Guruprasanna Rajukannan Suresh

+1 (548) 333-0491 | guruprasannasuresh@gmail.com | LinkedIn: GuruprasannaRS | GitHub: GuruprasannaRS

SUMMARY

AI/ML Engineer with solid experience delivering end-to-end solutions, from data pipelines to production-ready models. Specialized in machine learning, deep learning, and LLMs with strong expertise in MLOps for scalable AI systems, plus additional skills in agentic workflows, software development, and analytics.

TECHNICAL SKILLS

Programming: Python, SQL, C++
ML & Data: PyTorch, TensorFlow, Scikit-Learn, OpenCV, NumPy, Pandas, PySpark, Matplotlib, Seaborn
LLM & NLP: Hugging Face Transformers, LangChain, LlamaIndex, Crew AI, OpenAI API, SpaCy, NLTK, Ollama, llama.cpp, vLLM, Unsloth
Infra & Deployment: Docker, Kubernetes, FastAPI, MLflow, Redis, AWS, Azure AI, CI/CD, Celery

EXPERIENCE

UN-Habitat - Quality of Life Initiative
Jul 2024 – Present
AI Software Engineer — *Waterloo, ON, Canada*

- Architected ReAct agent toolkit executing dynamic SQL and semantic retrieval with **sub-30s response times**
- Built LangGraph multi-agent system with **Llama-3.3-70B** routing between case-study, law, and web-search nodes
- Created WebSocket streaming interface showing real-time LLM agent reasoning for 20+ UN admins
- Built ingestion pipeline with PGVector and MapReduce processing 3000-character chunks with keyword extraction
- Implemented Redis-Celery pipeline on AWS ECS reducing document processing from **1h → 15min**
- Built LLM-driven pipeline extracting entities/relationships from case studies into AWS Neptune for knowledge graph

HolisticMindAI
Apr 2024 – Jul 2025
AI Developer (Volunteer) — *Kitchener, ON, Canada*

- Implemented speaker diarization with **WhisperX** large-v3 and **PyAnnote**, separating speakers using 1.5s threshold
- Led backend with FastAPI endpoints and AWS Bedrock **Mistral-8x7B**, enabling scalable therapy platform delivery
- Engineered LlamaParse and spaCy pipeline, generating 384-dim embeddings for semantic document search
- Built RAG pipeline with LlamaIndex hybrid search, recall improved from **68% → 92%**

X-Care
Sep 2023 – Dec 2024
Machine Learning Engineer (Vounteer) — *Waterloo, ON, Canada*

- Built two-stage fracture detection system: **EfficientNet** transfer learning (94.3% accuracy) + **YOLOv8** localization
- Fine-tuned YOLOv8 for fracture localization achieving **IoU >0.85** on X-rays, 0.75–0.78 on MRIs
- Optimized PyTorch training pipeline using **Optuna** hyperparameter tuning, reducing training time by **60%**
- Implemented MLflow/DagsHub versioning pipeline managing 10+ model runs with automated deployment tracking
- Preprocessed 15K+ medical images with augmentation and balanced sampling for robust model training

Region of Waterloo (Smart Waterloo Region Innovation Lab)
Sep 2023 – Aug 2024
MLOps Engineer (Co-op & Contractor) — *Kitchener, ON, Canada*

- Fine-tuned **Mistral 7B** with LoRA and deployed GGUF model to Hugging Face for lightweight inference
- Built LLM-powered Q&A system using RAG for school policy management, deployed across 5+ school boards
- Implemented function calling and agentic workflows achieving **BLEU score 0.85** for Q&A accuracy
- Engineered automated CI/CD pipeline with dynamic ports, auto-versioning, and containerized deployments

Multiversal Automation Private Limited
Jan 2022 – Oct 2022
Automation Engineer — *Chennai, India*

- Built real-time data pipeline across 10 stations, enabling multi-sensor traceability and improving efficiency
- Applied **XGBoost** on torque/load data for predictive maintenance, extending average tool life by **25%**
- Developed a time-series anomaly detection model on press/flow/torque curves, reducing false rejects by **15%**

Tube Investments Optoelectronic Solutions
Feb 2020 – Dec 2021
Project Engineer & Project Trainee — *Sricity, India*

- Designed ETL pipelines to ingest and manage machine data (50+ sources) via TCP/IP into relational database
- Prototyped a full-stack web app for real-time visualization, delivering 24/7 insights and **cut reporting by 40%**
- Built Tableau dashboards turning 200K+ data points into executive KPIs, **boosting decision-making by 25%**

PROJECTS

Legal QA Assistant with LLM Fine-Tuning

- Preprocessed legal documents with PySpark and a **T5-base QG model**, creating 15k QA-ready chunks.
- Fine-tuned **Llama-2-7B** using QLoRA with Unsloth, enabling training on 8GB VRAM.
- Achieved **72.4% EM** and **74.1 F1**, demonstrating improved accuracy compared to the base model.
- Deployed as a quantized GGUF model, providing optimized, low-latency local inference through **Ollama**.

MCP Finance Assistant: Excel Analytics & Web Search

- Developed two **MCP servers** on FastAPI for Excel-to-SQL analytics and web search with **Tavily API**.
- Automated Excel-to-SQL ingestion with natural language **text-to-SQL** for finance insights.
- Integrated **real-time web search** to link SQL results with market trends.
- Deployed via **Claude MCP client**, enabling seamless financial Q&A in a unified workflow.

Azure-Powered Wikipedia Chatbot

- Built a conversational AI with **Azure Prompt Flow** and OpenAI GPT-4o for real-time Wikipedia search.
- Created tools in Python, Jinja2, BeautifulSoup for query extraction, scraping, and processing.
- Deployed a cloud-native pipeline on **Azure Container Instances** with memory and error handling.

Translating NLI: Cross-lingual Classification Analysis

- Translated 25k English–French FEVER pairs using a **Seq2Seq model** for cross-lingual NLI tasks.
- Implemented robust **NLTK preprocessing** and fine-tuned BERT, RoBERTa, ELECTRA, XLNet.
- Achieved **BLEU 0.60** and maintained semantic similarity of **0.77** across translations.
- Evaluated with **68.5% F1**, and **70% recall**, highlighting RoBERTa's robustness on translated data.

EDUCATION

Master of Engineering, Electrical and Computer Engineering Co-op **Jan 2023 — Aug 2024**
University of Waterloo, ON, Canada ***GPA: 3.9 / 4***

Key Courses: Computational Intelligence, Data Modeling & Deep Learning, NLP, Image Processing, Statistical Modelling

CERTIFICATIONS

- PCAP – Certified Associate in Python Programming *Python Institute*
- AI on Microsoft Azure *Microsoft*
- Computer Vision in Microsoft Azure *Microsoft*
- Crash Course on Python *Coursera*
- Microsoft Azure ML *Microsoft*
- Python for Machine Learning & Data Science Masterclass *Udemy*
- Azure AI Studio (AI Foundry) : Prompt Flow, LLMOps & RAG *Udemy*