

IN-DEPTH GUIDE

AI FOUNDATIONS

LLM, agent, RAG: the vocabulary of AI decoded for business people



GXN

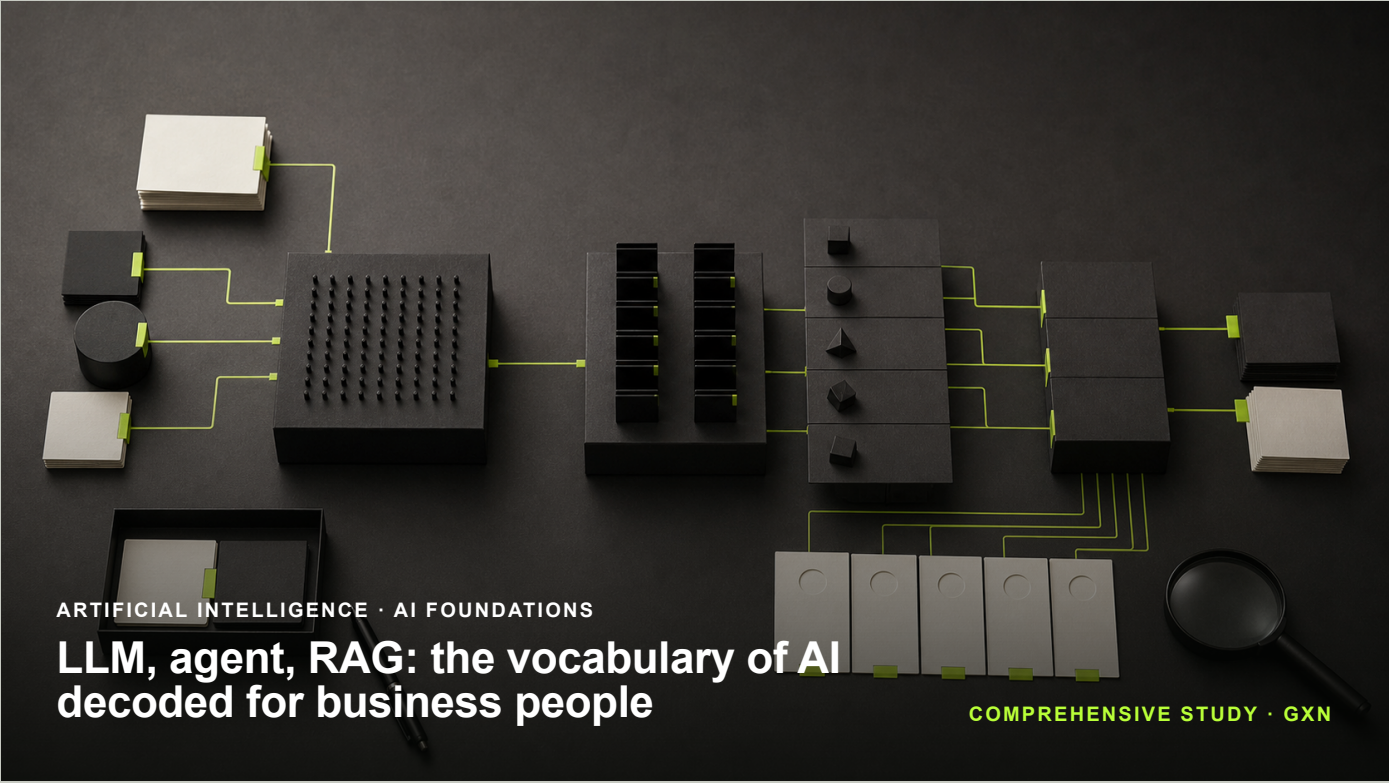
Senior expertise · Digital governance

Direct response

Limits, risks and decisions

July 30, 2026

Updated



ARTIFICIAL INTELLIGENCE · AI FOUNDATIONS

LLM, agent, RAG: the vocabulary of AI decoded for business people

COMPREHENSIVE STUDY · GXN

TL;DR

The answer in one minute

- One **LLM** is an engine that predicts text. Everything else is built on top.
- One **agent** is an LLM who has been given tools and the right to complete steps alone.
- The **RAG** allows the template to respond from your documents rather than its general brief. This is what makes AI useful in business.
- One **hallucination** is an invention delivered with confidence. This is not a temporary fault, it is an operating characteristic.
- One **jeton** is the billing unit. Your costs are calculated in tokens, not in questions asked.
- The **fine tuning** is almost always the wrong answer for an SME. The RAG solves the same problem for much less money.

