Aiden Le

Backend & Platform Engineer | AI Infrastructure

Austin, TX | aidendle94@gmail.com | aidenle.com | GitHub | LinkedIn

Backend and platform engineer with five years at H-E-B Digital. Cofounder of Tabi and Vexly. Builds cloud platforms, coding agents, and GPU inference systems, owning work from architecture through production.

Professional Experience

H-E-B Digital | Software Engineer II

Oct 2021 – Present

Search & Restaurant Platforms | Jun 2025 – Present

- Architected an in-house semantic recipe-search platform using embeddings and pgvector, replacing a licensed vendor and leading client migrations and legacy-service retirement.
- Built a shared restaurant catalog for H-E-B, Central Market, and Joe V's, with event-driven ingestion and catalog APIs.
- Architected an AI-powered service that converts recipe text into structured ingredients, quantities, units, and detailed preparation and cooking methods.

Internal AI Platforms | Oct 2025 – Present

- Built Buddybox, a multi-user agent platform with isolated execution and GitLab integration; onboarded 17 repositories at launch.
- Enhanced the internal AI code reviewer with inline suggestions, reusable review guidelines, and prompt caching.
- Built and distributed Buddy Code, a fully custom internal coding agent with job scheduling, swarm orchestration, and integrations across internal tools and services.

Delivery Platforms | Oct 2021 – Jun 2025

- Led expanded-delivery architecture across capacity planning, availability APIs, and administrative workflows.
- Extracted geocoding into a standalone service and migrated delivery availability to ECS Fargate.
- Replaced scheduled order-processing jobs with Kafka and SQS consumers; built APIs supporting Central Market catering and holiday meals.

Leadership & Operations: Mentored two interns, reviewed approximately 1,900 merge requests, and led production incident response. Established internal agent-development communities and presented at the AI guild.

Education

Texas Tech University | B.B.A., Information Systems | Cum Laude | 2019

Independent Product Engineering

JustStrings | Cofounder, Tabi & Vexly

Vexly | AI Agent Workspace

- Built an AI workspace spanning web, Slack, and Discord, with connected tools, scheduled tasks, and file ingestion.
- Implemented role-based permissions for agent tools and authenticated messaging workflows.
- Built Stripe subscriptions and usage-based credits, including plan limits and trial-abuse prevention.
- Developed agent evaluation and tracing tools to compare models and diagnose execution failures.

Tabi | iOS & Android Comic Collection App

- Built a Flutter comic collection app with 10,000+ Android installs and a 4.6-star iOS rating.
- Developed photo identification using multimodal image embeddings and concurrent GPU inference.
- Built asynchronous AI grading pipelines that assess page condition and generate overall comic grades.
- Delivered collection valuation using grade-adjusted market prices, cached pricing APIs, and historical value tracking.

Open-Source AI Infrastructure

DGX Spark Inference Runtimes | Maintainer

- Published open-source SGLang and vLLM deployments for DeepSeek-V4.1-Flash across four DGX Spark nodes, supporting 524K-token contexts.
- Optimized NVMe-backed model data access, CUDA graph execution, and RDMA integration; measured 3,068 tokens/s prefill and 331 tokens/s aggregate decode across 16 streams.
- Fixed prefix-cache invalidation in coding-agent sessions, reducing follow-up time to first token from about a minute to under a second.
- Quantized Engram tables from 203 GB to 111 GB using NVFP4, with measured token agreement matching the FP8 reference’s self-agreement baseline.

Technical Skills

Languages: Java, Python, TypeScript, JavaScript, Dart, SQL, C, Bash

Backend & APIs: Spring Boot, FastAPI, Node.js, Express, Pydantic, SQLAlchemy, Uvicorn, asyncio, Boto3/aioboto3, REST, GraphQL, gRPC, OAuth 2.0, RBAC, microservices, event-driven architecture

Web & Mobile: React, Vue, Redux Toolkit, Tailwind CSS, Vite, Flutter, Provider, Sembast, Firebase Auth, Firestore, Firebase Cloud Messaging, RevenueCat, Stripe

AWS: Lambda, SAM, ECS/Fargate, ECR, EC2, EKS, SageMaker, Bedrock, DynamoDB, S3, SQS, Cognito, RDS, VPC, IAM, KMS, Secrets Manager, ALB/NLB, CodeDeploy

Data & Search: PostgreSQL, pgvector, Kafka, Redis, Valkey, CloudSearch, SQLite, vector embeddings, semantic search, caching, asynchronous processing

Agents & Applied AI: LangChain, LangGraph, LangSmith, MCP, Claude, Gemini, coding agents, swarm orchestration, job scheduling, tool calling, agent evaluation, prompt caching, DINOv2, SigLIP, TEI

GPU & Inference: PyTorch, Hugging Face Transformers, SGLang, vLLM, FlashInfer, Triton, DeepGEMM, b12x, CUDA Graphs, CUDA streams, tensor/expert/context parallelism, speculative decoding, KV caching, RDMA, RoCE, NCCL, DGX Spark

Quantization: FP8, NVFP4, MXFP4, MXFP8, GPTQ, mixed-precision inference, block-scale optimization, NVMe model offloading

Infrastructure, Testing & Operations: Terraform, Docker, Kubernetes, GitLab CI/CD, GitHub Actions, Datadog, Envoy, pytest, pytest-asyncio, Moto, Testcontainers, Vitest, React Testing Library, Flutter Test, Mocktail, Ruff, mypy