

Choisir un modèle d'IA sans religion



Mario

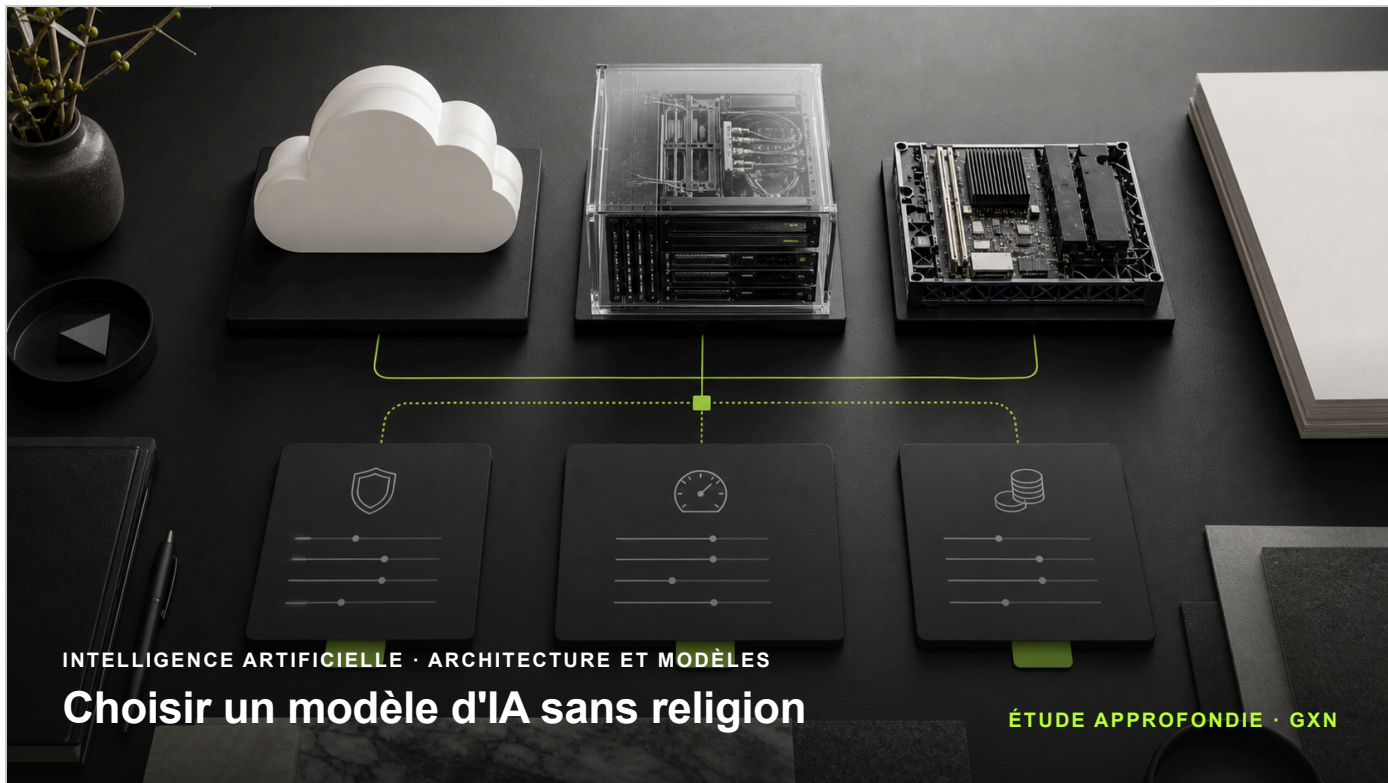
Expert senior · GXN

Maîtrise en intelligence artificielle

Plus de vingt ans en implantation technologique

30 juillet 2026

Publié et mis à jour



INTELLIGENCE ARTIFICIELLE · ARCHITECTURE ET MODÈLES

Choisir un modèle d'IA sans religion

ÉTUDE APPROFONDIE · GXN

TL;DR

La réponse en une minute

- Le choix d'un modèle est une décision d'ingénierie, pas une préférence de marque.
- Trois critères tranchent : la confidentialité exigée, la performance minimale nécessaire, et le coût à l'usage sur douze mois.
- La plupart des entreprises surachètent. Un modèle plus petit et bien encadré suffit pour l'extraction, le tri et le classement.
- Héberger un modèle chez vous coûte rarement moins cher. Ça s'impose pour des raisons de confidentialité, pas d'économie.
- Bâissez pour pouvoir changer de modèle. Dans dix huit mois, le meilleur choix aura changé.

Le mauvais débat

« Quel est le meilleur modèle » est la question que tout le monde pose et c'est la mauvaise.

Il n'existe pas de meilleur modèle dans l'absolu, pas plus qu'il n'existe de meilleur véhicule. Il existe un modèle adapté à une tâche, à une contrainte de confidentialité et à un volume. Les fournisseurs qui vous vantent une marque en particulier vous parlent de leur habitude, pas de votre besoin.

