

Position: Forward Deployed AI & Innovation Engineer
Department: AI & Innovation
Period: Permanent
Reports to: Lead Technical Product Manager, AI & Innovation
Location: Cape Town, Bellville (Hybrid)

Purpose

Pepkor NexTech is a dynamic IT company that delivers value-driven IT solutions to the Pepkor group. We are focused on creating value - simplified. Our work lives revolve around our DNA. We take pride in our work, treat our colleagues, partners, and our country with respect. We exercise collaborative decision making and believe in work-life balance. We achieve this all while having fun and enjoying what we do.

About the Role

We are seeking a highly adaptable, commercially minded **Forward Deployed AI Engineer (FDE)** to architect and execute AI-driven solutions that solve complex retail challenges and unlock net-new business capabilities. In this role, you will not build static models in a vacuum. You will operate as a high-velocity technical operative—embedding with our diverse retail operating companies (PEP, Ackermans etc), unpacking ambiguous operational realities, and driving MVPs through to production-ready states.

You are part software engineer, part solutions architect, and part visionary builder. You thrive in the unstructured reality of enterprise data, love writing production-grade code, and are obsessed with executing AI safely and cost-effectively at massive retail scale.

Responsibilities

- **Active Problem Solving & Innovation:** Sit with retail stakeholders to discover operational friction (supply chain, POS, merchandising etc) and identify "Blue Sky" opportunities for new AI innovations.
- **Agentic Engineering & Harness Building:** Write the actual orchestration logic that turns static LLMs into autonomous agents. Build the deterministic software "harnesses"—tool calling, API routing, and state management—that allow AI to safely read, reason, and execute actions across Pepkor's legacy retail systems.
- **MVP to Production:** Build functional, high-quality MVPs in days, not months. Own the end-to-end integration layer, seamlessly connecting Google Cloud AI primitives (Gemini, Vertex AI) with broad, complex legacy systems.
- **Scalable Innovation:** Identify recurring technical hurdles across different OpCos. Extract and codify reusable architectural patterns, standard API gateways, and AI integration templates to build our department's internal library, accelerating future PoCs.
- **LLMOps & FinOps:** Implement AI evaluations ("evals") to prevent model drift and hallucinations. Enforce strict token-optimization and caching layers to ensure AI economics remain viable for low-margin, high-volume retail.

Qualifications and Experience:

- Bachelor's degree in Computer Science, Software Engineering, Mathematics, or equivalent practical and demonstrable experience.
- **8+ years** of experience in backend or full-stack software engineering, with a proven track record of shipping and maintaining production systems.
- **1.5+ years** of hands-on experience specifically designing and deploying Generative AI, RAG pipelines, or Agentic workflows into production environments.
- **5+ years** of experience building scalable, cloud-native applications (Google Cloud Platform preferred).
- Deep proficiency in Python, TypeScript, and cloud-native software engineering (CI/CD, microservices, containerization).
- Proven experience with the Google Cloud ecosystem (BigQuery, Cloud Run, Vertex AI).
- Hands-on experience building with LLM APIs, vector databases, and semantic caching. Deep understanding of agentic orchestration (e.g., function/tool calling, ReAct loops, LangGraph, or building native autonomous execution loops from scratch).
- Demonstrable integration portfolio showing:
 - Building RAG pipelines over unstructured or messy enterprise data.
 - Writing secure API wrappers (using MCP) or middleware to connect modern AI logic to older on-premise or legacy retail systems.
 - Deploying agentic/multi-agent workflows that take autonomous actions based on system triggers.
 - Taking an unstructured prototype and stabilizing it with proper logging, fallback logic, and error handling for a live production environment.
- Domain Advantage (Bonus): Previous experience working in retail, supply chain, logistics, or dealing with legacy enterprise systems (e.g., legacy ERPs, WMS, POS).

Competencies and Skills:

- Ability to create order out of vague product ideas, define technical paths, and execute independently.
- Ability to translate probabilistic AI outcomes and ROI metrics clearly to non-technical retail executives (CFOs, Brand Heads).
- Building systems designed to be easily maintained by other IT departments, focusing on deterministic fallbacks, self-healing code, and clean documentation.
- High agency technical execution and commercial mindset.
- Strong problem-solving, analytical thinking, and decision-making skills.

How to apply

If you are interested in the above position and meet the requirements indicated above, please submit your CV to vacancies@pepkorit.com REF: Forward Deployed AI Engineer + Name

Alternatively, you may apply directly on our current careers website: <https://pepkorit.com/careers>. Please note the new website link, <https://pepkor-nextech.com/job-listings/forward-deployed-ai-innovation-engineer> is currently in progress for direct application submission.

Pepkor NexTech strives for equal opportunity in terms of its employment equity guidelines. By submitting your application, you are giving Pepkor NexTech implicit consent to the storage and processing of your personal information. If you do not hear from us within 3-4 weeks of the closing date of this position, please regard your application as being unsuccessful.

Closing date for applications: **23 October 2026**