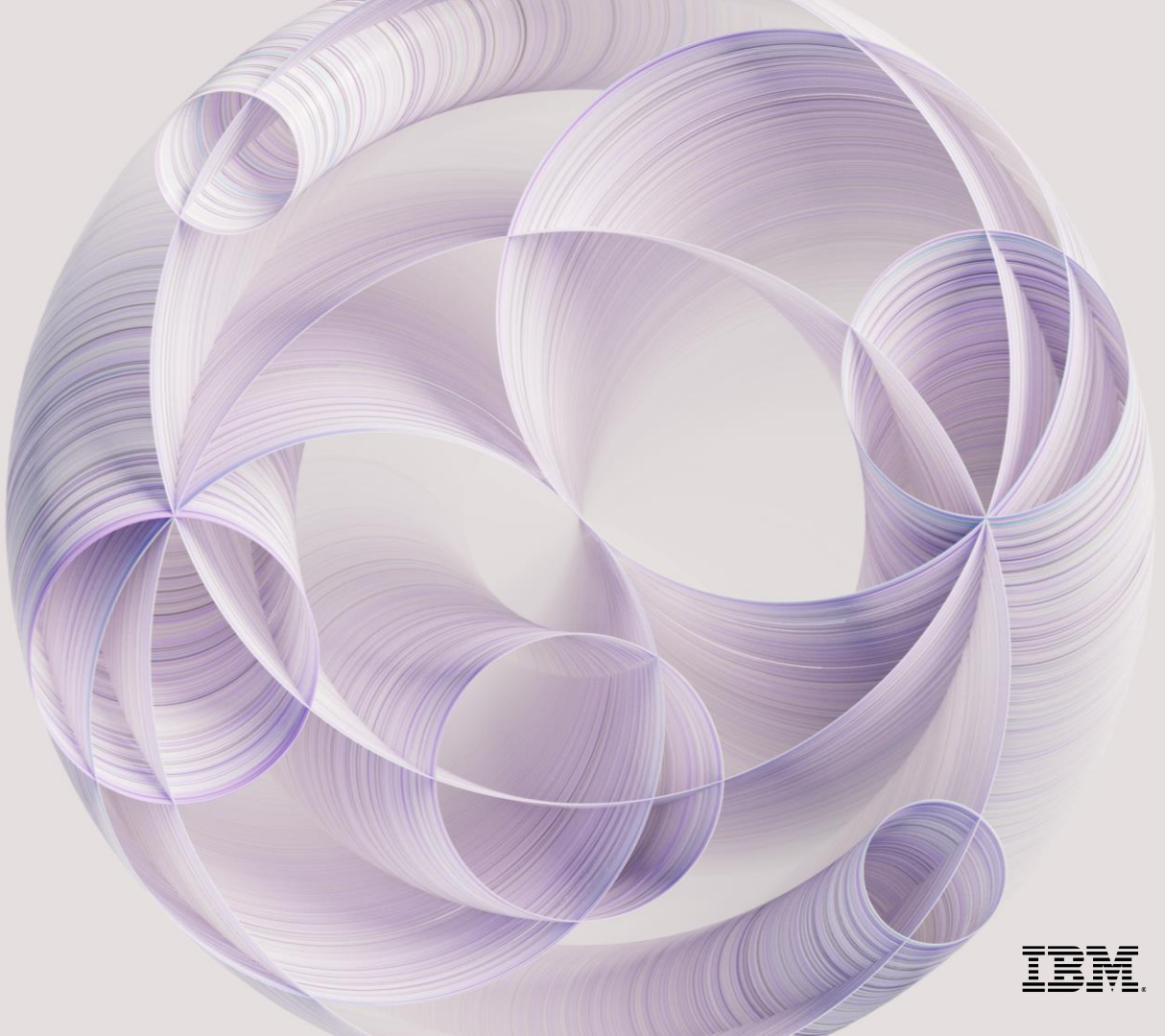
watsonx.ai knowledge sharing

Randy Phoa
Client Engineering Leader



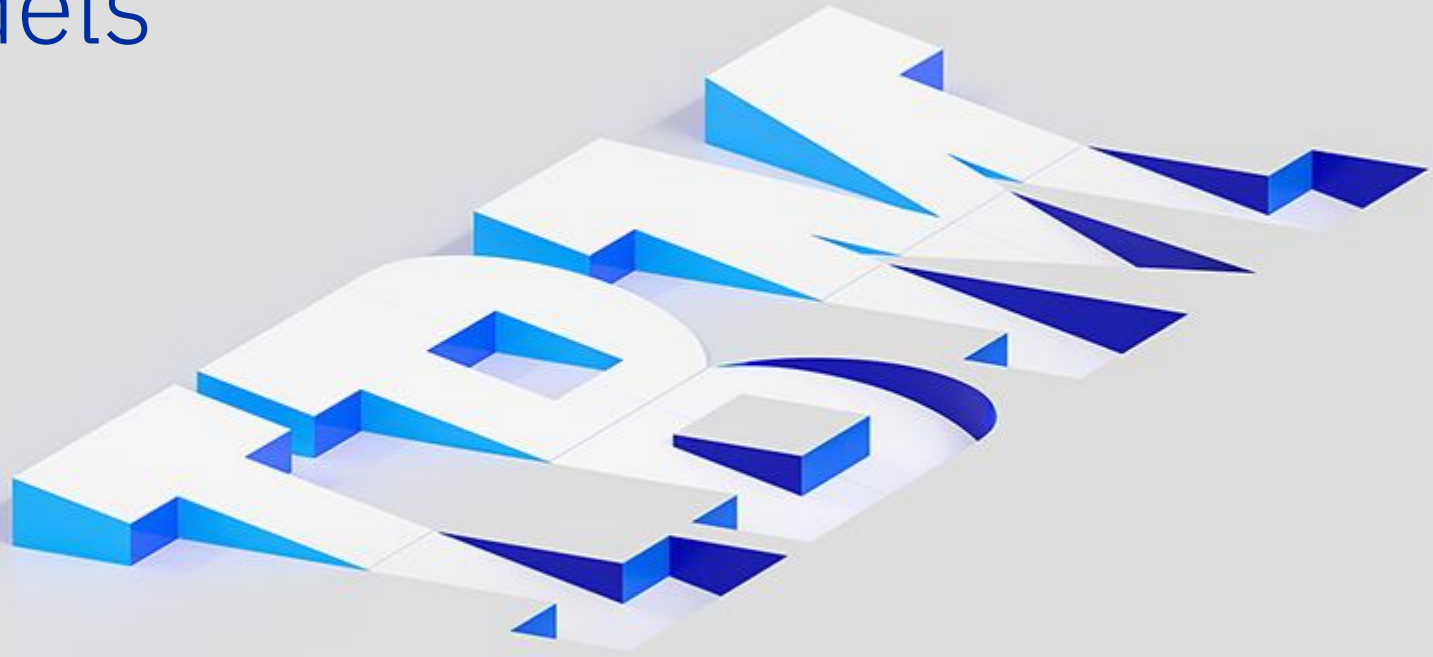
watsonx.ai

IBM

Agenda

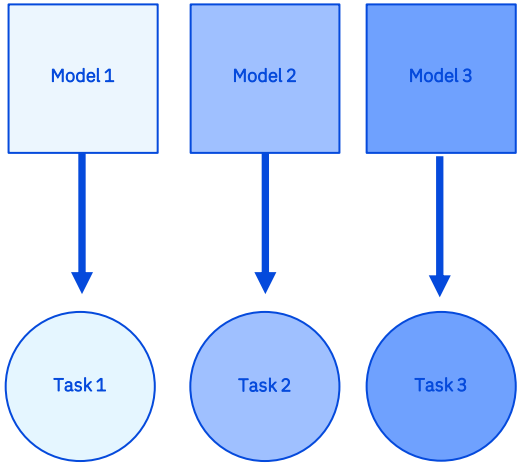
1. Foundation models
2. watsonx.ai platform
3. Retrieval Augmented Generation

