



VOTRE PARTENAIRE D'INTÉGRATION IA

L'intelligence artificielle, intégrée dans votre SI.

Nous installons, orchestrans et connectons vos modèles IA au cœur de votre système d'information : une IA branchée sur vos outils, vos documents et vos bases métier, de l'infrastructure GPU jusqu'à vos écrans. **Vos données ne quittent jamais votre réseau.**

100%

DONNÉES SUR VOTRE RÉSEAU
INTERNE

0

ABONNEMENT SAAS, ZÉRO
TIERS

5

COUCHES D'INTÉGRATION, DU
SERVEUR GPU À VOS ÉCRANS



VOTRE PARTENAIRE D'INTÉGRATION IA

Des ingénieurs qui branchent l'IA sur votre réalité métier.

Fondée en 2026 à Lyon, Confluence Partenaires est une PME d'ingénieurs spécialisée dans l'intégration d'IA souveraine. Notre métier commence là où s'arrête le « POC » : dimensionner l'infrastructure GPU, orchestrer les modèles, développer les connecteurs, serveurs MCP et API qui branchent l'IA sur vos applications métier, puis industrialiser les usages réels. **Le tout chez vous, sans dépendance cloud ni externalisation de vos données — y compris de santé.**

LA SOUVERAINETÉ, PAR CONCEPTION

Modèles open source (Llama, Mistral, Qwen) quantifiés et servis en local via vLLM / Ollama, fine-tunés (LoRA / QLoRA) sur votre documentation. Embeddings et vector store internes, inférence derrière votre pare-feu, journalisation complète des prompts et actions. Conformité RGPD native, hébergement compatible données de santé.

NOTRE SOCLE TECHNIQUE

01 Développement back-end

Architecture microservices, sécurité applicative, tests automatisés et industrialisation.

JAVA · SPRING BOOT · REST · OAUTH JWT · POSTGRESQL

02 Développement front-end

Interfaces métier sur mesure, refonte et montée en version d'applications existantes.

ANGULAR · TYPESCRIPT · DESIGN SYSTEM

03 Ingénierie IA

Serving de modèles, RAG, fine-tuning, évaluation et supervision de la qualité.

PYTHON · VLLM · OLLAMA · LANGCHAIN · PGVECTOR/QDRANT · LORA

04 Infrastructure & DevOps

Dimensionnement GPU, système & réseau, déploiement physique, exploitation et infogérance.

LINUX · DOCKER/K8S · GITLAB CI · RÉSEAU & FIREWALL · SUPERVISION 24/7

L'ÉQUIPE

CV

Corentin Venot

COGÉRANT · BIZDEV

Relation client, cadrage des besoins, réponses aux marchés publics.

MA

Marcellin

TECH LEAD

Java / Spring Boot / Angular. Architecture back-end & API.

BA

Bastien

FULLSTACK · IA

Angular / React / Python. RAG, data science & modèles.

JO

Jordan

INFRA · DEVOPS

Serveurs, GPU, réseau, CI/CD et infrastructure IA interne.

NOS RÉALISATIONS EN EXÉCUTION

Deux projets structurants, en production.

Nos équipes sont engagées sur deux projets structurants qui mobilisent toute notre chaîne technique : IA embarquée dans un logiciel métier manipulant des données de santé, et industrialisation d'un référentiel client chez un assureur.

EN EXÉCUTION

ÉDITEUR LOGICIEL · SANTÉ & SÉCURITÉ AU TRAVAIL

Plateforme de prévention & IA sur données de santé

Refonte et montée en version d'une plateforme de prévention des risques professionnels, incluant l'intégration d'une brique d'IA opérant sur des données de santé au travail. Toute la chaîne d'inférence est maîtrisée en interne : aucune donnée n'est transmise à un service tiers.

- **Modernisation du socle** : montée de framework, refonte backend, dette technique.
- **Brique IA** : classification des déclarations d'incidents, détection de signaux faibles.
- **RAG** sur référentiels réglementaires et documentation de prévention.
- **Inférence 100% locale**, pseudonymisation, cloisonnement, audit trail.

