AI Engineer

AI / ML | Startup | SMB | Enterprise

We are hiring an AI Engineer to design and build AI-powered features and applications at [Company Name] . This role focuses on leveraging large language models (LLMs), retrieval-augmented generation (RAG), and modern AI tooling to create intelligent products that deliver real value to users. You will work at the intersection of software engineering and applied AI, building systems that are reliable, performant, and production-ready.

Key Responsibilities

- Design and build AI-powered features using LLMs (GPT-4, Claude, Gemini, open-source models) through prompt engineering, fine-tuning, and RAG architectures
- Develop and maintain RAG pipelines including document ingestion, chunking strategies, embedding generation, vector storage, and retrieval optimization
- Build evaluation frameworks to measure AI feature quality including accuracy, latency, cost, and user satisfaction metrics
- Implement guardrails, content filtering, and safety mechanisms to ensure AI outputs are reliable and appropriate
- Integrate AI capabilities into existing product features through well-designed APIs and SDKs
- Optimize AI system costs by selecting appropriate models, caching strategies, and prompt optimization techniques
- Stay current with rapidly evolving AI landscape and evaluate new models, tools, and techniques for product applicability

Required Skills & Experience

- 3+ years of software engineering experience with at least 1 year building LLM-powered applications
- Strong Python programming skills and experience building production APIs (FastAPI, Flask, or similar)
- Hands-on experience with LLM APIs (OpenAI, Anthropic, Google) and LLM orchestration frameworks (LangChain, LlamaIndex, or custom)
- Understanding of RAG architectures including vector databases (Pinecone, Weaviate, Qdrant, pgvector), embedding models, and retrieval strategies
- Experience with prompt engineering techniques: few-shot learning, chain-of-thought, structured outputs, and system prompting
- Knowledge of AI evaluation methodologies and the ability to build systematic evaluation pipelines
- Solid software engineering fundamentals: version control, testing, CI/CD, and production deployment

Nice-to-Have

- Experience fine-tuning open-source LLMs (Llama, Mistral) using techniques like LoRA or QLoRA
- Familiarity with AI agent frameworks and multi-step reasoning systems
- Experience with multimodal AI (vision, audio, or video models)

- Knowledge of model quantization and optimization for edge or cost-constrained deployment
- Background in traditional ML or NLP is valuable but not required

Tech Stack

Python | FastAPI | LangChain | OpenAI API | Anthropic API | Pinecone | pgvector | Redis | Docker | PostgreSQL | TypeScript | Next.js

What We Offer

- Competitive salary and equity at [Company Name]
- Access to latest AI models and generous API budgets for experimentation
- Fast-moving team at the forefront of applied AI
- Comprehensive health, dental, and vision insurance
- Flexible remote-first work with optional in-person collaboration
- Opportunity to shape the AI strategy and product direction at [Company Name]

Interview Process

1. Recruiter phone screen (30 min) — background, AI experience, and role expectations
2. Technical screen (60 min) — Python coding, LLM concepts, and a prompt engineering exercise
3. System design interview (60 min) — design a RAG-based feature or AI agent for a realistic product scenario
4. Take-home or live coding (2-3 hours) — build a working AI feature using LLM APIs, demonstrate evaluation approach
5. Team and hiring manager conversation (60 min) — collaboration style, AI philosophy, and career goals
6. Reference checks and offer