Generative AI Engineer

Location: Hybrid / Remote | Experience: 2 - 8 Years

About the Role

We are looking for a Generative AI Engineer to design, develop, and deploy AI-powered applications leveraging large language models (LLMs), retrieval-augmented generation (RAG), and agentic AI frameworks. You will build intelligent solutions that automate workflows, enhance decision-making, and create new value for our clients across industries.

Key Responsibilities

- Design and develop Generative AI applications using LLMs (Claude, GPT, open-source models).
- Build and optimize RAG (Retrieval-Augmented Generation) pipelines with vector databases and embeddings.
- Develop agentic AI systems and multi-agent workflows for complex task automation.
- Fine-tune and customize foundation models for domain-specific use cases.
- Integrate AI solutions with enterprise systems, APIs, and cloud infrastructure (AWS, Azure, GCP).
- Implement prompt engineering strategies, guardrails, and evaluation frameworks for LLM outputs.
- Build scalable AI APIs and microservices for production deployment.
- Evaluate model performance, conduct A/B testing, and iterate on AI solutions based on feedback.
- Stay current with rapid advancements in Gen AI, LLMs, and AI safety/alignment research.
- Collaborate with data engineers, product managers, and business teams to deliver end-to-end AI solutions.

Required Qualifications

- Bachelor's or Master's degree in Computer Science, AI/ML, or a related field.
- 2+ years of experience in software engineering, ML engineering, or AI development.
- Strong proficiency in Python and experience with AI/ML frameworks (PyTorch, TensorFlow, Hugging Face).
- Hands-on experience with LLM APIs (Claude API/Anthropic SDK, OpenAI API) and prompt engineering.
- Experience building RAG systems with vector databases (Pinecone, Weaviate, ChromaDB, pgvector).
- Familiarity with LLM orchestration frameworks (LangChain, LlamaIndex, CrewAI).
- Experience with cloud platforms: AWS (Bedrock, SageMaker), Azure (OpenAI Service), GCP (Vertex AI).
- Understanding of ML fundamentals: NLP, transformers, embeddings, fine-tuning.
- Strong software engineering skills including API design, testing, and version control (Git).

Preferred Qualifications

- Experience with Claude API, Claude Agent SDK, and Anthropic's tool-use capabilities.
- Hands-on experience with Databricks for ML/AI workloads and MLflow for experiment tracking.
- Knowledge of AI safety, responsible AI practices, and LLM guardrail frameworks.
- Experience with Snowflake Cortex or similar in-database AI features.
- Familiarity with containerization (Docker, Kubernetes) and CI/CD for ML pipelines.
- Experience with multi-modal AI (vision, audio) and document AI solutions.
- Published research or open-source contributions in AI/ML.

Tools & Technologies

LLM Platforms: Claude API (Anthropic), AWS Bedrock, Azure OpenAI, GCP Vertex AI | Frameworks: LangChain, LlamaIndex, Hugging Face, PyTorch | Vector DBs: Pinecone, Weaviate, ChromaDB, pgvector | Data Platforms: Snowflake (Cortex), Databricks (MLflow, Feature Store) | Cloud: AWS (SageMaker, Bedrock, Lambda), Azure (ML, OpenAI Service), GCP (Vertex AI) | Languages: Python, TypeScript | Other: Docker, Kubernetes, FastAPI, Git, CI/CD tools

What We Offer

- Work at the forefront of Generative AI innovation across industries.
- Access to latest AI tools, models, and GPU infrastructure.
- Collaborative environment with AI researchers and engineers.
- Competitive compensation with performance-based incentives.
- Continuous learning: conference sponsorship, research time, and certification support.