Lead AI Engineer

About TresVista

TresVista is a global enterprise offering a diversified portfolio of services that enables its clients to achieve resource optimization through leveraging an offshore capacity model. TresVista's services include investment diligence, industry research, valuation, fund administration, accounting, and data analytics. TresVista has more than 1,800 employees across offices in North America, Europe, and Asia, providing high-caliber support and operating leverage to over 1,000 clients across geographies and asset classes, including asset managers, advisors, corporates, and entrepreneurs.

Overview:

We are looking for a hands-on, full-cycle AI/ML Engineer who will play a central role in developing a cutting-edge AI agent platform. This platform is designed to automate and optimize complex workflows by leveraging large language models (LLMs), retrieval-augmented generation (RAG), knowledge graphs, and agent orchestration frameworks.

As the AI/ML Engineer, you will be responsible for building intelligent agents from the ground up — including prompt design, retrieval pipelines, fine-tuning models, and deploying them in a secure, scalable cloud environment. You'll also implement caching strategies, handle backend integration, and prototype user interfaces for internal and client testing.

This role requires deep technical skills, autonomy, and a passion for bringing applied AI solutions into real-world use.

Key Responsibilities:

- Design and implement modular AI agents using large language models (LLMs) to automate and optimize a variety of complex workflows
- Deploy and maintain end-to-end agent/AI workflows and services in cloud environments, ensuring reliability, scalability, and low-latency performance for production use
- Build and orchestrate multi-agent systems using frameworks like LangGraph or CrewAI, supporting context-aware, multi-step reasoning and task execution
- Develop and optimize retrieval-augmented generation (RAG) pipelines using vector databases (e.g., Qdrant, Pinecone, FAISS) to power semantic search and intelligent document workflows
- Fine-tune LLMs using frameworks such as Hugging Face Transformers, LoRA/PEFT, DeepSpeed, or Accelerate to create domain-adapted models
- Integrate knowledge graphs (e.g., Neo4j, AWS Neptune) into agent pipelines for context enhancement, reasoning, and relationship modeling
- Implement cache-augmented generation strategies using semantic caching and tools like Redis or vector similarity to reduce latency and improve consistency
- Build scalable backend services using FastAPI or Flask and develop lightweight user interfaces or prototypes with tools like Streamlit, Gradio, or React
- Monitor and evaluate model and agent performance using prompt testing, feedback loops, observability tools, and safe AI practices
- Collaborate with architects, product managers, and other developers to translate problem statements into scalable, reliable, and explainable AI systems
- Stay updated on the latest in cloud platforms (AWS/GCP/Azure), software frameworks, agentic frameworks, and AI/ML technologies

Prerequisites:

- Strong Python development skills, including API development and service integration
- Experience with LLM APIs (OpenAI, Anthropic, Hugging Face), agent frameworks (LangChain, LangGraph, CrewAI), and prompt engineering
- Experience deploying AI-powered applications using Docker, cloud infrastructure (Azure preferred), and managing inference endpoints, vector DBs, and knowledge graph integrations in a live production setting
- Proven experience with RAG pipelines and vector databases (Qdrant, Pinecone, FAISS)
- Hands-on experience fine-tuning LLMs using PyTorch, Hugging Face Transformers, and optionally TensorFlow, with knowledge of LoRA, PEFT, or distributed training tools like DeepSpeed
- Familiarity with knowledge graphs and graph databases such as Neo4j or AWS Neptune, including schema design and Cypher/Gremlin querying
- Basic frontend prototyping skills using Streamlit or Gradio, and ability to work with frontend teams if needed

- Working knowledge of MLOps practices (e.g., MLflow, Weights & Biases), containerization (Docker), Git, and CI/CD workflows
- Cloud deployment experience with Azure, AWS, or GCP environments
- Understanding of caching strategies, embedding-based similarity, and response optimization through semantic caching

Preferred Qualifications:

- Bachelor's degree in Technology (B.Tech) or Master of Computer Applications (MCA) is required; MS in similar field preferred
- 7–10 years of experience in AI/ML, with at least 2 years focused on large language models, applied NLP, or agent-based systems
- Demonstrated ability to build and ship real-world AI-powered applications or platforms, preferably involving agents or LLM-centric workflows
- Strong analytical, problem-solving, and communication skills
- Ability to work independently in a fast-moving, collaborative, and cross-functional environment
- Prior experience in startups, innovation labs, or consulting firms a plus

Compensation:

The compensation structure will be discussed during the interview

---