

IN-DEPTH GUIDE

ARCHITECTURE AND MODELS

Choosing an AI model without religion



GXN

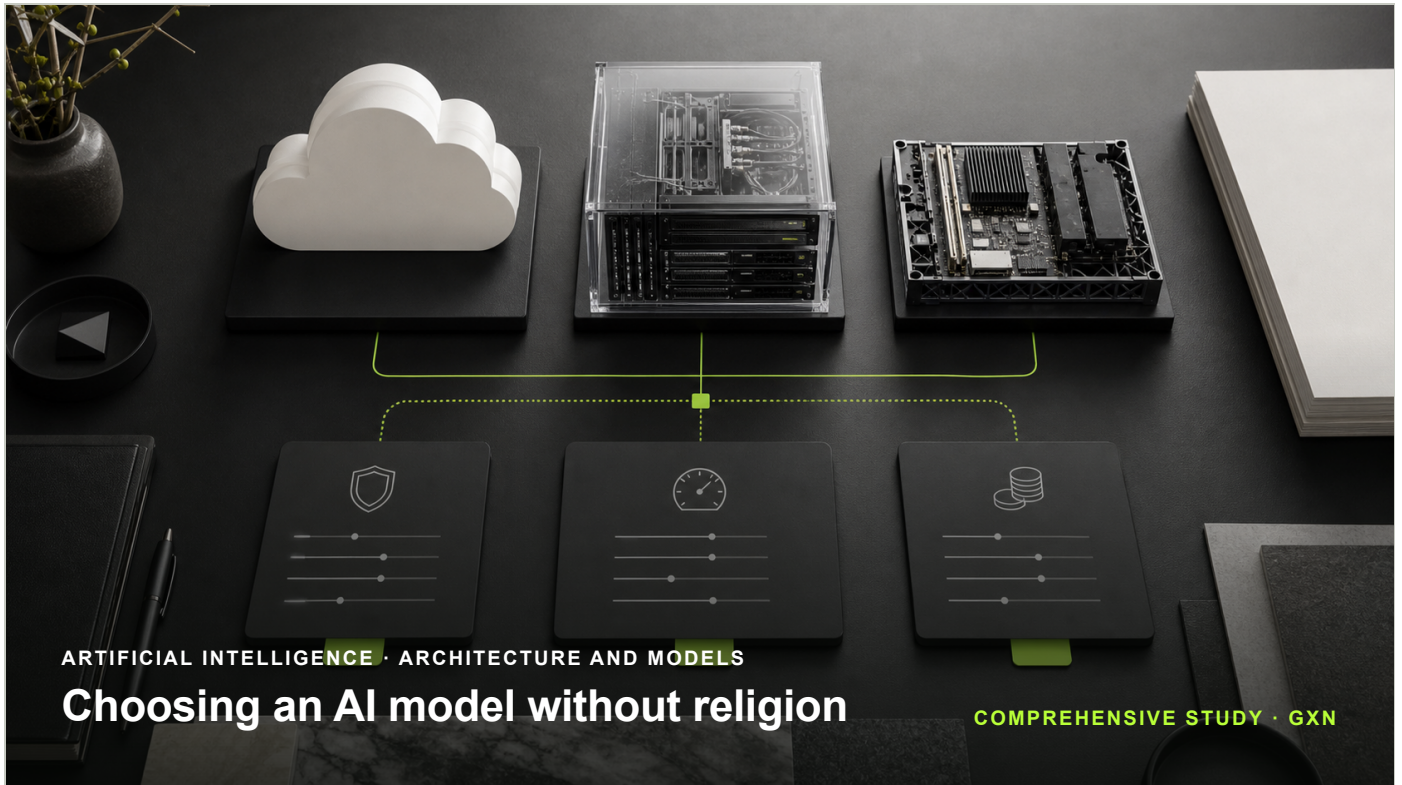
Senior expertise · Digital governance

Direct response

Limits, risks and decisions

July 30, 2026

Updated



ARTIFICIAL INTELLIGENCE · ARCHITECTURE AND MODELS

Choosing an AI model without religion

COMPREHENSIVE STUDY · GXN

TL;DR

The answer in one minute

- Choosing a model is an engineering decision, not a brand preference.
- There are three criteria: confidentiality required, minimum performance required, and cost to use over 12 months.
- Most companies overbought. A smaller, well-framed model is sufficient for extraction, sorting and classification.
- Hosting a model at home rarely costs less. This is necessary for reasons of confidentiality, not economics.
- Build to be able to change models. In eighteen months, the best choice will have changed.

