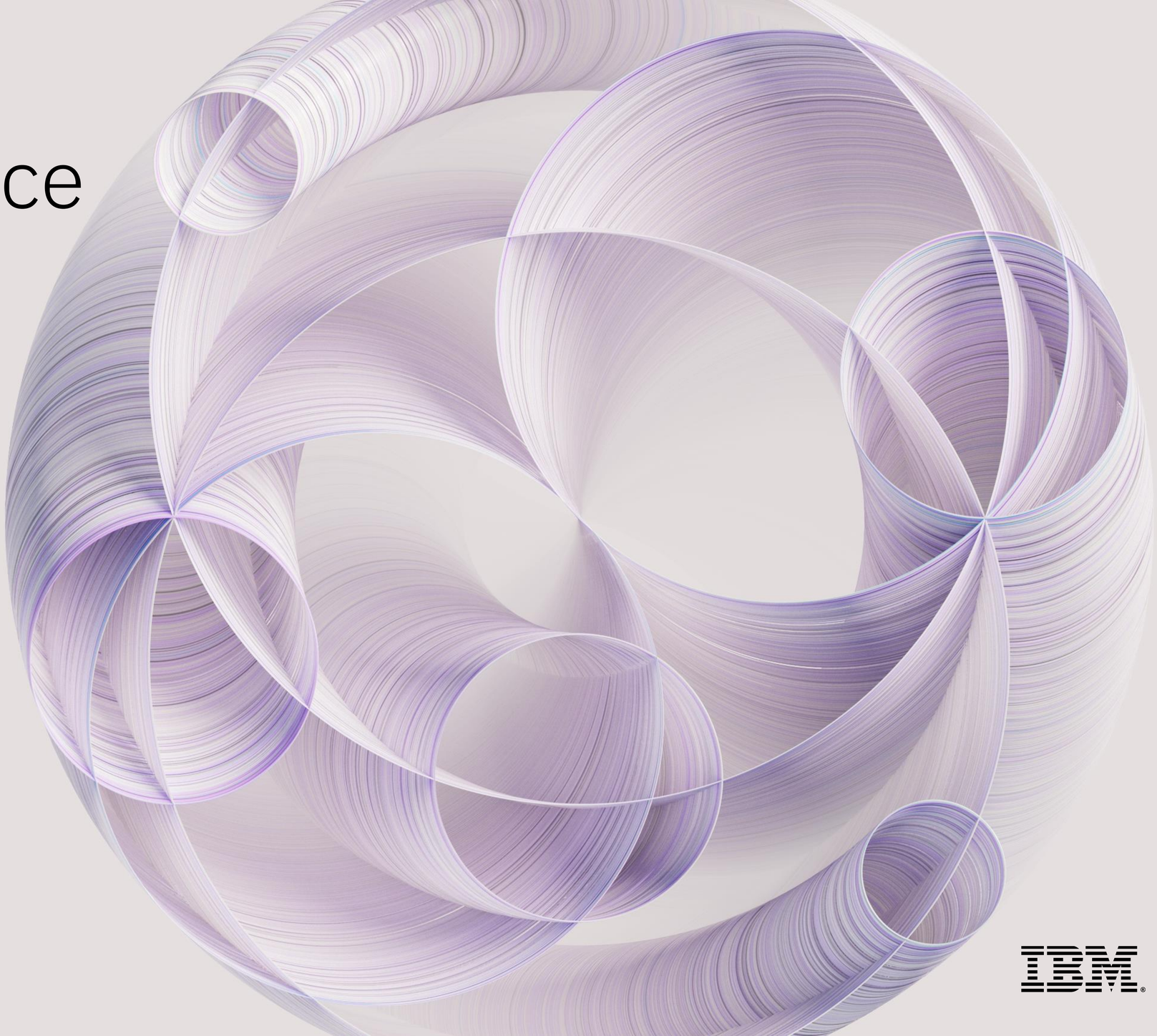
Watsonx.ai Proof of experience (PoX) reference

PoX architecture and other diagrams

Felix Lee
felix@ca.ibm.com



Seller guidance and legal disclaimer

IBM and Business Partner
Internal Use Only

Slides in this presentation marked as **"IBM and Business Partner Internal Use Only"** are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by client.

