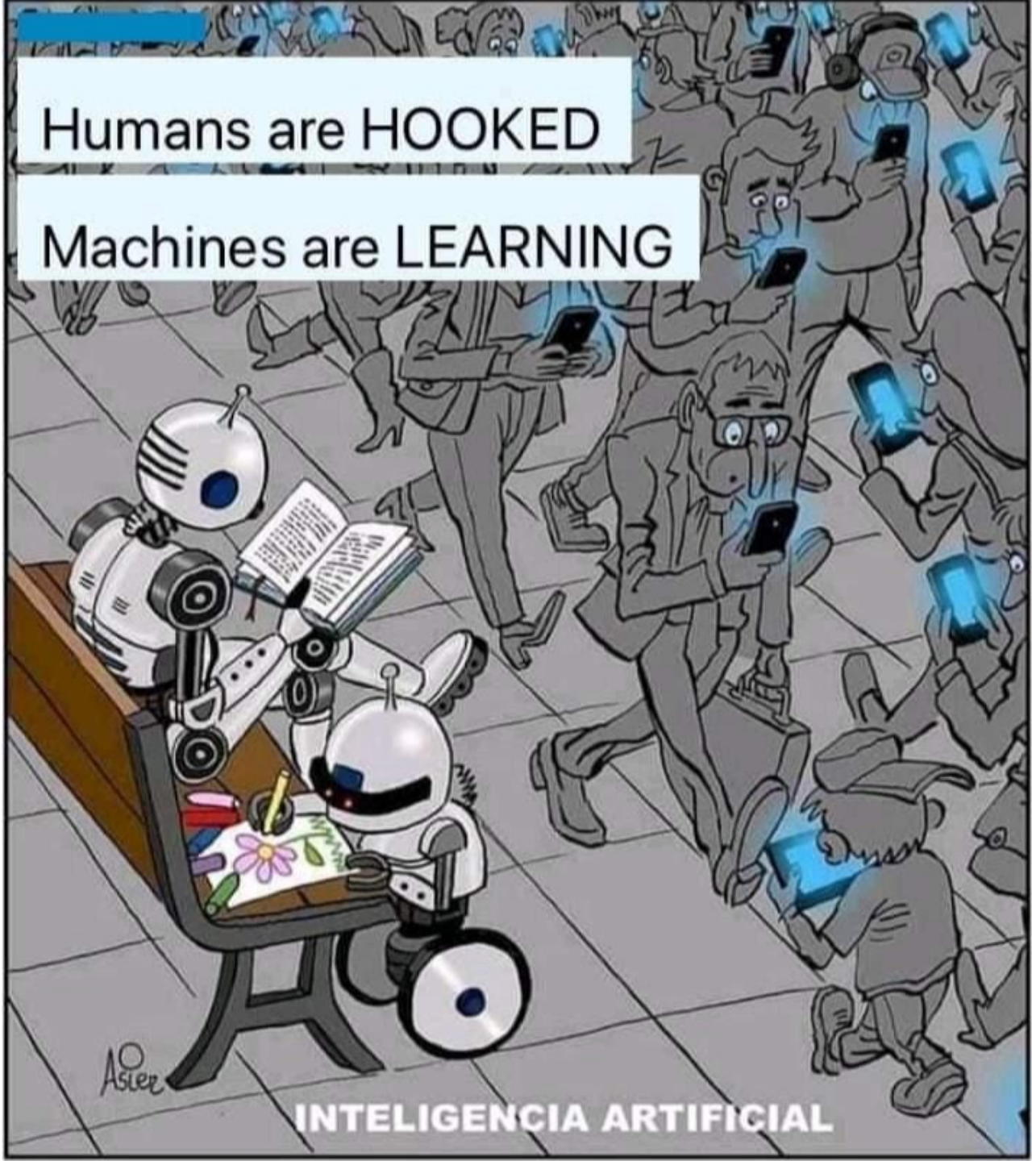


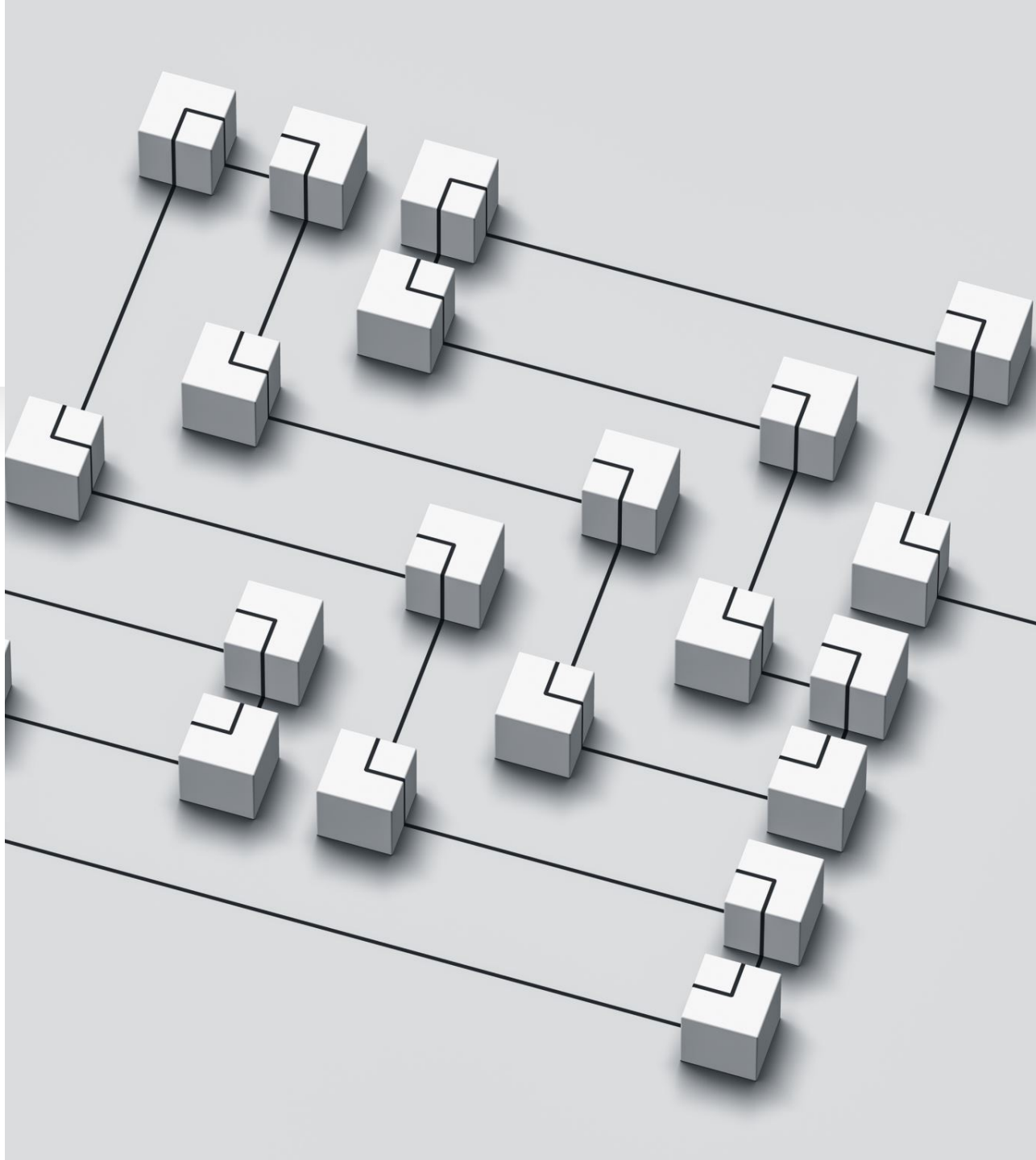
GenAI - RAG

Naveen Kumar Bhansali
Co-Founder BlitzAI | Adjunct Faculty @ IIM Bangalore


LLM Powered Application

Challenge with Foundation Models

- Foundation models are usually trained offline, making the model agnostic to any data that is created after the model was trained.
    - However, sometimes you need to work with newer or more current data.
- Additionally, foundation models are trained on very general domain corpora, making them less effective for domain-specific tasks.