Foundation models



Traditional AI Models

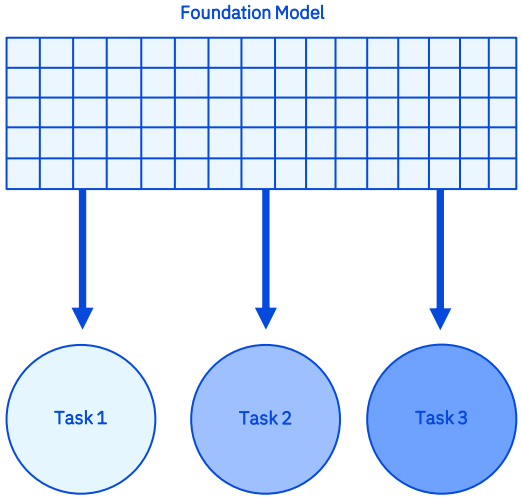
Each model is trained for a specific task



1,000s to 1,000,000s labeled data points per task

Foundation Models

One model that can address many tasks



0 to 1,000s labeled data points per task

watsonx.ai models

Model Categories

fm.language
fm.code

} Initial categories for watsonx launch

Note: language and code foundation models are typically referred to as LLMs (large language models). The categories fm.language and fm.code together represent LLMs which represent our primary focus for the launch.

Model Architectures

Encoder-only	Slate	Best cost performance trade-off for non generative use cases but require task-specific labeled data for fine tuning
Encoder-Decoder	Sandstone	Support both generative and non-generative use cases. Best cost performance trade-off for generative use cases when input is large but generated output is small.
Decoder-only	Granite	Designed explicitly for generative AI use cases; represents the architectures used in GPT-3 and other popular LLMs.

GA Models

IBM Foundation Models

Slate (encoder-only) Natural Language Processing Models

Slate
153 million params
multilingual distilled

Fine Tuning ***Required*** to support:

- Extract
- Classify

See next page for IBM fine-tuned versions

Open-Source Large Language Models

Encoder/decoder & decoder-only Large Language Models available in *Prompt lab*
(Tuning NOT required for most tasks)

flan-ul2-20b	gpt-neox-20b	mt0-xxl-13b	flan-t5-xxl-11b	mpt-instruct2-7b
20 billion params	*20 billion params*	*13 billion params*	*11 billion params*	*7 billion params*
encoder/decoder	decoder only	encoder/decoder	encoder/decoder	decoder only
Q&A	Q&A	Q&A	Q&A	Q&A
Generate	Generate	Generate	Generate	Generate
Extract		Extract	Summarize	
Summarize		Summarize	Classify	
Classify		Classify		

Note: Slate models can be fine-tuned via notebooks and APIs

Model variety to cover enterprise use cases and compliance requirements

Language Tasks

Q&A	Model responds to a question in natural language
Generate	Model generates content in natural language
Extract	Model extract entities, facts, and info. from text
Summarize	Model creates summaries of natural language
Classify	Model classifies text (e.g., sentiment, group)

LLM (flan-ul2) training data

https://github.com/google-research/FLAN/blob/main/flan/v2/flan_templates_branched.py

Physical Interaction

https://huggingface.co/datasets/piqa

How would someone go about accomplishing this goal?

How can I make chocolate ice cream?

In a medium mixing bowl combine 4 cups chilled whipping cream, 14 oz. can sweetened condensed milk, and 1 16 ounce can chocolate flavored syrup (Like Hershey's). Beat with an electric mixer until soft Peaks form. Fold in 3/4 cup toasted chopped peanuts. Transfer mixture to 8x8x2 inch baking pan. Freeze about 8 hours until firm. Serve any time after 8 hours freeze.

OpenBookQA

https://huggingface.co/datasets/openbookqa

Fact: Camouflage is used for hiding by animals from predators

Question: A mouse is being hunted by a Great Horned Owl. The owl watches the mouse carefully, and the mouse is unaware that it is being hunted. The mouse is unable to tell where the owl is, because

What is the answer?

