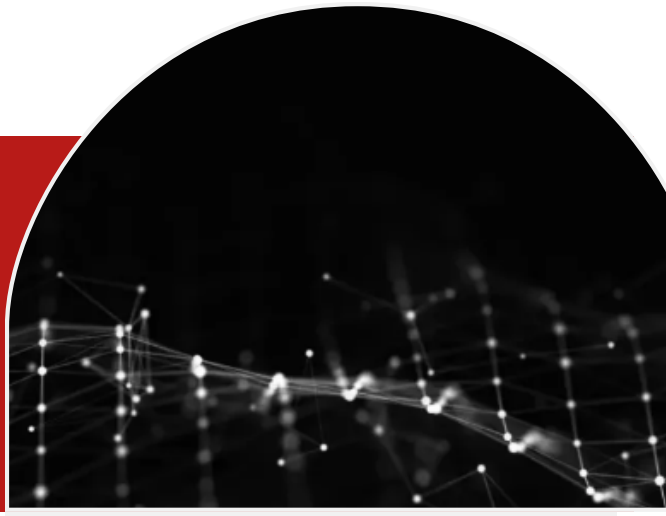
YEARLING AI

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.

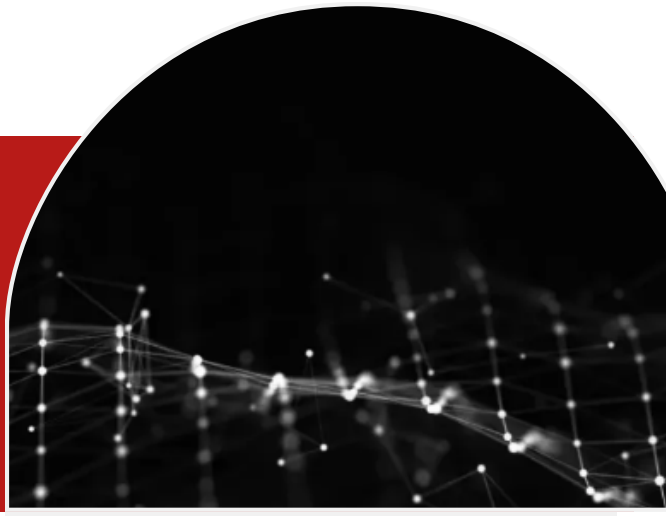
ABOUT YEARLING AI

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.

YEARLING AI

AUTOMATING RFP RESPONSE PROCESSES WITH MULTI-AGENT SYSTEMS

Revolutionizing Civil Engineering Consultancy with Intelligent Automation



INTRODUCTION

In the competitive world of civil engineering consulting, RFP responses are resource-intensive. RHC Engineering, a leader in water and irrigation design, faced slow, manual processes that limited efficiency and increased errors. Partnering with DreamAI, they implemented an AI-driven system that automates personnel matching, risk analysis, and document synthesis while keeping experts involved. The solution reduced response time by 150%, improved proposal quality, and allowed RHC to pursue 150-200% more projects annually.

CUSTOMER STORY

Industry: Civil Engineering Consulting (Water & Irrigation)
Pain Points:

- Time Constraints: Teams spent 8-10 weeks manually extracting RFP requirements, shortlisting resumes, and aligning past projects.
- Resource Bottlenecks: Limited qualified engineers led to frequent mismatches in personnel selection.
- Inconsistent Risk Analysis: Subjective scoring in the Go/No-Go phase caused missed opportunities or overcommitment.
- Document Fragmentation: Copy-pasting from 150+ documents led to errors and version issues.

Goals:

- Automate RFP requirement extraction and resume/project matching.
- Improve risk analysis with data-driven insights.
- Speed up response drafting while ensuring accuracy.
- Scale operations without increasing headcount.

TECHNOLOGIES

- AI Models: GPT-4, Claude 3, Gemini (ensemble for consensus-based outputs).
- Vector Database: LanceDB for fast RAG retrieval.
- Agents: Pydantic-AI workflows with memory retention.
- Frontend: React web app for real-time collaboration.

