

Rony Zeenaldeen

AI Systems Engineer | Backend, Infrastructure & Distributed Systems

ronizenalden@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

AI Systems Engineer (4+ years) designing and operating production-grade, event-driven distributed platforms spanning backend services, applied AI/ML, and agentic RAG systems at scale. Engineers polyglot infrastructure in Go and Python (Django, FastAPI), provisioning immutable Kubernetes platforms via Terraform and Ansible, and orchestrating Kafka-driven pipelines for high-throughput, fully observable systems. Proven impact: 40× faster queries, 60–70% API latency reduction via Redis caching, and 65% lower database load.

TECHNICAL SKILLS

Back-End: Go (Golang), Django, FastAPI, Node.js, TypeScript, JavaScript, REST API Design, GraphQL, WebSockets

AI / ML: Python, TensorFlow, PyTorch, Keras, Scikit-learn, XGBoost, EfficientNetV2, Facebook Prophet, Transformer Models

Databases: PostgreSQL, pgvector, Vector Search, SQL, Redis, MongoDB, MySQL, Firebase Firestore

Messaging & Streaming: RabbitMQ, Apache Kafka

Cloud & AWS: IAM, S3, RDS, SES, Route 53, Elastic Beanstalk

DevOps & Infrastructure: Terraform, Ansible, Docker, Kubernetes (Helm, k3s, Rancher-ready), NGINX, GitHub Actions, CI/CD, Linux

Monitoring & Reliability: Prometheus, Grafana, pgBackRest, pgBadger

Mobile: Flutter, Dart, GetX, BLoC, Firebase

PROFESSIONAL EXPERIENCE

AI & Backend Engineer (Independent)

Nov 2024 – Present

- **Enterprise AI Document Automation (Client – Austria / Remote):** Built an end-to-end Django + Claude Vision VLM pipeline converting scanned Arabic documents to formatted DOCX, collapsing a multi-stage OCR/LLM stack into one strict-JSON VLM call to eliminate error compounding — overriding python-docx limits via direct OpenXML table-cell (w:tc) manipulation and run-merging to preserve exact fonts and layouts.
- **Intelligent Human-in-the-Loop Routing:** Built a confidence-scored (0.0–1.0) routing workflow that auto-passes clean fields and routes uncertain ones (<0.85) to role-based reviewers (Employee/Reviewer/Admin), with full audit logging.
- **Production AI & Event-Driven Infrastructure:** Provisioned multi-node IaC environments (Terraform, Ansible) running decoupled FastAPI inference microservices across Kafka/RabbitMQ pipelines, with Prometheus/Grafana observability and pgBackRest point-in-time recovery.
- **High-Impact Applied ML Models:** Shipped a chest X-ray classifier (EfficientNetV2S, 91–92% accuracy / F1 0.95) via Django REST API, a Facebook Prophet sales-forecasting pipeline on 1M+ transactions, and an HR attrition model (87% accuracy, \$225K–\$450K est. annual retention savings).

Backend Systems Engineer (Independent)

Jan 2022 – Nov 2024

- **API Design & Integrations:** Built and shipped secure REST APIs (Django REST Framework, Go) with JWT/OAuth2 auth, versioning, and OpenAPI docs, plus GraphQL endpoints and Stripe/PayPal/Razorpay payment integrations with secure webhook handling.
- **Performance & Caching:** Reduced PostgreSQL query response times from 2000ms to <50ms, cut database load by 65%, and boosted API latency by 60–70% through strategic Redis caching.
- **Real-Time & Distributed Systems:** Designed real-time systems (WebSockets, Django Channels, Redis) for live ride-matching and location streaming; deployed Go microservices for latency-critical, high-concurrency components, with RabbitMQ/Kafka decoupling and pgBackRest/pgBadger strengthening PostgreSQL backup and query analysis.
- **Cloud & DevOps:** Deployed backend services using Docker, Kubernetes, NGINX, and AWS infrastructure for high availability and scalability.

KEY PROJECTS

NeuroSeek AI — Agentic AI Research Platform (LIVE) • 2026 – Present

[↗ Link](#)

Self-hosted research assistant that plans multi-step search queries, crawls and embeds web sources, and answers via hybrid RRF-fused RAG (pgvector semantic + keyword search) — streaming a fully cited report token-by-token.

