Sammy Junior

Senior AI Engineer

São Paulo, Brazil | +55 83 99106-5299 | samjrdev98@gmail.com | linkedin.com/in/sammyjdev | github.com/sammyjdev

SUMMARY

Senior AI Engineer with 6 years of distributed systems before the first model call. Verified results: 85.5% p50 token compression across 69 real context windows, 70% API latency reduction on a platform processing 1M+ daily requests, 10M+ government records consolidated with zero data loss, incident detection cut from 60 min to under 10. Java 21, Python, Spring AI, RAG, agents, MCP — across energy, fintech, e-commerce, and government in the US and Brazil.

SKILLS

Proficient: Java 21, Spring Boot 3, Kafka, AWS, Kubernetes, Docker, PostgreSQL, Redis, REST APIs, Microservices, Event-driven architecture, Clean Architecture, DDD, Python, Spring AI, LangChain4j, RAG pipelines, MCP, LLM orchestration, Qdrant, pgvector, Ollama, FastAPI

Intermediate: Kotlin, Rust, React, Angular, Azure, Terraform, Grafana, Prometheus, Elasticsearch, OAuth2/OIDC, JWT, ArchUnit, GitHub Actions, Neo4j, NetworkX, tree-sitter, mem0

EXPERIENCE

Senior AI Engineer

SJ AI Solutions | Mar 2026 – Present | Remote

- Identified the core problem of AI coding agents losing context across sessions and machines, and designed AXON to solve it: a self-hosted MCP context engine that captures state at git commits and session boundaries. Reduced input-token usage by 85.5% p50 / 78.8% mean across 69 real context windows, cutting cost and latency on every agent interaction.
- Recognized that standard LLM-as-judge eval frameworks produce pseudoreplication by treating queries as replication units instead of cases. Built GNOMON to fix it: bootstrap CIs bounded to [0,1], gating CI on the lower bound. Published as pip-installable library used in benchmarking across the stack.
- Built GLYPH as the retrieval layer feeding AXON: a knowledge-graph RAG library with graph, vector, and hybrid retrieval behind one port. Ran the benchmark honestly via GNOMON — graph thesis did not hold on code, published the result anyway.
- Validated the full stack in a production Java context with RPG Master AI: multilingual RAG in Java 21 and Spring AI serving EN/PT queries from one model, with ArchUnit enforcing hexagonal boundaries and 14 ADRs documenting every architectural decision.

Senior AI Engineer

Avangrid (Iberdrola Group) | Mar 2025 – Mar 2026 | United States · Remote

- Designed and presented an internal RAG agent in two stages (Python PoC to Spring AI production architecture) for onboarding and incident triage. Evaluated 6 models across 13 queries; calibrated similarity threshold from assumed 0.85 to measured 0.50 via cosine distance profiling.
- Architected a cross-service diagnostic framework in Python analyzing Kibana logs across 400+ microservices, reducing mean-time-to-investigate by 80%.
- Designed automated AWS CloudWatch monitoring cutting critical incident detection from 60 min to under 10 min. Drove SonarQube remediation elevating 10 legacy Java 8 services from grade E to A.

Senior Software Engineer

Yubico | Mar 2025 – Jun 2025 | United States · Remote

- Led PingOne OAuth2/OIDC integration from design to delivery, enabling enterprise customers to authenticate via a new identity provider within a 3-month window.
- Pushed automated test coverage from 40% to 90% via contract testing, integration tests, and OWASP scans. Enforced gates via GitHub Actions with Codecov on every pull request.

Senior Software Engineer

The Estee Lauder Companies | Jan 2025 – Mar 2025 | United States · Remote

- Cut API response times 70% (4s to 1.2s) on a platform processing 1M+ daily requests via PostgreSQL query optimization, N+1 elimination, and Redis caching on ElastiCache.
- Debugged Kafka consumer/producer flows on AWS MSK and resolved 10+ production incidents across 10+ microservices with full root cause documentation.

Senior Full Stack Engineer

TCU — Brazil's Federal Court of Accounts | Jan 2023 – Jan 2025 | Brasilia · Remote

- Reduced data access time from 2 min to 10 sec by consolidating 20+ fragmented SQL sources into a single Elasticsearch index with 10M+ records and zero data loss, unblocking a national-scale platform serving 30,000+ federal auditors.
- Built the platform full-stack: Kotlin backend with DDD and hexagonal architecture, React frontend with 8 pages, saved filters, and composite search, backed by REST APIs covering advanced search, document import/export, and audit operations.

EARLIER EXPERIENCE

CNH Industrial — Full Stack Engineer | 2022–2023 | *Texas, US · Remote* RBAC auth microservices for 60K+ employees across 180 countries. 30% critical defect reduction. Led Java 17 migration to AWS.

Banco do Brasil — Full Stack Engineer | 2020–2022 Quarkus microservices for 1,000+ daily legislative proposals. 60% efficiency gain, 40% CI/CD improvement.

TJPB — Software Engineer | 2019–2020 Java 8 legacy monolith for judicial document management. 40% maintenance efficiency improvement.

OPEN SOURCE PROJECTS

AXON | *Python · MCP · SQLite · Redis · Qdrant · mem0 · Claude API*

Architecture: SQLite as source of truth, Redis as graph cache, Qdrant for vector retrieval. Context captured at git commits and session boundaries via hooks. Task-based model routing (Haiku/Sonnet/Opus) with Redis circuit breaker and Ollama offline fallback. Pydantic v2 domain models throughout. 1157 tests, 22 decision records, TDD-gated release process.

GNOMON | *Python · pytest · Pydantic · Ollama · pip (gnomon-eval)*

Core design: percentile-bootstrap CIs bounded to [0,1] by construction, with CI lower bound as the gate metric rather than point estimate. Replication unit is the case, not the query — fixes the standard pseudoreplication failure mode. Fully offline via Ollama. 77 tests, 8 ADRs.

GLYPH | *Python · tree-sitter · NetworkX · Neo4j · pip (glyph-kg)*

Extraction pipeline: tree-sitter for Python and Java code, custom splitters for documents. NetworkX default graph backend; Neo4j adapter verified by 14 contract tests against real Neo4j 5 on Testcontainers. Port/adapter architecture isolates retrieval strategy from caller. Benchmarked via GNOMON. 281 tests, 98%+ coverage, 7 ADRs.

RPG Master AI | *Java 21 · Spring Boot 3 · Spring AI · LangChain4j · Qdrant · Ollama*

Chunking strategy: 856 segments from D&D 5e PHB with overlap tuned for rule interdependency. bge-m3 multilingual embeddings for cross-lingual retrieval without translation. Qdrant over gRPC (<1ms search). Dual retrieval ports (Spring AI + LangChain4j) behind one interface, boundary enforced by ArchUnit tests. 9.2x ingestion speedup via batch embedding API. 14 ADRs.

EDUCATION

Bachelor's, Computer Science | Faculdade Internacional da Paraiba, Brazil | 2022

CERTIFICATIONS

Scrum Foundation Professional | CertiProf EF SET English C2 | EF SET

LANGUAGES

Portuguese (Native) | English (C2)