BY **RAED** Consulting

OFFRE DE SERVICES · 2026

L'IA agentique en production — en toute confiance.

Conseil, déploiement et exploitation d'assistants IA autonomes & souverains. Vos données restent chez vous. Vos systèmes restent maîtrisés.

> *« Du serveur jusqu'à l'agent. »*

Raed Ben Youssef — Ingénieur infrastructure & IA · 7 ans de production critique · Paris / Île-de-France & Remote
contact@ben-youssef.com · +33 7 77 86 54 18

LE CONTEXTE

L'IA est partout. La mettre en production, c'est une autre histoire.

Tout le monde teste l'IA générative. Très peu réussissent à la transformer en **outil fiable, intégré et sécurisé** dans leur environnement réel. Quatre obstacles reviennent systématiquement :

Le POC qui n'avance pas

Une démo impressionnante… qui ne passe jamais en production faute de fiabilité, d'intégration et de gouvernance.

Vos données partent au cloud

Impossible d'envoyer des données sensibles (clients, RH, contrats, code) à un service tiers. RGPD & souveraineté bloquent.

La peur de l'autonomie

Un agent qui agit seul sur vos systèmes ? Le risque qu'il fasse une mauvaise manipulation freine toute adoption.

Des coûts imprévisibles

La facture de tokens cloud grimpe et devient difficile à maîtriser à mesure que l'usage se généralise.

EN UN MOT

Le frein n°1 à l'adoption de l'IA en entreprise n'est pas la technologie — c'est la **confiance** : « comment laisser un agent toucher nos systèmes et nos données sans prendre de risque ? ». **C'est précisément le problème que je résous.**

MA RÉPONSE

Des agents IA qui passent en production, en sécurité — et chez vous.

Je conçois, déploie et exploite des assistants IA autonomes **gouvernés** et **souverains**, capables de travailler au contact réel de vos systèmes — sans fuite de données et sans risque pour votre production.

Quatre engagements

Souveraineté

IA exécutable **100 % sur votre infrastructure** (on-premise). Vos données ne quittent pas vos murs.

Gouvernance

Chaque action de l'agent est classée et encadrée. Validation humaine sur les opérations sensibles.

Production-grade

Conçu par un ingénieur d'exploitation : supervision, fiabilité, sécurité — pas une démo jetable.

Autonomie

Build to Run : je vous transfère la maîtrise (documentation + formation). Vous n'êtes jamais captif.

> **Le résultat pour vous :** un assistant IA réellement utile, intégré à vos outils, qui respecte vos contraintes de sécurité et de confidentialité — et que vos équipes savent faire évoluer.

POURQUOI BY RAED CONSULTING

Un profil rare : l'ingénieur d'exploitation qui maîtrise les agents IA.

La plupart des spécialistes IA font des démos mais ne savent pas tenir une production. La plupart des experts d'infrastructure ne maîtrisent pas l'agentique. Je suis à l'intersection des deux.

Ce que j'apporte	Ce que ça change pour vous
7 ans de production critique — Ministère de l'Intérieur (haute sécurité), luxe international, retail 24/7, post-production VFX	Je sais ce qu'un incident coûte. Je conçois pour la fiabilité et la sécurité, pas pour la démo.
Un système multi-agents réellement en production — agents autonomes qui administrent une infrastructure complète 24/7	Je ne théorise pas l'agentique : je l'opère. Démonstration en live sur demande.
Une rigueur d'évaluation — méthode de benchmark reproductible pour choisir le bon modèle objectivement	Pas de choix « à la mode » : le modèle retenu est celui qui marche sur VOTRE cas, mesuré.
L'IA locale cross-platform — modèles open-weights exécutés sur Apple Silicon comme sur GPU AMD	Pas besoin d'un parc NVIDIA coûteux : je m'adapte à votre matériel.
La sécurité opérationnelle — segmentation réseau, gestion des accès, supervision, gestion d'incident	Votre agent est déployé dans un cadre durci, supervisé et auditable.

TRANSPARENCE

Mon expertise IA se concentre sur l'**orchestration, le déploiement et l'exploitation** d'agents et de LLM (pas l'entraînement de modèles). C'est 90 % du besoin réel des entreprises : non pas créer un modèle, mais **déployer des agents fiables**.

MES OFFRES

Du premier diagnostic à l'assistant IA exploité.

Cinq formats, du ponctuel au récurrent — combinables selon votre maturité. Tarifs indicatifs ; devis personnalisé après cadrage.

