Sai Krishna Vishnumolakala

krish.ms2023@gmail.com | linkedin.com/in/sai-krishna-vishnumolakala | github.com/SaiKrishna-KK

Summary

Software Engineer with experience across backend systems, ML pipelines, and cloud-native deployments. Skilled in designing scalable distributed systems, building real-time applications, and collaborating cross-functionally from design to deployment. Passionate about immersive platforms and leveraging new technologies (ML, LLMs, distributed infra) to enhance user experiences.

Technical Skills

- **Languages:** Python, Java, C++, SQL, Bash, C#
- **Data Engineering:** Spark, PySpark, Airflow, Kafka, DBT, ETL Pipelines
- **Databases:** PostgreSQL, MySQL, MongoDB, Snowflake, Redshift, Hive
- **Cloud & Infra:** AWS (ECS, EC2, S3, SageMaker, Lambda, Glue, IAM), GCP Vertex AI, Azure Synapse
- **DevOps/MLOps:** Docker, Kubernetes, Terraform, MLflow, GitHub Actions, CI/CD
- **Security & Compliance:** OAuth 2.0, JWT, RBAC, HIPAA, GDPR
- **Visualization/Testing:** Tableau, Power BI, Plotly, Selenium, PyTest
- **ML & DL:** PyTorch, TensorFlow, Scikit-learn, Transformers (BERT, GPT, RAG), YOLOv8, OpenCV
- **Voice & LLM Systems:** STT/TTS (Deepgram, Google, Azure), WebRTC (Daily, LiveKit), FastAPI, LangChain

Work Experience

Software Engineer | Intellectsoft (Internship) | USA **June 2025 – Current**

- Designed, developed, and deployed enterprise applications with embedded AI/ML models using Python, Flask, TensorFlow, and PyTorch, integrating fraud and anomaly detection features that improved accuracy by 22% and reduced false positives in financial systems.
- Built and maintained data pipelines with Apache Spark, Airflow, and Kafka to ingest and process 15TB+ of structured and streaming data daily, enabling real-time data flow into machine learning models and improving latency by 30% across analytical workflows.
- Created RESTful APIs and FastAPI microservices for embedding predictive ML models into production workflows, achieving high availability and scalability while reducing response times for transaction risk scoring by 20%.
- Integrated advanced NLP models (BERT, Hugging Face Transformers) into text analytics platforms and customer support chatbots, increasing query resolution accuracy to 85% and reducing customer service handling time by 40%.
- Designed and optimized data warehousing schemas in Snowflake and Redshift, writing optimized SQL queries and materialized views to accelerate reporting performance by 28% for business and compliance stakeholders.
- Containerized ML workflows with Docker and Kubernetes and orchestrated deployments through AWS SageMaker and Azure ML, cutting deployment timelines from 3 weeks to 5 days while ensuring reproducibility and model traceability.
- Collaborated with cross-functional teams of data scientists, DevOps engineers, and product owners to deliver robust MLOps pipelines using MLflow and Jenkins, increasing release frequency by 40% and enabling faster feedback loops.
- Created real-time dashboards in Power BI and Tableau, integrating KPIs for AI models, fraud detection results, and system health metrics, giving executives visibility into key financial insights and improving issue response times by 18%.
- Applied secure coding practices by implementing OWASP standards, encryption techniques, and access control policies, ensuring data security and compliance with HIPAA, GDPR, and SOX requirements in sensitive domains.
- Authored user stories, acceptance criteria, and technical documentation for AI/ML applications, aligning with Agile sprints, reducing delivery delays, and ensuring traceability of requirements from business need to technical delivery.

Machine Learning, Backend & iOS Engineer – Pillowtalk **Jun 2025 – Oct 2025**

- Reduced LLM pipeline latency from 10–15s to ~500–800ms using Pipecat with adaptive scheduling and parallel execution, enabling smooth real-time conversations and significantly improving user experience.

- Increased pipeline accuracy to 95% by refining STT/TTS alignment, tuning inference, and adding fallback providers, directly reducing failed responses and boosting user retention by 22%.
- Designed and deployed distributed STT → LLM → TTS pipelines powering real-time voice features for thousands of users, collaborating with product and design teams to improve user experience.
- Engineered full-stack features including iOS WebRTC clients and backend APIs, ensuring seamless end-to-end integration across platforms.
- Built and maintained AWS microservices with Kubernetes auto-scaling, ensuring scalable, fault-tolerant distributed systems in production.
- Integrated multi-provider STT/TTS with fallback logic and dynamic voice switching, improving robustness, personalization, and overall system reliability.

Data Analyst at DoSA, University of Maryland, Baltimore County **Jun 2024 - Jun 2025**

- Built predictive models with Scikit-learn and Statsmodels to forecast student engagement, applying statistical validation to optimize university resource allocation and improve long-term institutional planning efficiency.
- Designed Tableau dashboards with automated validation pipelines using Python, enhancing reporting workflows, boosting data integrity by **20%**, and improving overall trust in student performance analytics.
- Developed SQL-based ETL workflows for large-scale financial data processing, reducing reporting delays, improving query performance by **30%**, and enabling faster, more accurate decision-making for administrators.
- Built AI-powered analytics dashboards combining Python, Tableau, and predictive modeling, providing leadership with actionable insights into engagement trends and institutional financial forecasting.
- Collaborated with stakeholders, administrators, and developers to align analytics workflows with evolving requirements, improving adoption and bridging gaps between technical and non-technical decision-makers.
- Implemented secure authentication and role-based access in UMBC's digital infrastructure, enhancing security, compliance, and reliability of student and financial data systems.