- the owl is invisible
- the mouse is oblivious
- the mouse is blind
- the owl matches its background

the owl matches its background

Samsum

https://huggingface.co/datasets/samsum

Summarize the below text.

Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I'll bring you tomorrow :-)

Amanda baked cookies and will bring Jerry some tomorrow.

Math Dataset

https://huggingface.co/datasets/math_dataset

Solve 24 = 1601*c - 1605*c for c.

-6

Advance Functionalities

ReAct (Reasoning + Acting)

https://python.langchain.com/docs/modules/agents/agent_types/react

"Who is Leo DiCaprio's girlfriend? What is her current age raised to the 0.43 power?"

I need to find out who Leo DiCaprio's girlfriend is and then calculate her age raised to the 0.43 power.

- Thought: I need to find out who Leo DiCaprio's girlfriend is
- Action: Search
- Action Input: "Leo DiCaprio girlfriend"
- Observation: Camila Morrone

- Thought: I need to find out Camila Morrone's age
- Action: Search
- Action Input: "Camila Morrone age"
- Observation: 25 years

- Thought: I need to calculate 25 raised to the 0.43 power
- Action: Calculator
- Action Input: 25^0.43
- Observation: Answer: 3.991298452658078

Camila Morrone is Leo DiCaprio's girlfriend and her current age raised to the 0.43 power is 3.991298452658078.

Tools / Agent toolkits

https://python.langchain.com/docs/integrations/toolkits

SQL Database Agent

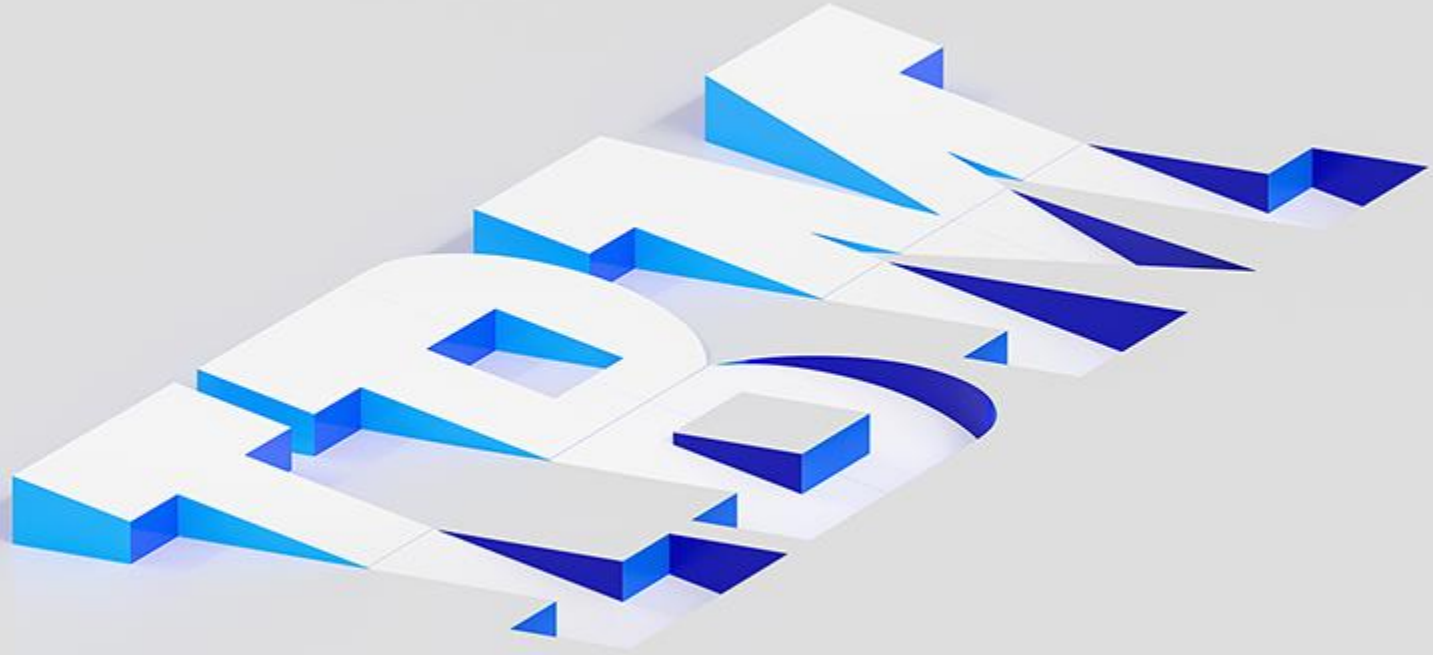
"List the total sales per country. Which country's customers spent the most?"

