

Robin Lasserre

Architecte IA agentique · CTO

robin.lasserre@elytre.dev 06 01 64 64 98 elytre.dev

Croix (59170) — Hauts-de-France [linkedin.com/in/robin-lasserre](https://www.linkedin.com/in/robin-lasserre) github.com/RobinLasserre

Je structure les équipes de développement autour de la migration vers l'IA agentique, avec des architectures hybrides IA + repli humain et un cas industriel medtech démontré. Directeur technique en prestation d'avril 2024 à août 2026 d'un dispositif médical numérique, sur une trajectoire de marquage CE.

Dossier de compétences

Architecte IA agentique · CTO — réalisations, technologies et indicateurs mesurés

Document établi le 16 septembre 2026. Il détaille, pour chaque expérience et chaque projet, ce qui a été fait, avec quelles technologies et avec quels indicateurs mesurés. Les projets personnels y figurent au même titre que les missions client, et leur statut est indiqué à chaque fois : ce qui tourne en production, ce qui est un prototype, et ce qui reste expérimental.

CHAMPS DE COMPÉTENCE COUVERTS

Python avancé, FastAPI, services IA · Python et FastAPI sur toute la chaîne : microservice de traitement 3D, API de revue humaine, serveurs MCP, outils de collecte et de qualification de données. Tests systématiques avec pytest, TDD rétroactif sur un existant de plus de 400 000 lignes.

Agentic AI, multi-agents, orchestration · C'est le cœur du métier exercé aujourd'hui : conception d'architectures d'agents, définition des rôles, coordination, reprise sur erreur, et surtout gouvernance des accès aux outils et aux données par serveur MCP. Workflows hybrides mis en production dans un contexte de dispositif médical, où l'agent propose et le repli humain décide.

LLM et orchestration, appel d'outils · Intégration de modèles propriétaires et de modèles ouverts exécutés localement. L'appel d'outils est exactement ce qu'un serveur MCP formalise : déclaration des outils, contrôle de ce que l'agent a le droit d'appeler, journal de ce qu'il a appelé.

RAG et recherche documentaire · Ingestion et classification documentaire à l'échelle en production sur le collecteur d'opportunités, qui est la première moitié d'une chaîne RAG. Architecture RAG d'entreprise conçue et documentée de bout en bout — magasin vectoriel et palier de bascule, embeddings, reranker, jeu d'évaluation de la recherche, citations obligatoires, coût par palier. La mise en production de la recherche elle-même reste à faire, et se dit telle quelle.

Évaluation des systèmes IA · Chaînes d'évaluation construites et maintenues : sur l'assistant e-mail, mesure des catégories, de l'urgence et de l'abstention, contre une référence déterministe et un corpus réservé. La mesure de non-régression est une pratique constante, pas un outil.

Industrialisation, MLOps et LLMops · Conteneurisation Docker et chaînes d'intégration continue sur tous les projets ; services exposés en API, composants réutilisables, versionnage et reprise après interruption. Services de production tenus sous systemd, avec relance automatique.

Intégration au SI, sécurité et traçabilité · Intégrations à des systèmes tiers en production : transport, relation client, outil de discussion d'équipe, logiciel métier. Séparation des droits, gestion des secrets, journalisation et traçabilité — le connecteur WhatsApp est un cas extrême assumé : base ouverte en lecture seule, aucune bibliothèque d'écriture chargée, périmètre d'accès couvert par des tests.

Machine learning et données · Segmentation par réseaux de neurones sur maillages 3D dans le pipeline orthodontique — PyTorch et scikit-learn, 17 169 lignes, 173 commits sur six mois. Pipelines d'ingestion et de réconciliation de six sources en production nocturne, et plus de trois ans de conception des indicateurs de pilotage d'une entreprise : finance, production, opérations, commercial.

Conformité et IA responsable · Audit EU AI Act et RGPD intégré au cadrage de migration, anonymisation des données, gouvernance des accès. Expérience d'un produit sous trajectoire de marquage CE, où la traçabilité d'une décision automatique n'est pas une option.

Cycle POC → MVP → production · Parcouru entièrement et plusieurs fois : le visualiseur 3D est passé du prototype au cœur du flux de production et y est depuis près de quatre ans ; le collecteur d'opportunités tourne en continu ; le pipeline orthodontique est resté au stade expérimental, et c'est dit ainsi.

EXPÉRIENCE PROFESSIONNELLE, EN DÉTAIL

Aligneurs Français — direction technique d'un dispositif médical numérique

Directeur technique en prestation · avril 2024 – août 2026 · Paris, à distance

Éditeur d'un dispositif médical numérique de traitement orthodontique par aligneurs sur mesure, engagé dans une trajectoire de marquage CE. Équipe de cinq développeurs et plus, produit en production continue chez des praticiens et des techniciens.