About the diagrams

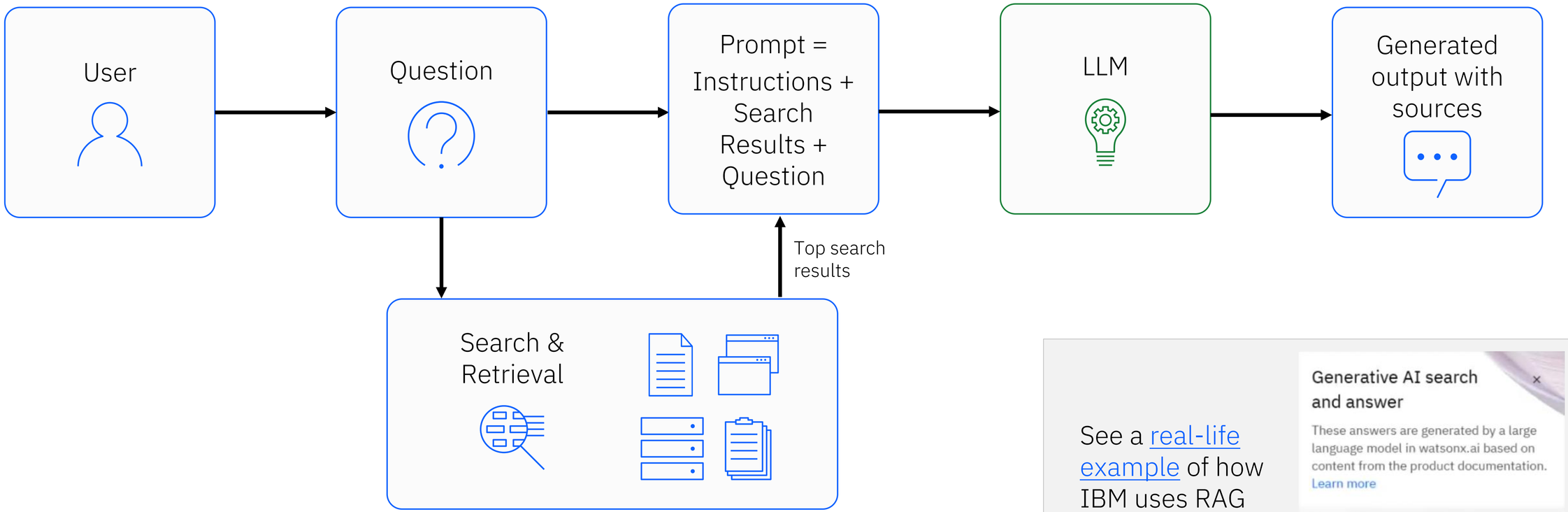
In a watsonx.ai PoX, clients will have many business requirements that the PoX intends to demonstrate. These may include:

- How watsonx components work together (especially watsonx.governance)
- How Watsonx.ai work with other IBM services such as:
  - Watson Discovery
  - watsonx Assistant
  - watsonx Code Assist

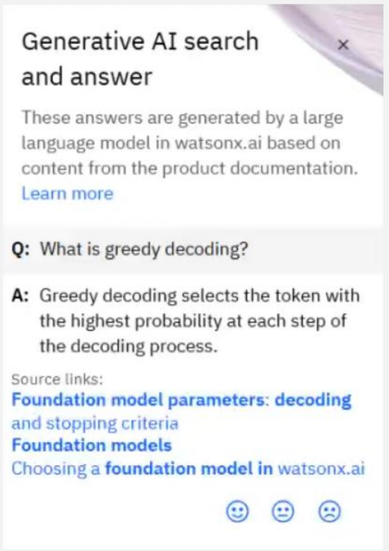
Sellers can leverage some of the diagrams included in this section.

- The diagrams are meant to be generic and provided as-is, as examples.
- Sellers can customize the diagrams to fit what a client is looking for in the PoX
- Sellers can use the diagrams as references to create new architectural diagrams to support a client’s use case
- The Seismic source is provided in the Speaker Notes. Note that some slides do not come from Seismic.

Typical Retrieval Augmented Generation process



See a real-life example of how IBM uses RAG to answer user questions about watsonx.ai in the product's documentation

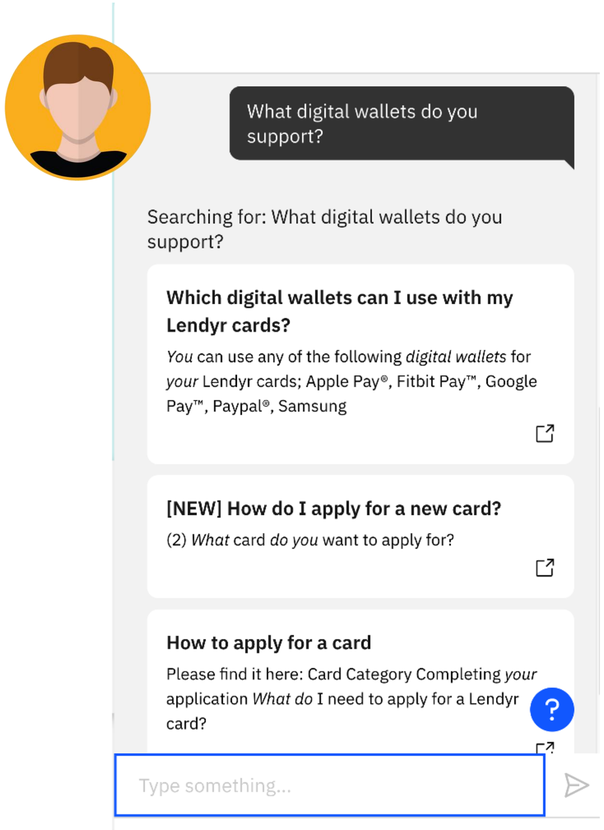


Retrieval Augmented Generation - Phase 1 – data preparation

Validated, filtered, and curated data is stored. These can be files/data in Cloud Object Storage, Box, or other data repositories.

