Rajesh Kumar Gupta

+1 (732) 421-2092 | rajesh21gupta@gmail.com | linkedin.com/in/rajeshgupt

Professional Summary

Principal Machine Learning Engineer with 16+ years building production-scale AI agents and intelligent automation systems for enterprise applications. Expert in architecting autonomous agents using LLMs, RAG, and orchestration frameworks (CrewAI, LangChain, LlamaIndex) that integrate deeply into HR and Financial workflows. Proven track record leading ML teams through complete product lifecycle from applied research to production deployment serving millions of users. Deep expertise in PyTorch, TensorFlow, cloud platforms (AWS/GCP), and building scalable ML infrastructure. Passionate about driving innovation in AI agent engineering, establishing continuous learning systems, and fostering collaborative team cultures focused on delivering transformative customer value.

Core Technical Competencies

AI Agent Engineering: Autonomous Agents, CrewAI, Multi-Agent Orchestration, Agent Workflow Design, Tool Integration
LLMs & Generative AI: GPT-4, Claude, LLaMA, Gemini, Fine-tuning, Prompt Engineering, RAG Architectures, RLHF
ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, Hugging Face Transformers
AI Orchestration: LangChain, LlamaIndex, Haystack, Agent Frameworks, Vector Databases, Knowledge Graphs
Production ML: Model Hosting at Scale, MLOps, A/B Testing, Monitoring, Drift Detection, MLflow, Airflow, Docker, Kubernetes
Cloud Platforms: AWS (SageMaker, Bedrock, EC2, S3, Lambda), GCP (Vertex AI), 10+ years experience
Deep Learning & NLP: BERT, Text Generation, Named Entity Recognition, Information Retrieval, Graph Neural Networks
Programming: Python (Expert - 16+ years), Java, SQL, REST APIs, Microservices Architecture

Professional Experience

AVP, Principal Machine Learning Engineer – AI Agent Systems *May 2023 – Present*
Barclays Investment Bank, New York, NY

- Led cross-functional team of 8 ML engineers architecting autonomous AI agent system for $2.8T Financial trading ecosystem, taking product from applied research through design, implementation, production deployment, and metric-based evaluation
- Architected multi-agent orchestration framework using CrewAI with specialized agents (researcher, analyzer, compliance checker, report writer) collaborating to automate financial workflows serving 500+ users with 85% efficiency improvement
- Implemented production-scale LLM integration using GPT-4, Claude, and LLaMA with RAG architecture processing 10M+ daily transactions, achieving 95% accuracy and sub-100ms response times
- Designed agent tooling strategy integrating enterprise systems (Kafka, databases, APIs) enabling agents to perform complex multi-step financial analysis and decision support
- Established monitoring, feedback loops, and continuous learning mechanisms with automated A/B testing and drift detection improving agent performance by 40% over 12 months
- Built scalable ML infrastructure on AWS using SageMaker, Bedrock, and ECS hosting 15+ production models with 99.9% uptime through automated scaling and rollback capabilities
- Mentored 5 junior ML engineers on AI agent best practices, production ML systems, and lifecycle management fostering culture of innovation and continuous improvement
- Owned sprint planning and development lifecycle for AI features collaborating with product, engineering, and data science teams
- Applied deep learning frameworks (PyTorch, TensorFlow) for custom agent behavior models and fine-tuned LLMs for financial domain

- Collaborated with cloud providers (AWS, GCP) and open-source communities to integrate cutting-edge AI technologies

Senior Machine Learning Engineer – AI Platforms & Knowledge Systems *Dec 2022 – May 2023*
S&P Global – Aloka Platform, New York, NY

- Designed and deployed intelligent agent system for manufacturing analytics using LangChain and graph neural networks processing 1M+ component relationships with 90% accuracy
- Built production ML infrastructure on AWS with automated CI/CD pipelines, model versioning, and monitoring using MLflow and CloudWatch
- Implemented RAG system combining vector embeddings with knowledge graphs for context-aware retrieval and semantic search
- Applied PyTorch and TensorFlow for deep learning models enabling intelligent chatbot and automated recommendation systems
- Led technical discussions with software engineers and data scientists establishing AI integration patterns and best practices
- Stayed current with AI advancements implementing latest techniques in LLMs, autonomous agents, and orchestration frameworks

Machine Learning Engineer – Healthcare AI & Intelligent Automation *Dec 2019 – Dec 2022*
Johnson & Johnson – Enterprise Knowledge Platform, USA

- Architected AI-powered automation system for pharmaceutical research workflows using multi-agent architecture and LLM integration reducing manual processing time by 85%
- Built production ML platform using PyTorch, TensorFlow, and Hugging Face Transformers for medical text analysis extracting entities from 50,000+ clinical trial PDFs with 90% F1 score
- Designed autonomous agent workflows for clinical trial matching, patient screening, and drug safety signal detection achieving 90% matching accuracy
- Implemented end-to-end MLOps on AWS SageMaker with automated model training, validation, deployment, and monitoring serving global research teams
- Applied graph neural networks and knowledge graphs for drug discovery and safety assessment with complex relationship analysis
- Led technical design reviews mentoring 3 junior ML engineers on production ML best practices and AI system integration
- Collaborated with product and engineering teams to implement AI-based solutions enhancing healthcare operations
- Established feedback loops and continuous learning mechanisms improving model performance through data-driven enhancements
- Shipped production Python code and models with robust testing, monitoring, and version control using Git and CI/CD

