

Guy Flenner

AI Engineering Leader · Hands-On Technical Lead | Agentic Orchestration & LLM Evaluation | Distributed Systems Netanya, Israel | +972 (54) 7179878 | guyflenner@gmail.com | [LinkedIn](#)

PROFESSIONAL SUMMARY

AI Engineering Leader with 15+ years building production systems at the intersection of **AI agent orchestration and scalable backend architecture** — equally effective **leading at scale and building hands-on**. I led a 12-engineer squad (hiring and **promoting multiple engineers into leadership**) and today work as a **hands-on technical lead**, delivering a production AI agent platform end-to-end alongside the **LLM-quality evaluation** that guards it and the **org-wide AI-development adoption** it drove across R&D. Earlier, I re-architected a legacy Sidekiq/Ruby system into **event-driven microservices** on Kafka at **hundreds of millions of events/day** (Affogata), while cutting cloud spend **30%**. **80% automation** of manual operations. Fluent English, native Hebrew.

KEY ACHIEVEMENTS

- **Agentic AI Platform (2025–26): 80% automation of manual operational tasks** for insurance users — delivered by a production multi-tenant orchestration layer linking Salesforce to Claude (AWS Bedrock, multi-MCP tools) with layered LLM guardrails and cost-aware model routing.
 - **LLM Quality & Evaluation:** Built the evaluation layer guarding the platform — a property-based **golden-question regression system** (structural/factual/cost/safety) wired as a CI gate with automated LLM root-cause diagnosis, plus a **quality-gated RAG ingestion** pipeline admitting new knowledge only after deterministic + LLM-as-judge checks, benchmarked across **500+ verified support questions**.
 - **AI-Adoption Leadership:** Drove the R&D org's shift to agentic development — authoring a shared catalog of tested, installable Claude Code agent skills adopted across R&D, and running the enablement program that onboarded engineers and team leads onto AI-assisted development.
-

CORE COMPETENCIES

- **AI & Agentic Systems:** Multi-Agent Orchestration, Model Context Protocol (MCP), AWS Bedrock (Claude), LangGraph, LangChain, RAG (ChromaDB, sentence-transformers), ReAct patterns, cost-aware model routing, Ollama (local), FastAPI, Python 3.13, pytest, ruff, mypy, uv
 - **LLM Quality, Safety & LLMOps:** LLM evaluation (golden-question harness, property-based scoring, LLM-as-judge), guardrails, **PII redaction & masking**, **prompt-injection defense**, Bedrock Guardrails, regression/drift detection, tracing & correlation IDs, LangSmith, token-cost & latency optimization
 - **Architecture & Distributed Systems:** Event-Driven Architecture, Microservices, Domain-Driven Design (DDD), Kafka, Kubernetes (Affogata), KEDA, Schema Registry, DLQ, idempotency, Bloom filters, transactional outbox
 - **Cloud & Infrastructure:** AWS (ECS Fargate, DynamoDB, RDS, S3, OpenSearch, KMS, ALB, ElastiCache), GCP, Docker, Terraform / Terragrunt (IaC), CircleCI, Jenkins; PostgreSQL, MySQL, Redis, ChromaDB (vector), SQLite, Elasticsearch
 - **Observability:** Datadog, Grafana, Prometheus, Kibana, CloudWatch, PagerDuty, SLOs, distributed tracing, structured logging, incident response
 - **Leadership & Delivery:** Team leadership (12–20 across roles), hiring & mentoring, promoting engineers into leadership, technical roadmap, cross-functional alignment (Product / Data Science / Design / executives), Agile (Scrum/Kanban), FinOps
-

