Derisk360

JOB DESCRIPTION

Forward Deployed Engineer (FDE)

Location: UK · Hybrid & client embed Type: Full-time Team: The Factory

Derisk360 is an enterprise AI services firm. We embed Forward Deployed Engineers inside regulated enterprises, primarily banking, insurance, and financial services, to implement governed agentic systems from discovery through production go-live. This role is for builders who ship outcomes in the client's environment, not advisors who leave after the pilot.

LLM & SLM architecture | RAG & knowledge graphs | Agentic AI & MCP | Eval & guardrails | VPC production | Financial services

The role

As an FDE at Derisk360, you embed with client teams (on-site or hybrid) to run structured accelerator programmes across the agentic stack: context and data engineering, MCP integration, knowledge graphs, multi-agent workflows, cloud and model engineering, evaluation, and managed operations. You are accountable for governed production go-live in the client's VPC, not proof-of-concept demos. You work closely with Forward Deployed Eval Engineers (FDEEs) on eval harnesses, red teaming, policy controls, and post-go-live quality.

What you will do

Delivery & embed

- Embed with business, data, platform, and risk teams to define use cases, success metrics, model-risk boundaries, and audit requirements.
- Lead technical workstreams within 4-layer delivery: plug-in (discover), ingest (context), build (agents), run (deploy & operate).
- Produce architecture decision records, runbooks, and handover material so client teams can operate systems after rollout.

Context, data & integration

- Design unified context layers: source inventory, field mapping, lineage, and governed access for agent workloads.

- Implement MCP servers and connectors to core banking, CRM, document stores, policy engines, and operational systems.
- Build and qualify knowledge graphs and structured retrieval surfaces that ground agent outputs in auditable enterprise data.

AI systems & agentic engineering

- Architect RAG pipelines (chunking, embedding models, re-ranking, hybrid search, citation grounding) tuned for accuracy and latency in production.
- Design LLM and SLM deployment patterns: model routing, right-sized models per task, cost/latency trade-offs, and fallback behaviour.
- Implement multi-agent orchestration (planner/worker patterns, shared state, tool calling, human-in-the-loop gates for high-risk actions).
- Integrate frontier and open models in a model-agnostic way (e.g. Claude, GPT, open-weight SLMs) within client security and procurement constraints.
- Harden prompts, context windows, and tool schemas for reliability; reduce hallucination and non-deterministic failure modes before go-live.

Evaluation, safety & production ops

- Build eval datasets and harnesses aligned to business KPIs and compliance scenarios; partner with FDEEs on regression and adversarial testing.
- Implement guardrails: policy engines, PII handling, output validation, escalation paths, and audit logging for every material agent action.
- Deploy via IaC/containers in client VPC; configure observability (tracing, token/cost metrics, quality dashboards) and incident runbooks for 24/7 operations.

Technical skills we expect

You combine strong software engineering with hands-on generative AI production experience. We do not expect every candidate to excel in every area below, but you should be deep in several and credible across the stack.

LLM & SLM architecture

- Transformer-based LLMs vs SLMs for edge/task-specific workloads
- Model selection, routing, and ensemble patterns
- Inference optimisation (batching, caching, quantisation awareness)
- Private / VPC deployment (API vs self-hosted trade-offs)
- Token economics, latency SLOs, and capacity planning

RAG & retrieval

- Embedding models and vector / hybrid retrieval
- Chunking strategies, metadata filters, re-ranking
- Graph-augmented RAG and knowledge graph query patterns
- Citation grounding and explainability for regulated use cases
- Evaluating retrieval quality (precision/recall, faithfulness)

Agentic systems

- Tool use, function calling, and MCP protocol integration
- Multi-agent workflows and orchestration frameworks
- State management, idempotency, and failure recovery
- Policy engines and human-in-the-loop design
- Agent memory vs stateless context engineering

Platform & data engineering

- Python and/or TypeScript production codebases
- AWS, Azure, or GCP (networking, IAM, secrets, private endpoints)
- APIs, event streams, ETL/ELT, and document pipelines
- SQL and operational data stores; data quality and lineage
- Docker, Kubernetes, or managed container patterns

Evaluation & MLOps / LLMOps

- Offline and online eval harnesses; golden datasets
- Red teaming, jailbreak resistance, and domain adversarial tests
- Observability: OpenTelemetry-style tracing, LLM logging
- Versioning prompts, models, and retrieval indexes
- Model-risk documentation for internal audit / compliance

Regulated delivery

- Banking, insurance, or financial services context (preferred)
- Security reviews, SOC-style controls, and audit trails
- Working with risk, legal, and model validation stakeholders
- Clear written communication under delivery pressure

Experience profile

- 5+ years in software, data, or ML engineering, with at least 2 years shipping LLM- or agent-based systems beyond notebooks or demos.
- Demonstrable production delivery: you can walk us through architecture, eval approach, and operational lessons learned from a live deployment.
- Comfort presenting trade-offs to CTO, CDO, and business sponsors in client workshops.
- Ownership mindset: outcome-scoped accelerators, not open-ended staff augmentation.

Nice to have

- Fine-tuning, LoRA/PEFT, or distillation experience (when RAG alone is insufficient).
- Familiarity with Anthropic, OpenAI, or open-weight model ecosystems and enterprise procurement patterns.
- ServiceNow, core banking, or policy-admin integration experience.
- Consulting or client-embed background with McKinsey, Big Four, or similar delivery discipline.

Why Derisk360

- Factory training on the full agentic stack: context engineering, MCP, knowledge graphs, agentic transformation, cloud & model engineering, evaluation & guardrails, and managed AI operations.
- Work on tier-1 financial services programmes where accuracy, audit, and uptime matter as much as model capability.
- Outcome-based culture: production agents with FDEE-led evaluation, not slideware or endless pilots.

How to apply

Email your CV and a short note describing one production LLM or agent system you architected (RAG, model choice, eval, and ops) to support@derisk360.com with subject line *FDE application*. Learn more at derisk360.com/factory.