- Thought: I should look at the schema of the relevant tables to see what columns I can use.
- Action: schema_sql_db
- Action Input: "Invoice, Customer"

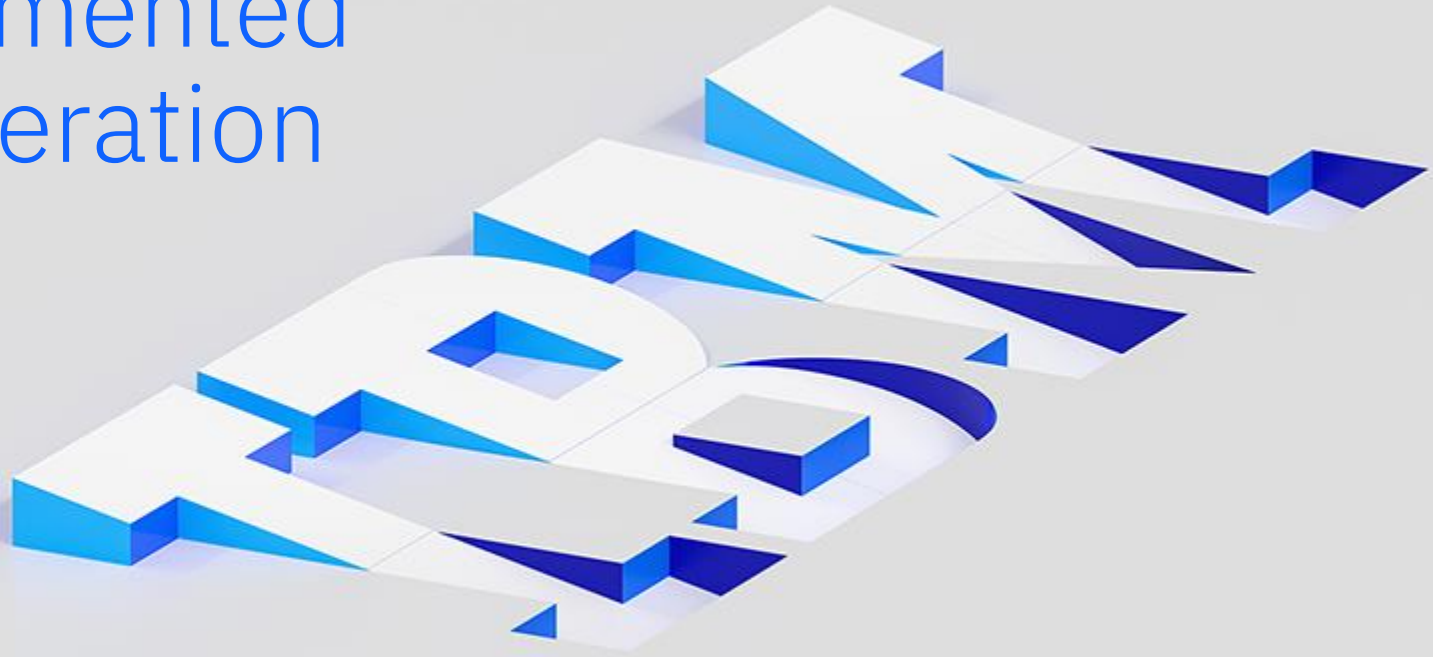
- Thought: I should query the Invoice and Customer tables to get the total sales per country.
- Action: query_sql_db
- Action Input: SELECT c.Country, SUM(i.Total) AS TotalSales FROM Invoice i INNER JOIN Customer c ON i.CustomerId = c.CustomerId GROUP BY c.Country ORDER BY TotalSales DESC LIMIT 10
- Observation: [('USA', 523.0600000000003), ('Canada', 303.9599999999999), ('France', 195.09999999999994), ('Brazil', 190.09999999999997), ('Germany', 156.48), ('United Kingdom', 112.85999999999999), ('Czech Republic', 90.24000000000001), ('Portugal', 77.23999999999998), ('India', 75.25999999999999), ('Chile', 46.62)]