Audit « Agents-Ready »

POUR : DÉMARRER DU BON PIED

- État des lieux de votre infrastructure & de vos données
- Cas d'usage priorisés + matrice de risques
- Feuille de route concrète & chiffrée

à partir de 3 500 € **1–2 sem.**

POC agent gouverné

POUR : PROUVER LA VALEUR VITE

- Un agent sur un cas d'usage réel
- Permissions + validation humaine intégrées
- Démonstration mesurable à vos équipes

à partir de 8 000 € **2–4 sem.**

Déploiement IA souveraine

POUR : PASSER EN PRODUCTION

- Assistant IA 100 % on-premise (LLM local + base de connaissances + connecteurs)
- Documentation technique + d'exploitation
- Formation & transfert à vos équipes

à partir de 15 000 € **4–8 sem.**

Benchmark & sélection de modèle

POUR : CHOISIR OBJECTIVEMENT

- Évaluation reproductible sur votre tâche
- Comparaison vitesse / qualité / coût / sécurité
- Recommandation chiffrée & argumentée

à partir de 4 000 € **1–2 sem.**

Assistant IA autonome — formule récurrente (exploité & maintenu)

TROIS DIMENSIONNEMENTS · TARIFS MENSUELS INDICATIFS HT

Fonctionnalité	Essentiel ~500 €/mois	Pro ~1 200 €/mois	Souverain ~2 500 €/mois
Nombre d'agents	1	Multi-agents	Multi-agents
Connecteurs (Teams, Slack, e-mail, métier…)	1–2	Étendus	Sur-mesure
IA 100 % locale / on-premise	✓	✓	✓
Supervision & reporting	—	✓	✓
Gouvernance (matrice de permissions)	Standard	**Avancée**	**Avancée**
Engagement de service (SLA)	Best-effort	SLA	SLA + astreinte
Mises à jour des modèles	—	Trimestrielle	**Continue**

+ Formation / ateliers : « Déployer un premier agent gouverné », « L'IA locale en entreprise », « AIOps : enrichir vos alertes par l'IA » — 800 à 1 500 € / session.

COMMENT ÇA SE PASSE

Une méthode en 5 étapes, du cadrage au transfert.

Approche progressive et sans engagement aveugle : chaque étape produit un livrable et une décision go/no-go avant la suivante.

1 **Cadrage** — GRATUIT

Échange de 30 min pour comprendre votre contexte, vos contraintes (données, sécurité) et identifier le cas d'usage à plus forte valeur.

2 **Audit « Agents-Ready »** — 1–2 SEM.

Diagnostic de votre infrastructure et de vos données, cas d'usage priorisés, matrice de risques, feuille de route chiffrée.

3 **POC gouverné** — 2–4 SEM.

Un agent réel sur le cas retenu, avec permissions et validation humaine. Vous mesurez la valeur sur pièce avant d'investir.

4 **Déploiement en production** — 4–8 SEM.

Mise en place durcie et supervisée, intégration à vos outils, IA locale si souveraineté requise, documentation complète.

5 **Run & transfert (Build to Run)** — RÉCURRENT

Exploitation, maintenance, mises à jour des modèles — et formation de vos équipes pour vous rendre autonomes.

> **Sans lock-in :** à chaque étape vous gardez la main. L'objectif final est de vous rendre **autonomes**, pas dépendants.

GOUVERNANCE & SÉCURITÉ

Un agent autonome ne veut pas dire un agent incontrôlé.

La clé pour adopter l'IA en confiance : encadrer ce que l'agent a le droit de faire. Chaque action est classée par niveau de risque et traitée différemment.

MATRICE DE PERMISSIONS — EXEMPLE DE CADRE APPLIQUÉ

Niveau	Type d'action	Comportement de l'agent
N1 — Lecture	Consulter un état, un log, une donnée	**Autonome**
N2 — Standard	Action non destructive à contexte clair	Exécute si le contexte est clair, sinon demande
N3 — Sensible	Modifier une configuration, exposer un service, toucher des données critiques	**Validation humaine OBLIGATOIRE avant**
N4 — Interdit	Supprimer des sauvegardes, couper des accès, agir en se faisant passer pour un humain	**Refus systématique**

🔒 Souveraineté

Modèles exécutés sur votre infrastructure. Aucune donnée envoyée à un cloud tiers (sauf choix explicite).

🧑‍💼 Human-in-the-loop

Les actions sensibles requièrent un feu vert humain, via vos canaux (Teams, e-mail…).

📋 Traçabilité

Chaque action de l'agent est journalisée et auditable. Vous savez qui a fait quoi, et quand.

🇪🇺 RGPD by design

La confidentialité est intégrée dès la conception, pas ajoutée après coup.