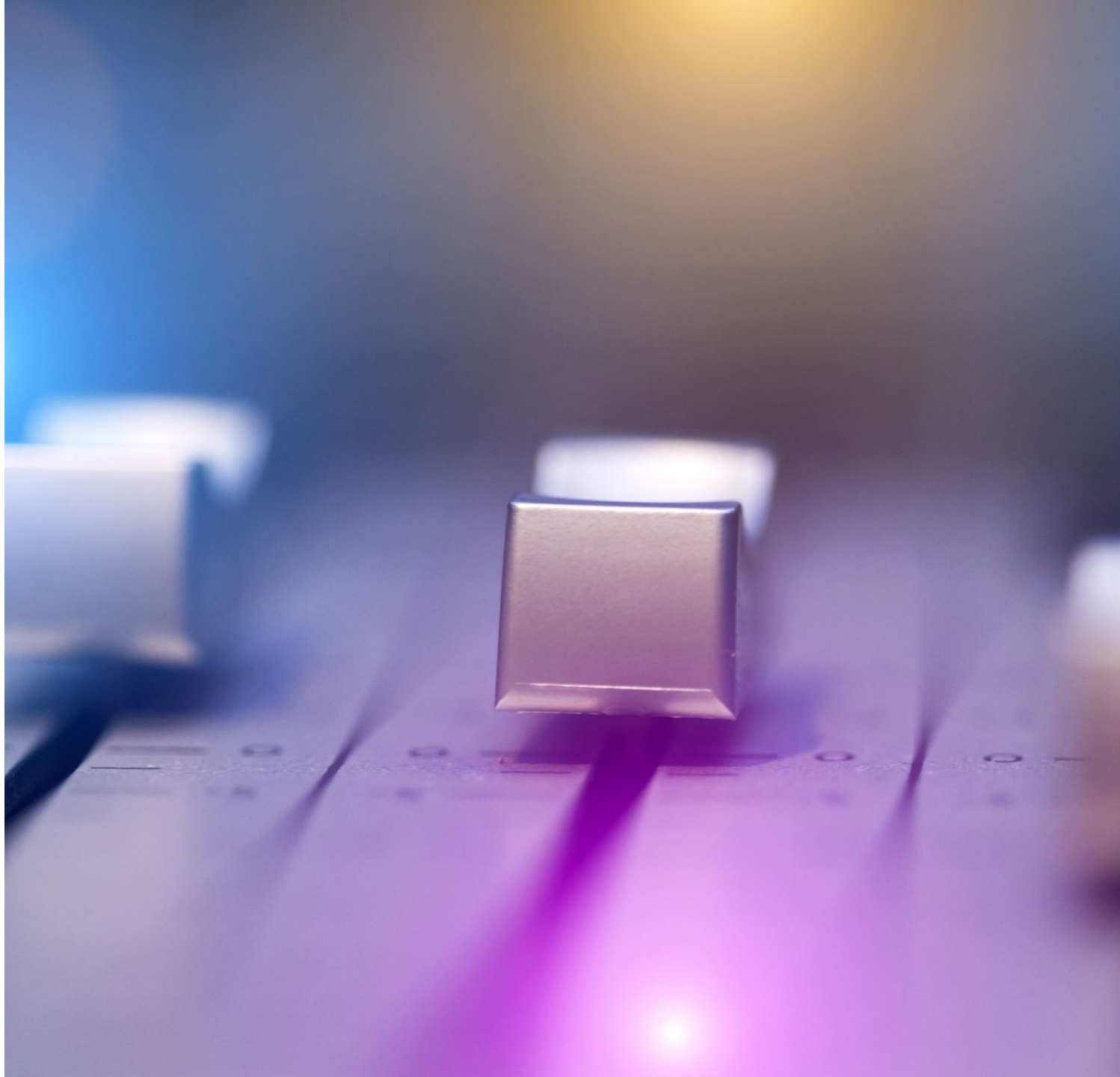
Approaches to address the challenge

- Fine-tuning or further training of the base model with new data.

