

# DB-Agent: Enabling Natural Language-Driven Data Retrieval with Multi-Model Support

Chandan Kumar

chandank@becloudready.com

## Abstract

DB-Agent is an AI-powered system designed to enable natural language interaction with databases. Unlike traditional tools that only generate queries, DB-Agent directly fetches data from databases based on user input, streamlining the data retrieval process for both technical and non-technical users. It leverages multiple inference backends, such as TGI and Ollama, and supports a variety of LLMs, including LLaMA, Gemma, Phi, ChatGPT, and Google Gemini, ensuring adaptability across diverse scenarios. In its long-term vision, DB-Agent aspires to evolve into a comprehensive AI agent capable of interacting with other AI agents, mirroring the service-oriented communication of microservice architectures. This advancement will enable DB-Agent to orchestrate complex workflows, providing actionable insights and holistic, AI-driven solutions for data-driven decision-making.

**Code** — <https://github.com/db-agent/db-agent>

## Introduction

DB-Agent is a system that bridges the gap between natural language and database interaction. By translating user queries into SQL or other database commands and fetching results directly, it simplifies access to data for users with varying levels of technical expertise. The platform leverages state-of-the-art inference backends, such as TGI and Ollama, and integrates multiple LLMs, including LLaMA, Gemma, Phi, ChatGPT, and Google Gemini. This multi-model support ensures that DB-Agent remains flexible and scalable across diverse hardware platforms, including laptops, CPUs, and GPUs.

In its current implementation, DB-Agent focuses on natural language-driven data retrieval. However, the long-term vision is to transform it into a fully-fledged AI agent capable of interacting with other AI agents to deliver results in a manner similar to how services interact in a microservice

architecture. This evolution aims to provide not just data retrieval but actionable workflows and integrated solutions for complex use cases.

## Architecture

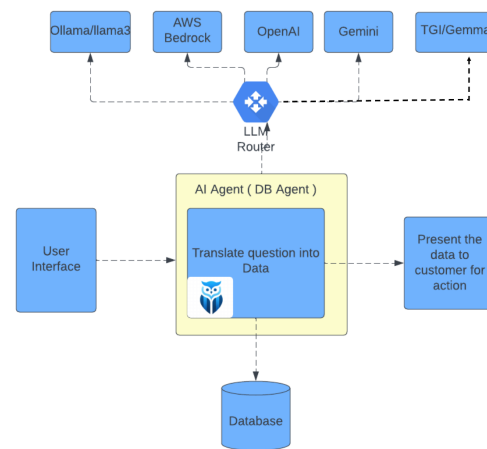


Figure 1: Architecture of the Application

- **AI Agent for Database and Application Interaction:** DB-Agent acts as a bridge between applications and databases, similar to how microservices operate. Unlike traditional tools limited to query generation, DB-Agent delivers actionable data to support workflow automation and decision-making.
- **LLM Router and Load Balancer:** The system employs a load-balancing and routing mechanism for inference backends, ensuring high availability and optimal performance. This design dynamically selects the most suitable LLM for a given task, enabling failover and switchover capabilities.

- **Actionable Insights Over Simple Retrieval:** While many AI tools focus solely on retrieving data, DB-Agent is designed to generate actionable insights. For example, it can recommend or initiate actions based on retrieved data, seamlessly integrating these insights into workflows.

## Demonstration Overview

The live demonstration highlights DB-Agent's capabilities and usability:

- **Deployment:** Demonstrating the deployment of DB-Agent using Docker Compose.
- **Database Interaction:** Connecting DB-Agent to a database and executing natural language queries for direct data retrieval.
- **Multi-Model Integration:** Showcasing the integration of various inference backends (e.g., TGI, Ollama) and their support for multiple LLMs (LLaMA, Phi, Gemma) across diverse hardware platforms, including laptops, CPUs, and GPUs.

## Open-Source Impact

As a completely open-source solution that leverages other open-source tools (libraries, models, and engines), it ensures compatibility with the rapidly evolving AI industry, which is still in its very early stages.

## Acknowledgment

The authors would like to acknowledge the use of the Denvr Dataworks platform for inferencing various models during the development and testing of DB-Agent. This platform's robust capabilities significantly contributed to the project's success.