Clients can update them whenever necessary, ensuring that only the latest is used.

Retrieval Augmented Generation - Phase 2 – data retrieval



User's question

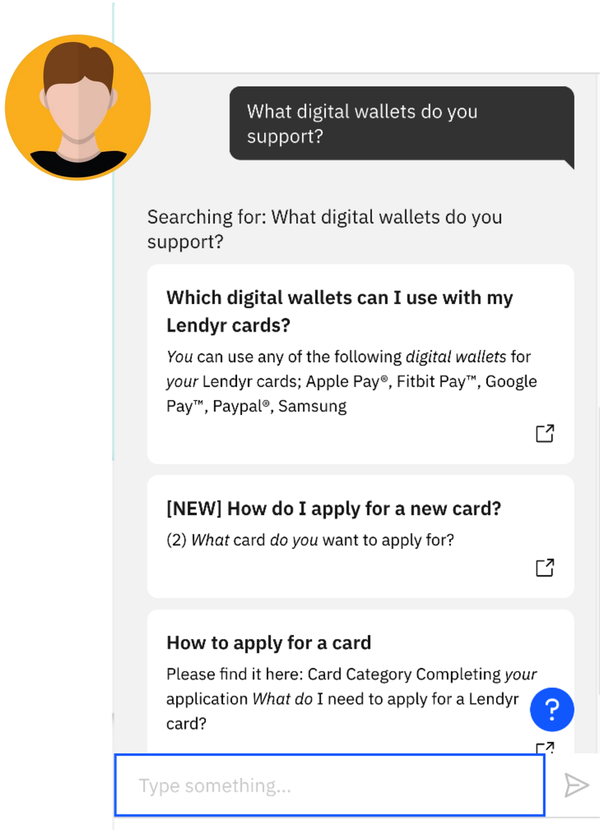
Few paragraphs related to the question

Can be data/information cataloged
Watson Discovery, or a vector database

Read the following paragraphs and answer the question:
[Paragraph 1]
[Paragraph 2]
[Paragraph 3]
Question: [User's question]
Answer:

