

SUVOJIT HORE

Senior AI Engineer

Email: suvojit.hore.1@gmail.com | Phone: +91 98041 25414

LinkedIn: [linkedin.com/in/suvojith](https://www.linkedin.com/in/suvojith) | Portfolio: suvojith.github.io

Location: Bangalore, India | Notice Period: 15 days



PROFESSIONAL SUMMARY

Suvojit is a Senior **AI Engineer with 6.6 years** of experience across Data Science, Machine Learning, and Generative AI, building and deploying production AI systems that create measurable business impact for enterprise clients.

His work spans across **large language models and agentic AI** orchestration, retrieval-augmented generation (RAG) and vector search, predictive modelling and statistical analysis, computer vision and natural language processing, using **Python** and full stack development across **Azure, AWS, Runpod, GCP and Databricks**. He has delivered **21 Agentic AI and Data Science applications deployed in production in real time**. He has also provided pre-sales and post-sales technical support directly to business stakeholders across oil and gas, insurance, banking, retail, healthcare, legal, and technology sectors, owning solutions end-to-end.

He holds a **Master's Degree in Artificial Intelligence** and Machine Learning from BITS Pilani, and has presented his work at Devcon, NIAT, Machine Learning Developers Summit (MLDS) and Nvidia Conferences. He is certified across Microsoft Azure, AWS, Databricks and NVIDIA in AI, DS and machine learning.

He currently works on developing and **productionizing agentic AI systems in Python for millions of users** in Australia, driving end-to-end automated insurance claims processing with production-grade monitoring, evaluation, and governance.

EDUCATION

Master's Degree - Artificial Intelligence and Machine Learning | **Birla Institute of Technology and Science** | Pilani, India (Online) | *Nov 2024*

Specialization: Natural Language Processing | CGPA: 8.8

Bachelor's Degree - Mechanical Engineering | **SRM Institute of Science and Technology** | Chennai, India | *May 2019*

Aggregate: 84%

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript, TypeScript, C++

Frameworks: LangChain, LangGraph, LlamaIndex, Autogen, Crew AI, Pydantic, PyTorch, TensorFlow, Transformers, React

AI/ML Concepts: Generative AI, LLMs, RAG, Agentic AI, Autonomous Agents, Prompt Engineering, Embeddings, Semantic Search, Predictive Modelling, Statistical Analysis, Time-Series Forecasting, Regression, Supervised & Unsupervised Learning, NLP, Computer Vision, Deep Learning, Multimodal AI

Tools & Platforms: Claude Code, Copilot, Azure Databricks, AWS, GCP, PySpark, MLflow, Neo4j, PostgreSQL, Faiss, ChromaDB, n8n, Microsoft Fabric, MCP, Docker, Kubernetes, Streamlit, Gradio, FastAPI, Jupyter, Git, VS Code, Cursor.

Practices: MLOps, AIOps, CI/CD, Evaluation Frameworks, Observability, Model Governance

WORK EXPERIENCE

Senior AI Engineer | **EXL** | Bangalore, India | *Aug 2024 - Present*

- Developed and deployed a real-time multi-agent claims-processing system in production for 1.5 million users in Australia using **Python, LangGraph, and PySpark on Azure Databricks with RAG** and vector databases (**Neo4j, PostgreSQL, and FAISS** vector databases), and MLflow lifecycle tracking, autonomously analyzing incidents and driving claims decisions with **Claude Opus 4.6**. Reduced time to claim finalization by 98% (2 months to 1 day) and cost of operation by 56%.
- AIOps and large-scale inference optimization** for **production LLM workloads** on Databricks Model Serving, including autoscaling, request batching, evaluation frameworks, and hallucination-rate monitoring.
- GenAI chatbot** with a **React frontend for 150,000+ Australian users**, backed by Neo4j graph databases and document intelligence with semantic search, enabling users to query insurance documents and file claims through n8n-automated workflows - OpenAI GPT-5.4, **finetuned LLM - Mistral**.
- Agent Toolkit, a reusable agent-building framework built with **LangGraph, LlamaIndex, and MCP**, adopted across all **agentic AI** initiatives, accelerating prototype-to-production time by 57%.
- PII deidentification tool leveraging LLMs with iterative feedback loops, **LLM as a judge**, and **MLflow tracing**, replacing manual masking cycles and accelerating prototype production by 70%.

