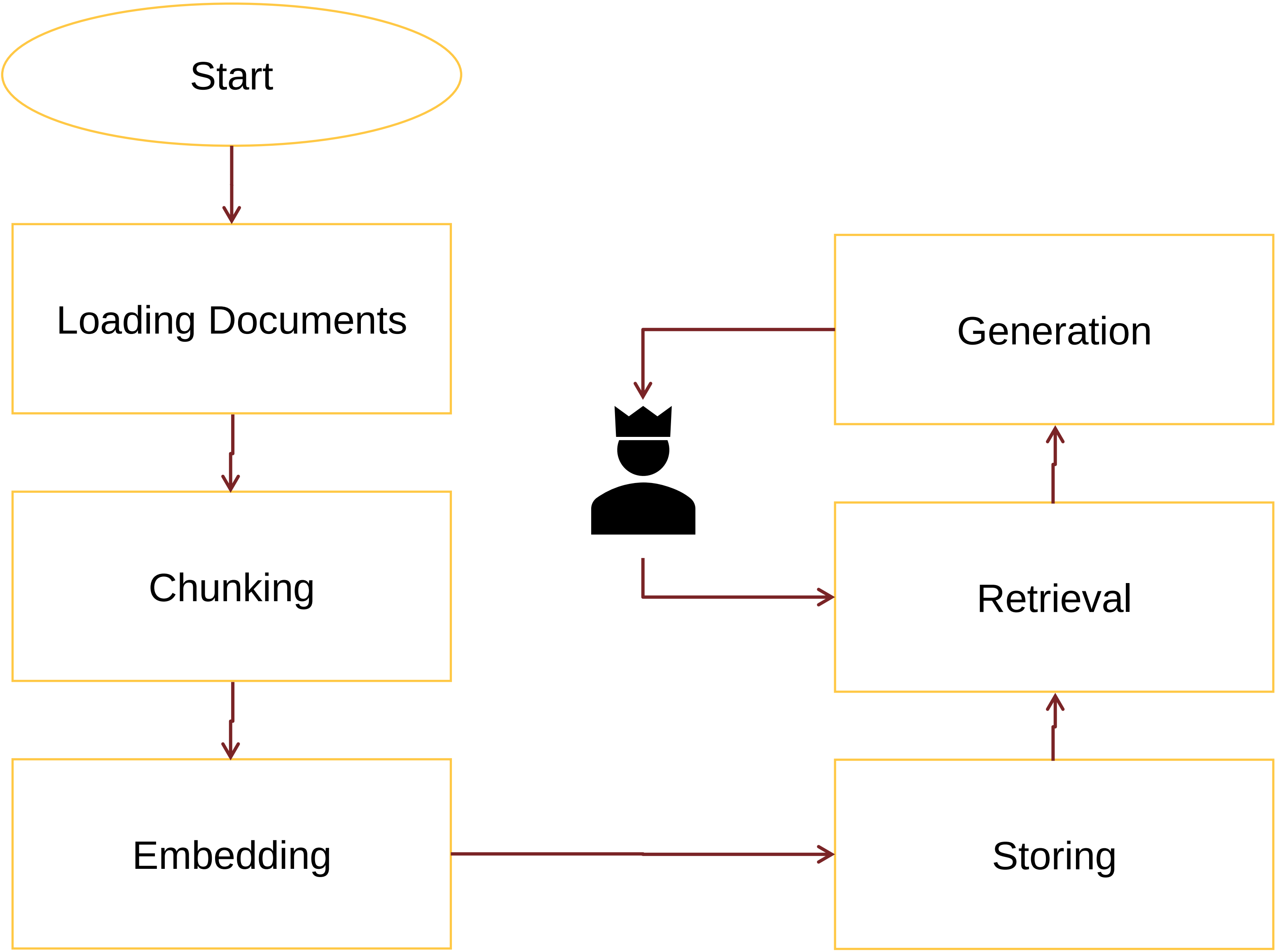
1. QA Strategy – Metric Driven Evaluation

Non-deterministic outputs, evaluating intelligence & reasoning, defining success metrics.

Instead of relying solely on exact output matching, adopt a metric-driven approach that defines and measures various aspects of AI quality.

This involves identifying relevant metrics for specific AI tasks (e.g., relevance, coherence, factual consistency, helpfulness, faithfulness)

Typical RAG process flow



Start

Loading Documents

Acquiring the knowledge source

Chunking

Dividing documents into smaller, semantically meaningful units.

Embedding

Converting each chunk into a vector representation.

Storing

Indexing and storing these embeddings in a vector store.