watsonx.ai

Retrieval Augmented Generation – Phase 3 – generate completion



User's question

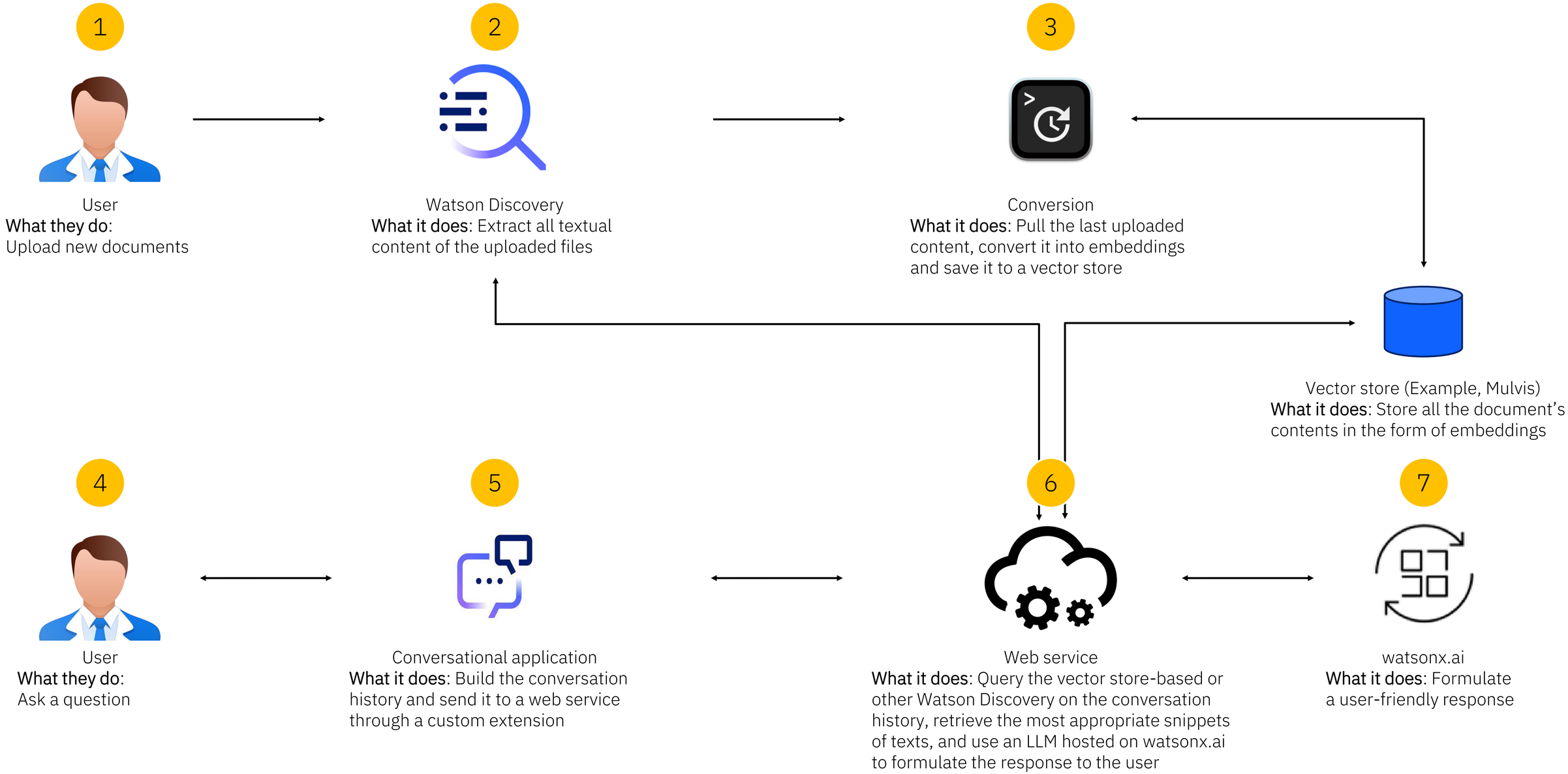
Few paragraphs related to the question

Can be data/information cataloged in Watson Discovery, or a vector database

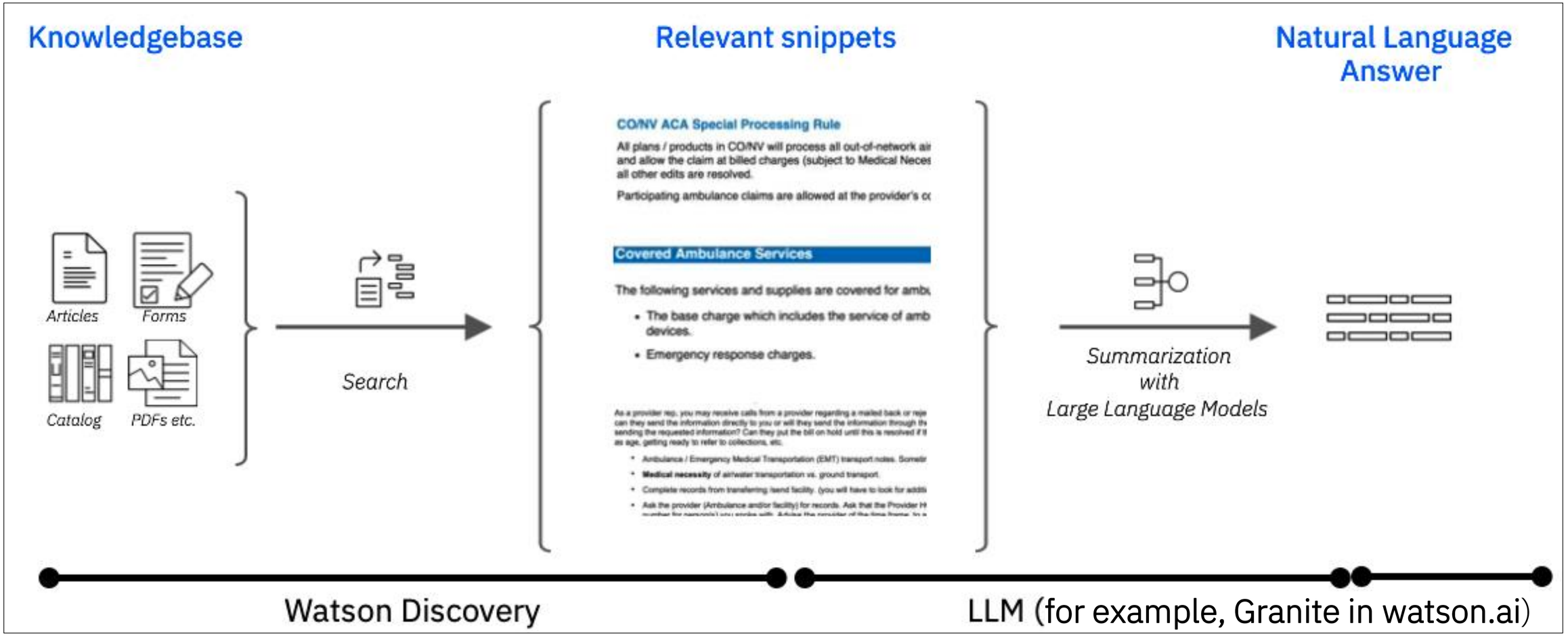
Read the following paragraphs and answer the question:
[Paragraph 1]
[Paragraph 2]
[Paragraph 3]
Question: [User's question]
Answer:

