

ENZO TRIPOLI

Senior Backend / Full-Stack Engineer - Tech Lead

Python • Go • PostgreSQL • React • gRPC • IA appliquée & LLM

Marignier, Haute-Savoie, France - à 25 min de Genève

Permis G (frontalier) - valide

+33 7 82 95 46 65 - enzotripoli21@gmail.com

Français (langue maternelle) | Anglais (professionnel)



PROFIL

Ingénieur Backend / Full-Stack Senior avec 9 ans d'expérience dans la conception, le développement et l'exploitation de systèmes à haute performance en production, avec une spécialisation sur l'intégration de LLM et les architectures RAG. Solide expertise **Python, Go et PostgreSQL**, avec un fort accent sur l'**optimisation des performances et des coûts**, les architectures scalables et les décisions techniques guidées par les enjeux métier. Tech Lead hands-on : décisions d'architecture, revues de code et accompagnement des développeurs, tout en restant quotidiennement dans le code. À l'aise de bout en bout, du cadrage jusqu'à l'exploitation en production.

Postes ciblés : Senior Backend Engineer • Senior Full-Stack Engineer • Lead Developer / Tech Lead hands-on • AI Engineer

COMPÉTENCES CLÉS

Backend & APIs - Python (profiling, optimisation CPU & mémoire) • Golang (concurrency, goroutines, channels) • FastAPI • APIs REST (conception, versioning, sécurité) • gRPC & Protobuf • Fondamentaux des systèmes distribués (timeouts, retries, idempotence, résilience)

IA appliquée & LLM - RAG (chunking, embeddings, recherche vectorielle pgvector, recherche hybride, reranking) • Intégration LLM (sorties structurées, function calling, prompt engineering, streaming) • Optimisation des coûts d'inférence et de la latence • Architectures hybrides réduisant la dépendance au LLM • Fiabilité (garde-fous, retry, fallback de modèle) • Évaluation (golden sets, LLM-as-judge) • Systèmes d'agents multi-tenant

Bases de données & Data - PostgreSQL avancé : optimisation de requêtes, stratégies d'indexation, plans d'exécution, vues matérialisées, triggers & procédures stockées, migrations • pgvector • Elasticsearch • Redis • Modélisation orientée performance

Frontend - React.js & React Native • Context API et patterns de gestion d'état • Architecture propre et séparation des responsabilités • Ownership full-stack avec approche backend-first

Cloud & Infrastructure - GCP : Cloud Run, Pub/Sub & Cloud Tasks, BigQuery, Cloud Logging & Monitoring, Cloud Vision • AWS : S3 • Infrastructure Linux bare-metal et VPS

DevOps & Production - Docker (multi-stage builds) • Kubernetes • Pipelines CI/CD • Observabilité (logs, métriques, debug en production) • RabbitMQ • SaltStack & Ansible • Fiabilité et mise en production

Méthodes & Collaboration - Workflows Git (PR, code reviews, stratégies de branches) • SemVer et gestion des releases • Conventional Commits • Cadrage technique à partir des besoins métier • Mentorat

EXPÉRIENCES PROFESSIONNELLES

ET Technology Consulting (SASU) - Tech Lead - Backend & AI

2024 - présent

Structure de conseil technique portant deux produits SaaS B2B en production : **NovoAgent** (agents IA conversationnels pour agences immobilières) et **TableAgent** (hôtellerie / restauration). Architecture multi-tenant, contraintes fortes de fiabilité, de conformité RGPD et de coût d'inférence.

- Conception et développement d'une **plateforme SaaS multi-tenant** d'agents IA conversationnels, de l'architecture à l'exploitation en production.
- Mise en place d'un **pipeline RAG complet** : ingestion documentaire multimodale, chunking, embeddings, recherche vectorielle sur PostgreSQL / pgvector, reranking et évaluation de la qualité de récupération.
- Conception d'une **architecture hybride réduisant la dépendance au LLM** : routage déterministe, cache sémantique, machine à états métier et appel au modèle réservé aux cas qui le justifient réellement.
- Intégration de LLM en production avec sorties structurées, garde-fous, stratégies de retry, fallback de modèle et suivi du coût par conversation.
- Développement backend **Python / FastAPI** : APIs REST, traitements asynchrones, isolation multi-tenant au niveau des données.
- Développement d'une **application mobile React Native** embarquant des fonctions IA non conversationnelles : speech-to-text, extraction et structuration de données à partir de la voix ou du texte, création automatique de tâches et de rappels, suggestions de mise à jour de profils et de fiches.
- Intégration de la **WhatsApp Cloud API** en tant que Tech Provider Meta : embedded signup, gestion multi-WABA, templates, webhooks idempotents, messages interactifs.
- Conception d'un **framework de test et de simulation d'agents** (scénarios versionnés, LLM-as-judge) évaluant fidélité, conformité et robustesse avant chaque mise en production.
- Mise en conformité **RGPD et EU AI Act (article 50)** : consentement opt-in, transparence IA, DPA article 28.
- Définition des standards de code, revues systématiques et cadrage technique pour une équipe de **10 développeurs** internes et externes.

