

SUAN TAY

Architecte GenAI & Lead Dev Python — Freelance

Agents IA (LangGraph) • RAG hybrides • Voix IA temps réel • MCP • LLMops • Docker • CI/CD

Clermont-Ferrand • Full remote ou 1 j/sem. Paris / Lyon • suan.tay@iafluence.fr • +33 6 65 19 76 33

[linkedin.com/in/suan-tay](#) • [github.com/suaniafluence](#) • Français • Espagnol bilingue • Anglais professionnel (B2 certifié)

PROFIL

Ingénieur informatique diplômé CTI, 17 ans de développement logiciel dont 6 en IA. Je conçois et mets en production des systèmes d'IA générative en Python : agents multi-étapes, RAG hybrides, voix temps réel et intégrations LLM sécurisées — du cadrage au déploiement. Développement augmenté par l'IA (Claude Code, Cursor) encadré par tests, revue de code et CI. Double culture technique et opérationnelle (ex-COO), habitué à arbitrer coût / latence / qualité.

COMPÉTENCES

GenAI & agents : LangGraph, LangChain, CrewAI, tool calling, mémoire, routage conditionnel, human-in-the-loop, prompt & context engineering, OpenAI, Anthropic, Ollama

RAG & recherche : Recherche hybride dense + BM25, chunking, embeddings (sentence-transformers), ChromaDB, Qdrant, FAISS, pgvector, filtrage explicable, déduplication

Backend : Python typé et asynchrone, FastAPI, Flask, Pydantic, REST, WebSockets, SSE, MCP, SQLAlchemy, PostgreSQL, SQLite, Redis

DevOps & qualité : Docker, Docker Compose, GitHub Actions (CI/CD, CodeQL, Dependabot), pytest, Cypress E2E, Nginx, systemd, Vercel, AWS, Azure

Sécurité : OAuth2 / JWT, HMAC, chiffrement Fernet, gestion des secrets, filtrage PII, rate limiting, RGPD, AI Act

LLMops : Estimation des coûts avant exécution, routage multi-modèles, workers asynchrones, checkpoint / reprise, observabilité

EXPÉRIENCE PROFESSIONNELLE

Architecte GenAI & Lead Dev Python | IAfluence (freelance)

Depuis mars 2024 • Remote

- Agents IA** : plateforme commerciale multi-agents LangGraph (classification, offre, négociation, supervision, CRM), API FastAPI avec OAuth2 Google, SQLite + Redis/Qdrant, Docker, CI/CD GitHub Actions avec CodeQL.
- Voix IA temps réel** : assistant vocal à audio bidirectionnel (OpenAI Realtime, WebSockets, Whisper, interruption vocale), tests pytest + Cypress E2E ; passerelle téléphonique Asterisk ARI/PJSIP.
- RAG full-stack** : EchoChat — scraping Playwright, indexation ChromaDB, réponses Claude ; FastAPI + Next.js, Docker Compose, déployé en production.
- Sécurité des LLM** : Warden, produit de médiation sécurisée entre agents et services (filtrage PII, rate limiting, OAuth2, Fernet, HMAC, connecteurs MCP/REST).

Architecte GenAI / Ingénieur IA — mission | QHSE Concept

Févr. – mai 2026 • Remote

- Conception et industrialisation de **Precision Pipeline**, moteur RAG réglementaire : fiche entreprise → référentiel applicable en 4 étapes, recherche hybride dense + BM25, filtrage explicable, déduplication, alertes.
- Indexation asynchrone par workers Redis parallèles avec checkpoint et reprise automatique après incident.
- Maîtrise des coûts : dry-run, ventilation par étape, routage multi-modèles — **coûts d'inférence divisés jusqu'à 5 à 10x** en mode économique.

Ingénieur IA & COO | ScreenACT

Mars 2022 – avr. 2024 • Riom

- Stratégie Data & IA ; développement et déploiement de solutions ML et LLM (Python, Databricks) pour l'inclusion de patients en essais cliniques.
- Pilotage d'une équipe pluridisciplinaire IT, Data Science et RH ; priorisation technique et mise en production.

Ingénieur développement décisionnel | Axege

Févr. 2019 – mars 2020 • Clermont-Fd

- Automatisation d'un système de reporting pour le suivi de patients.

Ingénieur logiciel | Accenture

Janv. – déc. 2018 • Clermont-Fd

- Logiciel de gestion de projet en environnement Agile ; revues de code et bonnes pratiques.

Développeur logiciel | Goodyear Dunlop Tires

Oct. 2008 – déc. 2017 • Riom

- Près de 10 ans sur une plateforme de gestion des ressources en environnement industriel automobile (Java, Python, SQL, UML, tests).

Enseignant vacataire en IA | ESC Clermont • Université Clermont Auvergne

Depuis 2024

- Cours d'IA générative, encadrement de projets M1/M2, jurys du titre professionnel CDA.

PROJETS OPEN SOURCE

AgenticSellerPOC — multi-agents LangGraph, FastAPI, CI/CD, CodeQL • [github.com/suaniafluence/AgenticSellerPOC](#)

voix_realtime — voix temps réel OpenAI Realtime, tests Cypress E2E • [github.com/suaniafluence/voix_realtime](#)

EchoChat — RAG full-stack, Docker Compose • [github.com/suaniafluence/EchoChat](#) • démo : [echochat.iafluence.fr](#)

FORMATION

Diplôme d'ingénieur en informatique (CTI) | CNAM Paris

2007 – 2017

Ingénieur IA / Data Scientist — titre RNCP niveau 7 (Bac+5) | CentraleSupélec × OpenClassrooms

2020 – 2021

MBA — stratégie, management, transformation numérique | ESC Clermont Business School

2023 – 2024