- Retrieval Augmented Generation (RAG)

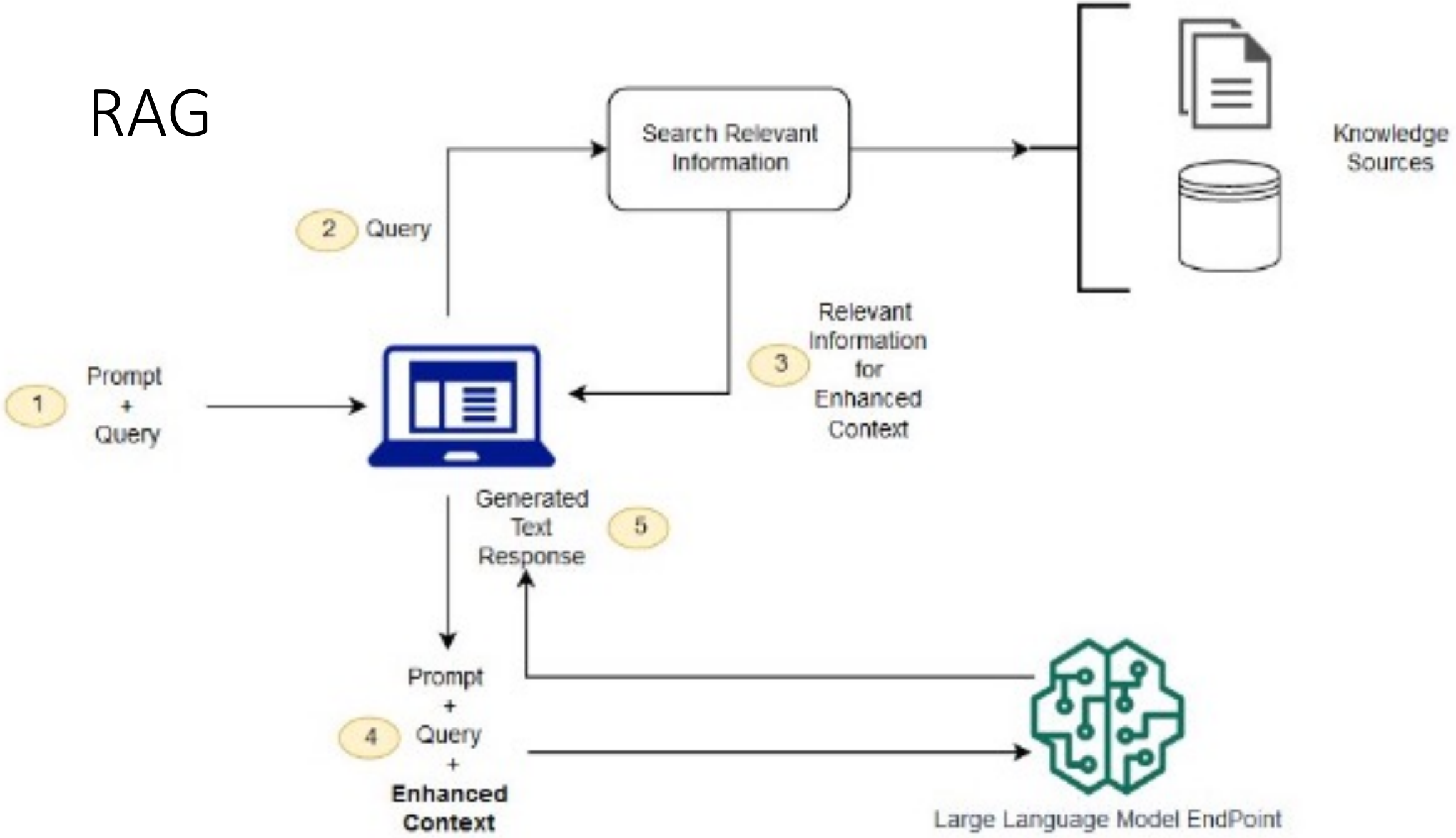
Fine-Tuning

It is suitable for continuous domain adaptation, enabling significant improvements in model quality **but** often incurring higher costs.