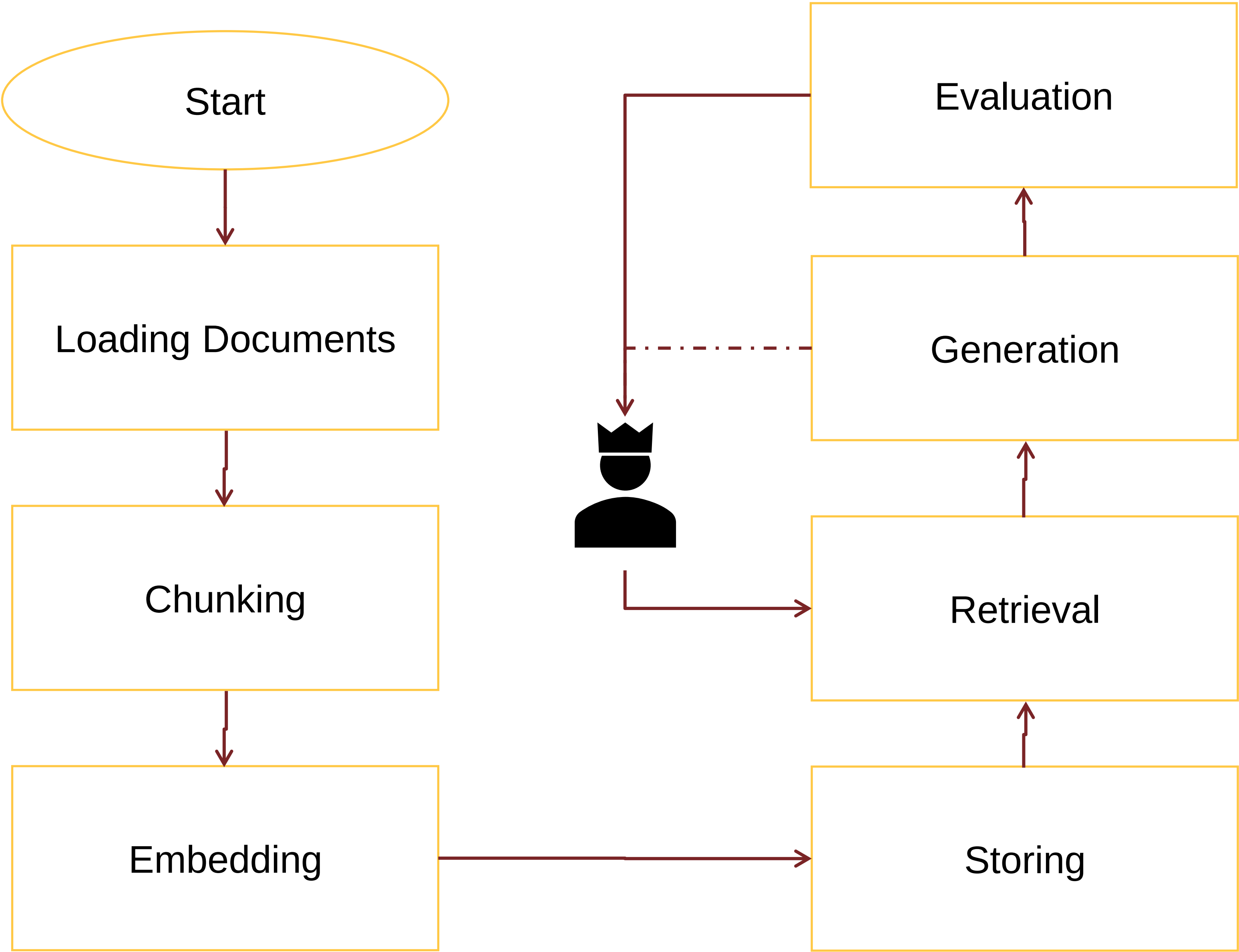
Retrieval

Embedding the user's query.
Searching the vector store for the most similar document embeddings.
Retrieving the corresponding text chunks.

Generation

Combining the retrieved context with the original user query.
Feeding this augmented input to a large language model to generate a relevant and informed response.

Typical RAG Evaluation process flow



Start

Loading Documents

Acquiring the knowledge source

Chunking

Dividing documents into smaller, semantically meaningful units.

Embedding

Converting each chunk into a vector representation.

Storing

Indexing and storing these embeddings in a vector store.

Retrieval

Embedding the user's query. Searching the vector store for the most similar document embeddings. Retrieving the corresponding text chunks.

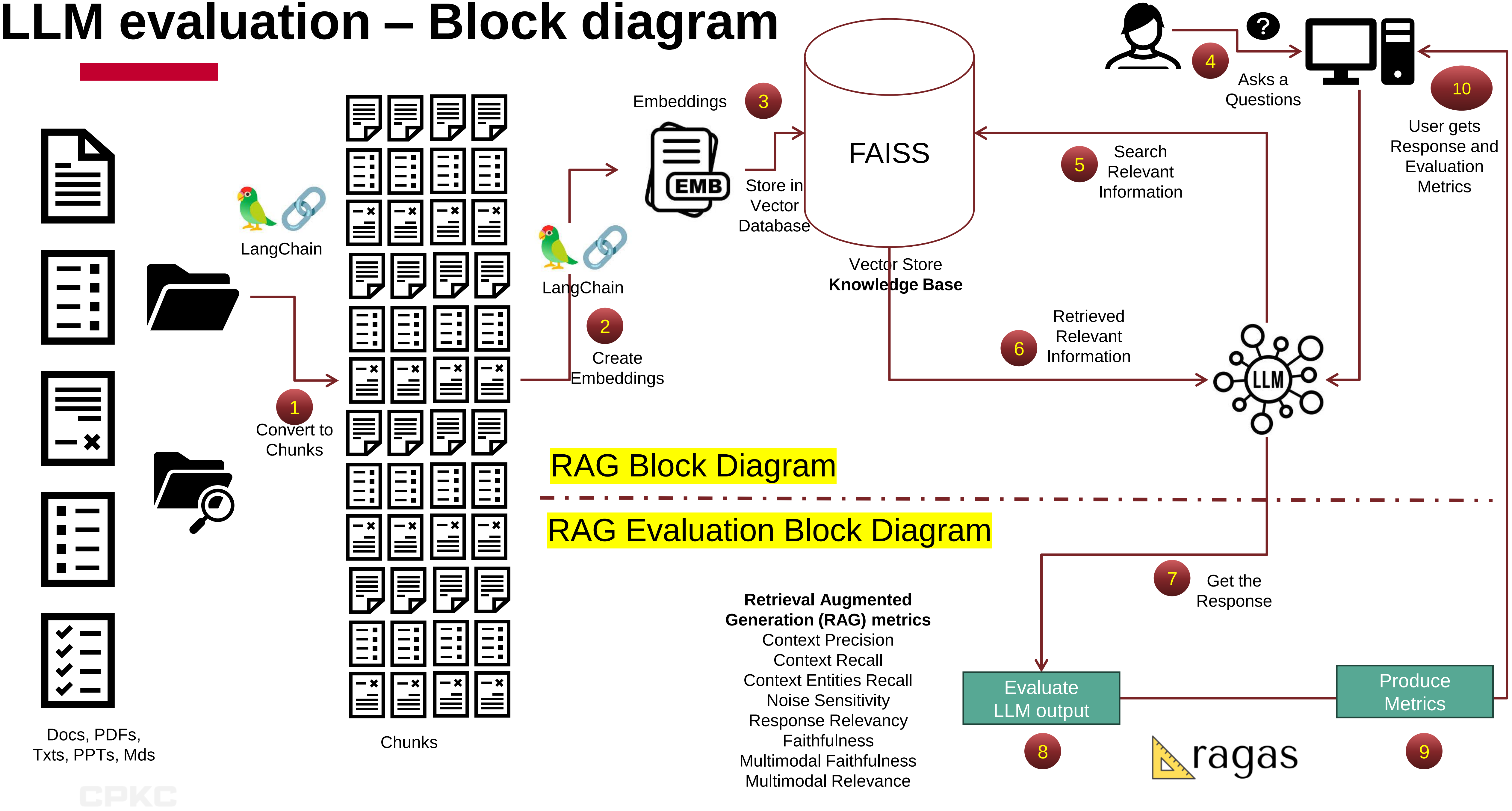
Generation

Combining the retrieved context with the original user query.
Feeding this augmented input to a large language model to generate a relevant and informed response.

