Manju Malateshappa

LinkedIn — Github — Portfolio — Publications
Vancouver, Canada

Email : mmanju2@gmail.com
Mobile : +1-604-721-8984

Summary

- **AI/ML Architect and Engineering Leader** with **12+ years** of experience delivering scalable AI systems and leading end-to-end solution development, specializing in **Generative AI**, **LLMOps**, and cloud-native ML platforms on **AWS**. Currently serving as **Machine Learning Architect** at Caylent, driving architecture, client engagement, and delivery of enterprise AI solutions.
- Proven track record of translating business requirements into high-impact AI solutions, enabling automation, improving decision-making, and delivering measurable business outcomes across **6+ enterprise client engagements**. Experienced in leading cross-functional teams, mentoring engineers, and driving technical strategy, while remaining hands-on in building **production-grade systems at scale (15M+ records)**.

Skills

- **Languages**: Python, TypeScript, JavaScript.
- **Generative AI & NLP**: Amazon Bedrock, Amazon SageMaker AI, Amazon AgentCore, LLMs (ChatGPT, Claude, and others), multimodal AI (text-to-speech, text-to-video, image-to-video), RAG, AI Agents, Prompt Engineering, AWS Comprehend.
- **MLOps / LLMOps**: Model lifecycle management, evaluation frameworks (text-to-SQL, RAG quality, agent reasoning), monitoring, CI/CD for ML systems.
- **AI & ML Frameworks**: PyTorch, Scikit-learn, MLFlow, LangChain.
- **Cloud & Infrastructure**: AWS (Lambda, Step Functions, EKS, S3, SQS, IAM), Terraform, Docker, Kubernetes.
- **Big Data & Orchestration**: Apache Spark, Kafka, HDFS, Apache Airflow.
- **Web Technologies**: ReactJS, Node.js, Django.
- **Databases**: SQL (MySQL, PostgreSQL), NoSQL (Cassandra, MongoDB).
- **Version Control & CI/CD**: GitHub, GitHub Actions, GoCD, Poetry.
- **Certifications**: Anthropic Claude Certified Architect — Foundations; HashiCorp Certified: Terraform Associate (004).

Experience

Caylent *Vancouver, Canada*

Machine Learning Architect *May 2026 – Present*

- Promoted to Machine Learning Architect; leading end-to-end architecture and delivery of scalable **GenAI** and **ML** solutions on **AWS** — from early discovery through production rollout — with a focus on **cost–performance trade-offs**, stakeholder alignment, and building robust, maintainable systems.

Senior Machine Learning Engineer *Dec 2023 – May 2026*

- Led end-to-end architecture, delivery, and client engagement across **6 enterprise engagements** (Fintech, Automotive, MarTech, MedTech, EdTech), translating business requirements into scalable, production-ready AI architectures.
- Designed and delivered **Generative AI** systems on **Amazon Bedrock**, including **RAG pipelines**, **vector-based retrieval architectures** (embeddings + vector stores), and LLM-powered automation workflows.
- Architected agent-based AI systems using **ReAct**, **planner–executor patterns**, and **multi-agent orchestration**, enabling scalable and reliable decision-making workflows.
- Established **LLM evaluation frameworks** (text-to-SQL, RAG quality, agent reasoning) and implemented **LLMOps** best practices including prompt versioning, model evaluation pipelines, observability, and production monitoring.
- Led architecture reviews and trade-off analysis (cost vs. latency, build vs. buy, model selection); authored **solution design documents** and **reference architectures** for Fintech and Automotive client engagements.
- **Co-authored Caylent's MLOps Solution Offering** — a reusable architecture blueprint standardising MLOps delivery patterns, adopted as a company-wide reference for client engagements.
- Mentored **2 engineers** on GenAI architecture and MLOps, enabling independent ownership of client deliverables.