01 · BACKGROUND GUIDE

The bad debate

"What is the best model" is the question that everyone asks and it is the wrong one.

There is no absolute best model, any more than there is a best vehicle. There is a model adapted to a task, a confidentiality constraint and a volume. Suppliers who praise a particular brand to you are talking to you about their habits, not about your needs.

The right question is in one sentence: what is the smallest configuration that solves my problem reliably, while respecting my data constraints.

Criterion 1. Confidentiality required

This is the criterion that decides first, because it eliminates options rather than comparing them.

Ask yourself in reverse order of what most people do. Not "what model I want," but "where my data have the right to go."

Three levels, in practice.

Current data. Marketing content, public documentation, general questions. No particular constraints. All options are open, take the most economical one that gets the job done.

Sensitive business data. Contracts, prices, customer files, financial information. Corporate offers from major providers are generally suitable, provided that you check three contractual clauses: no training on your data, retention period, and place of processing. These are checks, not guesses.

Regulated or highly sensitive data. Health information, employee records, professionally protected information, industrial secrets. There, architecture becomes the core of the project, and local or controlled infrastructure hosting comes into play seriously.

The Privacy Guide covers the exact questions to ask a vendor. Useful reminder: GXN offers technical implementation and operational support. The legal interpretation of your obligations should come from qualified legal counsel.

Criterion 2. Minimum necessary performance

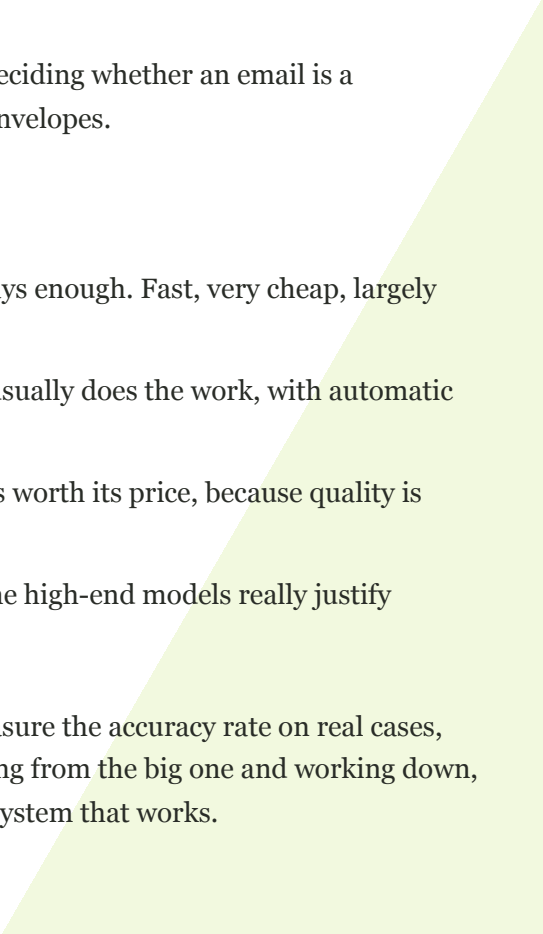
This is the part where businesses spend the most unnecessarily.

The most powerful model of the market is chosen for a task of deciding whether an email is a complaint or a bid request. It's buying a heavy truck to deliver envelopes.

A useful order of magnitude, drawn from real projects.

- **Classification and sorting.** A small model is almost always enough. Fast, very cheap, largely reliable when well framed.
- **Extraction of structured data.** An intermediate model usually does the work, with automatic checks over it.
- **Writing and synthesis for humans.** A stronger model is worth its price, because quality is seen.
- **Complex reasoning on long documents.** Only there, the high-end models really justify themselves.

The sound method: start with the smallest plausible model, measure the accuracy rate on real cases, and move up only if the numbers require it. The opposite, starting from the big one and working down, never happens in real life because no one wants to mess with a system that works.



Criterion 3. Usage cost over twelve months

The math that almost no bid presents, and that you should demand.

The principle is simple. You pay according to the quantity of text sent and received. So three variables matter: your monthly volume, the amount of context sent with each call, and the price of the chosen model.

What makes a bill explode is almost never the number of questions. This is the amount of context sent each time. A poorly designed system that sends thirty pages of documents with each query will cost dozens of times the price of a system that sends the relevant two paragraphs. Same result visible to the user. Completely different bill.

