As an AI Application Engineer at Rearc, you'll contribute to the technical excellence of our machine learning and artificial intelligence capabilities. Your expertise in Machine Learning, AI application development, and LLMOps will help design and deploy intelligent systems that are efficient, reliable, and production-ready. You'll work closely with cross-functional teams to build and operationalize GenAI solutions, pushing the boundaries of what AI is capable of in customer environments. Building strong partnerships with technical teams & stakeholders and aligning with business objectives will be essential as you support AI-driven initiatives and contribute to their successful execution.

What You Bring

2+ years of experience in AI engineering, machine learning (ML), or related fields, bringing valuable expertise in building and deploying intelligent systems.

Generative AI (GenAI expertise)

- Strong understanding of state of the art techniques in generative AI, including large language models (LLMs), text generation and other foundation models
- Familiarity with AI orchestration tools (e.g. LangGraph, CrewAI, Bedrock Agents, smolagents, etc)
- Experience in fine-tuning, prompt engineering or otherwise adapting

generative models for specific use cases
- Experience with AI model evaluation, including human-in-the-loop and LLM judge paradigms
- Familiarity with NLP libraries and frameworks

Knowledge Base & Tool Integration

- Hands-on experience in implementing Retrieval Augmented Generation (RAG) architectures and integrating retrieval systems with generative models
- Knowledge of at least one vector store or database (e.g. Opensearch, Pinecone, PostgreSQL with pgvector) and techniques for similarity search
- Familiarity with common data ingestion/ETL patterns for populating knowledge bases
- Experience with implementing LLM tool calling (either directly, via an orchestration framework, or using Model Context Protocol (MCP) clients)

Amazon Bedrock or Databricks AI experience

- Experience using Amazon Bedrock or Databricks Mosaic AI for deploying and managing generative AI models
- Ability to integrate Generative AI scripts with other services (such as AWS S3 or Lambda or Databricks Apps) to build scalable and secure AI solutions

Software & Tools Proficiency

- **Strong programming skills in Python (or similar languages)**
- **Familiarity with CI/CD pipelines and MLOps practices to ensure seamless integration, testing and deployment of AI models**

Nice-to-Haves

- **Experience with containerization and orchestration for deploying AI services**
- **Understanding of cloud security best practices and cloud architecture design in the context of AI workloads**
- **Experience writing MCP servers**
- **Experience training or fine-tuning ML models (classical ML or LLMs/diffusion models)**

What You'll Do

- **Collaborate with Colleagues – Work closely with colleagues to understand customers' business objectives and technical challenges, contributing to the design and development of effective GenAI solutions tailored to client needs.**
- **Apply GenAI Principles – Utilize modern tools and frameworks like LangGraph, to build scalable, reliable, and maintainable Compound AI systems. Leverage your understanding of AI fundamentals to ensure every project meets rigorous industry and ethical standards.**
- **Adapt to the latest Technologies & Patterns – continue to research, learn,**

and stay abreast of the most recent state of the art for AI application development.

- **Promote Knowledge Sharing –Bolster our culture of continuous learning by sharing knowledge about AI engineering best practices through blog posts, articles, and internal talks. Support a collaborative environment that fosters shared expertise and ongoing innovation across our community.**

About Rearc

At Rearc, we're committed to empowering engineers to build awesome products and experiences. Success as a business hinges on our people's ability to think freely, challenge the status quo, and speak up about alternative problem-solving approaches. If you're an engineer driven by the desire to solve problems and make a difference, you're in the right place!

Our approach is simple — empower engineers with the best tools possible to make an impact within their industry.

We're on the lookout for engineers who thrive on ownership and freedom, possessing not just technical prowess, but also exceptional leadership skills. Our ideal candidates are hands-on-keyboard leaders who don't just talk the talk but also walk the walk, designing and building solutions that push the boundaries of cloud computing.

Founded in 2016, we pride ourselves on fostering an environment where creativity flourishes, bureaucracy is non-existent, and individuals are encouraged to challenge the status quo. We're not just a company; we're a community of problem-solvers dedicated to improving the lives of fellow software engineers.

Our commitment is simple - finding the right fit for our team and cultivating a desire to make things better. If you're a cloud professional intrigued by our problem space and eager to make a difference, you've come to the right place. Join us, and let's solve problems together!

Your first few weeks at Rearc will be spent in an immersive learning environment where our team will help you get up to speed. Within the first few months, you'll have the opportunity to experiment with a lot of different tools as you find your place on the team.

Rearc is committed to a diverse and inclusive workplace. Rearc is an equal opportunity employer and does not discriminate on the basis of race, national origin, gender, gender identity, sexual orientation, protected veteran status, disability, age, or other legally protected status.