KENT

Applied AI Engineer

Portfolio: kentdoki.dev · **GitHub:** github.com/kendoki · kentdoky@gmail.com · LinkedIn · fourpoket.com

SKILLS

LLM & RAG: agent orchestration (supervisor/worker, approval-gated), tool/function calling, MCP servers, structured outputs (Pydantic, Zod), prompt versioning, context engineering, streaming (SSE), semantic caching, cost-aware model routing, hybrid search (pgvector + Postgres full-text), reciprocal rank fusion, cross-encoder reranking, section-aware chunking, RAG-vs-structured-query routing

Evaluation & Backend: versioned eval sets, LLM-as-judge (quorum/variance rules), CI regression gates, A/B and canary releases with auto-rollback; Python (FastAPI, Pydantic v2, SQLAlchemy, Celery), TypeScript/Node.js (Express, Prisma), PostgreSQL (pgvector, Row-Level Security, PostGIS), Redis, Docker, GitHub Actions, OpenTelemetry, Stripe

EXPERIENCE

Senior Applied AI Engineer · Etiqa Insurance Apr 2024 - Nov 2025

- Built an internal claims operations assistant that cut routine policy lookups ~90%, from minutes of cross-referencing four separate systems to a single grounded query with citations, run daily in production with a human in the loop.
- Designed the retrieval pipeline end to end: section-aware chunking, hosted embeddings, hybrid search (pgvector + Postgres full-text) with reciprocal rank fusion and cross-encoder reranking. Precision improved every delivery phase.
- Built a supervisor/worker agent layer with typed tools, persisted per-request state, and a hard step cap. Exposed structured lookup tools via an MCP server for reuse by other internal AI tools.
- Implemented two-stage routing: a classifier picks an off-the-shelf Q&A layer, the custom pipeline, or a direct structured lookup, then a model tier by complexity, keeping ~70% of traffic on a cheaper model with no quality regression.
- Stood up versioned prompts with A/B and canary rollout plus auto-rollback, LLM-as-judge scoring calibrated against human sampling, and a CI regression gate that blocks promotion on score drops.
- Enforced an augmentation-only boundary in code: a guardrail layer checks drafts for banned recommendation language, ungrounded figures, stray PII, and dead citations. The assistant never decides or writes back to records.
- Delivered multi-tenant isolation with Postgres Row-Level Security and onboarded a second business entity with scoped evals. Instrumented OpenTelemetry tracing, per-request cost tracking, and retrieval quality dashboards.

Senior Applied AI Engineer · Freelance Apr 2023 - Mar 2024

- Built and shipped an AI chatbot platform that answers questions over a company's own documents, from prototype to production, across both the Python AI pipeline and the TypeScript product layer.
- Built the full RAG path: ingestion, parsing, chunking, embeddings, and hybrid retrieval with rank fusion and cross-encoder reranking. Retrieval quality won head-to-head demos against cosine-only competitors.
- Cut per-query cost ~40% through multi-model routing and semantic caching, backed by token-level cost logging, per-customer quotas, and budget alerts that kept unit margins healthy.
- Shipped a streaming (SSE) embeddable Preact chat widget with inline citations, plus Slack, Intercom, Notion, and Drive integrations, the admin dashboard, and Stripe billing.
- Maintained multi-tenant isolation with zero cross-tenant leakage and reconciled billing. Built the eval harness early and used it to drive every retrieval and routing change.

Senior Software Engineer · Accenture Nov 2020 - Mar 2023

- **Document Intelligence Platform (logistics):** Took a document extraction product to production on a two-service architecture, a Node/TypeScript product layer with a Python OCR/NLP service, joined by typed contracts from one shared schema that caught drift at build time. Rebuilt the operator review screen keyboard-first, sharply improving review throughput.
- **Corporate Workflow and Approval System (team lead):** Rebuilt a travel and expense system as a composable workflow engine, with workflow and form templates configured as data and state machine transitions gated on form validation. Shipped across several business units and cleared the internal audit that flagged the prior process.

Software Engineer · Sunway Nov 2019 - Oct 2020

- Built the backend API and real-time ops dashboard for an events security operation: a constraint-based scheduling engine (SQL + PostGIS pre-filtering, weighted scoring) cut scheduling from hours to minutes and made double-bookings impossible at the database level.
- Delivered live attendance monitoring over Socket.IO with QR and GPS-verified check-in. One shared, documented API served the web dashboard and a separate mobile team.

Software Engineer · Corporate Services Firm Aug 2017 - Aug 2019

- Full-stack JavaScript web development.

PROJECTS

fourpoket · AI coding agent · live product Jan 2026 - Apr 2026

fourpoket.com · npm: fourpoket

- Designed, built, and shipped an AI coding agent CLI: multi-provider orchestration (DeepSeek, MiniMax, OpenRouter) behind one interface with retries, backoff, fallback, and per-call cost tracking, over a three-phase Read → Write → Verify pipeline on an AST engine with targeted reads and surgical writes.
- Shipped the full product: authentication, Stripe credit billing with webhook reconciliation and real-time balance enforcement, an admin dashboard, and production monitoring.