watsonx.ai

Retrieval Augmented Generation – 2 scenarios



Watson Discovery in RAG pattern



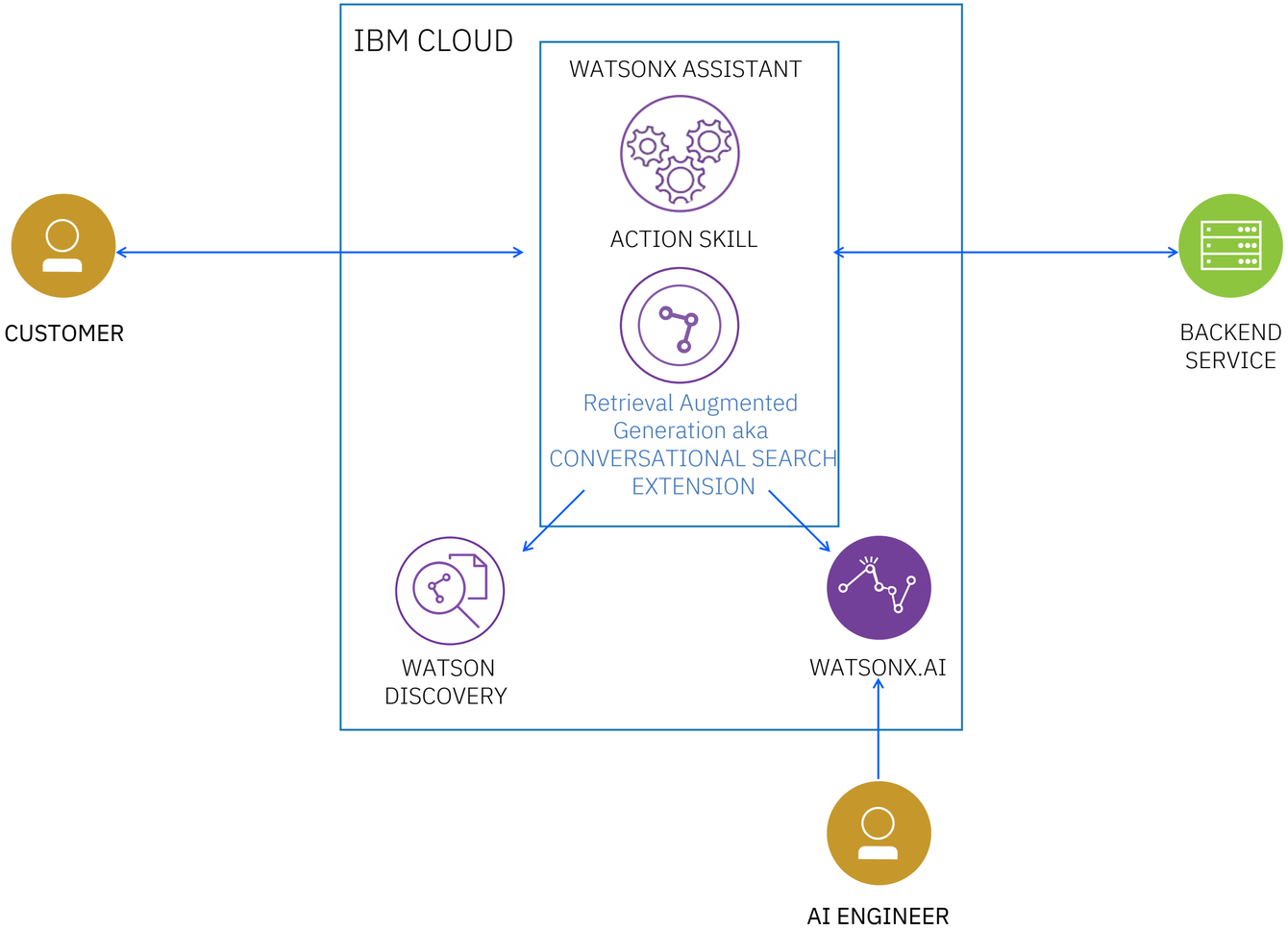
Native conversational search with RAG pattern

watsonx Assistant + Watson Discovery (Search) + watsonx.ai (LLM)

