Intelligence Artificielle Locale

Guide stratégique de mise en œuvre

Contexte stratégique

Enjeux business

- Protection des données sensibles et propriété intellectuelle
- Réduction des coûts opérationnels récurrents
- Conformité réglementaire (RGPD, ISO)
- Indépendance technologique

Bénéfices attendus

Sécurité

100% des données en interne

ROI

Élimination des frais d'abonnement

Personnalisation

Adapté aux processus métier

Infrastructure requise

Matériel

Configuration recommandée

Serveur / Workstation
CPU moderne, 16-32 Go RAM

GPU (optionnel)
NVIDIA 8-24 Go VRAM

Stockage
SSD 500 Go minimum

Logiciels

Stack technologique

Système d'exploitation
Linux (Ubuntu) / Windows Server

Runtime IA
Ollama / LM Studio / vLLM

Base vectorielle
FAISS / Qdrant / Chroma

Framework Python
PyTorch / Transformers

Feuille de route de déploiement

1

Phase d'analyse

Identification cas d'usage et audit données

2

Phase de préparation

Nettoyage et structuration corpus

3

POC (Proof of Concept)

Déploiement pilote RAG périmètre restreint

4

Déploiement production

Industrialisation et formation utilisateurs

Approches techniques

RAG

Retrieval-Augmented Generation

- Déploiement rapide (2-4 semaines)
- Maintenance simplifiée
- Mise à jour dynamique des données
- Coûts d'infrastructure modérés

Recommandation : Approche privilégiée pour MVP

Fine-tuning

Entraînement personnalisé

- Adaptation aux processus métier
- Contrôle du style de réponse
- Optimisation des performances
- Investissement initial plus élevé

Usage : Phase d'optimisation post-POC

Architecture RAG

1. Indexation

Conversion documents en vecteurs

↓

2. Recherche sémantique

Identification passages pertinents

↓

3. Génération réponse

Synthèse par LLM

Gestion des données

Sources documentaires

- Bases de connaissances internes
- Documentation technique et procédures
- Archives email et messagerie
- Rapports et études métier
- Données clients (anonymisées)

Traitement requis

- Normalisation et nettoyage
- Déduplication
- Anonymisation PII
- Chunking optimisé
- Enrichissement métadonnées

Sécurité et conformité

Conformité RGPD

Données hébergées en interne, traçabilité complète

Normes ISO

ISO 27001, 27701, documentation processus

Contrôles d'accès

Authentification, permissions, logs

Backup & DR

Sauvegardes quotidiennes, plan reprise

Analyse coûts-bénéfices

Investissements initiaux

Infrastructure	**15-40 k€**
Licences logicielles	**0-5 k€**
Développement	**20-50 k€**
Formation	**5-10 k€**
Total	**40-105 k€**

Économies annuelles

Support externalisé

-15k€/an

Gains productivité

+25k€/an

Risques et mitigation

Risque : Qualité réponses insuffisante

Mitigation : Tests avec utilisateurs pilotes

Risque : Performances limitées

Mitigation : POC dimensionnement adapté

Risque : Résistance au changement

Mitigation : Conduite du changement

Risque : Compétences manquantes

Mitigation : Recrutement ou partenariat

KPIs et métriques

Performance technique

Temps de réponse moyen

< 2 secondes

Précision des réponses

> 85%

Disponibilité système

99.5%

Adoption utilisateurs

Taux d'adoption

> 70% à M6

Requêtes quotidiennes

500+ par jour

Satisfaction NPS

> 40

Glossaire technique

LLM
Large Language Model - Modèle de langage à grande échelle

RAG
Retrieval-Augmented Generation - Génération augmentée par recherche

Embeddings
Représentations vectorielles de texte pour calcul de similarité

Vector Database
Base spécialisée pour recherche par similarité vectorielle

Chunking
Segmentation de documents en passages indexables

Fine-tuning
Spécialisation d'un modèle sur corpus spécifique

Inference
Phase d'utilisation du modèle en production

Quantization
Compression de modèle pour réduire empreinte mémoire

PII
Personally Identifiable Information - Données personnelles

NPS
Net Promoter Score - Indicateur de satisfaction

Recommandations

Approche recommandée

Démarrage rapide avec architecture RAG sur périmètre pilote

Phase 1

POC - 3 mois

Validation technique et métier

Phase 2

Déploiement - 6 mois

Industrialisation et formation

Phase 3

Optimisation continue

Fine-tuning et extensions