Evaluation

Assessing the quality of the generated response based on various metrics such as relevance, factual accuracy (groundedness in the retrieved context), coherence, and helpfulness. This step helps in understanding the performance of the RAG pipeline and identifying areas for improvement.

LLM evaluation – Block diagram



RAG Block Diagram

RAG Evaluation Block Diagram

Retrieval Augmented Generation (RAG) metrics
Context Precision
Context Recall
Context Entities Recall
Noise Sensitivity
Response Relevancy
Faithfulness
Multimodal Faithfulness
Multimodal Relevance

2. QA Strategy – Comprehensive Test Case Design

Handling vast input and output data spectrum, simulating real-world complexities.

Move beyond simple test cases to design a comprehensive suite that covers various input scenarios, including edge cases, adversarial inputs, and simulations of real-world complexities.

This requires understanding the AI's intended use, potential failure modes and greater degree of prompt engineering.

3. QA Strategy – Continuous Monitoring & Evaluation

Dynamic and evolving AI systems

Implement continuous monitoring of AI application performance in deployment.

Track relevant metrics over time to detect drift, degradation, or unexpected behaviors as the AI evolves with new data or model updates.

4. QA Strategy – Bias & Fairness Assessment

Bias and Fairness Assessment

Proactively identify and mitigate potential biases in the training data and the AI model's outputs.

This involves using bias detection tools, analyzing performance across different demographic groups, and incorporating fairness metrics into the evaluation process.

5. QA Strategy – Explainability & Interpretability Evaluation

Explainability and Interpretability Evaluation

If the AI system provides explanations, evaluate the quality and faithfulness of these explanations.

Assess if they are understandable, relevant, and accurately reflect the AI's reasoning (to the extent possible in a black-box setting).

6. QA Strategy – Human-in-the-Loop Evaluation

Evaluating intelligence & reasoning, ethical considerations and alignment.

Recognize that fully automated evaluation of complex AI, especially concerning ethical considerations and nuanced reasoning, often requires human judgment.

Incorporate human feedback loops into the evaluation process to assess aspects that are difficult to quantify automatically.

7. QA Strategy – Robustness & Adversarial Testing

Handling vast input and output data spectrum, dynamic & evolving systems.

Evaluate the AI's robustness to unexpected or adversarial inputs.

Test its behavior with out-of-distribution data, slightly out of scope inputs, and malicious prompts to identify vulnerabilities.

Tools and Technologies

Challenges	QA Strategies to adapt	Tools and Technologies
Non-deterministic outputs	Focus on statistical testing, metamorphic testing, and defining acceptable output ranges.	Statistical analysis tools (e.g., R, Python/SciPy), metamorphic testing frameworks.
Evaluating intelligence & reasoning	Use human evaluation, benchmark datasets for specific reasoning skills, and metrics beyond just accuracy.	Benchmark datasets (e.g., GLUE, ImageNet), human evaluation platforms, Explainable AI (XAI) toolkits (e.g., SHAP, LIME), DeepEval, Ragas (for evaluating reasoning in RAG).
Handling vast input/output data	Employ data sampling, boundary value analysis, synthetic data generation, and AI-powered test case generation.	AI-powered test generation tools, synthetic data generation tools, big data testing frameworks (e.g., Hadoop, Spark).
Bias and Fairness	Implement bias detection and mitigation throughout development, use fairness metrics, and apply debiasing methods.	Bias detection tools (e.g., AI Fairness 360, Fairlearn), fairness metrics libraries, adversarial training frameworks.
Explainability & Interpretability	Prioritize interpretable models and use XAI techniques to understand complex model reasoning. Integrate explainability into QA.	Explainable AI (XAI) toolkits (e.g., SHAP, LIME), inherently interpretable models, visualization tools for AI explanations.
Dynamic & evolving systems	Adopt continuous testing, automated retraining/evaluation, and model monitoring for performance and drift.	CI/CD tools (e.g., Jenkins, GitLab CI), model monitoring platforms (e.g., Arize AI, Fiddler AI), automated testing frameworks, **DeepEval** (for continuous evaluation), **Ragas** (for continuous RAG evaluation).
Simulating real-world complexities	Utilize sophisticated simulation tools, digital twins, and scenario-based testing.	Simulation software, digital twin platforms, scenario generation tools.
Defining success metrics	Set clear, measurable objectives aligned with business goals. Use diverse metrics (beyond accuracy) and A/B testing.	Performance monitoring tools, A/B testing platforms, user feedback collection tools, **DeepEval** (for defining and tracking custom metrics), **Ragas** (for its specific RAG metrics).
Ethical considerations and alignment	Incorporate ethical reviews, define ethical guidelines, and involve ethicists/domain experts in QA.	Ethical AI frameworks and guidelines, human review boards, tools for detecting harmful content, **DeepEval** (for evaluating toxicity and bias).