La bonne question tient en une phrase : quelle est la plus petite configuration qui règle mon problème de façon fiable, dans le respect de mes contraintes de données.

Critère 1. La confidentialité exigée

C'est le critère qui tranche en premier, parce qu'il élimine des options plutôt que de les comparer.

Posez vous la question dans l'ordre inverse de ce que font la plupart des gens. Pas « quel modèle je veux », mais « où mes données ont elles le droit d'aller ».

Trois niveaux, en pratique.

Données courantes. Contenu marketing, documentation publique, questions générales. Aucune contrainte particulière. Toutes les options sont ouvertes, prenez la plus économique qui fait le travail.

Données d'affaires sensibles. Contrats, prix, dossiers clients, informations financières. Les offres d'entreprise des grands fournisseurs conviennent généralement, à condition de vérifier trois clauses contractuelles : non entraînement sur vos données, durée de conservation, et lieu de traitement. Ce sont des vérifications, pas des suppositions.

Données réglementées ou hautement sensibles. Renseignements de santé, dossiers d'employés, information protégée par un ordre professionnel, secrets industriels. Là, l'architecture devient le cœur du projet, et l'hébergement local ou en infrastructure contrôlée entre en jeu sérieusement.

Le guide sur la confidentialité couvre les questions exactes à poser à un fournisseur. Rappel utile : GXN offre l'implantation technique et le soutien opérationnel. L'interprétation juridique de vos obligations doit venir d'un conseiller juridique qualifié.

Critère 2. La performance minimale nécessaire

Voici la partie où les entreprises dépensent le plus inutilement.

On choisit le modèle le plus puissant du marché pour une tâche qui consiste à décider si un courriel est une plainte ou une demande de soumission. C'est acheter un camion lourd pour livrer des enveloppes.

Un ordre de grandeur utile, tiré de projets réels.

- **Classement et tri.** Un petit modèle suffit presque toujours. Rapide, très peu cher, largement assez fiable quand il est bien encadré.
- **Extraction de données structurées.** Un modèle intermédiaire fait généralement le travail, avec des vérifications automatiques par dessus.
- **Rédaction et synthèse destinées à des humains.** Un modèle plus fort vaut son prix, parce que la qualité se voit.
- **Raisonnement complexe sur des documents longs.** Là seulement, les modèles haut de gamme se justifient vraiment.

La méthode saine : commencez par le plus petit modèle plausible, mesurez le taux d'exactitude sur des cas réels, et montez seulement si les chiffres l'exigent. L'inverse, partir du plus gros et redescendre, n'arrive jamais dans la vraie vie parce que personne ne veut toucher à un système qui fonctionne.

Critère 3. Le coût à l'usage sur douze mois

Le calcul que presque aucune soumission ne présente, et que vous devriez exiger.

Le principe est simple. Vous payez selon la quantité de texte envoyée et reçue. Donc trois variables comptent : votre volume mensuel, la quantité de contexte envoyée à chaque appel, et le prix du modèle choisi.

Ce qui fait exploser une facture, ce n'est presque jamais le nombre de questions. C'est la quantité de contexte envoyée à chaque fois. Un système mal conçu qui expédie trente pages de documents à chaque requête coûtera des dizaines de fois le prix d'un système qui envoie les deux paragraphes pertinents. Même résultat visible pour l'utilisateur. Facture complètement différente.

La question à poser à tout fournisseur : au volume que je vous ai donné, combien coûtera l'usage par mois, et sur quelle hypothèse de contexte moyen vous basez vous. Un fournisseur qui ne peut pas répondre ne comprend pas ce qu'il vend, ou ne veut pas que vous le sachiez.

Les trois critères qui tranchent

01

Confidentialité

Où les données peuvent aller

02

Performance

Le minimum réellement nécessaire

03

Coût annuel

Volume et contexte compris

**Configuration
défendable**

Infonuagique, local ou ouvert

Le vrai portrait, sans idéologie.

Modèle infonuagique fermé. Vous appelez le modèle chez son fournisseur. Aucune infrastructure à gérer, mises à jour automatiques, performance de pointe, coût variable. Vos données transitent chez un tiers, encadrées par contrat. C'est le choix par défaut raisonnable pour la majorité des PME québécoises.

Modèle ouvert hébergé chez vous ou dans votre infrastructure. Les données ne sortent pas de votre périmètre. Vous contrôlez tout. En échange : infrastructure à payer et à entretenir, compétences internes ou externes requises, performance généralement en retrait des meilleurs modèles commerciaux, et responsabilité complète des mises à jour.