CAS D'USAGE

Ce qu'un agent IA gouverné peut faire pour vous.

Assistant support interne

« Mes équipes perdent du temps à chercher l'info. »

Un agent qui répond sur vos procédures, contrats, documentation interne — sans que ces données quittent l'entreprise.

Agent d'exploitation (AIOps)

« Trop d'alertes, trop de bruit. »

Un agent qui trie et enrichit vos alertes, corrèle les incidents et prépare le diagnostic — l'IA au service de votre RUN.

Base de connaissance intelligente

« Notre savoir est éparpillé. »

Vos documents rendus interrogeables en langage naturel (RAG), avec réponses sourcées et à jour.

Automatisation de workflows

« Des tâches répétitives nous épuisent. »

Un agent qui orchestre des tâches multi-étapes (extraction, mise en forme, notification) en respectant vos règles.

Veille & reporting

« On rate des signaux importants. »

Un agent qui surveille des sources, résume et alerte — un rapport prêt chaque matin.

Sur-mesure

« Notre besoin est spécifique. »

Votre cas d'usage est unique ? On le cadre ensemble lors des 30 minutes d'échange initial.

PREUVES

Ce que j'ai déjà construit — et que je peux montrer en live.

Tout ce qui est présenté ici est réel et démontrable. Je n'opère pas sur de la théorie : mon propre système d'agents tourne en production continue.

Réalisation	Description	Résultat concret
Agent SRE autonome	Un agent qui administre une infrastructure complète (17 services, réseau segmenté) en continu	En production 24/7, gouverné, avec mémoire persistante & bascule multi-modèles
Bascule IA 100 % locale	Agent passé d'un modèle cloud à un modèle exécuté localement, sans coût d'usage	0 donnée envoyée au cloud · fonctionne sur Apple Silicon & GPU AMD
Méthode de benchmark	Banc d'essai reproductible pour choisir un modèle (qualité, vitesse, sécurité)	6 modèles évalués objectivement · sécurité des actions validée à 100 %
Infrastructure durcie	Réseau segmenté, accès maîtrisés, supervision complète, sauvegardes	~400 services supervisés · gestion d'incident réelle documentée

DÉMONSTRATION

La meilleure preuve est une démonstration : je peux vous montrer mon système d'agents en fonctionnement, en direct, lors de notre échange.

QUESTIONS FRÉQUENTES

Vos questions, sans détour.

Q. Mes données quittent-elles mon entreprise ?

Non, si vous choisissez le mode souverain : le modèle s'exécute sur votre infrastructure et vos données n'en sortent pas. Le recours au cloud n'est utilisé que si vous le décidez explicitement.

Q. Faut-il acheter du matériel NVIDIA coûteux ?

Pas nécessairement. Je déploie des modèles sur Apple Silicon (Mac) comme sur GPU AMD. On dimensionne selon votre matériel existant et votre cas d'usage.

Q. Combien de temps avant un premier résultat ?

Un POC sur un cas réel se livre généralement en 2 à 4 semaines. Vous mesurez la valeur avant tout investissement plus lourd.

Q. Et si je veux internaliser ensuite ?

C'est l'objectif. Ma méthode « Build to Run » inclut documentation et formation : vos équipes reprennent la main. Vous n'êtes jamais captif.

Q. Quels modèles d'IA utilisez-vous ?

Des modèles open-weights (type Qwen, Llama…) en local, ou un modèle cloud en secours selon le besoin. Le choix est fait objectivement, par benchmark sur votre tâche.

Q. Est-ce réservé aux grandes entreprises ?

Non. La formule « Essentiel » est pensée pour les PME/ETI. On commence petit, sur un cas d'usage, et on étend ensuite.

Q. Comment l'agent est-il empêché de faire une bêtise ?

Par la matrice de permissions (section Gouvernance) : les actions sensibles exigent une validation humaine, et certaines sont tout simplement interdites.

PARLONS-EN

Et si on commençait par votre cas d'usage ?

30 minutes, sans engagement, pour identifier où l'IA peut vous apporter le plus de valeur — et comment la déployer en sécurité. Repartez avec une première recommandation, même si nous ne travaillons pas ensemble.

Réservez 30 minutes →

CONTACT	**Raed Ben Youssef** — BY RAED Consulting (SASU)
E-MAIL	contact@ben-youssef.com
TÉLÉPHONE	+33 7 77 86 54 18
ZONE	Paris / Île-de-France & Remote
SPÉCIALITÉ	IA appliquée · Orchestration d'agents · IA souveraine (local-first)