Machine Learning Engineer – Predictive Analytics & ML Infrastructure *Jun 2018 – Nov 2019*
Scania AB, Sweden

- Built production ML system for predictive maintenance using PyTorch, TensorFlow, and ensemble methods (XGBoost, Random Forest) achieving 90% failure prediction accuracy on 50,000+ vehicles
- Designed scalable ML infrastructure with Docker/Kubernetes deployment on cloud platforms enabling real-time inference
- Implemented streaming ML pipeline with Kafka for continuous data ingestion and model updates
- Applied statistical analysis and supervised learning algorithms for vehicle behavior clustering and pattern recognition
- Collaborated with engineering teams integrating ML models into production IoT platform serving fleet management operations

Machine Learning Engineer – NLP & Automation Systems *May 2017 – May 2018*
AT&T, India

- Built end-to-end ML platform for IT operations automation using PyTorch, TensorFlow, and graph neural

networks
- Implemented NLP pipeline with BERT models for support ticket classification improving accuracy by 85%
- Developed production model hosting infrastructure on AWS with automated monitoring and alerting
- Applied deep learning for text analysis processing 100K+ tickets daily with sub-second inference latency

Machine Learning Engineer – Content Intelligence & Recommendation *May 2016 – May 2017*
AT&T, India

- Developed ML-based content categorization using deep learning achieving 90% classification accuracy
- Built recommendation engine using collaborative filtering and neural networks for personalized content delivery
- Shipped production Python code serving millions of users with scalable ML infrastructure
- Applied statistical analysis and unsupervised learning for content understanding and user behavior modeling

Project Lead – ML-Based Semantic Solutions *Mar 2014 – May 2016*
AT&T, India

- Led team of 4 ML engineers building automated metadata search platform using NLP and machine learning
- Implemented semantic indexing and ML-based keyword generation for enterprise search optimization
- Deployed production ML infrastructure with monitoring and continuous improvement mechanisms
- Fostered collaborative team culture taking ownership of development lifecycle and sprint planning

Software Engineer – ML & NLP Systems *Aug 2009 – Mar 2014*
Avaya & Alexis, India

- Developed ML-based semantic chatbot with natural language understanding achieving 80% QA accuracy
- Built automated ontology generation using machine learning and graph algorithms from text corpus
- Implemented NLP pipelines for entity extraction and semantic analysis using Python and ML frameworks
- Created production systems with CI/CD deployment and version control

Research ML Engineer – Medical AI Systems *Jun 2008 – Jul 2009*
Traditional Knowledge Digital Library (TKDL), Government of India

- Built semantic search system using ML algorithms and knowledge graphs for medical diagnosis recommendations
- Developed novel ontology merging algorithms using graph theory and machine learning
- Published IEEE paper on ML-based ontology integration methodology – awarded Best Paper

Leadership & Team Management

Team Leadership: 10+ years leading and mentoring ML engineering teams (5-8 engineers) across multiple organizations
Sprint Planning: Extensive experience owning development lifecycle, backlog prioritization, and agile methodologies
Cross-Functional Collaboration: Strong track record building relationships with product managers, software engineers, data scientists, and stakeholders
Culture Building: Fostered cultures of collaboration, transparency, innovation, and continuous improvement across teams
Technical Leadership: Led technical design reviews, architecture decisions, and established ML engineering best practices

Education & Certifications

Master of Technology in Software Engineering – National Institute of Technology, Allahabad, India, 2009
Bachelor of Engineering in Computer Science – Rajiv Gandhi Technological University, Bhopal, India, 2007
Certifications: AWS Data Analytics Specialty, CKAD (Certified Kubernetes Application Developer)

Publications & Thought Leadership

IEEE Publication: "An Instance Based Methodology for Merging Domain Ontology" – Best Paper Award (2009)
https://ieeexplore.ieee.org/document/5395380

Research Paper: "Road Safety Ontology - SafeOn: Overview & Design" – Transportation Research Procedia (2024)
https://www.sciencedirect.com/science/article/pii/S2352146524004630

Conference Speaker: "How Ontology, Knowledge Graph and GenAI solve Financial Industry Issues" – NextGen Banking (2024)
https://nexgenbanking.com/USEdition/

Technical Proficiencies

AI Agents & Orchestration: CrewAI, LangChain, LlamaIndex, Haystack, Autonomous Agents, Multi-Agent Systems
ML Frameworks: PyTorch (10+ years), TensorFlow (10+ years), Keras, Scikit-learn, XGBoost, Hugging Face
LLMs & GenAI: GPT-4, Claude, LLaMA, Gemini, Fine-tuning, RAG, RLHF, Prompt Engineering
Deep Learning: Neural Networks, Transformers, BERT, Graph Neural Networks, Text Generation, NER
MLOps & Production: Model Deployment at Scale, A/B Testing, Monitoring, MLflow, Airflow, Docker, Kubernetes
Cloud Platforms: AWS (10+ years - SageMaker, Bedrock, EC2, S3, Lambda), GCP (Vertex AI, Cloud Functions)
Programming: Python (Expert - 16+ years), Java, SQL, Shell Scripting, REST APIs, Microservices
Data Engineering: Apache Kafka, Spark, ETL Pipelines, Real-time Streaming, Vector Databases
Knowledge Systems: Neo4j, Knowledge Graphs, GraphRAG, Semantic Search, Ontology Engineering