# Rohan W

Bengaluru, India
+91 9840571093 rohan.w.charles@gmail.com linkedin.com github.com

## Experience

**Senior Machine Learning Engineer** **March 2025 – Present**
*Avalara* *Banglore, IN (remote)*

- Engineered and deployed an internal, multi-cloud LLM server that provided a secure, standardized OpenAI-compatible API across Azure, AWS Bedrock, and Google Vertex AI, ensuring model flexibility and strict data governance for internal teams.
- Developed and integrated full-featured API endpoints into the server that were fully compliant with the OpenAI specification for all critical internal use cases, including Chat Completion, Responses, Model Evaluation, and Fine-Tuning.
- Improved ML operational reliability and scalability by leading the infrastructure migration of long-running evaluation and prompt finetuning jobs from Celery to Temporal, enhancing workflow management and fault tolerance.

**ML Research Engineer - II (L5)** **April 2024 – March 2025**
*Amazon* *Banglore, IN*

- Engineered and deployed a massive-scale, high-accuracy attribute identification system. This involved fine-tuning a Small Language Model (Llama), integrating a multi-agent validation step leveraging the "More Agents Is All You Need" architecture for quality assurance, and deploying the system using a VLLM-based distributed setup (with page attention and dynamic batching) to accurately identify attribute values for millions of products while achieving high throughput and minimizing cost.
- Ideated and Developed a platform using LLMs (Claude 3.5 Haiku) to accelerate the speed at which Data Associates could add missing attributes to the Amazon catalog, significantly boosting their productivity for data backfilling.
- Developed and presented a Platform Proof-of-Concept for Automated Prompt Finetuning using DSPy to leadership. This system was projected to reduce manual prompt engineering effort and boost ML Engineer productivity by 140%, helping the team reach the Q4 goal faster and increase the goal for the following year.
- Enhanced the Amazon retail search experience by utilizing advanced prompting techniques, including Chain-of-Thought (CoT) and Meta-Prompting, across internal services.

**Senior Machine Learning Engineer** **Jul 2021 – June 2024**
*Rapid Acceleration Partners* *Chennai, IN*

- Led, mentored and guided a team of 4 members in the Machine Learning Team
- Was instrumental in pushing for and implementing changes to the training pipeline for our product (RAPFlow) to increase efficiency and usability
- Upgraded the in-house OCR solution resulting in a 48% reduction in inference time without compromising on inference quality
- Developed and implemented a chatbot powered by GPT-3.5 and LangChain to automate customer interactions and lead them to setup an appointment with the car dealership. The chatbot was able to answer customer questions, provide product information, and schedule appointments. As a result, the chatbot increased customer satisfaction, reduced drop off, and improved the sales funnel

**Machine Learning Engineer** **Feb 2020 – Jul 2021**
*Rapid Acceleration Partners* *Chennai, IN*

- Worked on developing NLP and Multi-Modal components for rapflow ( Low code AI Solutions Platform )
- Lead a team that built and deployed AI driven automation tools to optimize sales leads for US car dealerships as well as create new sales leads to drive revenue growth. In addition to that we developed a tool to optimize warehouse workflow
- Developed a document classification model to classify documents in the airplane parts space and a document understanding models to extract key value pairs from the same

**Data Science Associate** **June 2019 – Nov 2019**
*Zoomrx* *Chennai, IN*

- Developed data processing tool so as to ensure that data from different sources integrates seamlessly with the different modules that consume that data
- Developed a tool for validating and cleaning millions of rows of clinical trails data which is stored across multiple databases across multiple formats in real-time

**Data Science Intern** **Dec 2018 – May 2019**
*Zoomrx* *Chennai, IN*

- Worked on an automated annotation tool that parses medical data and extracts keywords for Ferma.ai
- Worked on an internal tool to QC parsed data and update them real-time

## Education

**SRM University** **July. 2015 – July 2019**
*B.Tech E.C.E* *Chennai, TN*

## OSS Contributions

- Active contributor to widely adopted open-source ML and systems libraries, with merged pull requests across core Python runtime, LLM infrastructure tooling, and research-driven optimization frameworks; representative selection shown below, with full contribution history available on GitHub.

- CPython • LiteLLM • TextGrad

## Certifications

**Deep Learning Specialization** **Nov 2019**
*Coursera*

**Game Theory** **Sept 2020**
*Stanford University*

## Awards

**Employee of the Quarter** **August 2023**
*Rapid Acceleration Partners*

**RISE** **Q4 2024**
*Amazon*

## Technical Skills

- **Languages**: Python, Typescript, C
- **Technologies/Frameworks**: Linux, Git, PyTorch, TensorFlow, Flask, FastAPI, Scikit-learn,
- **Deployment**: Docker, Docker-compose, Podman, K8, Helm
- **Data Manipulation**: Pandas, Numpy, SQL
- **Data Visualization**: Matplotlib, Seaborn, Plotly, D3.js
- **Cloud Computing**: AWS, GCP