

Farouk Ben Fraj

AI Engineer & Tech Lead, LLM, RAG and evaluation in production

Paris, France · +33 6 65 01 78 95 · farouk.ben-fraj@frajtech.com
fbenfraj.frajtech.com · linkedin.com/in/farouk-benfraj · github.com/fbenfraj · Freelance, available · French (native), English (fluent)

PROFILE

Software engineer, **8 years of experience**, specialised in putting LLM systems into production inside existing enterprise stacks. I design **RAG** pipelines together with the **evaluation harnesses** that go with them: a written definition of "correct", a measured baseline, a regression suite, rather than demos. **Tech Lead** on mission-critical systems (RATP: 2023 Rugby World Cup, 2024 Paris Olympics, zero production incidents) and on the rebuild of a legacy platform into a **multi-tenant AI SaaS**. Led teams of 3 to 6 developers.

TECHNICAL SKILLS

AI & LLM	RAG (Retrieval-Augmented Generation), OpenAI, Anthropic Claude, LangChain, prompt engineering, embeddings, vector search, reranking, function calling, LLM-as-a-judge, guardrails and controlled refusal
Evaluation	Evaluation harnesses, annotated gold sets, recall@k, MRR, citation F1, inter-annotator agreement (Cohen's kappa), regression suites, observability, cost and latency control
Languages & backend	TypeScript, JavaScript (ES6+), Python, SQL, Solidity ; Node.js, Express, NestJS, Fastify, REST APIs, GraphQL, microservices, Prisma, Sequelize, RabbitMQ
Frontend & data	React, Next.js, Electron, Tailwind CSS, design systems ; PostgreSQL, pgvector, MongoDB, Redis, schema design and migration
Cloud, DevOps & testing	AWS (EC2, S3, RDS, Lambda, SQS), Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Datadog ; Jest, Vitest, Cypress, TDD, ESLint
Architecture & leadership	Multi-tenant SaaS, modular architecture, CTO as a Service, product scoping, mentoring, Agile (Scrum, Kanban), GDPR

PROFESSIONAL EXPERIENCE

AI Tech Lead, FrajTech (freelance) April 2022 – present
Paris / Île-de-France, AI, RAG and technical leadership engagements

- Design and production rollout of **RAG systems** for B2B clients: corpus construction, vector search, reranking, guardrails and controlled refusal when the answer is not in the corpus.
- Versioned **evaluation harnesses** (annotated question sets, retrieval and generation metrics, regression suites) to tell an improvement apart from a coincidence.
- **CTO as a Service** engagements: product scoping, architecture decisions, roadmap arbitration, cloud cost optimisation and handover to internal teams.
- End-to-end delivery of MVP and SaaS projects (React, Node.js, PostgreSQL, AWS Lambda, Docker, Kubernetes), with Datadog monitoring and a 99.9% availability commitment.

Tech Lead, Staffdesk March 2024 – present
Paris, AI-augmented HR SaaS platform

- Legacy platform rebuilt into a **multi-tenant SaaS** on AWS (EC2, S3, RDS) ; **+30% MRR**.
- **LLM (OpenAI)** integration for automated HR analysis generation: **–60% time spent on reporting** on the client side.
- Node.js / Express microservices and normalised KPI computation in PostgreSQL, restoring **100% data consistency** across modules and across clients.
- CI/CD set up (GitHub Actions) and test coverage above **85%** (Jest).
- **Led 3 developers**: code reviews, quality standards and Agile ceremonies.

Fullstack Tech Lead, RATP Group June 2023 – January 2024
Paris, event management system, digital factory

- Tech Lead of a team of **5 to 6 developers** on the event system supporting the **2023 Rugby World Cup** and the **2024 Paris Olympics: zero production incidents**.
- Modular backend architecture (Node.js, Sequelize, PostgreSQL) supporting **over 1,000 concurrent users**.
- Platform migration to AWS (S3, SQS, RDS) and automated deployments on Kubernetes.
- End-to-end (Cypress) and unit (Jest) tests at **90% coverage** ; Datadog monitoring and **–60% incident response time**.
- Worked with the UX/UI team and the Product Owner ; **–15% churn** and rising user adoption.