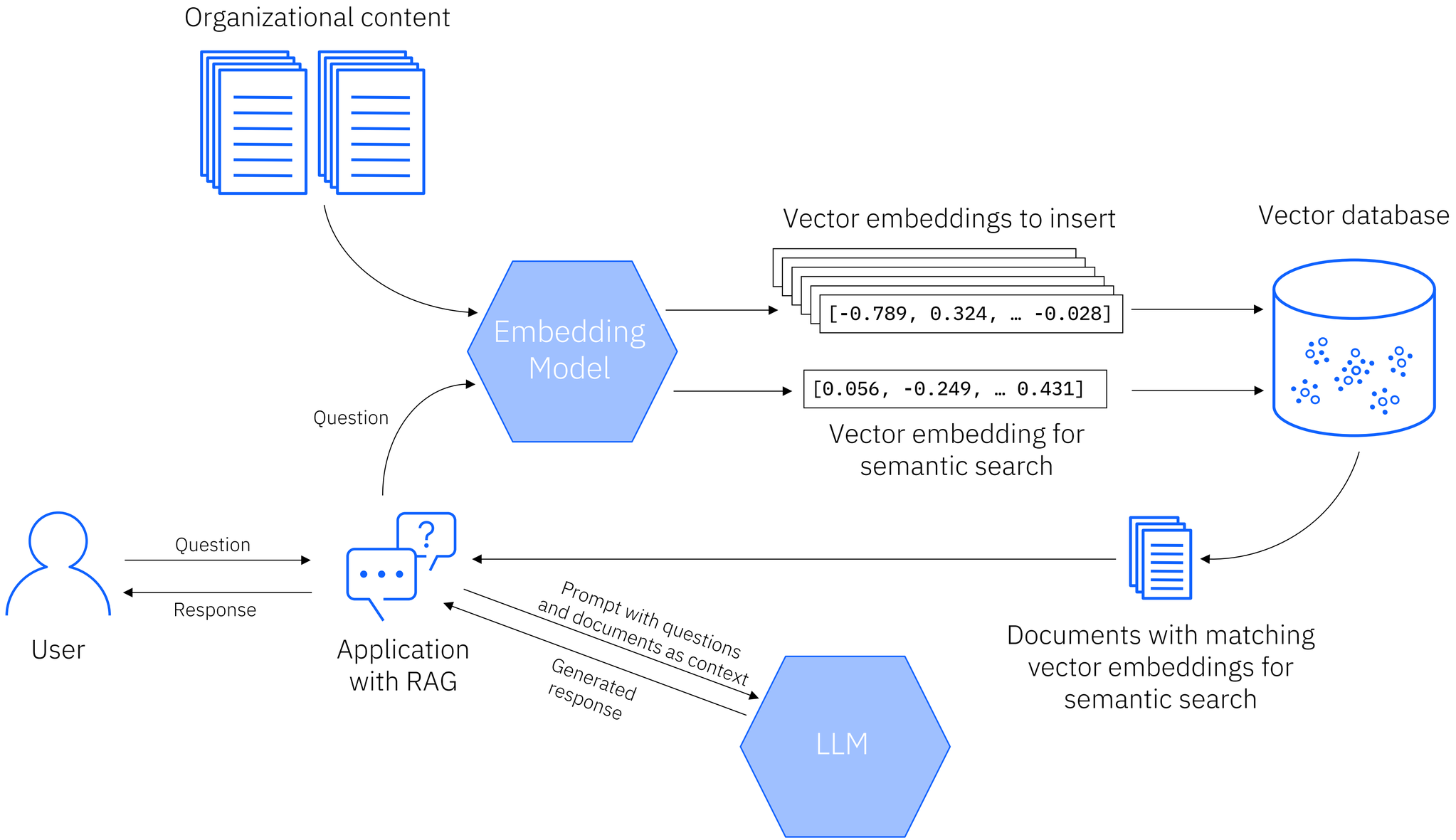
Conversational Search via Extensions

Use Watson Discovery for search and LLMs (running on watsonx.ai) to generate content-grounded conversational answers to customer and employee questions.

You can find sample code for RAG or bring your own. Then call it through watsonx Assistant extension to present the answer to the end user.



Retrieval Augmented Generation (RAG) with a vector database

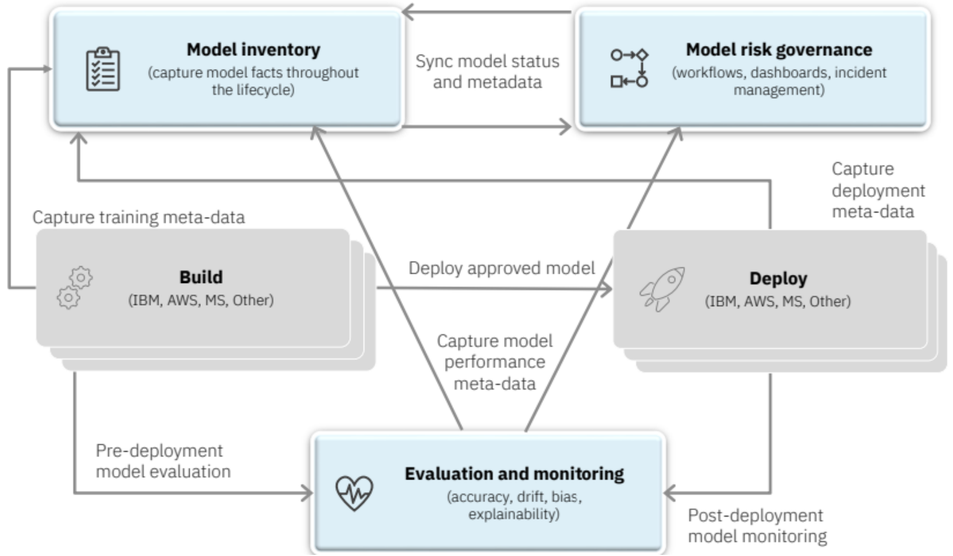


watsonx.ai and watsonx.governance

watsonx.ai

A next-generation enterprise studio for AI builders to train, validate, tune and deploy generative AI, foundation models, and machine learning capabilities

- Foundation Model Library with IBM and open-source models
- Prompt Lab to experiment with foundation models and build prompts for various use cases and tasks
- Tuning Studio to tune your foundation models with labeled data
- Data Science and MLOps to build machine learning models automatically with model training, development and visual modeling



watsonx.governance

Accelerate responsible, transparent and explainable AI workflows across the entire lifecycle.

