Deloitte's Three-Agent AI Architecture: Strategic Insights for BC Post-Secondary Institutions

Assisting Deloitte consultants in extracting strategic insights from their latest institutional documents while minimizing hallucination.

The Problems

The research process is time-consuming

The existing AI tool is too generalized to provide tailored responses

Hallucination exists in AI-generated answers

GPT not scalable for large document sets

No persistent retrieval or memory beyond the session

Why is our framework better than just using an OpenAI API?

We choose the embedding model, optimizing for accuracy and cost.

We define how documents are chunked, structured, and refined — which dramatically improves answer quality.

We can scale to thousands of documents and handle complex logic across multiple data types.

We can integrate semantic triples, institutional metadata, and advanced filtering — something no OpenAI assistant can currently do.

Project Workflow Overview

1

PHASE 1 – DATA INGESTION PIPELINE

2

PHASE 2 – BUILD KNOWLEDGE GRAPH

3

PHASE 3 – SEMANTIC SEARCH & EMBEDDING

4

PHASE 4 – CONSULTANT-FACING SEMANTIC INTERFACE

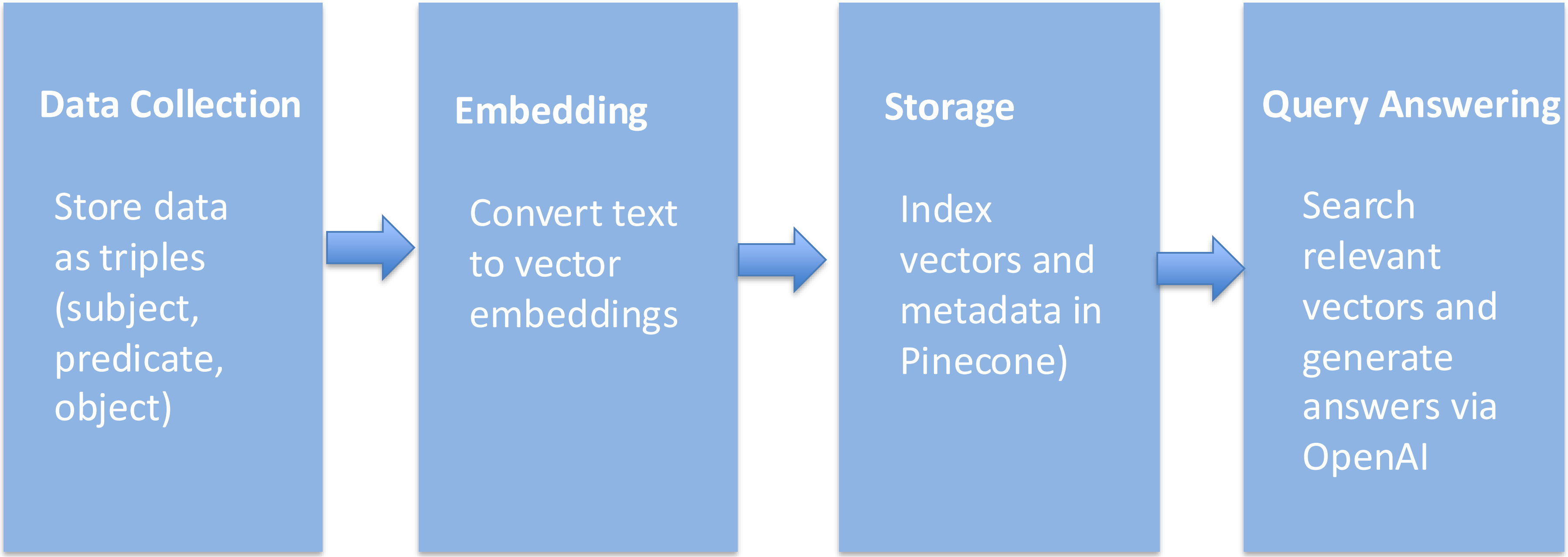
Pinecone-based Question Answering System Overview

System Components

-Pinecone vector database for similarity search

-OpenAI API for embeddings and answer generation

-Query processor handling user interactions



Data Collection

Store data as triples (subject, predicate, object)

Embedding

Convert text to vector embeddings

Storage

Index vectors and metadata in Pinecone)

Query Answering

Search relevant vectors and generate answers via OpenAI

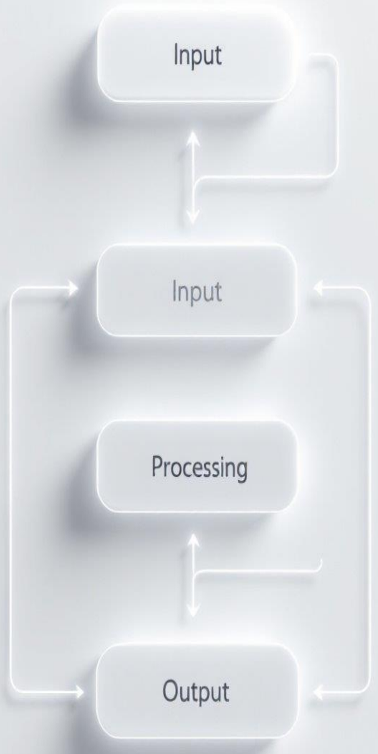
Using Pinecone Vector Database

Database Setup
Connect via Pinecone API key. Create and manage indexes for vector storage.