OUR SOLUTION

YearlingAI built an AI-driven data retrieval system, seamlessly integrated as a microservice on Google Cloud Platform (GCP). Users can access information via voice or text, eliminating manual database queries. AI-powered search, using NLP, Knowledge Graphs, and CypherQL, ensures accurate results. Automated email responses extract answers from past inquiries, reducing workload. With a microservice architecture, the solution integrates effortlessly into existing systems.

How It Works
The system converts voice or text queries into structured searches using speech-to-text and NLP. AI generates Cypher queries via Knowledge Graphs and Neo4j for precise data retrieval. Intent classification and entity recognition refine results, delivering real-time insights via text or speech. A fine-tuned BERT model automates email responses, streamlining repetitive inquiries.

CONCLUSION

By integrating AI agents into their RFP workflows, RHC Engineering transformed from a labor-intensive consultancy to a data-driven industry leader. The system's ability to synthesize complex requirements, align resources, and mitigate risks has positioned RHC as a benchmark for innovation in civil engineering.

Future Roadmap:

- Expand AI agents to automate permit applications and environmental impact assessments.
- Integrate predictive analytics for bid win probability.

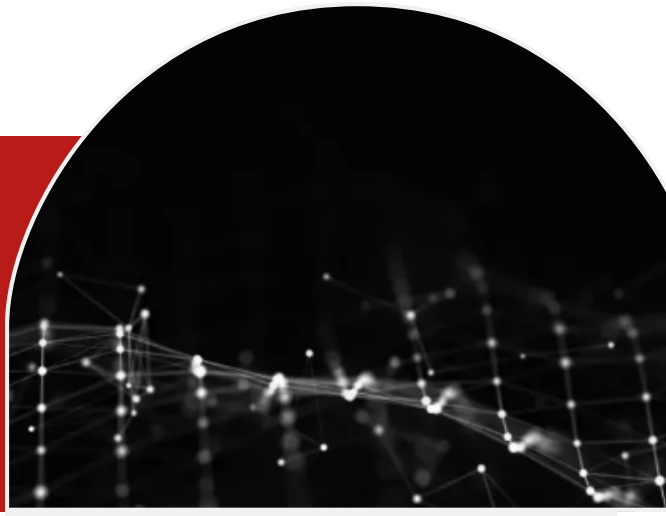
ABOUT YEARLING AI

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.

YEARLING AI

SCALABLE RESUME PROCESSING FOR ENTERPRISES

Smarter Decisions.
Faster Hiring.
Stronger Teams.



INTRODUCTION

Resume processing in bulk is a tedious, time consuming and difficult task for Human Resource people. Medium and Large enterprises receive thousands of resumes every month for a variety of job openings. Furthermore, job search portals like jobs.com, Monster.com and countless others also need to process resumes at scale. In this project, Yearling AI has implemented a scalable resume processing solution for enterprises that automates resume processing and mining data from them in bulk, and thus helps reduce the time and manual labor of extracting information from resumes. The solution developed in this project is a critical component of office process automation for enterprises and human resource companies.

CUSTOMER STORY

This solution was developed based on requirements from two customers via Yearling AI's consulting partner on Google Cloud Platform. Currently, it's being demonstrated to both clients, with a software license agreement in progress. The customers aim to eliminate manual resume processing for hundreds of thousands of resumes. Key requirements include:

- Parsing resumes and storing key terms for easy searchability.
- Extracting sections like Education, Experience, Skills, and Contact Info.
- Identifying named entities (e.g., universities, companies) for quick lookup.

OUR SOLUTION

We developed an end-to-end machine learning pipeline for resume processing with the following steps:

1. **Text Extraction:** Detect and extract text from PDF or DOC resume files.
2. **Embedding Creation:** Generate text embeddings for sentences/paragraphs using NLP-based vectorization.
3. **Clustering:** Use unsupervised learning to group text into sections like Education, Experience, and Contact Info.
4. **Named Entity Recognition:** Identify entities such as organizations and locations within each section.
5. **Data Storage:** Store original text, extracted information, and embedding vectors in a datastore.
6. **Search Indexing:** Build global and intra-resume search indexes to facilitate efficient keyword and phrase searches.