Ragas vs DeepEval

Metrics Category	Metric Name	Ragas	DeepEval	Description
RAG-Specific	Faithfulness	Yes	Yes	Measures the factual consistency of the generated answer with the retrieved context.
RAG-Specific	Answer Relevancy	Yes	Yes	Assesses how well the generated answer addresses the query.
RAG-Specific	Context Precision	Yes	Yes (Contextual)	Evaluates the signal-to-noise ratio of the retrieved context; relevance of retrieved documents.
RAG-Specific	Context Recall	Yes	Yes (Contextual)	Measures if the retrieved context contains all the necessary information to answer the question (compared to ground truth).
RAG-Specific	Context Relevancy	Yes	Yes (Contextual)	Gauges the relevancy of the retrieved context to the question.
RAG-Specific	Context Entity Recall	Yes	No	Measures the recall of entities present in both ground truths and retrieved contexts.
RAG-Specific	Noise Sensitivity	Yes	No	Measures the robustness of the RAG system to irrelevant information in the context.
RAG-Specific	Context Utilization	Yes	No	Measures how much of the retrieved context is actually used to generate the answer.

Technical Demo

Demo briefing – RAG Evaluation using RAGAs

RAG Evaluation using RAGAs

Environment Specifications

Local Development Environment:
Operating System: Windows
Python Version: 3.11
VS Code Version: 1.99.0

Technical Specifications

Python Dependencies:
tiktoken
pandas
streamlit
pytest
Dotenv

LLM Dependencies:
langchain
ragas
faiss-cpu
langchain-openai

LLM Access:
OpenAI API
(gpt-3.5-turbo, gpt-4)

Vector Database
FAISS

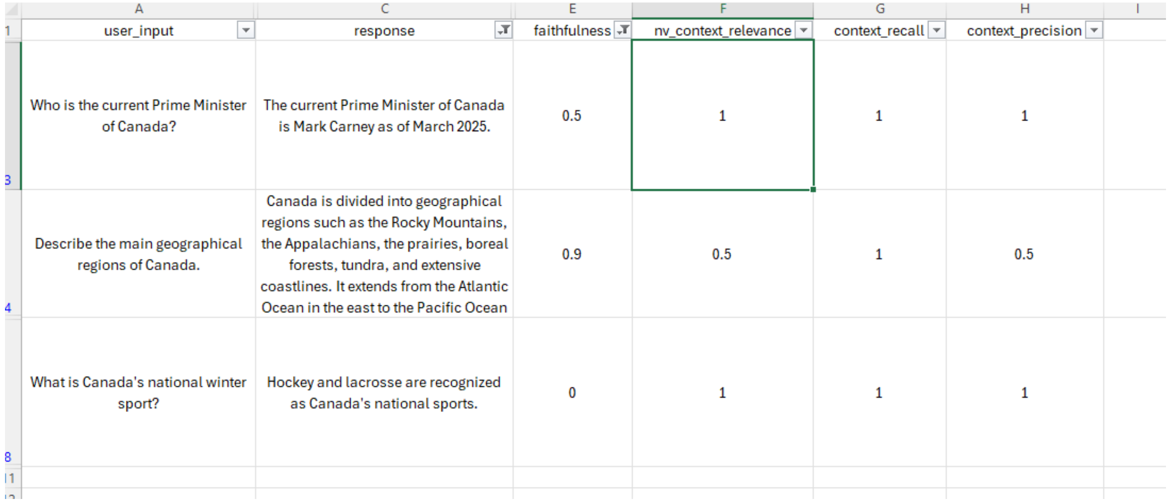
Test Scenario

Build RAG system:
Using bunch of documents, create a Knowledge base as a Vector Table.
Ask Questions to LLM:
Seek answers.
Evaluate the Response:
Explain the metrics

Demo Objectives

Primary Objective:
To demonstrate a practical application of the Ragas framework for evaluating the performance of Retrieval Augmented Generation (RAG) systems.
Secondary Objectives:
To provide a clear understanding of key RAG evaluation metrics (faithfulness, answer relevancy, context precision, context recall).
To highlight the importance of systematic evaluation in the development of RAG-based applications.

Review the Evaluation metrics



user_input	response	faithfulness	nv_context_relevance	context_recall	context_precision
Who is the current Prime Minister of Canada?	The current Prime Minister of Canada is Mark Carney as of March 2025.	0.5	1	1	1
Describe the main geographical regions of Canada.	Canada is divided into geographical regions such as the Rocky Mountains, the Appalachians, the prairies, boreal forests, tundra, and extensive coastlines. It extends from the Atlantic Ocean in the east to the Pacific Ocean	0.9	0.5	1	0.5
What is Canada's national winter sport?	Hockey and lacrosse are recognized as Canada's national sports.	0	1	1	1

How to improve the Faithfulness score?

What measures we must take to increase the faithfulness?

- Adjust the KB
- Add additional context about Canada Geography
- Update Context - Remove Winter

Key Take-aways

Key take-aways – we learned about

- Talked about
  - LLMs
  - AI Agents
  - Agentic AI, and their Capabilities
- Challenges in Traditional QA strategies and approaches

- Need for new QA strategies
  - Metric-Driven Evaluation
  - Comprehensive Test Case Design
  - Continuous Monitoring and Evaluation
  - Bias and Fairness Assessment
  - Explainability and Interpretability Evaluation
  - Human-in-the-Loop Evaluation
  - Robustness and Adversarial Testing

Key take-aways – we learned about

Demo’s Primary Objective:

- To demonstrate a practical application of the Ragas framework for evaluating the performance of Retrieval Augmented Generation (RAG) systems.

Demo’s Secondary Objectives:

- To provide a clear understanding of key RAG evaluation metrics (faithfulness, answer relevancy, context precision, context recall).
- To highlight the importance of systematic evaluation in the development of RAG-based applications.
- Discussed the sensitivity of Evaluation methods

Key take-aways – we learned about

- Issues with Automated Evaluation
  - String Matching Limitations
  - The "Semantic Gap“
  - Contextual Understanding

- Need efforts to Improve Reliability
  - Semantic Similarity Metrics
  - Fact Verification Models
  - Human Evaluation
  - More Sophisticated Evaluation Frameworks
  - Adversarial Testing