Selected Client Outcomes:

- **Fintech** – Built a GenAI document processing pipeline that reduced financial statement processing time from **30–60 minutes to ∼90 seconds** (97% faster), achieved **>95% extraction accuracy**, scaled throughput **10x** (50→500 documents), and cut per-document cost to **$0.36–0.47** (under $0.50 target), enabling same-day application reviews.
- **MarTech / CRM** – Fine-tuned and productionized **DistilBERT** (email classification: **99.3% F1**, **0.064s** inference, **$0.23**/10k records) and **Mistral-7B** (name extraction: **97%+** accuracy) across **beta, prod-us, and prod-eu** — among the first production deployments of fine-tuned open-source LLMs (Mistral-7B, released Sept 2023) for enterprise NLP, serving a platform used by **25,000+ organizations**.
- **Automotive** – Sole **Technical Leader** for a multimodal AI avatar evaluation; benchmarked **5 models** (OmniAvatar, HunyuanVideo, Wan2.2, Wan2.2-S2V-14B, Amazon Nova Reel) across quality (1–5 scale), generation time (2–35 min per 18s video), and cost ($1.38–$2.98/video) on AWS GPU infrastructure (p5.4xlarge, H100 80GB);

identified top performer and delivered a reusable evaluation blueprint for future model upgrades — customer verdict: *"A great success. Work of the highest quality."*

Alida Inc *Vancouver, Canada*
Senior AI Engineer *Sep 2022 – Jun 2023*

- Led development and productization of AI services, managing a **team of 4 engineers** and driving technical delivery aligned with product and business goals.
- Designed and implemented **AI pipelines processing 15M+ records**, directly securing a **multi-year enterprise contract** and establishing the platform as a core product offering.
- Built reusable text analytics frameworks (**PII masking**, **sentiment**, **taxonomy**) using **Node.js** and **Python**; scaled infrastructure using **Terraform** and **AWS EKS**.
- Managed **MLOps** workflows for **13+ microservices**, standardising CI/CD, testing, and deployment processes to increase release velocity and improve production stability.
- Contributed to growing the AI engineering team from **2 to 6 members** by conducting technical interviews and supporting recruitment; mentored engineers in NLP, MLOps, and cloud-native AI development, enabling independent ownership of production pipelines.

AI/ML Engineer *Apr 2020 – Sep 2022*

- Designed, deployed, and maintained **5 core NLP microservices** (text analysis, tagging, and translation) using **spaCy**, **NLTK**, **IBM NLU**, and **AWS Translate**; built supporting **CI/CD pipelines** with **Terraform** and **GitHub Actions**, improving deployment consistency and release reliability.
- Built **ML models (XGBoost)** and designed end-to-end ML pipelines including data collection, ETL, and training, deployed to production via **AWS EKS**.

Nagra Kudelski *Bangalore, India*
Software Engineer *Jan 2014 – Aug 2019*

Grew from individual contributor to **leading UI development** across multiple concurrent projects over 5 years, delivering embedded TV applications for telecom platforms across **6+ projects** serving **2M+ users** globally.

- Coordinated **4 engineers** across **3 concurrent projects** (Chivas TV, Euskaltel TV, TBC Taiwan), managing requirements, sprint planning, and production release sign-off.
- Key contributor at **FOXTEL Australia** — led core feature development including **Cloud DVR**, **content protection**, **parental lock**, and **Google Analytics-based recommendations**.
- Supported live **4K content** deployment during the **2016 Rio Olympics** for 2M+ users, resolving production issues on-ground in Brazil — securing a **2-year contract extension** with Nagra Kudelski Group.
- Mentored **6 engineers** in UI architecture, code quality, and software engineering best practices.

Key Achievements

- **Winner** – Alida Hackathon 2023: Automated PR generation and code reviews using Jira, Slack, Git diff, and LLMs via GitHub Actions.
- **Winner** – Alida Hackathon 2022: XGBoost speedster detection solution deployed on AWS SageMaker, integrated into the Alida platform.
- **SFU Course Project – Model Fairness & Transparency** (2020): Detected and mitigated racial bias in ML classification models (COMPAS dataset) using **IBM AIF360**, **Google What-If**, and **SHAP**; improved Disparate Impact from **0.7 to 0.91** using XGBoost with Reweighing pre-processing — published on Medium (SFU CS publication).

Education

- **Simon Fraser University** Burnaby, Canada
  Master of Science in Computer Science (Big Data and Machine Learning) *Aug 2019 – May 2021*
- **University Visvesvaraya College of Engineering** Bangalore, India
  Bachelor of Engineering (B.E.) in Computer Science *Aug 2009 – Jun 2013*