- Direction technique et tech lead : architecture, revue de code, arbitrages de dette technique, recrutement technique et conduite du changement.
- Refactoring assisté par IA agentique d'un back-end de plus de 400 000 lignes, avec TDD rétroactif et tests de bout en bout générés puis revus.
- Workflows agentiques hybrides en production : l'agent propose, le repli humain décide. Le choix du repli humain est une contrainte du domaine, pas une précaution de principe.
- Pipelines d'ingestion de facturation et de support : six sources hétérogènes réconciliées chaque nuit, en production continue.
- Conception des tableaux de bord et production des indicateurs de l'entreprise pendant plus de trois ans : finance, production, opérations, suivi par commercial.
- Automatisation logistique : regroupement des commandes par destinataire, génération d'étiquettes de transport et restitution dans l'outil de discussion de l'équipe.

Indicateurs · Back-end de plus de 400 000 lignes repris. • 38 % de frais de transport supprimés, sur 1 432 expéditions traitées. • Six sources de données réconciliées chaque nuit. • Équipe de cinq développeurs et plus.

Technologies · Python, Django, PostgreSQL, Redis, pytest, tests de bout en bout, Docker, CI/CD, Metabase, HubSpot, API FedEx, Slack, Three.js, WebGL, React, TypeScript

Elytre (EURL) — conseil en architecture IA agentique

Gérant, associé à 100 % · depuis mai 2025 · Croix (59), SIREN 945 053 023

Structure de conseil indépendante. Cadrage de migration vers l'IA agentique, conception d'architectures d'agents, accompagnement et formation d'équipes de développement.

- Cadrage de migration vers l'IA agentique, avec audit EU AI Act et RGPD intégré à la conception plutôt qu'ajouté à la fin.
- Formation et accompagnement d'équipes de développement à Claude Code et au protocole MCP : gouvernance des accès aux outils et aux données, garde-fous, mesure de non-régression.
- Serveurs MCP sur mesure : exposition contrôlée d'outils et de sources de données à un agent, avec séparation stricte entre lecture et écriture.
- Conception d'architecture RAG documentée pour une fabrique IA d'entreprise : choix du magasin vectoriel avec palier de bascule, modèle d'embeddings, reranker, jeu d'évaluation de la recherche, citations obligatoires et coût par palier. Document de conception, argumenté et chiffré.

Technologies · Python, FastAPI, MCP (Model Context Protocol), Claude Code, LLM locaux, orchestration multi-agents, évaluation et garde-fous, EU AI Act, RGPD, Docker

Activité indépendante — ingénieur développeur web 2D et 3D

Entreprise individuelle · août 2022 – septembre 2025 · Lille

Trois ans d'exercice indépendant en développement web et 3D temps réel, dont la conception du visualiseur qui est devenu le cœur du flux de production d'Aligneurs Français.

- Conception en autonomie complète du visualiseur 3D orthodontique à partir de 2022.
- Applications web et 3D temps réel pour plusieurs clients.

Technologies · React.js, Next.js, TypeScript, Three.js, WebGL, Python, Django

RÉALISATIONS EN MISSION CLIENT

Questions métier en langage naturel depuis l'outil de discussion

Réalisation client documentée · medtech

Commande de question-réponse adossée à un modèle de langage et à la base métier, appelée depuis l'outil de discussion de l'équipe. Le modèle reçoit la question en langue naturelle, produit l'interrogation de la base, et la réponse revient dans le fil. C'est de la traduction du langage naturel vers une requête, et non de la recherche augmentée par documents.

Technologies · Claude, Slack, intégration LLM, base métier

Modernisation du back-end medtech

Mission client · mandat de direction technique 2024-2026

Refactoring assisté par IA d'un back-end de plus de 400 000 lignes, avec TDD rétroactif et tests de bout en bout. Architecture et conduite du changement au sein de l'équipe.

Indicateurs · Plus de 400 000 lignes reprises.

Technologies · Python, Django, PostgreSQL, Redis, pytest, tests de bout en bout

Pipelines de facturation et de support

Mission client · medtech · en production continue

Réconciliation nocturne de six sources hétérogènes pour fiabiliser les données de pilotage : ingestion, normalisation, rapprochement, détection des écarts.

Indicateurs · Six sources réconciliées chaque nuit.

Technologies · Python, ingestion et réconciliation de données, ordonnancement

Expéditions automatisées

Mission client · medtech

Regroupement des commandes par destinataire, génération des étiquettes de transport et restitution dans l'outil de discussion de l'équipe.

Indicateurs · 38 % de frais de transport supprimés, sur 1 432 expéditions.