Text Embedding
Convert text into vectors using OpenAI models. The embed_text() function transforms text into numerical representations.

Query Search
Convert questions into vectors. Search for similar vectors in Pinecone to retrieve relevant information.

Query Answering System Workflow



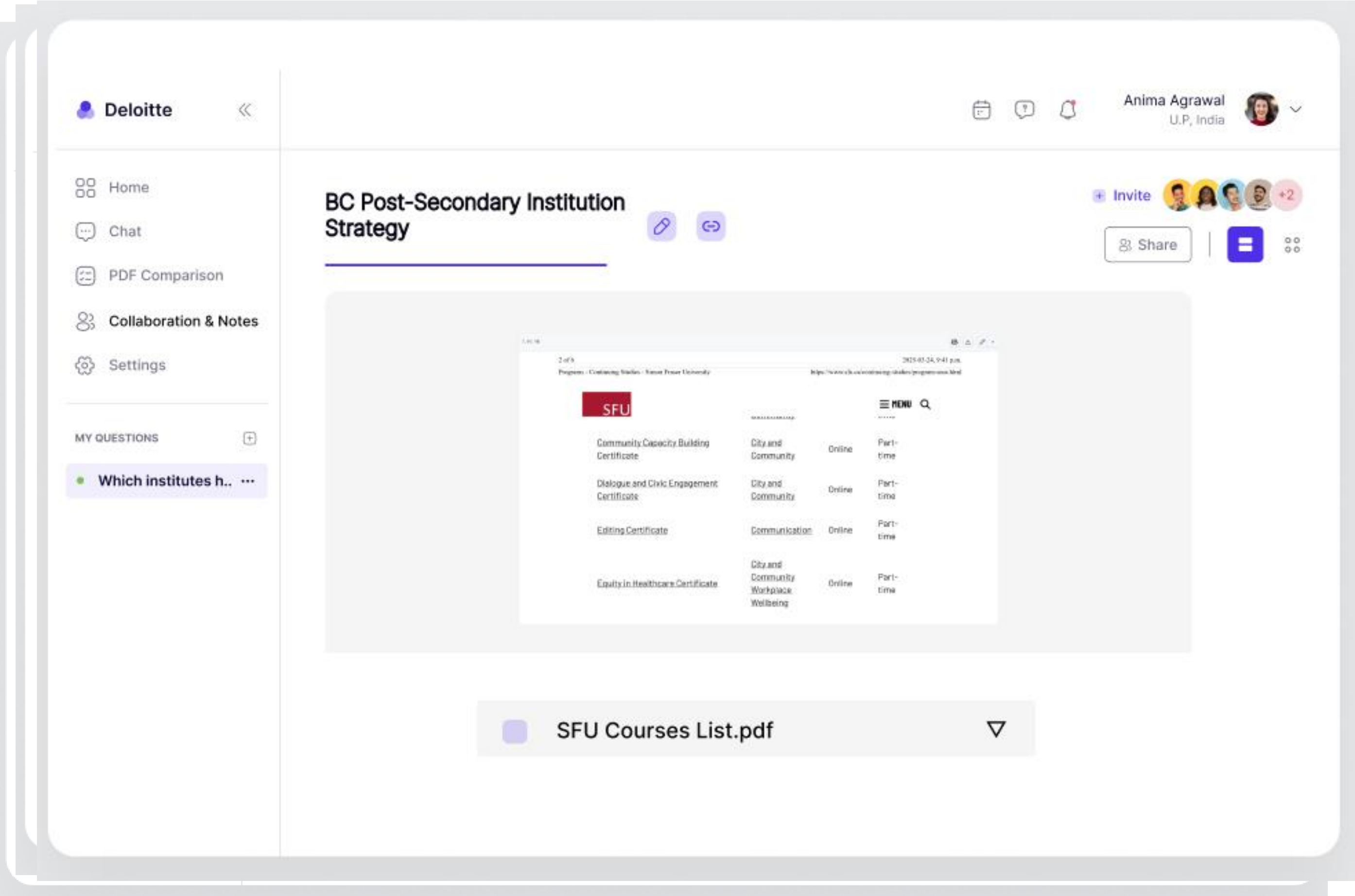
User Query Input
User enters a question like "Tell me about research achievements."

Query Embedding Generation
The get_query_embedding() function converts text to vector format.

Pinecone Similarity Search
The search_pinecone() function finds relevant information triples.

OpenAI Answer Generation
The ask_openai() function sends relevant context to GPT-3 for answer generation.