- AI Twin application with a React-based interface replicating an individual's appearance, voice, and speech using Agentic AI, RAG, and lip-sync; used by EXL CEO Rohit Kapoor in his showcase.

Senior Data Scientist | EY (Ernst & Young) | Noida, India | May 2024 - Aug 2024

- Multimodal LLM-based compliance solution verifying image, video, audio, and text in advertisements for computer hardware clients. **Finetuned Llava and Florence models, and trained Yolo** model on logo recognition.
- Fine-tuned open-source LLM chatbot to answer queries about India's new legal framework - Bhartiya Nyaya Sanhita.

Senior Applied Data Scientist | Dunnhumby | Gurgaon, India | Dec 2021 - May 2024

- **Recommendation engine** and revenue management system for the UK's largest retailer using **Python, PySpark, SQL, and Scikit-learn on Azure Databricks with MLflow** experiment tracking, driving a 315% sales increase through predictive modelling, dynamic pricing, and demand forecasting.
- **Fine-tuned open-source LLM chatbot** for business users with distributed training across multi-GPU Azure Databricks clusters, converting natural language to SQL and Python code; won Hackrush hackathon for this project.
- Customer family segmentation models using Python, PySpark, **SQL, Scikit-learn**, and XGBoost to drive personalized revenue management and dynamic pricing. Increased sales by 57% due to better targeted ads based on segment.
- Media campaign evaluation and revenue attribution using Python, PySpark, **random forest, and K-means** clustering for retail and ecommerce clients across UK and Norway. This helps in planning future advertisements in retail stores.
- **Time-series demand forecasting** models using Python, PySpark, **PyTorch, XGBoost, and FastText**, improving product placement and category revenue growth by 23%.
- RAG-powered **dashboard and coding assistant** bot using LLMs and vector databases to assist data scientists with coding and debugging.

Data Scientist | Infosys R&D | Bangalore, India | Dec 2019 - Dec 2021

- Pix2Pix C-GAN using **TensorFlow and Keras** to generate synthetic images of meat, deployed in production for a large US food processing company. This is used to create synthetic datasets using a particular set of manually created textures as input to the C-GAN. The C_GAN was trained and finetuned on a set of Input texture to actual Meat Images.
- RASA chatbot (TensorFlow-based NLU) for a British telecommunications company. This is a rule based bot (2020) with intents, entities context used by internal stakeholders of the company to query internal policy information.
- **Cluster analysis and topic modelling** dashboard for an American bank. This is fully automated for the users to cluster, model and auto-analyze the content across unstructured chat, messages and email communications (anonymized).
- Real-time classification of quality of meat with **Convolution Neural Network (CNN)** on a conveyor belt for an American food processing company. This is trained on a combination of actual and synthetic images created with my previous app, achieving substantial precision, recall and F1 scores. Reduced operation cost by 43% and manual verification time by 97.5% (2 minutes to ~3 seconds).
- Automated **form and receipt OCR using OpenCV and Tesseract**, used across multiple projects by other teams.

RESEARCH PUBLICATIONS & PERSONAL PROJECTS

Beyond the Mirror: AI Twins Using Large Language Models | MLDS 2025

- Presented at MLDS 2025. Won Early Trailblazer Award at EXL for the same project.

Breaking the Language Barrier: Natural Language to SQL Using LLMs | MLDS 2024

- Presented at MLDS 2024. Won Hackrush hackathon at Dunnhumby for the same project.
-

CERTIFICATIONS & ACHIEVEMENTS

Microsoft Azure Data Scientist & AI Engineer | Microsoft

AWS Machine Learning Specialty & AI Practitioner | Amazon Web Services

NVIDIA Generative AI LLMs | NVIDIA

Databricks LLM for Production | Databricks

Early Trailblazer & Emerging Analyst Award | EXL | 2024 - 2025

- Delivered the AI Twin project for EXL CEO within 3 months of joining (2024); spearheaded client innovation with Agentic AI within 5 months of joining (2025).

Hack Rush Winner & GMS Star Award | Dunnhumby, Infosys & Indian Statistical Institute | 2014, 2021 - 2023

- Hackrush hackathon winner (Dunnhumby, 2023); GMS Star Award (2022); Award for Excellence (Infosys, 2021).