Technologies · API FedEx, Slack, intégrations et traçabilité

Tableaux de bord et indicateurs de pilotage

Mission client · pilotage de l'entreprise · plus de trois ans

Conception des tableaux de bord et production des indicateurs : finance, production, opérations et suivi par commercial. Alimentation par des robots depuis l'outil de relation client, centralisation et restitution.

Technologies · Metabase, HubSpot, indicateurs, tableaux de bord

PureControl · visualiseur 3D orthodontique

Mission client · en usage depuis près de quatre ans

Visualisation et édition des plans de traitement : animation des dents étape par étape, translations et rotations, IPR, taquets, comparaisons. Contacts occlusaux, simulation de gencive, affichage CBCT ; diffusion en flux des modèles et scripts de préparation et de contrôle des maillages. Échange de setups entre techniciens et praticiens.

Technologies · React, TypeScript, Three.js, WebGL, Python, Django, glTF/STL

Pipeline orthodontique

Développement métier · génération automatique expérimentale

Microservice Python : scans STL et prescription en entrée, segmentation dentaire, puis préparation et export d'un paquet exploitable par le logiciel métier. API de lancement de tâches, suivi d'avancement et récupération des résultats ; contrôles de maillages et rapports qualité. L'export est techniquement intégré ; la validation clinique de la génération automatique n'est pas établie.

Indicateurs · 17 169 lignes de code, 57 modules • 173 commits entre le 16 mars et le 29 août 2026 • segmentation dentaire par réseaux sur maillages 3D, avec affinage des frontières, classification des dents et diagnostic de segmentation

Technologies · Python, FastAPI, PyTorch, scikit-learn, NumPy/SciPy, trimesh, uvicorn, pytest, CLI

PROJETS PERSONNELS

Projets menés hors mission, sur fonds et temps propres. Les chiffres qui suivent ont été relevés sur les dépôts le 16 septembre 2026.

Assistant e-mail — tri d'une boîte fonctionnelle par modèle local, avec chaîne d'évaluation

Projet personnel · septembre 2026 · en cours

Service qui lit en continu une boîte fonctionnelle, applique des règles écrites avec le client, pose des étiquettes, signale l'urgent, résume, prépare des brouillons, et range dans « à trier » tout ce dont il doute. Il n'envoie jamais rien. Le graphe de décision est orchestré avec LangGraph et un point de reprise SQLite ; les sorties du modèle sont validées par schéma. Trois évaluateurs mesurent la catégorie, l'urgence et le passage en « à trier » contre une vérité terrain, et produisent un rapport daté. Le traçage externe est câblé mais désactivé par défaut : les messages d'un client ne partent pas chez un tiers pour faire joli dans un tableau de bord.

Indicateurs · 16 651 lignes de code • 32 commits entre le 7 et le 9 septembre 2026 • 5 services conteneurisés (tableau de bord, moteur, inférence locale, serveur de courrier de test, webmail) • vérité terrain de 120 messages répartis en 9 catégories

Technologies · Python 3.12, LangGraph, LangChain, Ollama, pydantic 2, FastAPI, Docker Compose, IMAP, SQLite, pytest

Collecteur d'opportunités multi-sources, exposé en serveur MCP

Projet personnel · en production continue depuis juillet 2026

Chaîne complète d'ingestion de données : dix-neuf connecteurs de sources publiques hétérogènes, normalisation, dédoublonnage par empreinte, qualification paramétrable, stockage, et exposition à un agent par serveur MCP avec six outils déclarés. Deux collectes par jour, tenues par une unité de service système. La qualification a été corrigée sur mesure : sur un échantillon de 652 offres, la part de faux positifs est passée d'environ 69 % à environ 4 %. Ce chiffre est le point important du projet, parce que c'est celui qui a été mesuré avant, puis après.

Indicateurs · 17 137 lignes de code, 19 connecteurs de sources • 63 commits entre le 29 juillet et le 7 septembre 2026 • 15 563 offres et 1 386 sociétés distinctes en base, 116 collectes journalisées • 6 outils MCP exposés, 29 fichiers d'autotest

Technologies · TypeScript 5.7, Node 20, SDK Model Context Protocol, better-sqlite3, zod, systemd

SARTOR — pipeline agentique à rôles séparés

Projet personnel · août 2026

Système multi-agents où chaque rôle est un prompt versionné et verrouillé par empreinte : un concepteur, un critique, un spécialiste matière, un rédacteur. Le transport vers le modèle est injectable, et la sortie est contrainte par un schéma JSON que le SDK valide et fait corriger au modèle en cas d'écart. Bac à sable d'exécution, portes qualité avec seuils verrouillés dans un fichier, et décisions d'architecture consignées une par une. Le tout est exposé à la fois en ligne de commande, en API et en serveur MCP, les trois étant de fines couches sur la même fonction.