AI/ML Research Intern with Dr. Sobin CC, SRM University **Aug 2022 - Jun 2023**

- Authored **3 IEEE publications** on emotion recognition, classroom engagement, and AI-assisted research.
- Developed real-time classroom pipeline (FER2013 + eye tracking), boosting engagement detection accuracy.
- Designed GPT-4 research companion with retrieval, guardrails, and prompts for literature review support.
- Implemented preprocessing, training, and inference pipelines, reducing latency in live student analytics.
- Conducted robustness checks on YOLOv8 and CNN baselines, improving model generalization across settings.
- Presented research at **IEEE EDUCON/FIE** conferences, gaining recognition for applied AI in education.

Software Engineer | TCS | India **Sep 2021 – July 2023**

- Modernized full-stack enterprise apps (Python/Django/Flask, React.js, Java), migrating monoliths to microservices for retail and banking clients, boosting scalability and maintainability.
- Built ETL pipelines (PySpark, Hive, Talend, SQL) integrating ERP/CRM/transaction data into Snowflake, cutting reporting timelines by **40%** and improving accuracy.
- Automated ML pipelines (TensorFlow, Scikit-learn) for sales forecasting, improving accuracy **15%** and optimizing demand planning.
- Designed real-time ingestion with Kafka for IoT and financial logs, scaling to **50K+ events/sec** with minimal latency, enabling fraud/anomaly detection.
- Optimized SQL/PLSQL queries in Oracle/PostgreSQL, reducing runtime **25%** and improving reporting efficiency.Built test automation frameworks (Selenium, JUnit, PyTest), raising coverage **85%** and shortening QA cycles.
- Delivered Tableau/Power BI dashboards for KPIs and ML performance, accelerating business decisions by **20%**.
- Deployed ML-enabled apps to Azure/GCP (Synapse, Vertex AI), reducing infra costs **18%**.Collaborated in Agile teams (PMs, BAs, QA), achieving **95%** sprint completion.
- Authored SOPs, playbooks, and ML guidelines, standardizing adoption and reducing onboarding time **30%**.

Projects

ModelPort (Python Lib) | 2025

- Designed ModelPort, a framework-agnostic tool to export ML models (e.g., PyTorch) to ONNX for portable deployment across x86, ARM, and Apple Silicon.

- Integrated Apache TVM for optional native compilation, enabling zero-dependency execution and GPU/CPU acceleration.
- Developed a unified inference workflow supporting both ONNX Runtime
- Compiled libraries via Python API & CLI.
- Built benchmarking utilities to evaluate model performance and optimize deployment efficiency.

Git: *https://github.com/SaiKrishna-KK/model-port*
Pypi: *https://pypi.org/project/model-port/*

TopicMind | 2025

- Built TopicMind, a topic extraction and summarization platform that processes long-form text (articles, Reddit threads, forums) into concise, structured summaries.
- Implemented a multi-stage NLP pipeline combining BERTopic for topic modeling, semantic refinement, and a two-pass BART summarizer for high-quality outputs.
- Developed an interactive Streamlit web UI and REST API for seamless user interaction and system integration.
- Built scalable summarization APIs with caching and provenance tracking, supporting multi-user load while reducing latency.

Git: *https://github.com/SaiKrishna-KK/topic-mind*

XR DaaS - AI Powered Chest X-Ray Diagnostics | UMBC, 2024

- YOLOv8 based abnormality detection integrated with GPT for narrative reports.
- Deployed Dockerized services on AWS EC2 with S3 encryption, OAuth/SAML, and HIPAA-compliant access.
- React and Flask app with clinician friendly UX and role specific access.
- **35% latency cut** in chest X-ray inference via quantization, pruning, and batch scheduling on AWS GPUs.
- **C++ Flask services** for high-throughput image processing and vector-backed GPT with sub-second responses.
- On device privacy first models that eliminated network hops and improved local response by about 40 percent.

Git: *https://github.com/SaiKrishna-KK/xr-daas.git*

Education

University of Maryland, Baltimore County **June 2025**
Master of Science in Computer Science

- Relevant Coursework: Distributed Systems, Analysis of Algorithms, Operating Systems, Advanced Data Structures

SRM University **June 2023**
BTech. in Computer Science and Engineering Specialization in Machine Learning | Minor in English Literature

Research Publications

- System for Emotion and Engagement Recognition in Education (SEERE) - Presented and Published FIE2024
- Patent: "Analysing Emotions and Concentration Levels of Students" (App No.202341026695)
- IEEE Conference Presentation: "iSEEDS" at IEEE EDUCON 2023
- Springer's Book Chapter: "Deep Learning Models in Finance" ARC - AI Based Research Companion
- Research Presentation: "Commodity Price Prediction" at MIND 22

Achievements and Certifications

- Patent applicant: Analyzing Emotions and Concentration Levels of Students, App No. 202341026695.
- AWS Certified Solutions Architect.
- TensorFlow Developer Certificate - Google.
- AI Programming with Python Nanodegree - AWS and Udacity.
- BuiDL for Web3 Winner - prize 5,000 dollars.