Last updated in October 2025

Aakash Kaushik

✉ kaushikaakash7539@gmail.com ☎ +917575885094 kausky in kaushikaakash7539 kausky.com

Education

SRM University (Sri Ramaswamy Memorial University) Jul 2019 – Jun 2023
BTech in Computer Science Engineering; **CGPA: 9.55/10**

Skills

Languages: Golang (3+), Python (5+), C++, Bash, SQL

Technologies: GCP, AWS, Azure, Kubernetes, Docker, Pulumi, Terraform, CI/CD, Git, Ray, Baseten, PyTorch, TensorFlow, OpenVINO, gRPC, REST APIs, FastAPI, Fiber, Redis, MySQL, PostgreSQL, ClickHouse, PubSub, Prometheus, Grafana, OpenTelemetry, SQLAlchemy, Alembic, Protocol Buffers, Cloud Storage (S3, GCS, Azure Blob), SFTP, Jinja2, Prompt Engineering, Catch2

Key Expertise: Machine Learning and Generative AI (LLM/VLM/Image Gen), MLOps, Scalable ML Systems, ML Observability, High-Throughput Data Pipelines, Distributed Systems, Cloud Infrastructure, Microservices Architecture, Performance Optimization, API Gateway Development, Infrastructure as Code (IaC), Event-driven Architecture, Database Schema Design, Open-Source Contribution, Multi-cloud Integration

Experience

Lyric Feb 2025 – Present
Backend Software Engineer, ML

- Led an MLflow refactoring to scale time series logging from 10K to over **10 million**, reducing millions of MLflow runs to just two per execution by implementing a batched, Parquet-based storage strategy.
- Re-architected the time series data ingestion pipeline for Ray, leveraging ClickHouse for on-demand data joining and partitioning to eliminate I/O bottlenecks and enable efficient parallel data access for workers.
- Migrated foundational time series models (TimesFM, Chronos) to Baseten, resolving cold-start issues and implementing a gRPC serving layer to handle high-throughput predictions for over **1 million** time series.
- Deployed Kubernetes ServiceMonitors for Ray jobs, enabling automated metric collection via Prometheus and creating Grafana dashboards for real-time observability of ML training and inference workloads.

Tune AI Jul 2023 – Feb 2025
Software Engineer 3

- Engineered a high-throughput distributed proxy server handling **over 1 million** requests/day for various LLM providers (OpenAI, Anthropic, OpenRouter) with low latency, token limits and authorization.
- Built a multimodal document information extraction pipeline processing **100K docs** (**3.8 million** pages) daily with **95%** precision/recall, reducing processing cost by **54%**.
- Led development of backend services for a Generative AI platform supporting various data validation, fine-tuning jobs (LoRA, QLoRA), and flexible LLM deployment scenarios (BYOC, managed).
- Implemented platform wide billing, advanced configurations for OpenAI, Anthropic, Gemini agents, and support for multi-modality features including VLMs in the platform.

Document Processing Pipeline

- Implemented a robust decoupled event-driven architecture with fault tolerance and retry mechanisms to process documents at scale while avoiding duplicate processing.
- Designed and implemented batch inference to reduce system latency by **25%** compared to real-time inference and achieve significant cost savings within budget constraints.
- Built a custom multi-modal page classifier to label document pages, reducing overall processing load by **70%** and decreasing processing costs.
- Developed monitoring dashboards to track system health and document processing status using comprehensive logging and telemetry.

Infrastructure and Data Management

- Developed a flexible, scalable infrastructure engine using Pulumi and Kubernetes to manage cloud resources across AWS, GCP, and Azure.

- Created a high-performance file system server that manages files and logs on multiple cloud providers with robust CRUDL operations.
- Implemented event-driven, worker-queue architecture (Pub/Sub) backed by MySQL for document processing with fault tolerance and retry mechanisms.
- Built monitoring dashboards for tracking system health and document processing status, with advanced logging and OpenTelemetry integration.

Tune AI Oct 2020 – Jun 2023

Software Engineering Intern

- Developed a sidecar server for cloud VMs to provide fully managed Generative AI development space with a single click.
- Created a cloud-agnostic file management system (Relics Server) using Python, FastAPI, and gRPC, supporting file operations across AWS S3, Azure Blob Storage, and Google Cloud Storage.
- Built the Infrastructure Creation Engine (ICE) for BYOC functionality, enabling users to connect their own Kubernetes clusters, reducing infrastructure costs by up to **80%**.
- Implemented a custom code generation tool using Jinja2 templating to automate API client creation, reducing development time by approximately **25%**.

Google Summer of Code Mlpack May 2021 – Aug 2021

Developer

- Implemented MobileNetV1 and ResNet model builders in C++, integrating pre-trained weights to reduce training time by **40%**.
- Contributed to mlpack/mlpack: Added Mean Absolute Percentage Error (MAPE) and Softmin Activation function with backward implementation and migrated test files from boost to catch2.
- Spearheaded the migration of approximately **60%** of core testing suite from Boost to Catch2, resulting in improved test execution time and maintainability.
- Addressed over **100** static code analysis warnings and style issues, improving code quality and reducing potential bugs in the codebase.

OptimEyes.ai Nov 2021 – Jun 2022

AI/ML Intern

- Architected and deployed ML regression models for cloud workload security scoring, improving threat detection accuracy by **30%** and reducing inference latency by **19%**.
- Engineered feature extraction pipelines that processed **200+** cloud workload metrics, reducing false positives by **25%**.
- Implemented automated CI/CD pipelines for model deployment, reducing release cycles from 2 weeks to 3 days and cutting engineering overhead by **40%**.

Mavoix Solutions May 2020 – Aug 2020

Deep Learning Engineering Intern

- Engineered text recognition and image classification models on medical images to prescreen patients achieving up to **95%** accuracy on standard benchmarks.
- Developed Flask APIs for model deployments and optimized codebase to improve performance.

Publication and Certifications

- mlpack 4: A fast, header-only C++ machine learning library: DOI 10.21105/joss.05026
- Machine Learning Data Lifecycle in Production
- Introduction to Machine Learning in Production
- Deep Learning Specialization
- Machine Learning.
- Google Cloud Platform