The customers from the USA spent the most, with a total of $523.06.

watsonx.ai platform

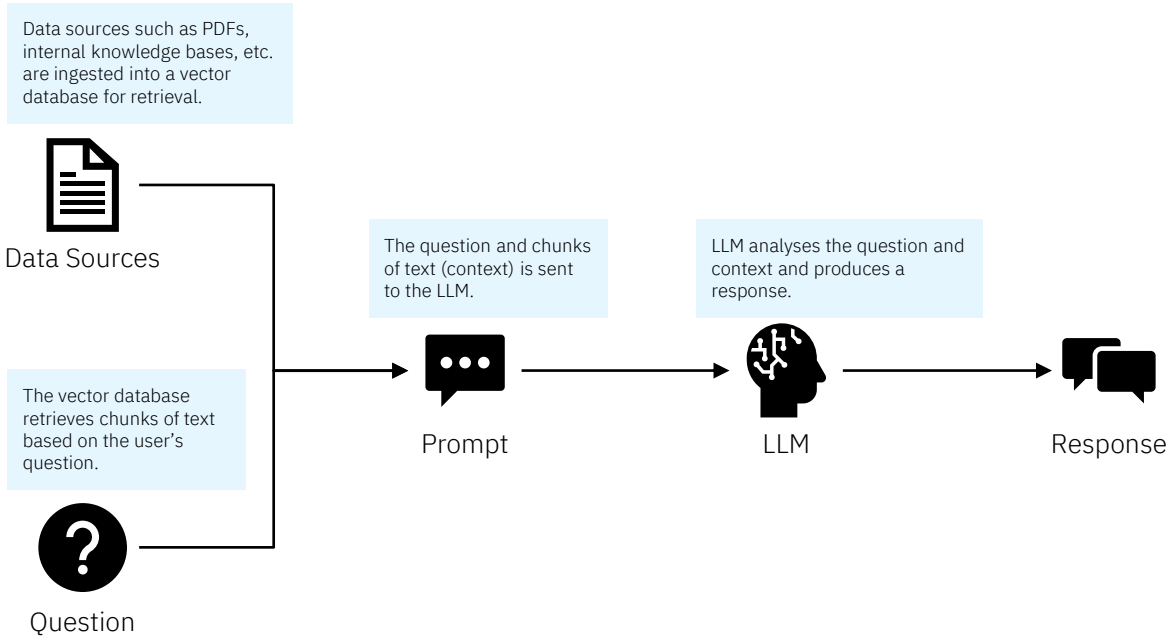


Retrieval Augmented Generation

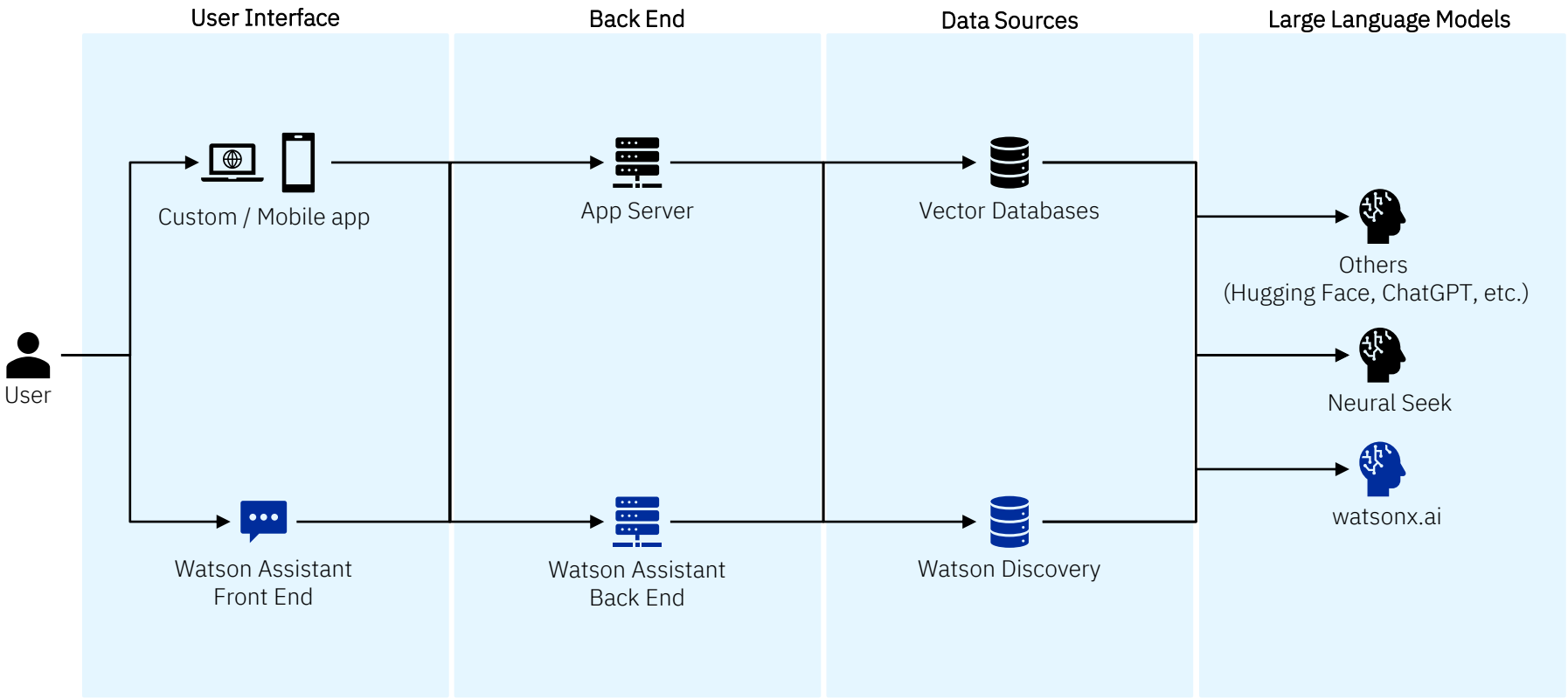


Retrieval Augmented Generation

The main objective of the RAG is to leverage LLM capabilities on a data source based on questions from a user.



RAG Overview



Value Proposition

User Interface

- Watson Assistant covers most of the basic functionalities for a conversational UI
- Many templates available to get started quickly
- Integrations and debugging are relatively simple

Back End

- Readily available out of the box

Data Sources

- Watson Discovery has Smart Document Understanding
- CP4BA has workflows and document processing capabilities like ADP

Large Language Model

- On-premise offering
- AI guard rails
- Curated and cleansed dataset

4 steps in Retrieval Augmented Generation

1. Chunking

Documents are split into chunks to overcome tokens limitation.

2. Embedding

Texts are tokenized and converted into embeddings (word vector space) that captures language semantics.

3. Prompting

Prompts help direct the behavior of Large Language Models to achieve specific tasks.