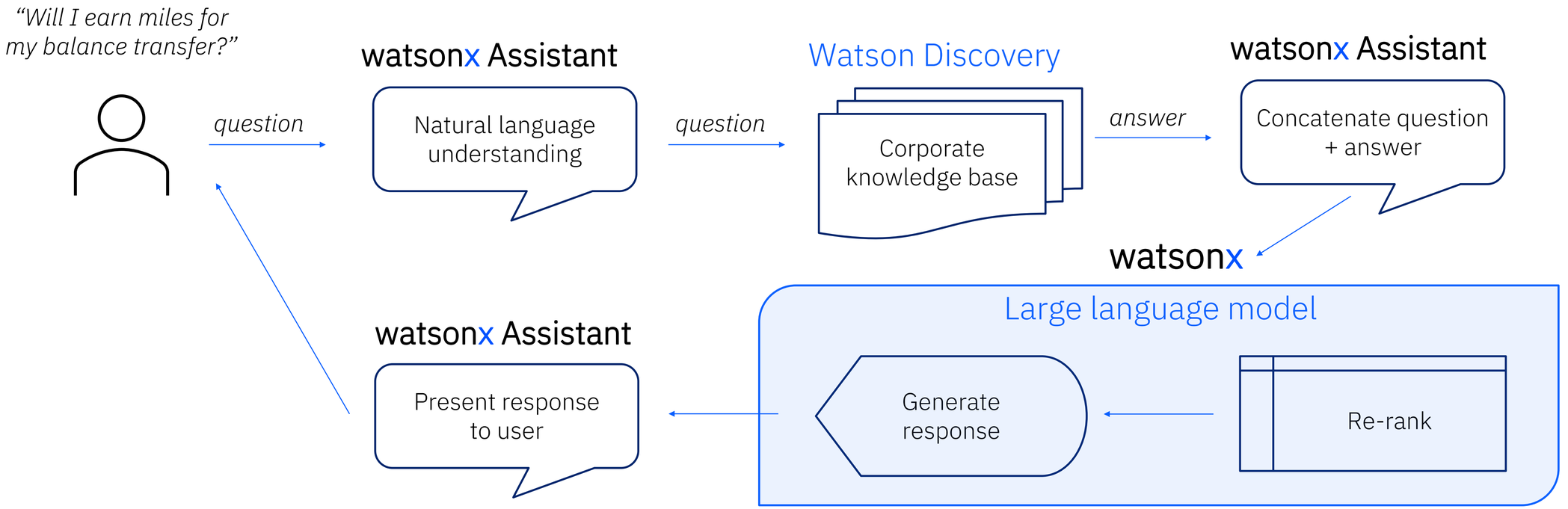
- Automate and consolidate tools, applications, and platforms
- Govern ML models, including those from 3rd parties and generative models
- Manage risk and protect reputation-setting tolerances to proactively detect bias & drift
- Capture metadata and document lineage throughout the model lifecycle
- Improve adherence to AI regulations such as the proposed EU AI Act, internal policies and industry standards
- Improve collaboration & communication with customizable dashboards & reports