JAVA / SPRING BOOT · ANGULAR · PYTHON · MODÈLE LOCAL · RAG · PSEUDONYMISATION · CI/CD · AUDIT TRAIL

EN EXÉCUTION

ASSURANCE · RÉFÉRENTIEL & DONNÉES CLIENTS

Référentiel client unifié pour un assureur

Application de centralisation des référentiels clients : unifier des données dispersées entre plusieurs systèmes historiques en un référentiel unique, fiable, gouverné et exploitable par les métiers comme par les applications aval.

- **Modélisation du référentiel pivot** et des règles de gouvernance (golden record).
- **Moteur de rapprochement et dédoublonnage** : résolution d'identité, arbitrage.
- **Connecteurs & API** d'ingestion, synchronisation, exposition aux applications.
- **Socle prêt pour l'IA** : un référentiel propre conditionne toute IA fiable.

JAVA / SPRING BOOT · ANGULAR · POSTGRESQL · API REST · ETL & CONNECTEURS · MATCHING / DÉDUPLICATION · RGPD

ILS NOUS FONT CONFIANCE

Apicil · Lefebvre Dalloz · Alteca · Previsoft

NOTRE OFFRE D'INTÉGRATION IA

L'intégration IA on-premise, de la couche 1 à la couche 5.

Cinq couches que nous concevons, déployons et maintenons pour brancher l'IA sur vos systèmes existants : ERP, GED, bases métier, applications internes. **Chaque couche s'appuie sur celle du dessous** — de vos écrans métier (couche 05) à la fondation serveur installée chez vous (couche 01).

05	Applicatif & usages métier Interfaces métier sur mesure, workflows automatisés, intégration dans vos outils. L'IA devient un usage quotidien, pas une démo.	ANGULAR · DESIGN SYSTEM · SSO · TESTS E2E
04	API d'inférence d'entreprise Serveur et API d'inférence exposées à vos applications : authentification, droits, quotas, journalisation. Le socle qui rend l'IA callable partout dans le SI.	SPRING BOOT · REST · OAUTH2/JWT · RATE LIMITING · AUDIT LOG
03	Agents & serveurs MCP Serveurs MCP et connecteurs vers vos systèmes, agents outillés capables d'agir (lire, écrire, déclencher) sur vos données de façon contrôlée et traçable.	MODEL CONTEXT PROTOCOL · TOOLS & SCOPES · CONNECTEURS ERP/GED/SQL
02	Orchestration, RAG & fine-tuning Routage des modèles, RAG sur vos bases, chunking et stratégie d'embeddings, fine-tuning sur votre doctrine, évaluation et supervision de la qualité.	VLLM · OLLAMA · LANGCHAIN · PGVECTOR/QDRANT · LORA/QLORA
01	Infrastructure GPU souveraine Serveur GPU dimensionné selon vos volumes et votre latence cible, serveur d'inférence, durcissement réseau et déploiement physique sur votre site. La fondation souveraine.	GPU NVIDIA · CUDA · DOCKER/K8S · VLAN & CLOISONNEMENT

BRANCHÉ SUR VOTRE SI	ERP · GED · Bases SQL · Applications métier · Messagerie · Annuaire LDAP / AD
----------------------	---

CE QUE L'IA INTÉGRÉE FAIT POUR VOUS

Des usages concrets, branchés sur vos données.

Chaque cas s'appuie sur les couches d'intégration : modèle local, connecteurs vers vos données, API interne et automatisation.

Assistant de connaissance interne

Réponses instantanées sur vos procédures, en RAG sur votre GED, avec citation des sources et respect des habilitations.

Agents métier outillés

Des agents qui agissent : interroger l'ERP, créer un ticket, mettre à jour une fiche. Chaque action via un serveur MCP sécurisé.

IA as a service interne

Une API d'IA appelable par toutes vos applications, derrière votre pare-feu. L'IA comme service maison, sans envoi au cloud.

Traitement intelligent de documents

Saisie, classification, détection de doublons et d'anomalies dans factures, relevés et notes de frais, dès l'arrivée du document.

IA sur données sensibles & santé

IA sur données de santé ou personnelles : pseudonymisation en amont, inférence strictement locale, cloisonnement et traçabilité.

Référentiels & données clients

Centralisation, rapprochement et dédoublonnage de vos référentiels. Un socle de données propre, condition d'une IA fiable.

