Abdullah Waheed

San Jose, CA | 408-674-0311 | abdullahw888@gmail.com | www.linkedin.com/in/abdullah-waheed0524/

Software Engineer focused on AI applications, backend systems, and cloud infrastructure. Built and deployed production-grade LLM agents, API gateways, Kubernetes systems, and observability stacks across AWS and Azure.

EDUCATION

San Jose State University **Aug 2021 - May 2024**
Bachelor of Science, Software Engineering *San Jose, California*

TECHNICAL SKILLS

Languages: Python · Java · JavaScript · SQL
AI/ML Frameworks: LangChain · PyTorch · OpenCV · TensorFlow · HuggingFace Transformers · Ollama
Cloud/AI Infrastructure: AWS (EKS, Aurora, EC2, S3, IAM) · Azure · Kubernetes · Docker · Terraform · GitHub Actions · Jenkins · Apache APISIX · NGINX
Databases: PostgreSQL · MySQL · NoSQL(MongoDB, DynamoDB)
Observability: Grafana · ELK Stack · OpenTelemetry

WORK EXPERIENCE

DefendAI, *Software Engineer (AI Applications and Infrastructure)* **June 2024 - Oct 2025**

- **Built and shipped LLM-powered agents** (CTO, PMO, Research, RAG workflows) using LangChain + FastAPI + Discord integrations, enabling automated document analysis and reporting across the team.
- Developed high-throughput **APISIX-based LLM gateway** implementing **rate limits, redaction filters, and policy enforcement** for 15+ services. Reduced security-related failures and misrouted LLM requests in system.
- Engineered autonomous **threat-hunting agents** that identified and mitigated 100+ emerging LLM vulnerabilities, **reducing mean-time-to-response** (MTTR) by 50%.
- Built and maintained **hybrid infrastructure** on AWS and Azure, using Kubernetes, Docker, GitHub Actions, and Jenkins, **streamlining CI/CD** and enabling **rapid AI experimentation**.
- Integrated **full observability dashboards** (Grafana, Prometheus, OpenTelemetry) into LLM pipelines, **reducing triage time by half** and increasing system reliability for production workloads.

Wortise, *DevOps Engineer (Contract)* **Sep 2023 - Dec 2023**

- Tuned and optimized the ELK stack for >1M daily log entries, improving indexing speed and query latency.
- Enhanced Elasticsearch Watchers for high-volume alerting, **reducing query latency by 25%**.
- Delivered real-time Kibana dashboards, **enabling clearer incident investigation for engineering teams.**.

Cloudvista, *Software Engineering Intern* **May 2022 - Aug 2022**

- Built a Django-based log ingestion tool deployed on Vercel and integrated with AWS S3 for scalable log storage.
- Automated data-processing scripts for internal monitoring and analytics workflows..

PROJECTS

Wozway - (GitHub | Demo **)** **Nov 2024 - May 2025**

- Architected LLM security gateway for scaling to 10k+ req/s with regex, anonymization, and phishing-policies.
- Supported scaling to high request volumes via APISIX + Kubernetes + FastAPI.
- Integrated with foundational models and OpenWebUI with audit logs and real-time metrics.

AttendanceAI (GitHub | Demo**)** **Aug 2023 - May 2024**

- Built a real-time facial recognition system with OpenCV+TensorFlow on Jetson Nano, optimized for real-time inference on embedded hardware.