

SAI KUMAR M

Ph: [+1\(980\)689-0147](tel:+19806890147) | Email: msk05@microsoft.com | [LinkedIn](#) | [GitHub](#)

SUMMARY

AI Engineer with 4+ years building real-time financial AI and **multi-agent** systems. Delivered fraud detection at scale across 30M+ transactions per day, agentic platforms using **LangChain**, **LangGraph**, and **MCP**, and low-latency model serving on AWS. Strong in **PyTorch** and TensorFlow, PySpark and end-to-end MLOps with MLflow, Docker, Kubernetes, Terraform, and CI/CD, with governance-ready explainability and drift monitoring.

PROFESSIONAL EXPERIENCE

- AI Engineer, McKinsey & Co.** Sep 2023 - Present
- Designed and developed multi-tiered enterprise applications that improved data processing efficiency by **35%**, supporting consulting analytics workflows.
 - Designed and implemented a **multi-agent AI platform** using **LangChain** and **LangGraph** to automate **trading, risk management, compliance, and customer analytics** for large-scale banking operations.
 - Architected **specialized agent workflows**, enabling collaborative decision-making via **Trading Agent, Risk Agent, Compliance Agent**, and **Customer Intelligence Agent** modules, mirroring expert advisory teams.
 - Integrated **Model Context Protocol (MCP)** for seamless agent access to **financial APIs, regulatory databases, live market feeds, credit scoring tools, and AML screening systems**, supporting dynamic data-driven decisions.
 - Developed **agent memory systems** combining real-time conversational context with **persistent knowledge bases**, ensuring continuity and personalization across **1,000+ user sessions**.
 - Engineered **inter-agent communication protocols** for **distributed cognition**, enabling agents to collaboratively solve complex **cross-domain banking problems**.
 - Built **scalable agent infrastructure** on **AWS Lambda** with auto-scaling, maintaining **sub-2-second response time** and supporting **50,000+ financial queries daily**.
 - Collaborated with cloud infrastructure teams to deploy **large language models (LLMs)** on **AWS Trainium/Inferentia** instances, improving inference throughput by **35%**.
 - Developed **internal ML tooling** to profile and debug accuracy-performance tradeoffs across hardware accelerators.
 - Partnered with data science teams to tune model parallelism and batching techniques for enterprise-scale inference workloads.
 - Contributed to benchmarking frameworks measuring **end-to-end model performance** and **resource utilization** across multi-node deployments.
- ML Engineer, PNC Bank.** Oct 2019 - Jun 2021
- Developed core backend services for financial transaction processing, improving system throughput by **45%** and reducing latency by **30%**.
 - Built **real-time fraud detection models** for **30M+ daily transactions**, improving **precision to 98%** and reducing **false positives by 35%**.
 - Built **ensemble fraud models** combining **gradient boosting (XGBoost)** and **deep neural networks**, achieving **94% precision** and **87% recall** while reducing **false positives by 52%**.
 - Developed a **streaming feature store** with **PySpark**, delivering **200+ behavioral and merchant features**.
 - Deployed **PyTorch** and **TensorFlow** models on **AWS SageMaker** with **autoscaling**, maintaining **99.9% uptime** across peak loads.
 - Accelerated inference using **quantization** and **pruning**, cutting **latency by 30–45%** and lowering **compute cost by 40%**.
 - Implemented end-to-end **ML CI/CD** with **GitHub Actions, Docker, Terraform, and Helm**, reducing deployment cycles from days to **under three hours**.
 - Set up **MLflow** experiment tracking and a **model registry** with approval gates for **risk and compliance**, ensuring full **lineage and auditability**.
 - Added production **monitoring** for **data quality, drift, and prediction skew** using **Prometheus** and **Grafana**, with **on-call alerts** and runbooks.
 - Delivered **FastAPI** gateways and **TensorFlow Serving** endpoints, plus **Kafka** and **REST** contracts that enabled **real-time case management** and **cut investigation time by 45%**.

PROJECTS

- Developed a **scalable large language model inference platform** using PyTorch, FastAPI, and AWS Bedrock. Focused on optimizing latency and throughput via **distributed serving, caching, and model parallelism**. Integrated with **Kubernetes and Docker** for containerized deployment and MLOps automation, achieving significant performance and cost improvements.
- Built an **AI-powered assistant** to automate engineering workflows like code review summarization, documentation generation, and knowledge retrieval. Designed a **cloud-native API platform** using Python, LangChain, and AWS Lambda that handled thousands of daily requests with sub-second latency, boosting developer productivity and accelerating GenAI adoption internally.

SKILLS

- Programming:** Python, C++, Java, Bash
- ML/DL Frameworks:** PyTorch, TensorFlow, Keras, ONNX, Hugging Face
- AI/GenAI:** NLP, LLMs (GPT, Llama, Claude), LangChain, LangGraph Bedrock, SageMaker
- MLOps & Deployment:** Kubernetes, Docker, Terraform, MLflow, CI/CD, GitHub Actions
- Backend & APIs:** FastAPI, Flask, REST, GraphQL, AsyncIO, Redis
- Cloud & Infrastructure:** AWS (ECS, Lambda, API Gateway, DynamoDB, CloudWatch, ECR, S3)
- Monitoring & Security:** Datadog, Prometheus, IAM, RBAC, Cloud Logging
- Other Tools:** Git, Postman, Agile, Open Telemetry

EDUCATION

M.S. Applied Statistics & Decision Analytics | Western Illinois University