01 · BACKGROUND GUIDE

Why this Guide

You're in contact with a supplier. It follows: agent, orchestration, RAG, context window, fine setting, crazy guard. You're nodding your head. You cannot judge what he proposes.

This guide fixes this problem. Each term receives a simple definition, a concrete image, and above all the only thing that really interests you: what it changes for your business and your invoice.

Read it once. Keep it on hand for your next meeting.

02 · BACKGROUND GUIDE

Basic terms

Artificial intelligence

Umbrella term that covers almost everything, which makes it almost useless in a meeting. When a supplier says "our solution uses AI," he said nothing. Ask him what model, to do what, on what data.

What it changes for you: Nothing until the next question is asked.

Machine learning

A system that improves its results from examples rather than from rules written by a human. It's the technical family. Your spam filter has been one of them for 20 years.

What it changes for you: It reminds you that AI did not appear in your company in 2023. She was already there, discreetly.

Large Language Model, or LLM

The engine behind ChatGPT, Claude, Gemini and company. He read an immense amount of text and became very good at predicting which sequence of words answers a query well.

The useful image: imagine a brilliant intern who has read everything, who writes quickly, who remembers nothing from one meeting to the next, and who never admits that he ignores an answer.

What it changes for you: it's excellent for reading, writing, classifying and retrieving. It's not a database, not a calculator, and not a source of truth about your business.

Model

A precise LLM, with its version and features. There are dozens of them, from very big and expensive to small and fast. The choice of model is an engineering decision, not a brand preference.

What it changes for you: a supplier who uses only one model for everything, whatever the problem, works out of habit and not by analysis. A smaller and well framed model often beats a large model poorly integrated, for a fraction of the price.

Token

The text cutting unit. About three quarters of a word in French. Everything you send to the model and everything it responds to is counted in tokens, and that's exactly what is charged.

What it changes for you: this is the key to your usage costs. A system that sends 30 pages of context to each question costs a hundred times more than a system that sends the relevant paragraph. Two solutions that do the same thing may have completely different monthly bills. Ask the question.

Context window

The amount of text the model can have in front of them at one time. Beyond that, he no longer sees the beginning.

The image: a work desk. You can display a certain number of documents there. The rest stays in the binder.

What it changes for you: this limitation explains why we cannot simply “give all our documents to the model”. This is precisely the problem that the RAG solves.

Prompt, or prompt

Instructions given to the model. In a business system, the prompt is not a question typed by a user, it is a constructed, tested, versioned text, which contains the rules, the tone, the expected format and the prohibitions.

What it changes for you: it is a serious part of the implementation work. A vendor that treats the prompt as a detail will deliver an unstable system.

Hallucination

When the model produces false information, formulated with the same assurance as true information. An invented date, an article of law that does not exist, a plausible but false amount.

Important point, and often misunderstood: it is not a bug in the process of being corrected. It's the very way these systems work. The rate drops with the best models and good architecture. It doesn't fall to zero.

What it changes for you: We don't settle it by trust, we settle it by design. Human verification where error is expensive, citing sources, format constraints, and never an irreversible automatic decision.

03 · BACKGROUND GUIDE

Architectural terms

RAG, or generation increased by recovery

The method that allows the model to respond from your documents to you. The system first searches for relevant snippets in your documentation, then gives them to the model and asks it to respond with just that.

The image: instead of asking your intern what he remembers, you place the three correct files on his desk before asking the question.

What it changes for you: it's the technology behind internal enterprise search and the majority of useful assistants. This also reduces hallucinations the most, because the model can cite where its response comes from. For an SME, this is almost always the right approach, and almost always cheaper than the alternatives.