APPENDIX

Metrics and Measurement Framework

Retrieval Augmented Generation metrics

Context Precision

Context Recall

Context Entities Recall

Noise Sensitivity

Response Relevancy

Faithfulness

Multimodal Faithfulness

Multimodal Relevance

Nvidia / Agents or tool use cases

Answer Accuracy

Context Relevance

Response Groundedness

Topic adherence

Tool call Accuracy

Agent Goal Accuracy

CPKC

Natural Language Comparison

Factual Correctness

Semantic Similarity

Non LLM String Similarity

BLEU Score

ROUGE Score

String Presence

Exact Match

Others

Summarization

Aspect critic

Simple Criteria Scoring

Rubrics based scoring

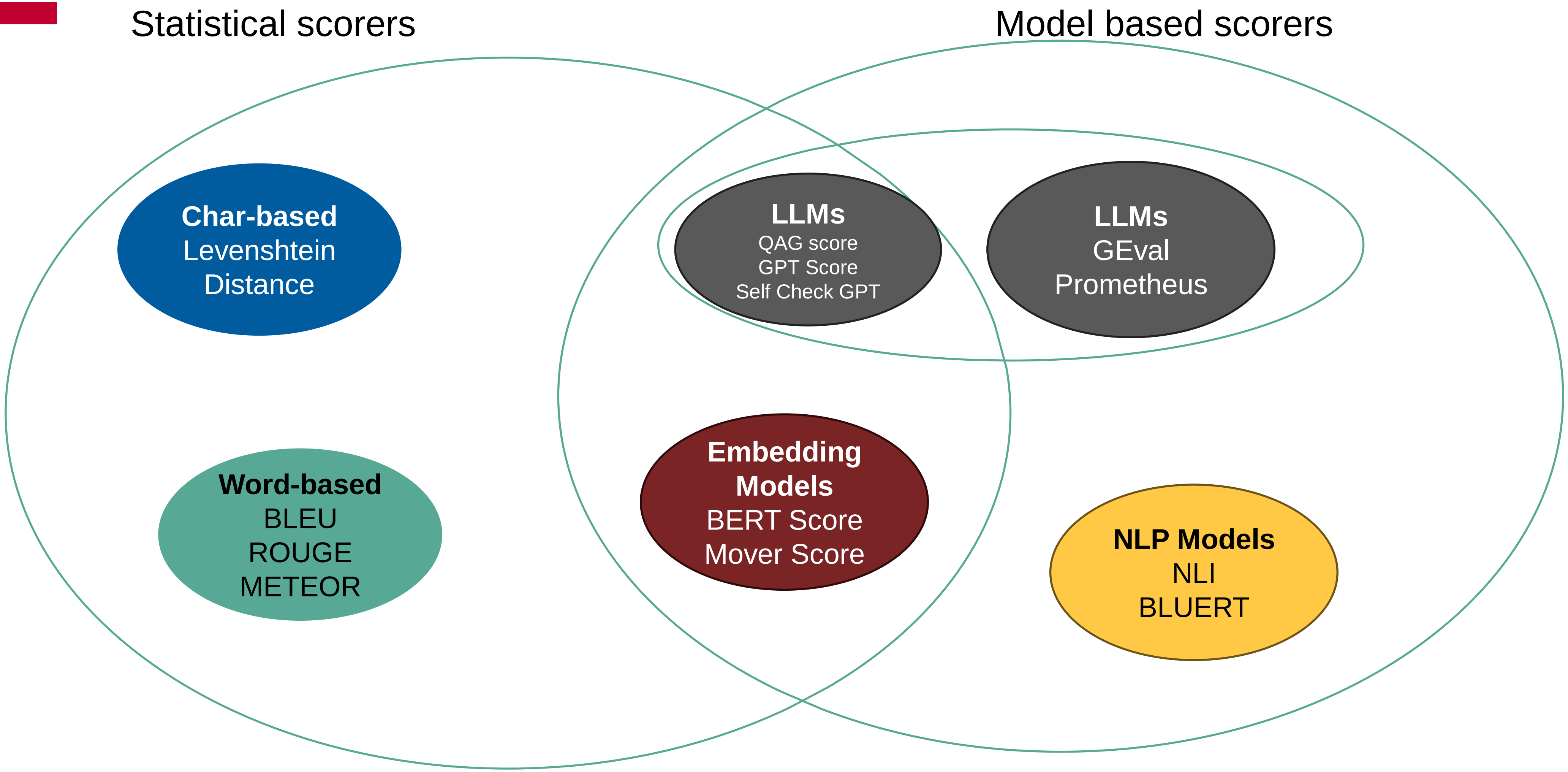
Instance specific rubrics scoring

Execution based Datacompy Score

SQL query Equivalence

CPKC

Complex scorers and evaluation frameworks



Statistical scorers

Model based scorers

Char-based
Levenshtein
Distance

Word-based
BLEU
ROUGE
METEOR

LLMs
QAG score
GPT Score
Self Check GPT

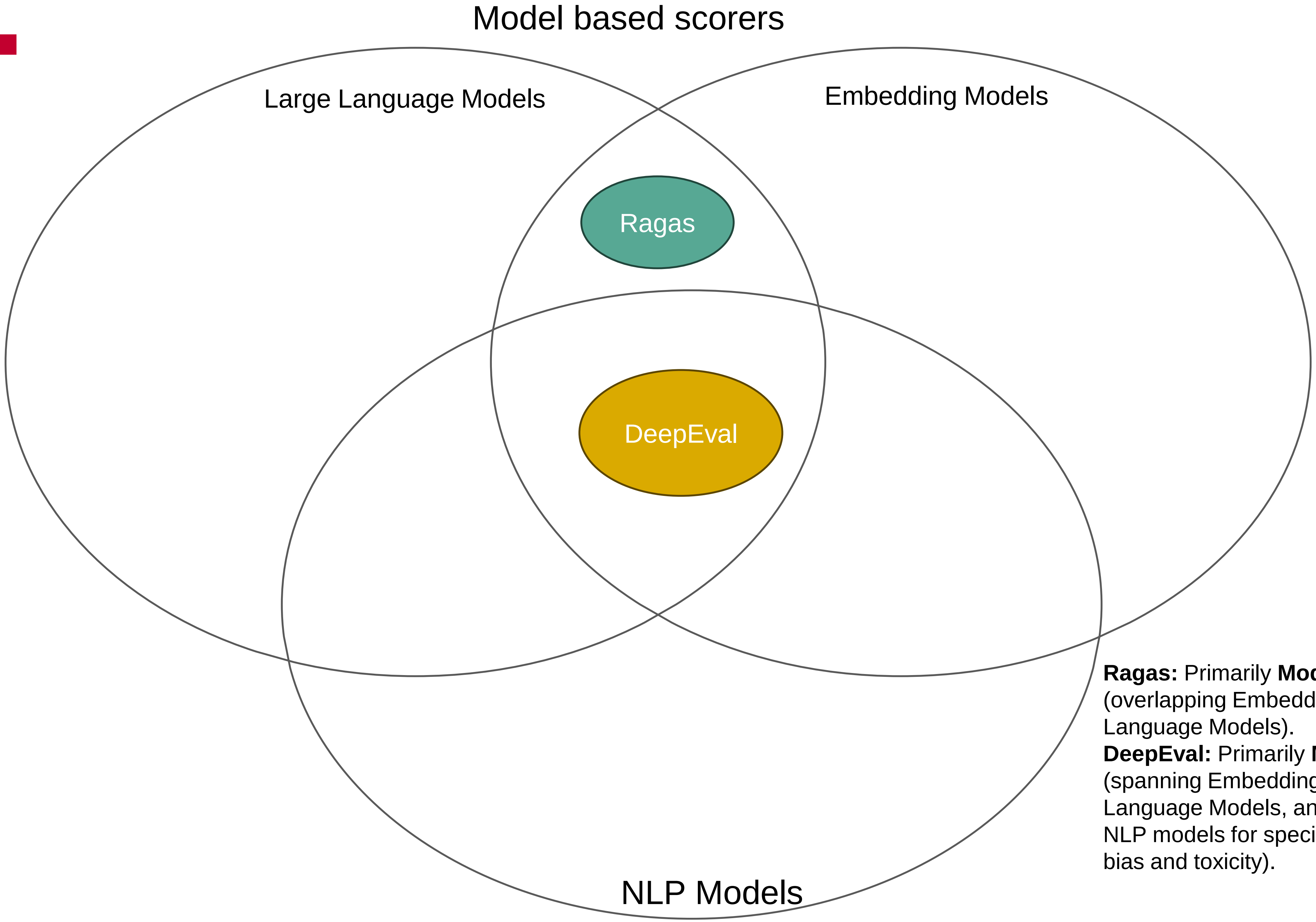
LLMs
GEval
Prometheus

Embedding Models
BERT Score
Mover Score

NLP Models
NLI
BLUERT

Ragas: Primarily **Model-Based Scorers** (overlapping Embedding Models and Large Language Models).
DeepEval: Primarily **Model-Based Scorers** (spanning Embedding Models, Large Language Models, and potentially Other NLP models for specific evaluations like bias and toxicity).

Complex scorers and evaluation frameworks



Ragas: Primarily **Model-Based Scorers** (overlapping Embedding Models and Large Language Models).
DeepEval: Primarily **Model-Based Scorers** (spanning Embedding Models, Large Language Models, and potentially Other NLP models for specific evaluations like bias and toxicity).

Ragas vs DeepEval

Metrics Category	Metric Name	Ragas	DeepEval	Description
Answer Quality	Answer Semantic Similarity	Yes	No	Assesses the semantic resemblance between the generated answer and the ground truth.
Answer Quality	Answer Correctness	Yes	No	Gauges the accuracy of the generated answer when compared to the ground truth.
General LLM Output	G-Eval	No	Yes (Custom)	A framework to create custom LLM-as-a-judge metrics based on specific criteria.
General LLM Output	Hallucination	No	Yes	Measures the extent to which the LLM generates information not present in the context.
General LLM Output	Toxicity	No	Yes	Evaluates the toxicity level of LLM outputs.
General LLM Output	Bias	No	Yes	Measures the presence of bias in LLM outputs.
General LLM Output	Summarization	Yes	Yes	Evaluates the quality of summarization tasks.
General LLM Output	JSON Correctness	No	Yes	Evaluates the correctness of JSON outputs.

Ragas vs DeepEval