Impacts clés

- **Coûts d'inférence divisés par 50** entre la première version et l'architecture optimisée, à qualité de réponse constante.
- Déploiements en production chez une **vingtaine de clients** (agences immobilières haut de gamme, France et Suisse romande).
- Automatisation du traitement des demandes entrantes sur WhatsApp, email et SMS, avec escalade vers l'humain sur les cas hors périmètre.
- Réduction significative du délai de première réponse et de la charge opérationnelle des équipes clientes.

Reveeld - Senior Backend / Full-Stack Engineer - Tech Lead

2024 - juin 2026

Application de rencontre comptant plus de 80k utilisateurs actifs, sur une architecture cloud-native GCP. Fortes contraintes de scalabilité, de performance, de sécurité et de fiabilité en environnement grand public.

- Conception et mise en œuvre d'une **architecture cloud-native orientée functions** sur GCP, plutôt qu'une API monolithique classique.
- Développement de services backend en **Python et Go** et d'**APIs REST et gRPC** pour la communication inter-services, sur des projets conçus from scratch.
- Arbitrage des choix d'architecture en fonction des performances attendues, des compétences de l'équipe et du time-to-market.
- Développement d'un **backoffice complet en React.js** (modération, opérations, support, analytics) et contribution au frontend React Native.
- Optimisation **PostgreSQL** : requêtes, schémas, stratégies d'indexation avancées et vues matérialisées.
- Mise en place d'une brique data/analytics via BigQuery et de workflows IA (Cloud Vision) pour la modération de contenu.
- Absorption des pics de trafic via traitements asynchrones (Pub/Sub, Cloud Tasks) et scaling adaptatif de Cloud Run.

Impacts clés

- Backend stable à **80k+ utilisateurs actifs**, avec anticipation des pics de charge sans explosion des coûts d'infrastructure.
- Throughput de la modération amélioré de **60-70%** via automatisation et workflows assistés par IA.
- Efficacité des campagnes de push notifications (delivery / CTR) augmentée de **30-40%**.

Yacast - Senior Backend Engineer → Tech Lead

2019 - 2024

Environnement data intensif : catalogues musicaux, reconnaissance de contenus TV/Radio et reporting automatisé pour des clients entreprise.

- Conception et livraison de projets backend from scratch, avec arbitrage entre **APIs REST et architectures gRPC** selon les besoins de performance et d'intégration.
- Pilotage technique de la **refonte progressive d'un monolithe vers des microservices**.
- Conception et optimisation de **pipelines ETL traitant des millions de lignes** issues de catalogues musicaux et de flux broadcast.
- Développement de services de **reconnaissance musicale TV et Radio** fonctionnant en continu (24/7) en environnement critique.
- Mise en place de **queues distribuées RabbitMQ** pour orchestrer les traitements asynchrones et parallèles.
- Optimisation avancée de **PostgreSQL** (schémas, requêtes, vues matérialisées), **Elasticsearch** pour la recherche concurrente et caches **Redis**.
- Exploitation d'une ferme d'environ **50 serveurs physiques** (≈180 cœurs virtuels chacun) et automatisation via SaltStack et Ansible.

Impacts clés

- **Latence API réduite de 30-40%**, certaines requêtes SQL critiques améliorées d'un **ordre de grandeur** après indexation et analyse des plans d'exécution.
- Services critiques de reconnaissance musicale maintenus **24/7** sur une infrastructure on-premise à grande échelle.
- Livraison fiable de rapports automatisés pour plusieurs clients entreprise.

PROJETS PERSONNELS

IpOp Tools (Suisse) - Application web full-stack (Python, PostgreSQL, React) conçue pour un client suisse : architecture backend, couches API, optimisation base de données, déploiement et exploitation. Intègre une brique IA embarquée (RAG sur la base documentaire, assistant guidant l'utilisateur dans l'expression et la structuration de son besoin).

FORMATION

Master 2 (Bac+5) - Informatique, spécialisation Machine Learning, Intelligence Artificielle et Data Science (2019)

ESIEA - École Supérieure d'Informatique, Électronique et Automatique, Paris

Références disponibles sur demande.