

Generative AI Made Simple



An AI assisted system to respond to common queries

KEY FEATURES

- Point and Click Data Loading of all data types - pdf, image, video, documents, database ..
- Automatic Data Chunking
- Automatic Embedding and Vectorizing Data
- Automatic Saving of Prompts and Ontology Construction
- Text-Based, Charts and Tabular Output
- Feedback Feature
- Registration Utility or Single Sign-on (SSO) Integration

KEY BENEFITS

- Frictionless User Experience of Leveraging AI techniques within their line of business
- Improved Operation Efficiency & Improved Strategic Decision Making
- Deployment of AI-assisted chatbot system for the given documents
- Documentation of the system's functionality
- Documentation of the implementation details

OVERVIEW

Large language models (LLMs) have significantly revolutionized how we engage with information. These models go beyond traditional uses like passing qualification exams or writing academic papers. They now delve into generating text, crafting SQL queries, coding, creating art, and enhancing product support through generative AI - activities that were once considered challenging. This technological progress has sparked interest among corporate leaders, who recognize the potential for generative AI to drive productivity and boost revenues. Nonetheless, these advancements are restrained by limitations in terms of the scope of tasks they can handle. Standard LLMs are restricted to the data they have been trained on, falling short when confronted with queries outside their training parameters. To overcome this obstacle, Retrieval Augmented Generation (RAG) based LLM tools bridge this gap and enhance the applicability of LLMs to specific data sets. By utilizing generative AI models fine-tuned and enhanced with a company's proprietary data and knowledge assets, businesses can produce outputs tailored to the organization's specific requirements in a way only an internally trained model can achieve.

WHAT DO WE DELIVER

- a) A Proof of Concept provides: A framework of processes (an Accelerator) that fast-tracks development, training, testing, tuning and implementation of an AI system that serves as an advisor on selected procedures / policies of the Company
 - Creating a chatbot to answer common queries of business users
 - The chatbot shall provide responses supported by information available in PDF, PPT, Excel, Word, JSON, XML, Image, Audio, Video even hand-written notes, databases provided by the business community.
 - The queries will be in English. It would have the provision to save the query prompt and create ontology and FAQ over time.
 - Access to the chatbot would be via a browser
 - User testing of the product before final deployment - based on simulated random queries by subject matter experts from the business community
- b) Setting up the required infrastructure based on customer's preferences and requirements

Generative AI Made Simple



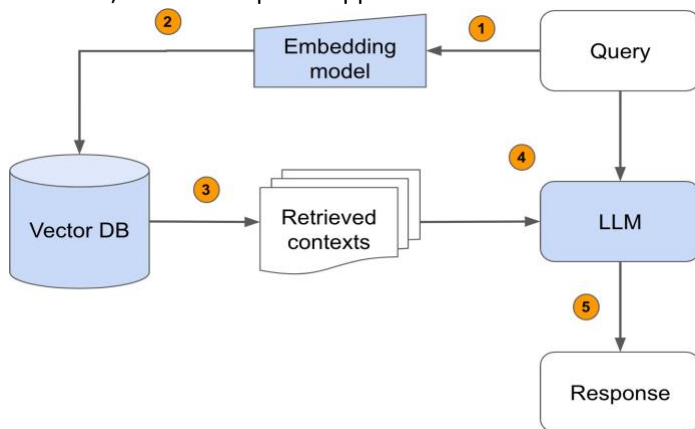
INSIGHT DRIVEN BUSINESS INTELLIGENCE FOR INDUSTRIES

An AI assisted system to respond to common queries

HOW WE DELIVER

eBIW GENAI PROCESS STEP BY STEP

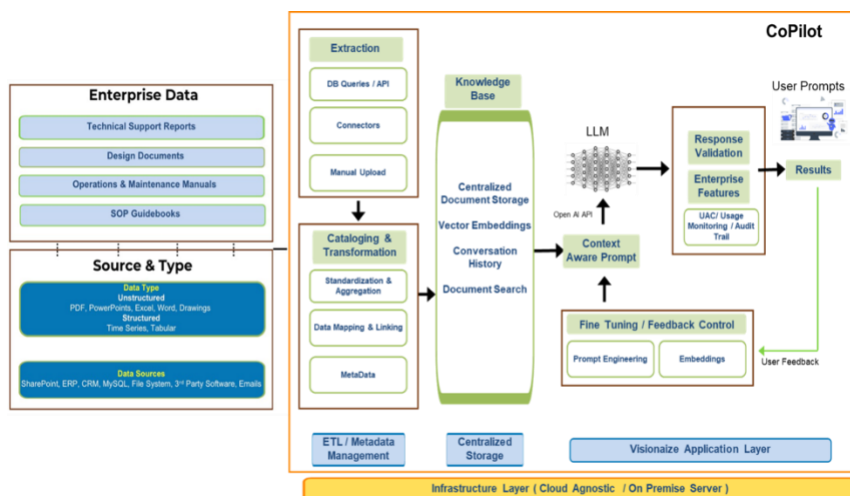
Besides just building LLM application, eBIW GenAI Accelerator focuses on scaling and transitioning models into production. While scale can become a bottleneck for LLM applications due to large datasets, complex models, compute intensive workloads, we develop our application to be able to handle any scale as the world around us continues to grow. The process to facilitate performant inquiries includes:



1. Pass the query to the embedding model to semantically represent it as an embedded query vector.
2. Pass the embedded query vector to our vector DB.
3. Retrieve the top-k relevant contexts – measured by distance between the query embedding and all the embedded chunks in our knowledge base.
4. Pass the query text and retrieved context text to LLM.
5. The LLM will generate a response using the provided content.

IMPLEMENTATION ARCHITECTURE WITH GENAI ACCELERATOR

eBIW GenAI initiative provides the necessary architecture to deliver relevance in understanding and responding to enterprise-specific queries:



1. **Connecting and Curating Data Sources** - allows users to connect and curate various enterprise IT, OT and ET (Emerging Technology) data sources. The robust metadata management layer ensures data consistency by standardizing data formats and providing contextual information necessary for effective operation.

2. **Simple Point & Click Data Loading** - offers an easy to use interface to upload any types of data from sources like - PDF, Excel, Word, Hand-written Notes, Images, Audio files, Video files, Website, Sensor data, XML, JSON, Relational Database load datasets in such a way that we can perform operations at scale (embed, index, etc.).

3. **Data Chunking** - Automatic dividing data into chunks prior to storage, so that each chunk can be inspected for relevance to the input query during a search.

4. **Centralized Storage or Data Lake** - a centralized knowledge base is created to store the ingested documents. This knowledge base serves as a repository for the various types of data, enabling efficient document search functionality.

5. **Vector Embeddings** - Text or Images are transformed into numerical vectors using embeddings, encoding similarity in information as proximity in a multi-dimensional space. By embedding enterprise data and comparing it to the prompt's embedding, the most similar pieces of information are identified and provided to the model as context.

Generative AI Made Simple









INSIGHT DRIVEN BUSINESS INTELLIGENCE FOR INDUSTRIES

An AI assisted system to respond to common queries

6. **Vector Database Creation** - All documents get vectorized and stored in VIZI vector database facilitating semantic search and similarity analysis. Additionally, conversation history is stored, allowing for the tracking and retrieval of previous interactions.
7. **Vector Search** - Vector Search is a way of finding the closest match in a vector database to some input vector, e.g. the vector embedded query passed to our RAG model.
8. **Creation of Context-Aware Prompts:** -To enhance the relevancy and accuracy of responses, context-aware prompts are created. This is achieved by prepending queries with context information derived from the knowledge base.
9. **Response Generation & Validation** - We use the context to generate a response from our LLM. Without this relevant context that we retrieved, the LLM may not have been able to accurately answer the question. As data sets grow, we embed and index new data to be able to retrieve it to answer questions.
10. **Building Ontology with Saved Prompts** - Ability to save prompts and build FAQ and Ontologies over time. This enables non IT users who can simply point and click saved prompts from dropdown lists and get the desired answers to their queries.
11. **Evaluating** - Because a GenAI solution has many moving parts, it needs to perform both unit/component and end-to-end evaluation. eBIW Accelerator provides many techniques to ensure the highest levels of response quality.
12. **User Feedback & Fine-Tuning** - Allows the model to understand enterprise-specific concepts without including them in each prompt, making the inference process more cost-effective.

HOW DOES CUSTOMER BENEFIT

A working LLM model provides:

-  **Develop** retrieval augmented generation (RAG) based LLM application
-  **Scale** (load, chunk, embed, index, serve, etc.) across multiple workers with different compute resources.
-  **Evaluate** different configurations of our application to optimize for both per-component and overall performance
-  **Implement** a hybrid agent routing approach to create the most performant and cost effective application.
-  **Serve** the application in a highly scalable and available manner.
-  **Learn** how methods like fine-tuning, prompt engineering, lexical search, reranking, data flywheel, etc. impact our application's performance

CONTACT US

FOLLOW US

blogs.ebiw.com/ebiw • facebook.com/ebiw • Linked In.com/ebiw • ebiw.com

For more information about **eBIW Generative AI** visit ebiw.com or call to speak to an EBIW person