KEY TECHNOLOGY & FRAMEWORK

Technology

- Optical Character Recognition (OCR) for Text Detection
- OCR for Text Recognition
- Natural Language Processing
- Clustering

Framework

- Pytorch-Lightning
- Huggingface Transformers
- Jupyter Notebooks
- Pandas
- Numpy
- FastAPI
- Google Cloud Storage
- Google Kubernetes Engine

CUSTOMER BENEFITS

This solution is a critical component of office process automation for enterprises and human resource companies. It can reduce human engagement in the tedious process of extracting useful information from potentially hundreds of resumes daily. This leads to significantly more efficient office operations and saves valuable time for the Human Resource personnel.

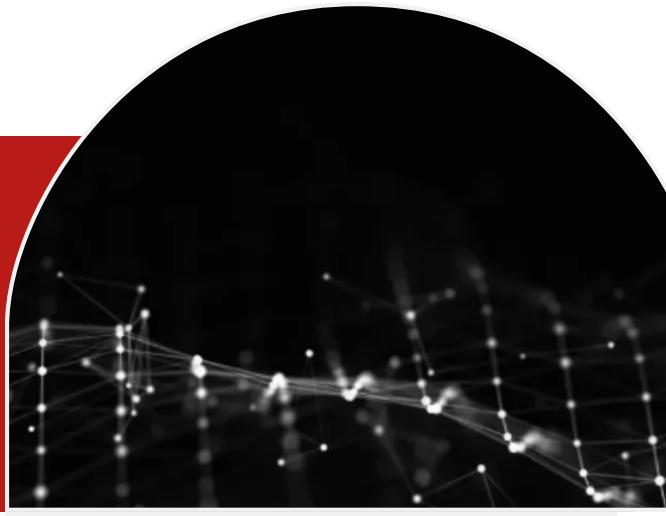
ABOUT YEARLING AI

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.

YEARLING AI

REAL-TIME VIDEO FEEDS ENHANCE AVIATION OPERATIONAL SAFETY

Stay Alert.
Stay Ahead.
Stay Secure.



INTRODUCTION

Ensuring the safety of aircraft and ground operations is essential for maintaining smooth and efficient airport logistics. This project utilizes advanced computer vision technology to analyze real-time video feeds from strategically placed cameras across key airport areas, including docking tunnels, bridges, moving vehicles, stairs, luggage belts, and other critical zones. By detecting potential safety hazards in real time, the AI-powered system generates immediate alerts, enabling ground crews to respond quickly and prevent incidents before they occur.

CUSTOMER STORY

A leading AeroTech solution provider, specializing in Ground Support Equipment and Gate Equipment, wanted to automate equipment movement and logistics during aircraft docking. Traditionally, these operations were managed manually, but the widespread presence of airport cameras presented an opportunity to leverage AI/ML for real-time safety monitoring.

The challenge was to develop an AI/ML solution that could detect safety-critical scenarios in real-time video feeds and generate alerts to prevent damage to equipment, aircraft, and personnel.

OUR SOLUTION

Use Case 1: Aircraft Door Safety
- Scenario: Aircraft docks at the airbridge, and the door opens.
- AI Actions: Detect key objects, measure distances, and alert if the door touches the safety shoe.
- Objective: Prevent door-floor contact.

Use Case 2: Airbridge-Engine Collision Avoidance
- Scenario: Aircraft docking and door opening.
- AI Actions: Track engine and airbridge movement, measure proximity, and issue alerts when too close.
- Objective: Prevent collisions.

Use Case 3: Stairway Safety
- Scenario: Movement of stairs for luggage and personnel access.
- AI Actions: Detect objects, track movement, and measure distances.
- Objective: Ensure safe stair and object movement.

KEY TECHNOLOGY & FRAMEWORK

Technology

- Object Detection: Bounding boxes to identify objects.
- Predicting Segmentation Masks: Assigning each pixel to a class.
- Depth Estimation:
- Monocular (single-camera-based)
- StereoVision (stereoscopic cameras capturing the same scene from different angles)
- Optical Flow Maps: Predicting pixel direction and magnitude across frames.
- Kalman Filters: Velocity estimation of moving objects.

