Job Description: Senior AI Engineer – Computer Vision, Edge AI & Agentic Systems

Location: Gurugram

Experience: 3+ Years

Department: AI / Engineering

Company: Virtualyyst

Role Overview

Virtualyyst is looking for a Senior AI Engineer who can build production-grade AI systems for real-world environments.

This is not a research-only or model-training-only role. We need a hands-on AI engineer who can design, build, deploy, optimize, and maintain AI systems across computer vision, edge AI, industrial automation, IoT devices, video intelligence, and agentic LLM workflows.

You will work on AI systems that run in factories, warehouses, logistics environments, industrial sites, safety workflows, and enterprise platforms. These systems must perform reliably under real-world constraints such as poor lighting, camera movement, latency, limited edge compute, noisy data, occlusion, changing environments, and data drift.

The ideal candidate is a strong builder who understands that AI in production is not just about accuracy, but also latency, reliability, observability, scalability, explainability, retraining, and failure handling.

Key Responsibilities

1. Computer Vision & Video Intelligence

Design and develop advanced computer vision systems for industrial and enterprise use cases.

You will work on:

- PPE detection
- Crowd detection
- Restricted-area entry detection

- Object detection and tracking
- Worker safety monitoring
- Anomaly detection
- Pose estimation
- Temporal action recognition
- Small-object detection
- Multi-camera video analytics

Build AI models and pipelines that can operate in challenging environments with issues such as low light, dust, motion blur, occlusion, poor camera angles, and class imbalance.

---

2. Scalable Inference Architecture

Design scalable inference pipelines for high-throughput video and sensor workloads.

Responsibilities include:

- Building inference pipelines for multiple camera streams
- Optimizing models for low-latency response
- Designing edge vs cloud inference strategies
- Handling real-time and near-real-time video processing
- Supporting multi-camera deployments
- Building scalable model-serving architectures
- Ensuring sub-second or near real-time inference where required

You should be comfortable designing systems that can scale from pilot deployments to large enterprise rollouts.

---

3. Edge AI & Model Optimization

Own the deployment and optimization of AI models on edge and cloud environments.

You will work on:

- TensorRT optimization
- OpenVINO optimization
- Model quantization
- ONNX conversion

- GPU inference
- CPU inference
- Edge device constraints
- Docker-based AI microservices
- NVIDIA DeepStream
- Triton Inference Server
- Model benchmarking and profiling

The role requires strong understanding of how to make AI models production-ready, efficient, and reliable.

---

4. Agentic AI & LLM Workflows

Build agentic AI systems that assist in enterprise decision-making, operations, monitoring, and automation.

You will work on:

- LLM-based agents
- Multi-agent workflows
- Tool-calling systems
- Retrieval-Augmented Generation (RAG)
- Vector databases
- Long-term memory for agents
- Workflow automation
- AI assistants for dashboards and operations
- Natural-language querying over enterprise data
- Human-in-the-loop approval flows

The candidate should be able to build LLM systems that are reliable, auditable, and suitable for business use cases.

---

5. Industrial AI & Sensor Fusion

Work on AI systems that combine video intelligence, IoT telemetry, and time-series sensor data.

You will help build intelligence layers for:

- Industrial safety
- Manufacturing monitoring
- Predictive maintenance
- Device health monitoring
- Digital twins
- Operational anomaly detection
- Site-level intelligence dashboards

You should be able to connect AI models with real-world data sources such as cameras, IoT devices, sensors, APIs, logs, and enterprise systems.

---

6. MLOps & Production AI Lifecycle

Own the complete AI lifecycle from experimentation to production.

Responsibilities include:

- Dataset preparation and annotation workflows
- Model training and validation
- Model evaluation and benchmarking
- Deployment pipelines
- Model monitoring
- Drift detection
- Automated retraining workflows
- Versioning of models and datasets
- A/B testing of models
- Observability for AI systems
- Failure analysis and continuous improvement

You should ensure AI models become stable production systems, not just demos.

Required Technical Skills

AI / ML Frameworks

Strong hands-on experience with:

- PyTorch or TensorFlow
- OpenCV
- YOLO models such as YOLOv8, YOLOv10, YOLO-NAS, or similar
- Object detection
- Image classification
- Segmentation
- Tracking algorithms
- Pose estimation
- Video analytics pipelines

LLM & Agentic AI Stack

Hands-on experience with one or more of:

- LangChain
- LangGraph
- LlamaIndex
- RAG pipelines
- Vector databases
- Tool calling
- Multi-agent orchestration
- Prompt engineering
- Function calling
- Open-source LLMs such as Llama, Mistral, Gemma, or similar

Good understanding of:

- RAG optimization
- Parent document retrieval
- Hybrid search
- Embeddings
- Re-ranking
- Fine-tuning
- Guardrails
- Hallucination mitigation

AI Infrastructure & Deployment

Experience with:

- NVIDIA DeepStream
- Triton Inference Server
- TensorRT
- OpenVINO
- ONNX
- Docker
- REST APIs
- Microservices
- GPU-based deployment
- Cloud or on-prem deployment
- Edge AI deployment

Real-World AI Problem Solving

The candidate should have experience solving practical AI challenges such as:

- Occlusion
- Low-light video
- Motion blur
- Small-object detection
- False positives and false negatives
- Class imbalance
- Poor-quality datasets
- Data drift
- Latency bottlenecks
- Model degradation after deployment
- Hardware compute limitations

Good to Have

- Experience in industrial automation, manufacturing, logistics, safety-tech, surveillance, or IoT
- Experience working with embedded or edge devices
- Experience with camera pipelines, RTSP, WebRTC, GStreamer, or video streaming systems
- Experience with time-series sensor data
- Experience with predictive maintenance or digital twins
- Experience building AI dashboards or AI-powered enterprise workflows
- Understanding of cloud platforms such as AWS, GCP, or Azure

---

Ideal Candidate Profile

You should be someone who:

- Loves building real AI products, not just notebooks
- Can take a model from experiment to production
- Understands both AI accuracy and system performance
- Can debug model failures in real-world environments
- Can work across AI, backend, edge, device, and product teams
- Can design systems that are scalable, observable, and maintainable
- Can handle ambiguity and convert it into working solutions
- Has strong ownership and problem-solving ability