Point important, souvent mal compris : héberger localement coûte rarement moins cher. À faible volume, c'est presque toujours plus cher, parce que vous payez une infrastructure qui dort. Le calcul bascule seulement à très gros volume constant. Si quelqu'un vous propose l'hébergement local en invoquant l'économie, demandez le calcul.

Le choix se fait donc sur la confidentialité et le contrôle, pas sur le prix. Quand vos contraintes de données l'exigent, l'hébergement contrôlé se justifie pleinement, prix ou pas.

Approche hybride. Fréquente et souvent la meilleure : un modèle local pour les données sensibles, un modèle infonuagique pour le reste. Ça demande une architecture pensée dès le départ, mais ça évite de payer partout le prix de votre cas le plus contraignant.

COMPARER LES ARCHITECTURES SANS PRÉFÉRENCE DE MARQUE		
Architecture	Contrôle	Effort opérationnel
Infonuagique fermé	Contractuel	Faible
Ouvert hébergé	Maximal	Élevé
Hybride	Adapté par flux	Intermédiaire

La décision qui compte plus que le modèle

Concevez pour pouvoir changer.

Les modèles évoluent tous les quelques mois. Les prix baissent. Les gagnants d'aujourd'hui ne sont pas nécessairement ceux de l'an prochain. Une application construite autour d'un modèle précis devra être réécrite. Une application construite autour d'un besoin, avec le modèle comme pièce remplaçable, traversera plusieurs générations sans drame.

Concrètement, ça veut dire une couche d'abstraction entre votre logique d'affaires et le modèle, des invites versionnées et testables, et une suite de tests qui permet de comparer un nouveau modèle sur vos propres cas en quelques heures.

C'est une décision d'architecture qui se prend le premier jour. Après, ça coûte cher à rattraper.

Ce que je vois sur le terrain

Deux réflexes coûteux reviennent constamment.

Le premier, c'est le réflexe de marque. « On veut du ChatGPT » ou « on veut du Claude », comme on choisirait un fournisseur de café. Ce sont des interfaces et des familles de modèles avec des forces différentes selon les tâches. Choisir avant d'avoir défini la tâche, c'est mettre la charrue devant les bœufs.

Le second, c'est la surqualité par confort. On prend le modèle le plus puissant pour ne pas avoir à mesurer. Ça marche, ça coûte cinq à dix fois trop cher, et personne ne s'en rend compte parce que la facture reste modeste au début. Puis le volume monte.

Mon test rapide sur une soumission : si le modèle proposé est le même pour toutes les tâches du projet, personne n'a fait l'analyse.

Quand ce guide ne s'applique pas

Si votre projet est petit et jetable, ne surinvestissez pas dans cette décision. Prenez une option raisonnable, mesurez, et corrigez si nécessaire. L'analyse détaillée se justifie quand le volume est significatif, quand les données sont sensibles, ou quand le système devient critique pour vos opérations.

QUESTIONS ET RÉPONSES

Questions fréquentes

01 Quel modèle recommandez vous.

Aucun par défaut. Je recommande après avoir vu la tâche, le volume et les contraintes de données. Une recommandation donnée avant ces trois éléments est une opinion, pas un conseil.

02 Faut-il héberger l'IA au Canada.

Ça dépend de la nature de vos données et de vos obligations. Quand c'est pertinent, des options canadiennes existent et nous documentons où résident les données. Pour l'interprétation de vos obligations légales, référez-vous à votre conseiller juridique.

03 Un modèle ouvert est-il moins performant.

Sur les tâches courantes de classement et d'extraction, l'écart est souvent négligeable. Sur le raisonnement complexe, les meilleurs modèles commerciaux gardent une avance. L'écart se referme continuellement.

04 **Combien coûte l'hébergement d'un modèle chez nous.**

Infrastructure, entretien et compétences. Pour une PME, ça se justifie rarement par le prix seul. Faites faire le calcul complet avant de vous engager.

05 **Peut on changer de modèle après le lancement.**

Oui, si l'architecture le prévoit. Non, si l'application a été construite autour d'un modèle précis. C'est pourquoi cette question doit être posée avant la première ligne de code.

MÉTHODE ÉDITORIALE

Guide de fond écrit par Mario pour GXN (Gouvernance numérique), une division de MD79. Révision prévue selon le rythme documenté pour cette grappe.



À PROPOS DE L'AUTEUR

Mario

Expert senior directement accessible, avec plus de vingt ans d'expérience en implantation technologique et une maîtrise en intelligence artificielle.

[Voir le profil de Mario ↗](#)

DIAGNOSTIC IA - 750 \$ À 1 500 \$

Vous voulez la configuration la plus économique qui règle vraiment votre problème.

Le diagnostic IA compare les options sur vos données et votre volume réel, avec le coût d'usage estimé sur douze mois. De 750 \$ à 1 500 \$, crédité sur votre projet lorsque pertinent.

Pas d'appel obligatoire. Pas de séquence de vente.