Framework

- PyTorch-Lightning
- OpenCV
- Jupyter Notebooks
- FastAPI
- Optuna
- Kubeflow
- Google Cloud Platform

CUSTOMER BENEFITS

Our AI-driven solution enhances safety, reduces costs, and improves efficiency by:
- Preventing hazards and costly aircraft or equipment damage.
- Automating real-time monitoring for smoother operations.
- Enhancing safety for personnel and logistics.
- Reducing repair costs, delays, and disruptions

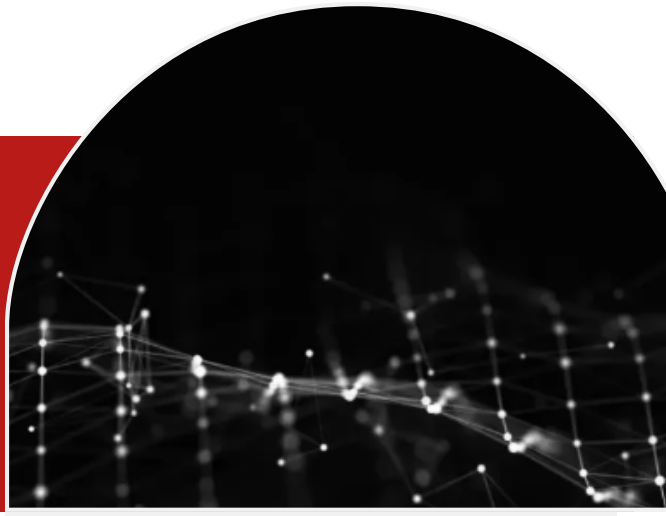
ABOUT YEARLING AI

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.

YEARLING AI

AI-POWERED Q&A BASED ENTERPRISE DATA RETRIEVAL

Elevate Intelligence.
Deliver Impact.



INTRODUCTION

Modern enterprises accumulate vast amounts of structured and unstructured data, including databases, emails, messages, and documents. Retrieving relevant information quickly is crucial for daily operations but remains a complex challenge. Traditional query implementation is time-intensive, requiring custom-built interfaces and backend APIs that connect different applications.
Our AI-powered solution allows users to ask questions in natural language—through text or speech—and instantly retrieve relevant information.

CUSTOMER STORY

A leading grocery supplier struggled with inefficient data retrieval, slowing sales operations. A senior sales rep noted, "I spent hours manually pulling data from multiple systems for simple inquiries—it was frustrating and time-consuming." Key challenges included:

- Slow access to critical business data (orders, suppliers, inventory, pricing).
- Dependence on manual database queries, requiring ongoing development support.
- Sales reps lacked real-time insights for quick decisions.
- Administrators manually processed over 100 daily emails, repeatedly searching past records for responses.

OUR SOLUTION

YearlingAI built an AI-driven data retrieval system, seamlessly integrated as a microservice on Google Cloud Platform (GCP). Users can access information via voice or text, eliminating manual database queries. AI-powered search, using NLP, Knowledge Graphs, and CypherQL, ensures accurate results. Automated email responses extract answers from past inquiries, reducing workload. With a microservice architecture, the solution integrates effortlessly into existing systems.

How It Works
The system converts voice or text queries into structured searches using speech-to-text and NLP. AI generates Cypher queries via Knowledge Graphs and Neo4j for precise data retrieval. Intent classification and entity recognition refine results, delivering real-time insights via text or speech. A fine-tuned BERT model automates email responses, streamlining repetitive inquiries.

KEY TECHNOLOGY & FRAMEWORK

Technology

- Knowledge Graphs & Ontologies
- Speech-to-Text
- CypherQL (Query Language for Graphs)
- Intent Classification & NER Models

Framework

- Neo4j (Graph Database)
- Whisper AI (Speech-to-Text)
- Hugging Face Transformers (Intent Classification, NER, Email Q&A Models)
- Nvidia Nemo (Text-to-Speech AI)

CUSTOMER BENEFITS

- Instant Access to Over 50 Query Types – Users can retrieve diverse business insights in real-time.
- Significant Time Savings – Eliminates the need for manual database queries and custom UI development.
- 80% Reduction in Email Handling Time – AI-generated responses streamline customer communication.
- Enhanced Decision-Making – Sales representatives receive timely insights without technical intervention.

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.