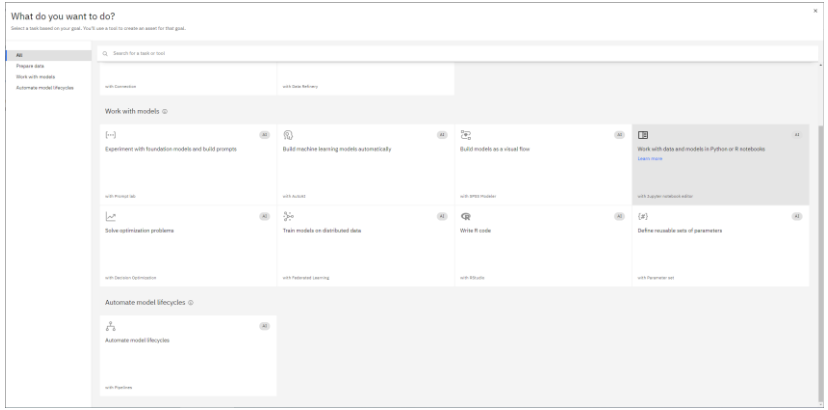
4. Generation

Get response from the Large Language Model with the given prompt.

Hands-on Exercise 1

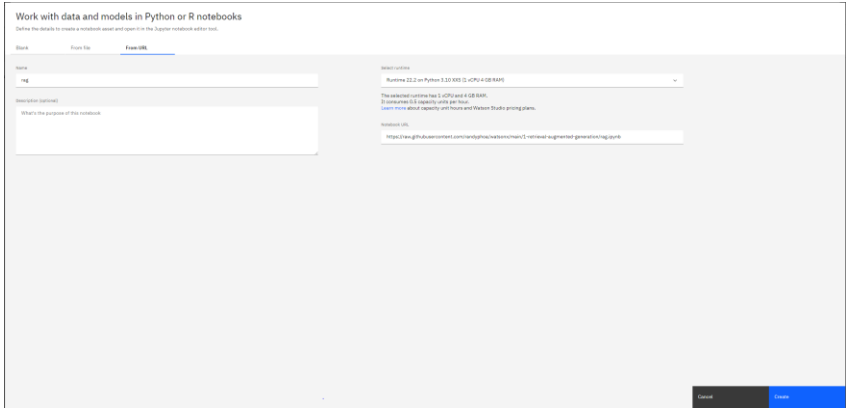
https://github.com/randyphoa/watsonx

Create new Jupyter Notebook



Add Notebook from URL:

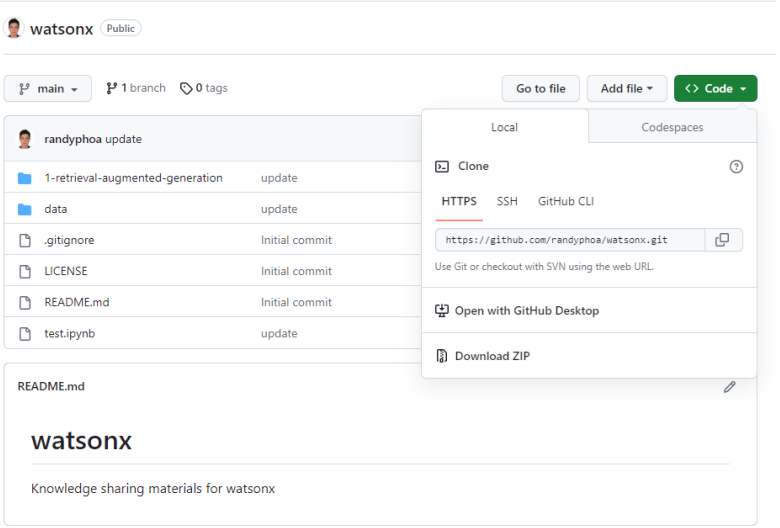
https://raw.githubusercontent.com/randyphoa/watsonx/main/1-retrieval-augmented-generation/rag.ipynb



Hands-on Exercise 2

https://github.com/randyphoa/watsonx

Clone / download repository



Install dependencies and run streamlit.

python –m streamlit run main.py