Frontend Tech Lead, **Raisponse**

November 2022 – October 2023

Paris, LLM-powered real-time writing assistant

- Electron desktop application and Chrome extension built on the **OpenAI API** for real-time reply suggestions: **–50% drafting time**.
- Modular React design system (hooks, Context API) ensuring a consistent experience across platforms.
- **GDPR** compliance and AES encryption of user data.
- Frontend performance work (code-splitting, lazy-loading): **FCP under 1.5 s**.

Fullstack Web3 Developer, **Gateway**

May 2022 – November 2022

New York (remote), certification platform for creators

- POC then MVP of a Next.js / TypeScript dApp with Solidity smart contracts ; adopted by **more than 500 early users**.
- Node.js, NestJS, Hasura, Prisma and PostgreSQL backend on AWS ; real-time notifications through Redis and SQS (99.5% SLA).

Backend Developer, **Trusk**

October 2020 – January 2022

Saint-Ouen, on-demand logistics

- REST microservices in Express / TypeScript orchestrated with Docker Compose, with RabbitMQ queues and real-time data streaming.
- AWS deployments automated with Terraform: **–70% provisioning time**.

Fullstack Developer, **Capgemini / Sogeti**

December 2018 – September 2020

Paris, enterprise engineering and innovation

- Design of an **NLP search engine in Python** for Institut Pasteur, queryable in natural language.
- SEO-optimised static sites (GatsbyJS, React) powered by a headless CMS.

Systems Administrator, **Ministère de l'Intérieur**

December 2017 – August 2018

Paris, security-cleared work

- Active Directory, GPO, GLPI and Exchange administration for **over 2,000 users**.
- Definition and enforcement of cybersecurity policies (MFA, encryption): **–25% vulnerabilities**.

AI PROJECTS

Le Bon Article, **measured RAG on French labour law**

2025

Solo project · Corpus, retrieval, evaluation harness, LLM judge, public report

- RAG system answering labour-law questions from retrieved articles, and **refusing to answer** when the corpus does not hold the answer. Corpus of **3,030 articles**, gold set of **142 questions**.
- **Six named configurations** compared, every result file versioned: no score can be quietly revised.
- Recall@5 raised from **84.2% to 90.8%**, MRR from **0.572 to 0.823**, citation F1 from **58.7% to 69.7%** ; repealed law in the top 5 brought down from **66.7% to 0.0%** of queries.
- Refusal accuracy of **100%**, always published beside its false-refusal rate (8.3%) ; answer correctness published as a range (44.2%–80.0%) rather than a single figure the evidence does not support. Judge/human agreement measured (**Cohen's kappa 0.489**), cost of **0.0045 EUR per query**.

Elyra, **visual identity generation pipeline**

2025

Sole engineer · Generative pipeline, evaluation loop, cost control

- Pipeline turning a written brief and a Spotify profile into cover art calibrated for Spotify, Apple Music and social networks, with no designer in the loop.
- **Vision model as judge**: every candidate scored on a weighted five-criteria rubric ; below the 0.7 threshold the pipeline **rewrites its own prompt** and re-runs generation, up to three times, then falls back on the best batch rather than failing.
- Multi-provider routing (OpenAI, Anthropic, fal.ai) with health checks and a circuit breaker ; spend caps verified *before* every billable call. **56 test files**, including a suite that checks the observability itself.

Keytt, **AI platform for artists**

2024, Co-founder & CTO

Architecture, backend, frontend and AI integration

- React / Node.js / PostgreSQL platform turning an artist's audience data into actionable recommendations, with an LLM-powered content planning assistant.
- **Complete MVP delivered in under 6 weeks** ; containerised AWS infrastructure, modular architecture ready to scale.

EDUCATION & LANGUAGES

Master of Science, Software Engineering, EPITECH, Paris

Bachelor of Science, Computer Science, Université de Paris

French: native speaker · English: fluent in a professional setting (remote engagement for a New York company)