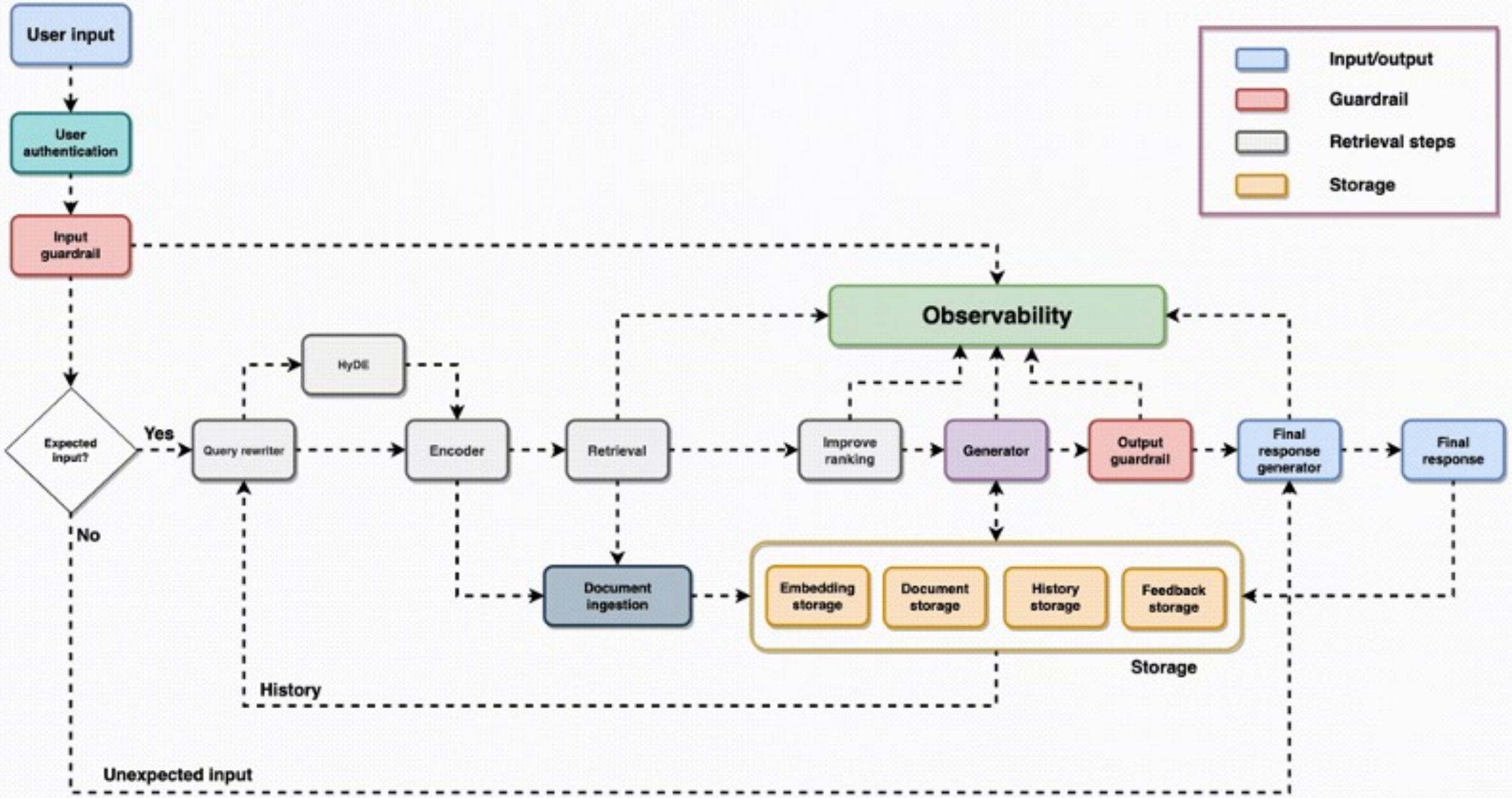


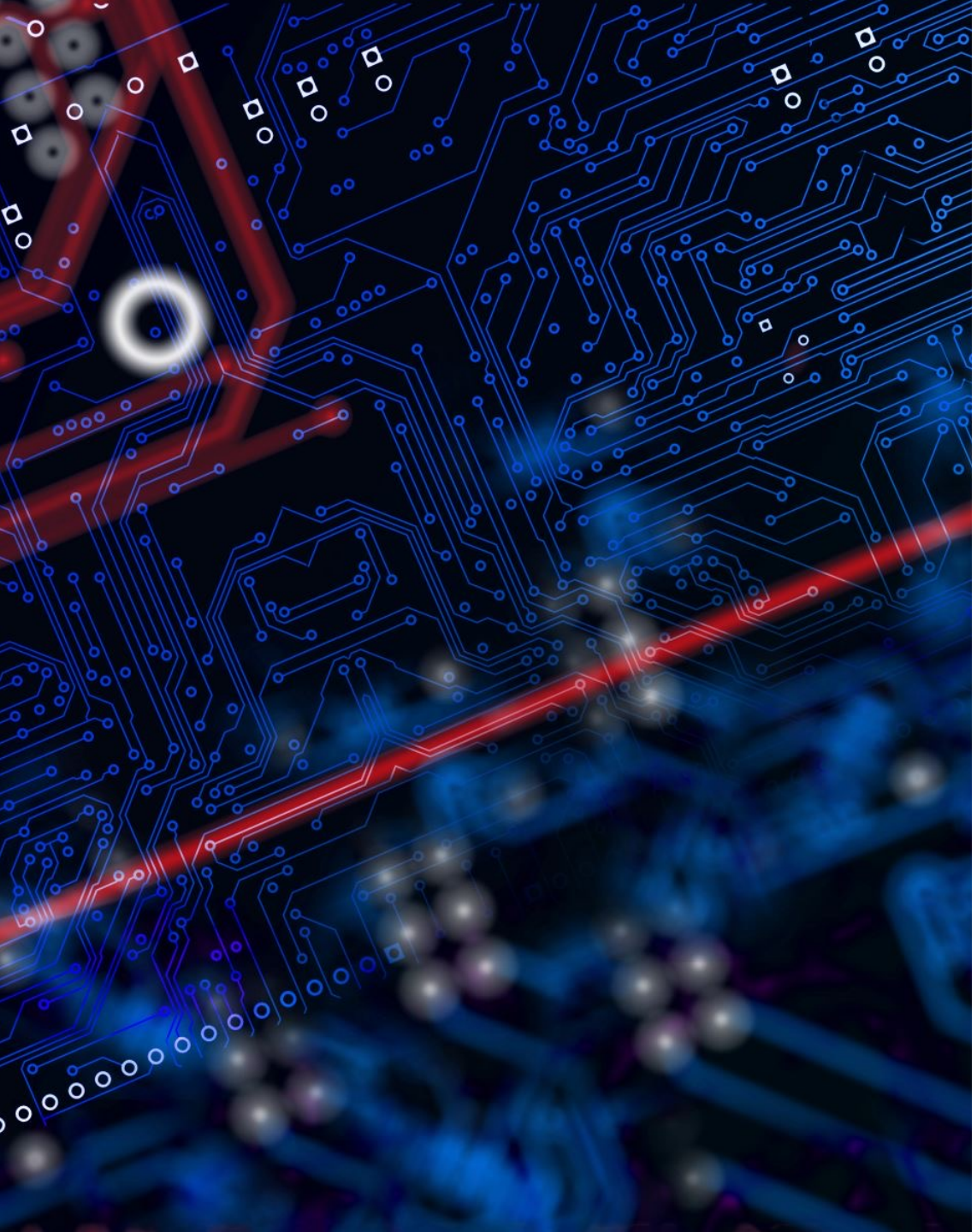

Let us learn how to RAG

RAG


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response
5
Prompt + 4 Query + Enhanced Context
Large Language Model EndPoint

Architecture For Enterprise RAG





RAG

- RAG allows the use of the same model as a reasoning engine over new data provided in a prompt.
    - This technique enables **in-context learning** without the need for expensive fine-tuning, empowering businesses to use LLMs more efficiently.
- RAG combines an information retrieval component with a text generator model.