Metrics Category	Metric Name	Ragas	DeepEval	Description
Agent/Tool Use	Tool Correctness	Yes	Yes	Checks if agents use the right tools with the correct parameters.
Agent/Tool Use	Task Completion	Yes	Yes	Determines if an agent successfully achieves its goal.
Agent/Tool Use	Topic Adherence	Yes	No	Measures if the agent's response stays on the topic.
Agent/Tool Use	Tool Call Accuracy	Yes	No	Measures the accuracy of the agent's tool calls.
Agent/Tool Use	Agent Goal Accuracy	Yes	No	Measures if the agent achieves the specified goal.
Conversational	Conversational G-Eval	No	Yes	Evaluates the overall quality of a conversation using an LLM judge.
Conversational	Knowledge Retention	No	Yes	Measures how well the LLM retains information across conversation turns.
Conversational	Role Adherence	No	Yes	Evaluates how well the LLM adheres to its assigned role in a conversation.
Conversational	Conversation Completeness	No	Yes	Measures whether the conversation addresses all aspects of the user's request.
Conversational	Conversation Relevancy	No	Yes	Evaluates how relevant each response is to the ongoing conversation.

Ragas vs DeepEval

Metrics Category	Metric Name	Ragas	DeepEval	Description
Traditional NLP	Factual Correctness	Yes	No	Measures the factual accuracy of the generated text (can be without context in some Ragas metrics).
Traditional NLP	Semantic Similarity	Yes	No	Measures the semantic similarity between two pieces of text (e.g., answer and ground truth).
Traditional NLP	Traditional NLP Metrics	Yes	No	Includes metrics like BLEU, ROUGE, Exact Match, String Presence (often without LLM as judge).
SQL Evaluation	Execution based Datacompy Score	Yes	No	Evaluates the correctness of generated SQL queries by executing them and comparing results.
SQL Evaluation	SQL Query Equivalence	Yes	No	Evaluates if two SQL queries are semantically equivalent.
Other/Custom	Aspect Critique	Yes	No	Allows for evaluating specific aspects of the generated text based on defined criteria.
Other/Custom	Rubrics based scoring	Yes	Yes (Custom)	Scores outputs based on predefined rubrics. DeepEval's G-Eval can be used for this.
Other/Custom	Instance specific rubrics	Yes	Yes (Custom)	Scores individual outputs based on rubrics tailored to that specific instance. DeepEval's G-Eval can be used for this.
Other/Custom	Custom Metrics	Yes	Yes	Both frameworks allow for the creation of user-defined evaluation metrics. DeepEval provides G-Eval and a base class for this.
RAGAS Holistic Metric	RAGAS Metric	Yes	Yes (via wrapper)	A composite metric in Ragas that averages Faithfulness, Answer Relevancy, Context Precision, and Context Recall. DeepEval offers a wrapper.