AI-Powered User Interface



BC Post-Secondary Institution Strategy

2 of 6 2025-03-24, 9:41 p.m.

Programs - Continuing Studies - Simon Fraser University https://www.sfu.ca/continuing-studies/programs.html

SFU

Community Capacity Building Certificate	City and Community	Online	Part-time
Dialogue and Civic Engagement Certificate	City and Community	Online	Part-time
Editing Certificate	Communication	Online	Part-time
Equity in Healthcare Certificate	City and Community Workplace Wellbeing	Online	Part-time

SFU Courses List.pdf

Potential Use Cases for Deloitte Consultants

- Rapid benchmarking of institutional goals, programs, and metrics

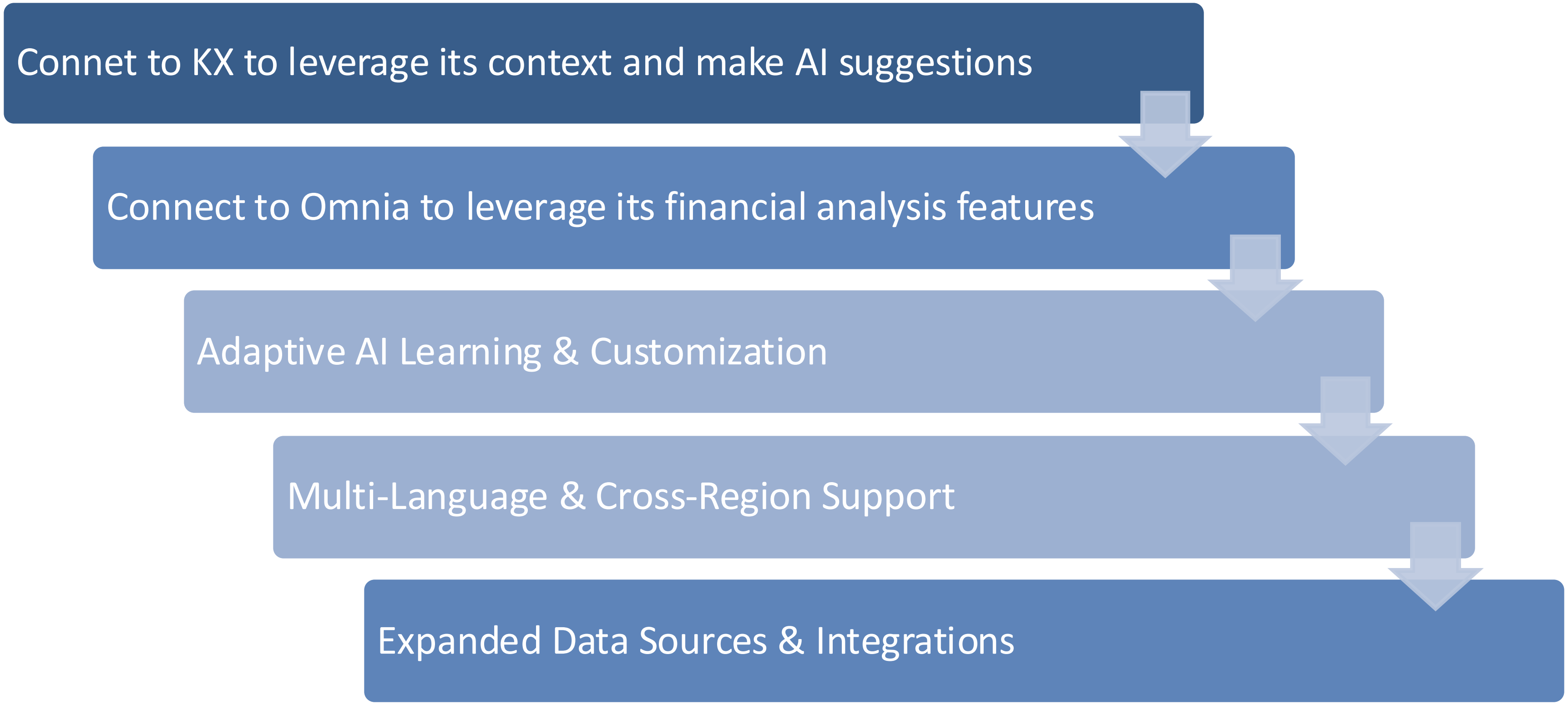
- Identification of curriculum sharing or partnership opportunities

- Automated extraction of KPI-relevant insights from mandate letters and strategic plans

- Accelerated response to RFPs by surfacing aligned institutional capabilities

- Detection of trends across institutions (e.g., sustainability, AI, Indigenous initiatives)

Future Advancements



Connet to KX to leverage its context and make AI suggestions

Connect to Omnia to leverage its financial analysis features

Adaptive AI Learning & Customization

Multi-Language & Cross-Region Support

Expanded Data Sources & Integrations

Thank you