

Yosri Khiari

Software Engineer | AI & Full-Stack LLMs, RAG & Distributed Systems

 Kelibia, Nabeul, Tunisia

 +216 22-323-039

 khiariyosri02@gmail.com

 github.com/yosrikhiari

 linkedin.com/in/yosrikhiari

 yosrikhiari.vercel.app

Skills

AI & LLMs

RAG, Self-verifying RAG loops, GraphRAG, Semantic Search, Hybrid Retrieval (BM25 + kNN), Vector Embeddings, Prompt Engineering, AI Agents, Multi-Agent Orchestration (LangGraph), LLM Evaluation, OCR/NER, OpenAI / Anthropic / Gemini / Grok / Ollama

Programming

Python, TypeScript, C#, Go, Java, SQL

Backend & Architecture

ASP.NET Core (.NET 8/10), Spring Boot, Go, Microservices, Event-Driven (RabbitMQ, Kafka), REST, SSE, SignalR

Frontend

Vue 3, React, Angular, Tailwind CSS, Vite, TipTap, Chrome Extensions

Cloud & DevOps

Docker, Kubernetes (k3s, Kustomize, Rancher), Traefik, GitHub Actions → GHCR, Azure, OpenStack, OpenTelemetry, Prometheus, Grafana, Loki, Tempo

Databases

PostgreSQL (RLS), OpenSearch, Qdrant, Redis, MongoDB, pgvector, SQL Server

Certifications

Network Security

Cisco — Apr 2026

AWS Cloud Operations

Amazon Web Services — Nov 2025

Hashgraph Developer

Hashgraph / Hedera — Nov 2025

CCNA: Intro to Networks

Cisco — May 2024

Languages

Arabic: Native

French: B2

English: B2

Interests

Creative: Writing, Reading

Gaming: Video Games, Chess

Software Engineer with 13 months of internships across three companies, specialized in AI/LLM systems and full-stack development. Proven track record building production-grade RAG pipelines, multi-tenant SaaS platforms and ML forecasting systems in Python, TypeScript, C#/.NET and Go. Lead developer of a semantic search engine (6 custom services on a 29-container stack) with hybrid retrieval, GraphRAG and full observability. Autonomous in designing resilient, event-driven architectures from database schema to Kubernetes deployment.

Experience

- Neoledge MEA

Software Engineer Intern (PFE)

Feb 2026 – Aug 2026

 - Semantic Search Engine:** Lead developer of a **6-service platform on a 29-container stack** combining **RAG** and **GraphRAG** — hybrid BM25 + kNN vector retrieval (OpenSearch HNSW, 1024-dim embeddings) fused by Reciprocal Rank Fusion, with ACL-filtered chunk retrieval, reranking and grounded answer generation with citations.
 - Document Ingestion Pipeline:** Async **Apache Tika** → Tesseract OCR → **GLiNER NER** → format-aware chunking → OpenSearch indexing over **RabbitMQ** with transactional outbox, idempotent uploads and dead-letter retry.
 - NLP Query Parser:** Multilingual parser covering **300+ patterns** in French, English and Arabic at sub-millisecond latency, enabling natural-language filtering by date, author, amount and type.
 - Observability & Evaluation:** **OpenTelemetry** with Prometheus, Grafana, Loki, Tempo — 7 dashboards, distributed traces, and an **867-case evaluation harness** in CI gating merges on quality thresholds.
- HorizonLux

Full-Stack Developer Intern

Sep 2025 – Jan 2026

 - Expense Management:** Shipped end to end — approval workflows, **receipt OCR** with automatic vendor creation, FODEC withholding, split payments, bulk operations and accountant handoff queues in a multi-tenant ERP (.NET 10, React, PostgreSQL RLS).
 - Tax Compliance Engine:** Tunisian TEJ withholding-declaration exports with rectificative support, FODEC calculations, VAT rules and a fiscal-obligation calendar, replacing manual spreadsheet filings.
 - Multi-Tenant Isolation:** Proved **5-layer tenant isolation** (PostgreSQL RLS, fail-closed EF Core filters) in CI, making cross-tenant data leakage a build failure. Managed **Kubernetes** (k3s, Kustomize, Rancher) deployments across dev/staging/prod.
 - Coppler (Self-verifying RAG):** LinkedIn content-repurposing **Chrome extension** + .NET modules — Grok / OpenAI / Anthropic / Gemini adapters with fallback, few-shot retrieval from an embedding store, cosine-similarity dedup with regeneration, LinkedIn OAuth publishing on Hangfire.

- L-Mobile

(2 internships)

ML / Full-Stack Developer Intern

Jul 2024 – Sep 2025

 - ML Forecasting:** Built a hybrid **Prophet + XGBoost** system for SKU-level demand planning with Bayesian hyperparameter tuning, **Redis** caching and **Prometheus/Grafana** monitoring dashboards, deployed via Azure Container Registry.
 - Warehouse Routing:** Developed a forklift routing app (**Angular, Spring Boot, MySQL**) in an Agile team, improving logistics efficiency by **25%**.

Projects

- AgentOps — Self-Hosted LLM Operations Gateway

github.com/yosrikhiari/AgentOps

Go, PostgreSQL 16 + pgvector, Prometheus / Grafana, k6, MCP, Docker, GitHub Actions

 - LLM Gateway:** OpenAI-compatible chat API with SSE streaming, prompt classification routing, cloud fallback, virtual API keys with rate limits — **0/2,725 requests failed** at 50 RPS (p99 = 70ms).
 - Evaluation & Drift:** Claim-level faithfulness scoring via pgvector retrieval and LLM-as-judge — scores **0.972** on a 72-pair golden set with scheduled drift detection.
- Versatile — Offline-First AI Writing Platform

github.com/yosrikhiari/Versatile

Vue 3, Pinia, TipTap, Dexie/IndexedDB, LangGraph.js, .NET 10, Ollama / OpenAI / Anthropic / Gemini / Groq

 - Multi-Agent Writing on LangGraph:** Writer, Critic and Editor as separate agents on GPU / CPU lanes (the judge is never the author), checkpointed per step, every call traceable through an AgentOps gateway — **29/30 scenes** cleared on a full 10-chapter run on an 8 GB consumer GPU.
 - Offline-First:** Dexie/IndexedDB persistence with sync to optional .NET 10 API; provider-agnostic AI layer; **~3,100 tests** (Vitest + Playwright); exports to DOCX, EPUB and PDF.

Education

Engineering Degree in Computer Science — ESPRIT, Tunis

2023 – 2026

Preparatory Diploma in Maths & Physics — ISSATMA, Mahdia

2021 – 2023