Indicateurs · 11 891 lignes pour le paquet, 22 098 pour le dépôt • 45 commits entre le 5 et le 24 août 2026 • 32 décisions d'architecture consignées, 37 fichiers de tests • 4 rôles d'agent versionnés et verrouillés

Technologies · Python 3.12, pydantic 2, FastAPI, FastMCP, claude-agent-sdk, JSON Schema draft-07, PyTorch (CPU), NumPy, shapely, pytest, hypothesis, ruff, mypy

Connecteur WhatsApp en lecture seule, par serveur MCP

Projet personnel · août – septembre 2026 · en service

Pont vers un compte personnel, en deux processus volontairement séparés : l'un se connecte et écrit dans une base, l'autre lit cette base et l'expose à un agent. La base est ouverte en lecture seule et aucune bibliothèque capable d'écrire n'est chargée côté agent. Le périmètre d'accès est couvert par des tests. C'est un exercice de conception d'accès : rendre une capacité dangereuse structurellement impossible plutôt que simplement interdite.

Indicateurs · 3 900 lignes de code • 11 commits entre le 20 août et le 15 septembre 2026 • 7 outils MCP exposés, tous en lecture

Technologies · TypeScript, SDK Model Context Protocol, SQLite, Baileys, vitest

Dictée vocale locale, avec banc de mesure

Projet personnel · depuis février 2026 · dépôt public

Transcription par inférence locale sur carte graphique, puis mise en forme par un modèle exécuté localement lui aussi : rien ne quitte la machine. L'intérêt du projet pour une mission d'ingénierie IA n'est pas la dictée, c'est le banc de mesure qui l'accompagne : onze combinaisons de modèle et de précision comparées sur le même corpus, avec pour chacune la mémoire occupée, le temps de chargement, le facteur temps réel et le taux d'erreur de mots, avec et sans biais de vocabulaire. Les scores de la partie générative sont moyennés sur trois tirages, parce qu'un modèle n'est pas déterministe même à basse température et qu'un tirage unique se déplace d'un point.

Indicateurs · 8 394 lignes de code, 23 commits depuis le 9 février 2026 • corpus de mesure de 30 clips, 166 secondes d'audio • 11 combinaisons de modèle et de précision comparées • 3 tirages par cas sur la partie générative

Technologies · Python, faster-whisper, CTranslate2, CUDA 12, Ollama, SQLite

Majordome — poste de pilotage à agents

Outil personnel · en usage quotidien depuis août 2026

Orchestration de sessions d'agents qui travaillent en parallèle sur des tâches délimitées, s'écrivent entre elles, et rendent compte. Chaque agent reçoit un périmètre, des interdictions explicites et une consigne de rendu ; les règles apprises d'une erreur réelle sont écrites dans une doctrine que tous les agents lisent, avec la mesure qui les a produites. C'est la mise en pratique quotidienne de ce qu'on demande à une plateforme d'agents : cloisonnement, journal, reprise, et refus de rendre un verdict quand l'outil n'a pas pu mesurer.

Indicateurs · 84 fichiers Python écrits sur mesure, 13 797 lignes • agents adressables entre eux, avec périmètre et journal

Technologies · Python, MCP, orchestration de sessions, systemd, tmux, SQLite

Pile d'inférence auto-hébergée

Installation personnelle · en service

Pile locale de modèles de langage : service d'inférence, interface, modèles personnalisés par fichier de définition, moteur de recherche auto-hébergé, le tout isolé sur un réseau interne, avec des chiens de garde qui relancent le service et surveillent la carte graphique.

Technologies · Docker Compose, Ollama, Open WebUI, SearXNG, scripts de supervision

Laboratoire d'industrialisation Java

Projet de formation · août 2026

API en architecture hexagonale et interface web, conteneurisation en plusieurs étages, chaîne d'intégration continue et tests d'intégration lancés contre de vrais services démarrés en conteneur. Projet monté explicitement pour travailler l'industrialisation et le déploiement, pas pour l'IA.

Indicateurs · 4 923 lignes de code, 52 fichiers Java • 17 commits entre le 13 et le 18 août 2026

Technologies · Java 21, Spring Boot, Vue 3, Docker multi-étages, JUnit 5, Mockito, Testcontainers, intégration continue

FORMATION, LANGUES ET CADRE D'INTERVENTION

Diplôme d'ingénieur, Polytech Lille — informatique, micro-électronique et automatisme, 2017-2022. Baccalauréat scientifique mention très bien.

Français langue maternelle. Anglais B2, TOEIC 850 en 2022.

Intervention en prestation par Elytre (EURL), SIREN 945 053 023, gérant unique et associé à 100 %. Basé à Croix (59170), métropole lilloise.

Une partie des travaux décrits est couverte par des engagements de confidentialité : les noms et les chiffres repris ici sont ceux qui sont communicables. Les dépôts personnels peuvent être ouverts et commentés en entretien technique.