Build – train – validate – tune – deploy

manage – monitor – retrain – document facts

Conversational search

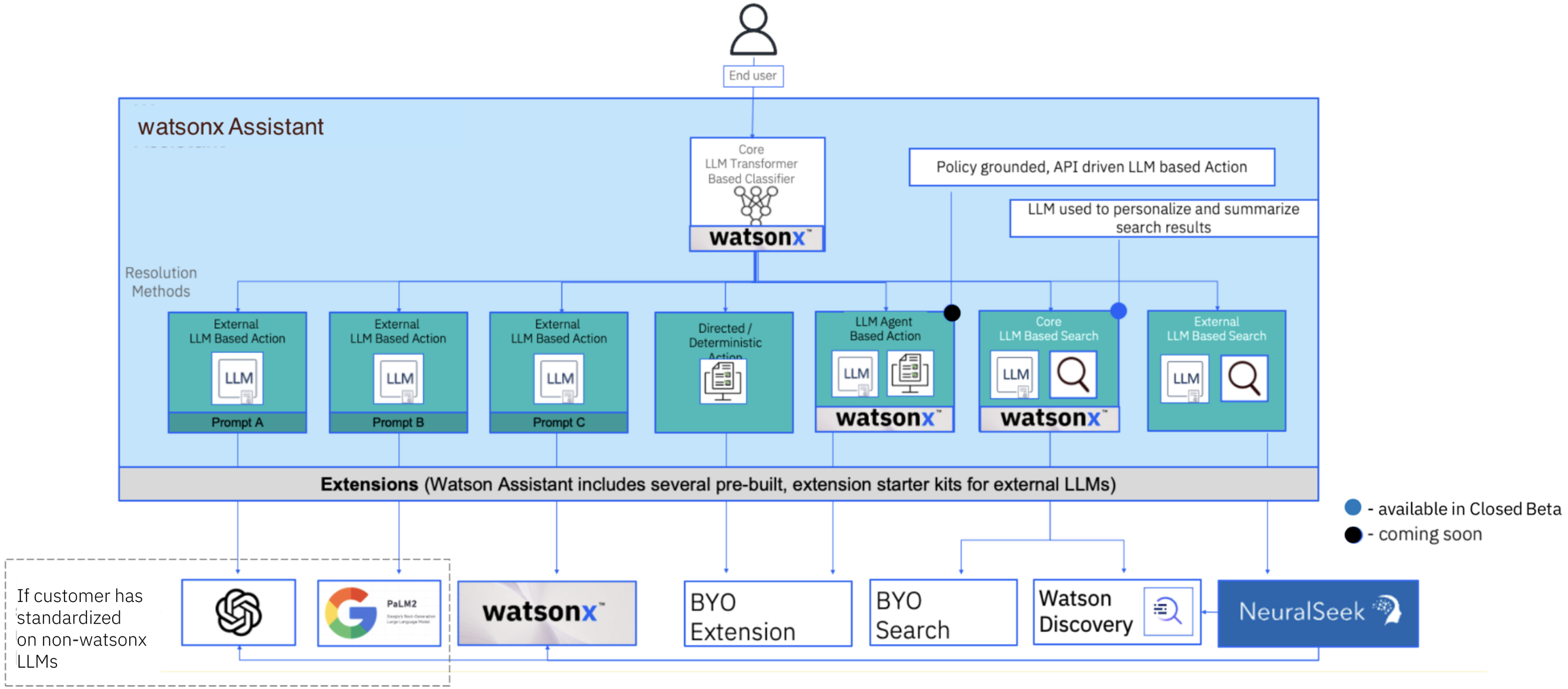
order of operations and the value add of watsonx Assistant



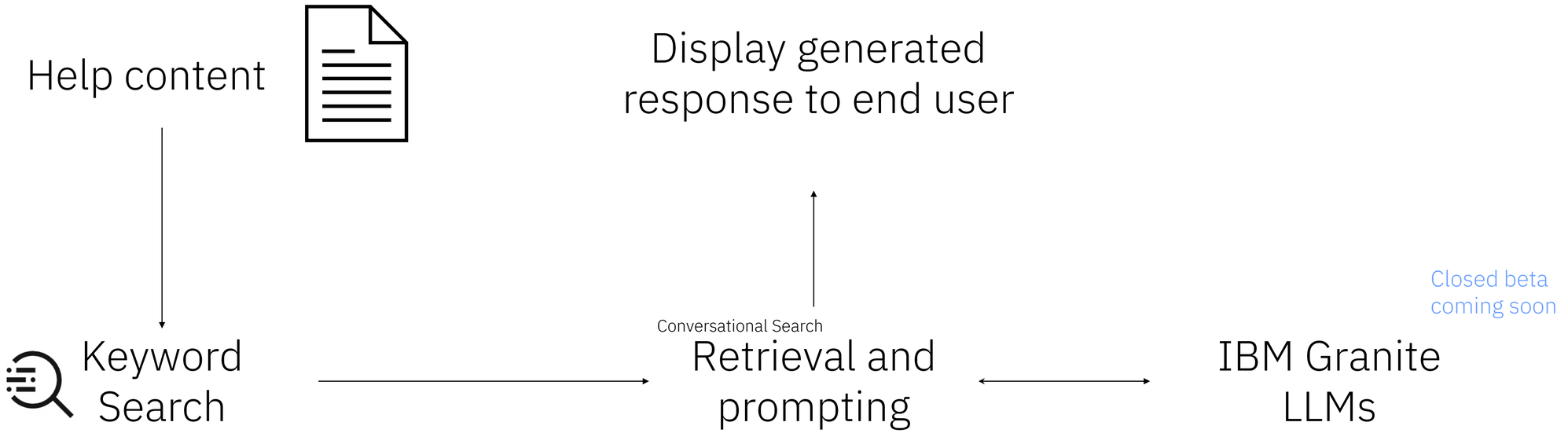
Conversational Search

Architectural patterns

watsonx Assistant connects natively to customized watsonx LLMs and company-specific content



Conversational search architecture for SaaS



Customers purchase blue box

Watson Discovery + watsonx Assistant

IBM Cloud

Data and watsonx (focus is on data)