- Implemented an agentic “gap-driven” re-search loop: the system audits its own answer coverage, identifies gaps, and autonomously re-searches before finalizing a report.
- Engineered a polyglot backend (Django 5 + DRF + Celery/Redis; dedicated Go WebSocket streaming hub and Kafka consumer) on immutable infrastructure — Terraform/Ansible, Docker Compose/Kubernetes (Helm, k3s/Rancher-ready), Kafka (KRaft), Prometheus/Grafana, pgBackRest (PITR)/pgBadger — hardened with SSRF protection, JWT rotation/blacklisting, and rate limiting.

Tech: Python • Django • DRF • Celery • Redis • Go • PostgreSQL • pgvector • Apache Kafka • Ollama • Next.js • TypeScript • Terraform • Ansible • Docker • Kubernetes • Prometheus • Grafana • pgBackRest

PulmoSight — Doctor-in-the-Loop Chest X-Ray AI (LIVE) • 2026 – Present

[↗ Web App](#) • [↗ Mobile App](#)

Self-hostable clinical AI system that classifies 18 chest pathologies with Grad-CAM heatmaps and drafts a structured radiology report grounded strictly in the classifier’s confirmed findings.

- **System Architecture:** A polyglot backend (Django 5 + DRF, FastAPI/PyTorch inference, Go Telegram bot) now shipping a role-aware native Flutter mobile app (v1.1.0) alongside the doctor/reviewer web workflow, with an on-device camera pipeline (auto-deskewing, EN/AR voice dictation, prior-study diffs).
- **On-Device Privacy & Delivery:** Hardware-backed keystore tokens, AES-encrypted offline storage, screenshot blocking, and biometric locks harden the client; e-signed PDFs move through expiring, dual-salted capability tokens via a RabbitMQ-decoupled, PIN-verified Go bot with nightly backups.

Tech: Python • Django 5 • DRF • FastAPI • Go • Flutter • Dart • BLoC • RabbitMQ • PostgreSQL • PyTorch • Grad-CAM • Docker • Telegram Bot API

ROTECHAI — AI + Software Studio Platform (LIVE) • 2026 – Present

[↗ Link](#)

Self-hosted, multilingual (EN/AR/NL/DE) marketing site paired with a live AI product — real-time text classification and a RAG chat assistant for services, pricing, and process questions.

- Right-sized the AI layer to a single ~400MB multilingual embedding model (fastembed/ONNX Runtime, no PyTorch, lazy-loaded) with hybrid cosine + keyword RAG over an in-memory knowledge base — no GPU or paid LLM API required; backend split as Django 5 + DRF with an independently restartable FastAPI inference microservice, decoupling AI-service restarts from site uptime.
- Built a Next.js 16 (App Router) + React 19 + TypeScript frontend across 4 locales with full RTL Arabic support and SEO best practices (structured data, hreflang, sitemaps); automated NGINX edge deployment and environment hardening via Ansible — HSTS, CSRF/cookie hardening, brute-force lockout, rate limiting, fail-closed secrets.

Tech: Python • Django • DRF • FastAPI • PostgreSQL • Redis • Next.js • React • TypeScript • Ansible • Docker • NGINX

OPEN SOURCE CONTRIBUTIONS

- [FastAPI \(PR #12066\)](#): Reproduced a dependency-teardown deadlock under concurrent load (AnyIO threadpool exhaustion); verified the maintainer’s fix and identified a residual gap when the resource is acquired pre-yield.
- [pydantic-settings \(PR #920\)](#): Root-caused and fixed a generic-subclass detection bug breaking Secret-type subclasses on env-var loading, with a full regression suite added; ruff/mypy clean. Also scoped a design question on pydantic’s method-detection internals directly with the maintainer team ([Issue #10177](#)).
- [langchaingo \(PR #1513\)](#): Fixed the Anthropic-integration handler dropping tool-call content on multi-part AI messages, with unit tests; go vet/golangci-lint clean.

EDUCATION

B.Sc. Informatics Engineering — Software & Applications Development

Sep 2020 – Jun 2025

Al-Shahba Private University, Aleppo, Syria

CERTIFICATIONS

- Artificial Intelligence Essentials V2 — IBM (2025) [\[Link\]](#)
- Data Engineering Professional — IBM (2025) [\[Link\]](#)
- Strategic Planning Professional Certificate — Canada [\[Link\]](#)
- Project Management Professional (PMP v6) — Project International (2021) [\[Link\]](#)

LANGUAGES

- Arabic — Native
- English — B2-C1 (Proficient)