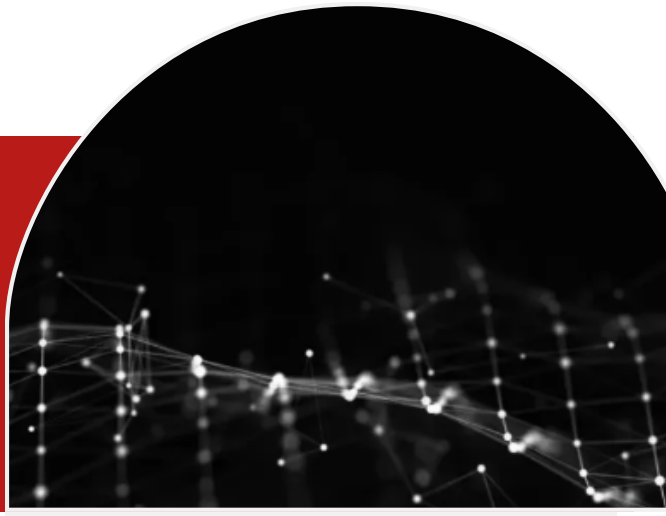
YEARLING AI

# ENTERPRISE LLM BENCHMARKING FRAMEWORK

Put LLMs to the Test—Before They Touch Your Data!



## INTRODUCTION

A leading enterprise technology company partnered with us to develop a sophisticated solution for evaluating Large Language Models (LLMs) for integration into their data ecosystem. The client required a framework that could objectively measure how well different LLMs retrieve, process, and analyze data across multiple departments while maintaining strict governance controls.

## CUSTOMER STORY

Our client needed to determine the best LLM for their enterprise, but faced several challenges in evaluating the many commercial and open-source options available.

Pain Points:

- Data spread across multiple systems made it difficult to assess an LLM's ability to retrieve and integrate information.
- An AI solution that adheres to strict role-based access controls.
- Lacked a standardized way to measure LLM accuracy, response time, and reasoning ability.
- Diverse business units had unique data access and query requirements.
- A way to compare LLMs and open-source alternatives

Goals:

- A framework to rigorously test LLMs in realistic enterprise scenarios.
- Reduce the risk of compliance violations, inaccurate insights, and poor user experiences.
- Ensure the chosen LLM could integrate with data, adhere to governance requirements, and perform well across all business units.

## TECHNOLOGIES

- AI Models: Multiple LLMs, including Claude, OpenAI, Gemini, DeepSeek, Llama 3, Mistral, and others.
- Backend & Data: PostgreSQL, DreamFactory, Python 3.12
- AI & Agent Tech: MCP, Pydantic AI, vLLM
- Evaluation & Monitoring: Langfuse, Pandas/NumPy
- Deployment: Heroku, Docker, Git

## OUR SOLUTION

The comprehensive solution we developed evaluates LLMs on their ability to access enterprise data through APIs, respect role-based access controls, and provide accurate insights across varying levels of complexity. The framework successfully benchmarked both open-source and commercial LLMs, providing detailed performance metrics that enabled the client to make informed decisions about which models to deploy in their production environment.

**How It Works**

The benchmarking framework evaluates LLMs in an enterprise setting. It features a multi-layered architecture with six components and a progressive challenge design with role-based testing. A comprehensive scoring system balances accuracy, response time, and errors.

## CONCLUSION

The Benchmarking Framework met our client's LLM evaluation needs and laid the groundwork for ongoing AI governance and optimization. As LLM technology advances, this framework will facilitate the adoption of new capabilities while upholding standards.

**Future Roadmap:**

- Enhance the framework with vector database integration, multi-modal support, monitoring, and workflow automation.

## ABOUT YEARLING AI

We build AI that works. At YearlingAI, we bring deep technical expertise to solve complex problems with machine learning, natural language processing, and generative AI. From intelligent automation to custom LLM agents, we design, build, and deploy solutions that drive results. As a Google Cloud partner, we specialize in cloud-native development—but also support AWS, Azure, and hybrid environments. Whether you're a growing startup or a global team, we deliver practical AI solutions that scale with your needs.