PROFESSIONAL EXPERIENCE

Novidea | Software Engineering Team Lead | Oct 2024 – Present

- **Architected and shipped a production multi-tenant AI agent platform** (Python/FastAPI, ~47K LOC, 800+ tests): Salesforce LWC ↔ Claude via **AWS Bedrock**, **multi-MCP orchestration** (Salesforce, Atlassian, custom Code Graph), per-user **JWT/RS256 MCP pool**, **cost-aware model routing**, adaptive retry with **partial-result fallback**, and **layered guardrails** (input PII masking, defense against prompt injection, an MCP wrapper exposing only allow-listed tools), with **GDPR-by-design** 3-tier erasure (KMS CMK) on 13-module Terraform IaC (EFS, ElastiCache, scheduled scale-to-zero) — **80% automation** of manual insurance operations.
- **Built the LLM-quality system guarding it**: a **qa-sentinel** regression harness scoring answers against golden questions via deterministic property assertions (structural/factual/cost/safety) as a **CircleCI** gate with LLM root-cause diagnosis; a **RAG eval & ingestion** pipeline (sentence-transformers + ChromaDB, 13 collections / ~2,500 chunks, LLM-as-judge; benchmarked Cognee as an alternative) admitting new knowledge **only after passing the quality gate**; and deterministic **cited answers** over LLM generation across a 500+ verified-Q&A corpus.
- **Built a custom Salesforce Code Graph MCP server** (FastMCP, **14 tools across 23 Apex domains**) for AI-driven impact, call-chain, and test-coverage analysis — **migrated FalkorDB → a zero-dependency read-only SQLite model** with a golden-question CI gate; hybrid Python tree-sitter + Node.js parser.
- **Led the R&D org's AI-development adoption**: built a **company-wide Claude Code plugin marketplace** (installable skills/agents/MCP configs adopted across R&D; authored ADR + agentic-SDLC skills), ran an **org-wide AI-tooling enablement session** for R&D and team leads, and authored a native **multi-agent SDLC orchestrator** (10 agents, repair loop, HITL) replacing an earlier LangGraph prototype. **Hired FE/BE engineers, promoted a senior engineer to tech lead**; drive a **DDD strangler-fig refactor** of the legacy claims platform with architects, product, and executives.

Affogata | Engineering Manager | Oct 2021 – Nov 2024 (3 years)

- Led re-architecture of a production **Sidekiq/Ruby pipeline** into **event-driven Python microservices** on **Kafka (AWS MSK) + Kubernetes** — processing **hundreds of millions of transactions/day** for real-time social-data sentiment analysis, with zero-downtime migration of a live system.
- Designed a fault-tolerant consumer topology (**DLQ, idempotency, Schema Registry**); owned production **incident response via PagerDuty on-call rotation**.
- **Reduced AWS/GCP cloud spend by 30%** via cost-per-service analysis, reserved instances, and Terraform-managed right-sizing.
- Led a cross-functional team of **20 engineers** (backend, data, infra) — hiring, mentoring, **promoting engineers into senior/lead roles**, performance management, roadmap, and delivery; technical liaison across Product, Data Science, and Engineering.

Seeking Alpha | Senior Full Stack Developer → Team Lead | Dec 2014 – Oct 2021 (7 years)

- Built and maintained the core **content platform** serving millions of readers (high-concurrency financial articles, market data, analysis), plus the **CMS pipeline** and **premium paid products** (Billing, Market Data).
- Built **React.js SPAs** and **RESTful APIs** on **Ruby on Rails** with **Redis/Sidekiq**; integrated Getty Images and Braintree/Zuora/Stripe. Grew from Senior Developer to **Team Lead** of the paid-products squad.

NCR / Retailix | Solutions Architect / Senior Software Engineer | 2010 – 2014

- Designed large-scale enterprise retail POS applications (C#, ASP.NET, SQL Server); led international on-site integrations from POC to production.

Israel Defense Forces (IDF) | Senior CRM & PLM Programmer (Major) | 2005 – 2010

- Developed enterprise CRM/PLM systems (ASP.NET, Oracle/SQL Server) for large-scale defense logistics; coordinated multi-system integrations.

EDUCATION & LANGUAGES

- **M.B.A. in Business Administration** | Ramat-Gan College | *Graduated with Distinction*
- **B.Sc. in Computer Science** | Tel-Hai College · **Languages**: Hebrew (Native) | English (Fluent)