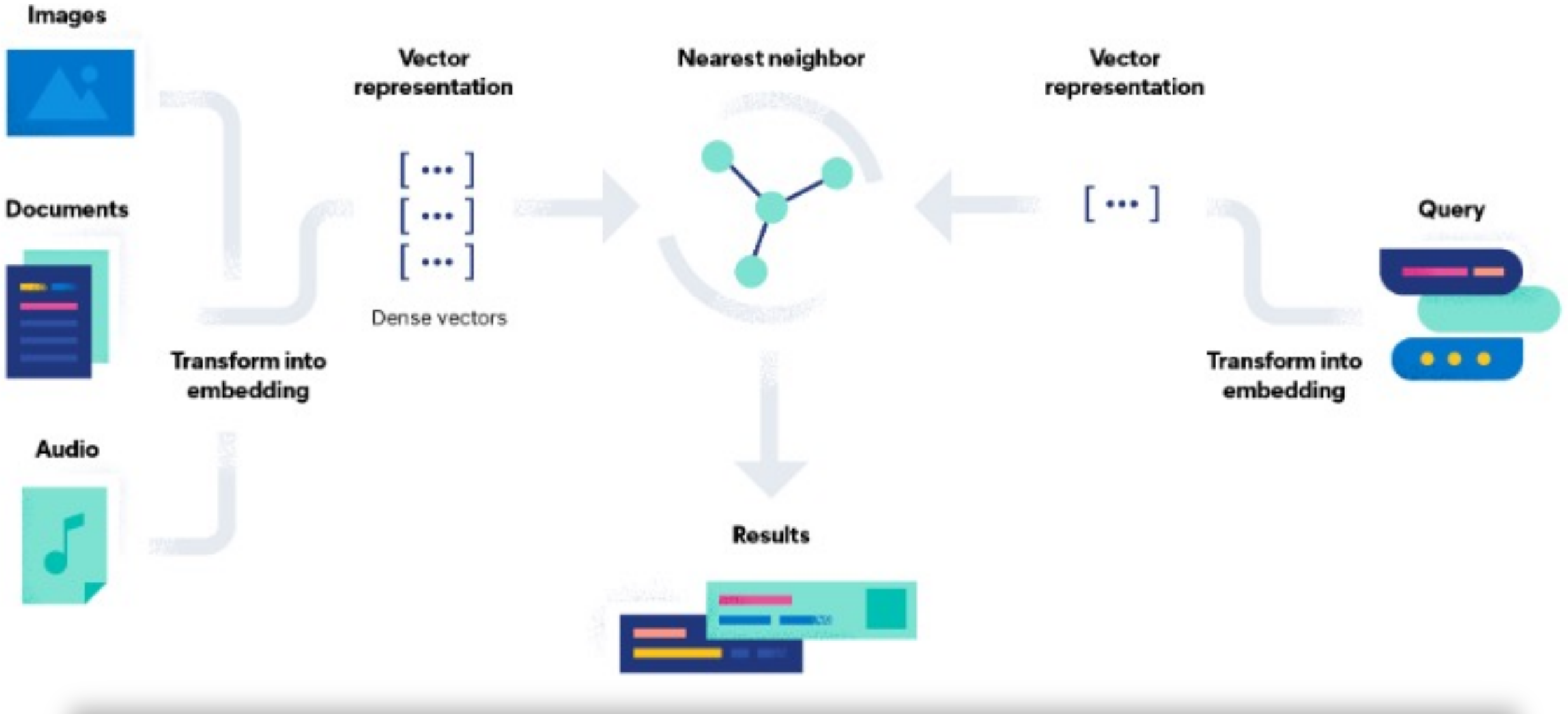
RAG - Information retrieval component

- The data are broken down into manageable chunks.
- Then they are converted into vector embeddings.
- The database indexes the embeddings for rapid retrieval.
- When a user queries the database, it computes similarity metrics between the chunk vectors and returns matches.
- Vector databases then precompute certain similarities between the vectors to further speed up the queries.

Vector Databases

- A vector database is a database that stores information as vectors, which are numerical representations of data objects, also known as vector embeddings.

- A vector database is different from a vector search library or vector index: it is a data management solution that enables metadata storage and filtering, is scalable, allows for dynamic data changes, performs backups, and offers security features.



Vector database pipeline

How does a vector database work?

- A vector database works by using algorithms to index and query vector embeddings.
- The algorithms enable approximate nearest neighbour (ANN) search through hashing or graph-based search.
- To retrieve information, an ANN search finds a query’s nearest vector neighbor.
- Less computationally intensive than a kNN search (known nearest neighbor, or true k nearest neighbor algorithm), an approximate nearest neighbor search is also less accurate.

Pre-processing & Post-processing

- Post-processing: The final step in a vector database pipeline is sometimes post-processing, or **post-filtering**, during which the vector database will use a different similarity measure to re-rank the nearest neighbors.
  - At this stage, the database will filter the query's nearest neighbors identified in the search based on their metadata.

- Some vector databases may apply filters before running a vector search. In this case, it is referred to as preprocessing or **pre-filtering**.

RAG going wrong

- **Air Canada Loses Court Case After Its Chatbot Hallucinated Fake Policies To a Customer**
- The airline argued that the chatbot itself was liable. The court disagreed.

AIR CANAD