Assistant relation client

Résumé des dossiers avant rendez-vous, compte-rendu généré après chaque entretien, classé directement dans votre outil.

Automatisation de processus

Enchaînements pilotés par l'IA et déclenchés par vos événements : un document arrive, l'IA analyse, route, notifie et met à jour.

Infrastructure & DevOps

Mise en place et durcissement de vos serveurs, réseau et sécurité. Pipelines CI/CD, conteneurisation, supervision et sauvegarde.

Infogérance & MCO

Nous exploitons votre infrastructure IA au quotidien : mises à jour, sécurité, disponibilité et support. Vous restez sur votre métier.

NOTRE MÉTIER HISTORIQUE · DÉVELOPPEMENT SUR MESURE

Applications métier, plateformes B2B/SaaS full-stack, montée en version et modernisation (Angular / Spring Boot), microservices et CI/CD. **L'IA s'intègre à un applicatif que nous savons aussi construire de A à Z.**

NOTRE MÉTHODE

Un parcours d'intégration éprouvé

01

Cadrage

Audit de votre SI, des données et usages cibles. Périmètre, dimensionnement GPU et objectifs mesurables.

02

Infrastructure

Mise en place du serveur, des modèles et de la sécurité réseau, déployés sur votre site.

03

Intégration

Connecteurs MCP, API, agents et applicatif. Itérations courtes avec vos équipes.

04

Run & MCO

Maintenance, évolution, supervision et formation pour l'autonomie progressive de vos équipes.

NOS ENGAGEMENTS

Garantie de résultat

Résultat satisfaisant, ou nous trouvons une alternative cohérente sans surcoût.

Garantie de remplacement

Livrable ou solution non conforme ? Correction ou remplacement dans les plus brefs délais.

Garantie de suivi

Points réguliers pour vérifier la satisfaction et réévaluer le besoin en cours de route.

EXIGENCES TECHNIQUES QUE NOUS TENONS

Code livré, documenté, réversible

Dépôts, documentation d'architecture et procédures de reprise vous appartiennent. Aucun verrou propriétaire, aucune boîte noire.

Sécurité & traçabilité

Authentification, gestion fine des droits, journalisation des prompts et des actions des agents, pistes d'audit exploitables.

Parlons de votre projet d'intégration IA

AUDIT GRATUIT · RÉPONSE SOUS 12H · DEVIS PERSONNALISÉ

06 35 26 17 15

corentin.venot@confluence-partenaires.fr

NOS OFFRES

Des forfaits clairs, sans surprise.

Forfait ferme après audit. Matériel à prix coûtant. Prix hors taxes.

SOCLE

Installation sur site, configuration de l'IA locale et mise en service. Vos données ne sortent jamais.

à partir de

3 500 €

+ infrastructure

LE PLUS CHOISI

PLATEFORME

Le socle plus un cas d'usage intégré : assistant RAG, automatisation ou connecteur métier.

à partir de

10 000 €

+ infrastructure

ÉCOSYSTÈME

Plateforme, connecteurs MCP multiples, workflows et infogérance incluse la première année.

sur devis, à partir de

20 000 €

+ infrastructure

PRESTATIONS À LA CARTE — FORFAIT FERME APRÈS CADRAGE

Base de connaissance RAG	dès 6 600 €	Automatisation de processus	dès 4 200 €
Connecteur / serveur MCP	dès 3 000 €	Back-end / API d'inférence	dès 5 400 €
Architecture données sensibles	selon périmètre	Fine-tuning modèle	selon périmètre

Tarifs indicatifs de départ. Le forfait final dépend du périmètre fonctionnel, arrêté ensemble après cadrage. Base de valorisation : 600 € / jour-homme.

MAINTENANCE & INFOGÉRANCE (MCO) — ABONNEMENT MENSUEL

VEILLE Supervision, MAJ sécurité, support ouvert. dès 450 € / mois	SÉRÉNITÉ + interventions, SLA, sauvegardes vérifiées. dès 750 € / mois	CRITIQUE + astreinte, GTR courte, dispo renforcée. sur mesure
---	---	--

Note — Infrastructure (serveur GPU, poste secondaire) fournie à prix coûtant, sans marge cachée. Développements hors forfait facturés en régie à 600 € / jour. Devis ferme remis après audit.