Canada Knowledge Base – Q&A

Question	Ground Truth
What are the two official languages of Canada?	The two official languages of Canada are English and French.
Who is the current Prime Minister of Canada?	As of May 18, 2025, the current Prime Minister of Canada is Mark Carney.
Name three provinces in Western Canada.	Three provinces in Western Canada are British Columbia, Alberta, and Saskatchewan. (Manitoba is sometimes also considered part of the Prairie Provinces, which are often grouped with Western Canada).
What is the capital city of Canada?	The capital city of Canada is Ottawa.
Which national symbol of Canada is also a tree?	The maple tree is a national symbol of Canada.
What is Canada's national winter sport?	Canada's national winter sport is ice hockey.
What is the name of Canada's national anthem?	Canada's national anthem is "O Canada".
What is the largest city in Canada by population?	Toronto is the largest city in Canada by population.
In which year did Canada gain independence?	Canada gained independence in 1867 (Confederation).
What is the currency used in Canada?	The currency used in Canada is the Canadian Dollar (CAD)
Name the capital city of Canada and the provinces that border Quebec.	The capital city of Canada is Ottawa. The provinces that border Quebec are Ontario, New Brunswick, and Newfoundland and Labrador.
What are the national colors of the Canadian flag and what is the symbol at its center?	The national colors of the Canadian flag are red and white. The symbol at its center is a red maple leaf.
Who is the current head of state of Canada, and who is the head of government?	The current head of state of Canada is King Charles III. The head of government is the Prime Minister, Mark Carney.
Name three major industries in Canada and at least one major city associated with each.	Three major industries in Canada are the automotive industry (Windsor), the oil and gas industry (Calgary), and the technology sector (Toronto or Waterloo).
What are the official languages of Canada, and in which province is French the predominant language?	The official languages of Canada are English and French. French is the predominant language in Quebec.

Canada Knowledge Base – Q&A

Question	Ground Truth
Describe the main geographical regions of Canada.	Canada can be broadly divided into several main geographical regions, including the Canadian Shield, the Western Cordillera (Rocky Mountains), the Interior Plains (Prairies), the Appalachian Region, the Arctic Region, and the Great Lakes-St. Lawrence Lowlands.
What type of government does Canada have?	Canada has a federal parliamentary democracy under a constitutional monarchy.
What is a recognized national sport of Canada?	Hockey and lacrosse are recognized as national sports of Canada.
Based on the provided information, what are some key aspects of Canada's bilingualism policy and where is it most evident?	Canada's bilingualism policy recognizes English and French as official languages, ensuring their use in federal institutions and services. This policy is most evident in the province of New Brunswick, which is officially bilingual, and in federal government operations across the country, particularly in areas with significant French-speaking populations like Quebec and parts of Ontario.
Considering its geography and major industries, what are some significant economic strengths of Canada?	Canada's significant economic strengths include its vast natural resources (oil and gas, minerals, timber), its extensive coastlines facilitating trade, its fertile prairies supporting agriculture, and its growing technology and manufacturing sectors concentrated in major urban centers.