The question to ask any supplier: at the volume that I gave you, how much will the use cost per month, and on what hypothesis of average context you base yourself. A supplier who can't respond either doesn't understand what they're selling, or doesn't want you to know.

DECISION FRAMEWORK

The three criteria that decide

01

Privacy

Where data can go

02

Performance

The minimum really necessary

03

Annual cost

Volume and context included

**Setup
defensible**

05 · BACKGROUND GUIDE

Cloud, local or open

The true portrait, without ideology.

Closed cloud model. You call the model at its supplier. No infrastructure to manage, automatic updates, state-of-the-art performance, variable cost. Your data pass through a third party, framed by contract. This is the reasonable default choice for the majority of Quebec SMEs.

Open model hosted at your home or in your infrastructure. The data does not leave your perimeter. You control everything. In exchange: infrastructure to pay for and maintain, internal or external skills required, performance generally behind the best business models, and complete responsibility for updates.

Important point, often misunderstood: hosting locally rarely costs less. Low volume is almost always more expensive, because you pay for an infrastructure that sleeps. The calculation only changes to very large constant volume. If someone offers you local hosting invoking economy, ask for the calculation.

The choice is therefore made on confidentiality and control, not on price. When your data constraints so require, controlled hosting is fully justified, price or not.

Hybrid approach. Frequent and often the best: a local model for sensitive data, a cloud model for the rest. It requires an architecture thought out from the start, but it avoids paying the price of your most restrictive case everywhere.

COMPARE ARCHITECTURES WITHOUT BRAND PREFERENCE		
Architecture	Control	Operational Effort
Closed cloud computing	Contractual	Low
Open hosted	Maximum	High
Hybrid	Adapted by flow	Intermediate

The decision that matters more than the model

Design for change.

The models evolve every few months. Prices are falling. Today's winners are not necessarily next year's winners. An application built around a specific model will need to be rewritten. An application built around a need, with the model as a replaceable part, will last several generations without drama.

Concretely, this means a layer of abstraction between your business logic and the model, versioned and testable prompts, and a test suite that allows you to compare a new model on your own cases in a few hours.

It's an architectural decision that is made on the first day. Afterwards, it costs a lot to catch up.

What I see on the ground

Two costly reflexes keep coming back.

The first is the brand reflex. “We want ChatGPT” or “we want Claude”, like we would choose a coffee supplier. These are interfaces and families of models with different strengths depending on the tasks. Choosing before having defined the task is putting the cart before the horse.

The second is over-quality for comfort. We take the most powerful model so as not to have to measure. It works, it costs five to ten times too much, and no one realizes it because the bill remains modest at the start. Then the volume increases.

Quick quote test: if the proposed model is the same for all project tasks, no analysis was done.

When this guide does not apply

If your project is small and disposable, do not overinvest in this decision. Take a reasonable option, measure, and correct if necessary. Detailed analysis is justified when the volume is significant, when the data is sensitive, or when the system becomes critical for your operations.

Frequently asked questions

01 **What model do you recommend.**

None by default. I recommend after seeing the task, volume and data constraints. A recommendation given before these three items is an opinion, not advice.

02 **Should AI be hosted in Canada?**

It depends on the nature of your data and your obligations. Where relevant, Canadian options exist and we document where the data reside. For the interpretation of your legal obligations, refer to your legal advisor.

03 **An open model is less efficient.**

On routine classification and extraction tasks, the gap is often negligible. On complex reasoning, the best business models stay ahead. The gap is continually closing.

04 **How much does it cost to host a model with us.**

Infrastructure, maintenance and skills. For an SME, this is rarely justified by price alone. Do the full calculation before committing.

05 **Can be changed after launch.**

Yes, if the architecture does. No, if the application was built around a specific model. That is why this question must be asked before the first line of code.

EDITORIAL METHOD

Background guide prepared by GXN (Digital Governance), a division of MD79. Revision planned according to the rhythm documented for this cluster.

OFFICIAL SOURCES

References to be checked according to your situation

These references support the external rules and frameworks cited in this guide. They are not a substitute for legal or professional advice tailored to your organization.

NIST

Artificial Intelligence Risk Management Framework



ABOUT THE PUBLISHER

GXN

A directly accessible senior expertise, with more than twenty years of experience in technology implantation and a master's degree in artificial intelligence within the company.

[Discover GXN](#) ↗

AI OPPORTUNITY ASSESSMENT

You want the most economical configuration that really solves your problem.

The AI diagnosis compares the options on your data and your actual volume, with the estimated usage cost over twelve months.

No call required. No sales sequence.