Grille 2 — Architecture et coût total sur trois ans

ADVISORE · Webinaire du 8 décembre 2026 · livrable de l'exposé 2

> Objectif : choisir un modèle et une infrastructure pour des raisons de coût, de contrôle et de réversibilité — pas par effet de mode.

Partie A — Les six questions qui décident

Répondez dans l'ordre. La première réponse « non » arrête le raisonnement : elle désigne le modèle.

#	Question	Si OUI	Si NON
1	Vos données peuvent-elles **sortir** de votre pays ou de votre réseau (obligation client, secteur, contrat) ?	cloud ou hybride possible	hébergement local imposé
2	Acceptez-vous qu'un fournisseur puisse **modifier le modèle** sans votre accord ?	API propriétaire acceptable	modèle figé / open source
3	Avez-vous une **équipe** capable d'exploiter une infrastructure (GPU, sauvegardes, mises à jour) ?	on-premise envisageable	infogérance ou cloud
4	Le cas d'usage tolère-t-il une **latence** de 1 à 3 secondes ?	API distante acceptable	traitement local requis
5	Le **volume** justifie-t-il un investissement matériel ?	> 2 M tokens / mois : on-premise à étudier	facturation à l'usage plus saine
6	La **sortie** doit-elle être possible en moins d'un mois ?	open source, formats ouverts	acceptation d'une dépendance

Arbre de décision

Configuration	Quand la choisir	Ce que ça coûte vraiment
API propriétaire (SaaS)	cas d'usage standard, volume faible, pas d'équipe infra	abonnement + coût par token ; dépendance forte
Open source hébergé (cloud privé)	données sensibles, besoin de contrôle, équipe légère	GPU à l'heure + intégration + exploitation
Open source sur site	souveraineté, volume élevé, équipe infra existante	matériel amorti + MCO + compétences
Hybride	données sensibles en local, raisonnement externalisé	deux chaînes à exploiter : le plus complexe à tenir

Partie B — Modèle de coût total (TCO) sur 36 mois

Renseignez les cellules. Les trois scénarios se comparent **hors subventions** et **hors gains**.

Poste de coût	Scénario 1 — API	Scénario 2 — Cloud privé	Scénario 3 — Sur site
Licences / abonnement (36 mois)			
Consommation (tokens, GPU/heure)			
Matériel (amorti sur 36 mois)			
Intégration initiale (jours × TJM)			
Exploitation / MCO (jours/an × 3)			
Sauvegarde, supervision, sécurité			
Formation et acculturation			
Coût de sortie (reprise de données, réécriture)			
TOTAL 36 MOIS			
Coût par utilisateur / mois			

Les cinq coûts que les devis oublient

1. **La qualité de la donnée** : nettoyage, déduplication, annotation — souvent 30 à 50 % du projet.
2. **L'exploitation** : quelqu'un doit surveiller les coûts par appel, les dérives de prompt, les erreurs.
3. **La sécurité** : journalisation, cloisonnement, revue des accès — ce n'est pas inclus dans l'abonnement.
4. **La sortie** : reprendre ses historiques et ses intégrations coûte 10 à 20 % du projet initial.
5. **L'acculturation** : sans formation, les utilisateurs contournent l'outil et le gain disparaît.

Partie C — Fiche de décision — architecture

Cas d'usage retenu	
Contrainte n°1 (donnée, pays, latence)	
Modèle retenu	☐ API ☐ Open source cloud ☐ Open source sur site ☐ Hybride
TCO 36 mois	
Coût de sortie estimé	
Dépendance acceptée (fournisseur, format, compétence)	
Point de non-retour (date / volume)	
Décideur / date	

Le test des trois questions

Avant de signer, répondez à voix haute :

1. **« Que se passe-t-il si le fournisseur double son prix demain ? »** — si la réponse est « on paie », le choix n'est pas une architecture, c'est une dépendance.

2. **« Comment je récupère mes données et mes intégrations en 30 jours ? »** — sans procédure écrite, la réversibilité est un slogan.
3. **« Qui, dans mon équipe, sait faire tourner ça dans 18 mois ? »** — un modèle sans exploitant devient un actif orphelin.

ADVISORE — Conseil en ingénierie 4IR · www.advisoreng.com · Livrable du webinaire « L'IA agentique en entreprise », 8 décembre 2026. Les montants sont à renseigner : aucun chiffre générique n'est fourni, il serait faux pour votre organisation.