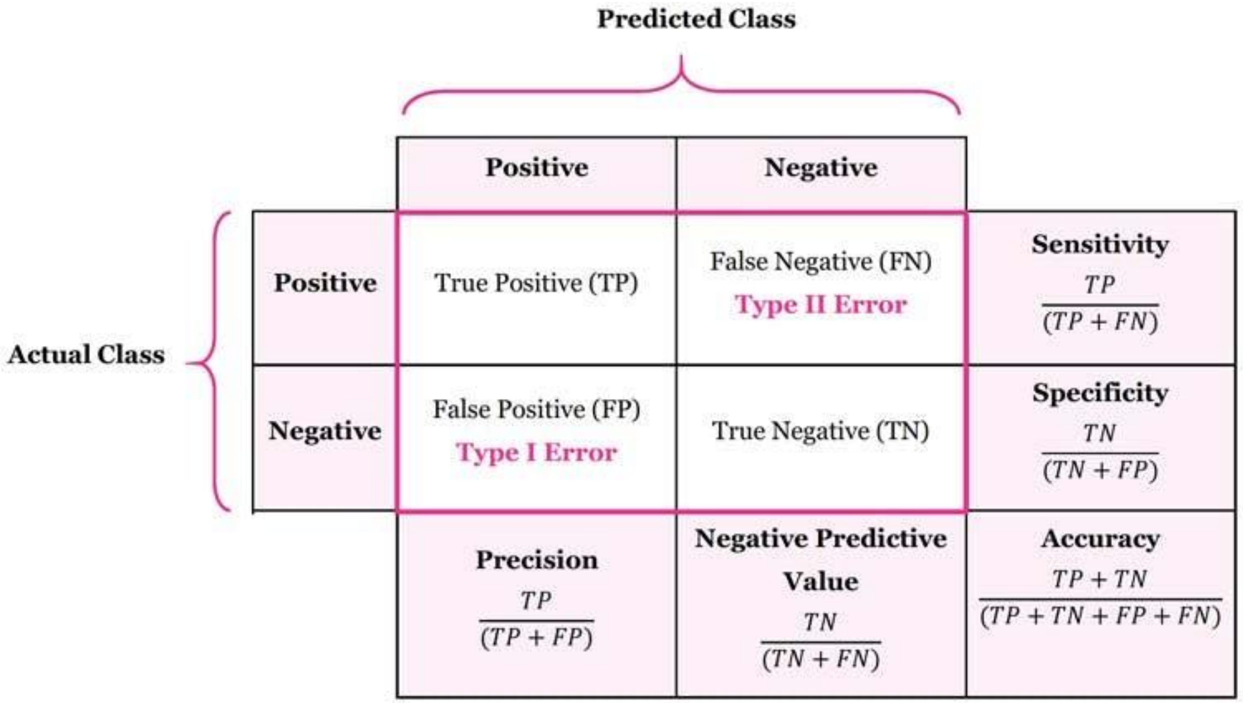
Canada Knowledge Base – Q&A

Question	Ground Truth
What was the weather like in Vancouver, British Columbia on March 15, 2024?	The weather in Vancouver on March 15, 2024, was likely mild and rainy, typical for that time of year. Specific details would include the temperature (around 8-12°C), precipitation (likely rain), and cloud cover. (Note: This requires historical weather data not present in the general knowledge base).
What are the current public opinions on the Canadian Prime Minister's latest policy regarding artificial intelligence?	Current public opinions on the Prime Minister's AI policy are mixed. Some groups praise its focus on ethical development and investment in research, while others express concerns about potential job displacement and the lack of specific implementation details. Recent polls indicate a 45% approval rate for the policy, with 30% disapproving and 25% undecided. (Note: This requires up-to-date public opinion data and analysis, which is not part of the general factual knowledge base).
What are some of the best new restaurants that opened in Montreal in the last six months?	Some of the best new restaurants in Montreal that opened in the last six months include "Le Petit Gourmet" (French fusion with excellent reviews), "Spice Route Kitchen" (authentic Southeast Asian cuisine), and "Urban Greens Bistro" (focusing on locally sourced, vegetarian options). These establishments have been praised for their innovative menus and ambiance. (Note: This requires very recent and subjective information about the restaurant scene, which is not in the general knowledge base).
What are the projected impacts of climate change on wheat production in the Canadian prairies over the next decade?	Projections suggest that climate change will have a complex impact on wheat production in the Canadian prairies. While warmer temperatures might initially extend the growing season, increased frequency of droughts, extreme weather events (like heatwaves and hailstorms), and changes in pest patterns are expected to negatively affect yields in the long term. Some studies predict a potential decrease of 10-20% in production by 2035 under certain climate scenarios. (Note: This requires specific climate modeling and agricultural forecasting data, which is beyond a general knowledge base).
What is the most popular Canadian television show currently streaming on major platforms?	The most popular Canadian television show currently streaming is likely "Northern Lights Mystery," a crime drama series that has gained significant viewership on "StreamFlix" in the past few months. It has been praised for its compelling storyline and strong performances. (Note: This requires real-time streaming data and popularity metrics, which are not part of the general knowledge base).

Canada Knowledge Base – Q&A

Question	Ground Truth	Primary Metric	Secondary Metrics
What are the two official languages of Canada?	The two official languages of Canada are English and French.	Faithfulness	Answer Relevancy, Answer Correctness.
Who is the current Prime Minister of Canada?	As of May 18, 2025, the current Prime Minister of Canada is Mark Carney.	Faithfulness.	Answer Relevancy, Answer Correctness.
Name three provinces in Western Canada.	Three provinces in Western Canada are British Columbia, Alberta, and Saskatchewan. (Manitoba is sometimes also considered part of the Prairie Provinces, which are often grouped with Western Canada).	Answer Relevancy	Faithfulness, Answer Correctness, Context Recall
What is the capital city of Canada?	The capital city of Canada is Ottawa.	Faithfulness.	Answer Relevancy, Answer Correctness.
Which national symbol of Canada is also a tree?	The maple tree is a national symbol of Canada.	Faithfulness	Answer Relevancy, Answer Correctness.
What is Canada's national winter sport?	Canada's national winter sport is ice hockey.	Faithfulness	Answer Relevancy, Answer Correctness.
What is the name of Canada's national anthem?	Canada's national anthem is "O Canada".	Faithfulness	Answer Relevancy, Answer Correctness.
What is the largest city in Canada by population?	Toronto is the largest city in Canada by population.	Faithfulness	Answer Relevancy, Answer Correctness.
In which year did Canada gain independence?	Canada gained independence in 1867 (Confederation).	Faithfulness	Answer Relevancy, Answer Correctness.
What is the currency used in Canada?	The currency used in Canada is the Canadian Dollar (CAD)	Faithfulness	Answer Relevancy, Answer Correctness.

Confusion Matrix



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

1. **Accuracy:** Accuracy quantifies the ratio of correct predictions (TP and TN) to the total number of predictions. While informative, this metric can be misleading when classes are imbalanced.
2. **Precision:** Precision evaluates the proportion of true positive predictions among all positive predictions (TP / (TP + FP)). This metric is crucial when the cost of false positives is high.
3. **Recall (Sensitivity or True Positive Rate):** Recall measures the ratio of true positive predictions to the actual number of positive instances (TP / (TP + FN)). This metric is significant when missing positive instances is costly.
4. **Specificity (True Negative Rate):** Specificity calculates the ratio of true negative predictions to the actual number of negative instances (TN / (TN + FP)). This metric is vital when the emphasis is on accurately identifying negative instances.
5. **F1-Score:** The F1-Score strikes a balance between precision and recall, making it useful when both false positives and false negatives carry similar importance.

Case: If patient have cancer or not	*What our model predicted*	*Conclusion 1*	*Conclusion 2*	*Combining both the conclusions*
have cancer	have cancer	**TRUE** prediction	**POSITIVE** prediction	True Positive (**TP**)
have cancer	doesn't have cancer	**FALSE** prediction	**NEGATIVE** prediction	Fasle Negative (**FN**)
doesn't have cancer	have cancer	**FALSE** prediction	**POSITIVE** prediction	False Positive (**FP**)
doesn't have cancer	doesn't have cancer	**TRUE** prediction	**NEGATIVE** prediction	True Negative (**TN**)

ACTUAL

If patient have cancer or not

PREDICTION

what our model predicted

	have cancer	doesn't have cancer
have cancer	number of **TP**	number of **FP**
doesn't have cancer	number of **FN**	number of **TN**