Vectorization and diving

The technique which allows you to find a document by the meaning rather than by the exact words. A search for “parental leave” brings up a document that talks about “absence at the birth of a child”.

What it changes for you: This is the internal mechanics of the RAG. You don't have to understand it in detail, but you should know that it explains why an AI-powered search finds things that your current search engine misses.

Agent

A LLM who has been given tools, as well as the right to sequence several steps without supervision between each. A classic assistant answers. An agent acts: he consults your system, calculates, writes to a database, sends an email.

What it changes for you: this is where the real potential lies, and also the real risk. An agent who makes a mistake doesn't just produce a bad sentence, he takes a bad action. Any conversation about agents must begin with a list of what they are allowed to do, and especially what they are not allowed to do alone.

In 2026, agents work well on delimited and well-defined tasks. Demonstrations of autonomous agents that manage complete processes are still, in the majority of cases, demonstrations.

Tool, or function call

What we plug into a model so that it can do something other than write: consult your inventory, create a customer file, calculate a tax.

What it changes for you: This is exactly where the integration work is located, and therefore a large part of the cost of a project. The pattern is the easy part. Connecting it correctly to your systems is the real job.

Orchestration

The coordination of several steps, models or tools in the same process. Extract, validate, classify, route, alert.

What it changes for you: A real process almost always requires orchestration. This is also what distinguishes a business project from a gadget connected to a single query.

Human in the loop

A mandatory stop point where a valid person before the system continues.

What it changes for you: it is your main risk control lever. My rule: human validation wherever an error affects a customer, commits the company, or modifies financial data. Full automation where the error is minor, detectable and correctable. This boundary is decided with you, explicitly, before conception.

Railings

The constraints that prevent the system from going out of its way: imposed formats, prohibited subjects, limits on amounts, automatic verification of responses.

What it changes for you: ask to see the guardrails provided. A supplier who hasn't thought about it hasn't designed for production, they designed for demonstration.

From motor to controlled gesture

01

LLM

Predicted text



02

RAG

Provides your sources



03

Agent

Use tools



04

Human

Take responsibility for the decision

Terms that are expensive

Fine tuning or fine tuning

Retrain a model on your own examples for a particular behavior or style.

What it changes for you: it is often suggested, rarely necessary. Nine times out of ten, what an SME really wants is for the model to respond from its documents, and it's the RAG that does that, for much less money and without having to start all over again when your documents change. If a supplier offers fine tuning from the outset, ask them why RAG would not be enough. His answer will tell you a lot.

Open model and closed model

An open template can be downloaded and hosted on your infrastructure. A closed model is used remotely at its supplier.

What it changes for you: the open model gives maximum control over data, with real infrastructure and maintenance cost. The closed model is simpler and often more efficient, but your data passes elsewhere. The choice is made on your confidentiality constraints, not on an ideological preference. A comprehensive guide covers this decision.

Inference

Rotating the model to get a response. Every inference has a cost.

What it changes for you: This is your variable cost. It grows with your use. Require a twelve-month estimate at the actual expected volume.

Temperature

A setting that controls the degree of creativity of responses. Low for data extraction, higher for writing.

What it changes for you: technical detail, but a teller. If an invoice extraction system gives you different results for the same document, this setting is one of the wrong things.

The translation table in one line each

- A LLM predicts text.
- A token is your unit of invoice.
- The RAG gives him your documents.
- An agent gives him hands.
- The guardrails give it limits.
- The human in the loop gives him a boss.
- A hallucination is the permanent risk that justifies all the above.

OPERATIONAL TRANSLATION OF KEY TERMS		
Term	Role	Business question
LLM	Text forecast	What task does it have to do?
RAG	Give your documents	Where does each answer come from?
Agent	Give hands	What actions can he take alone?
Human	Valid	Where does the company take over the decision?

What I see on the ground

The word “agent” is becoming what the word “cloud” was in 2012. It gets stuck on just about anything.

My chat test, and it still works: ask the provider what their bot can do without human validation. If the answer is vague or enthusiastic rather than specific, it’s not a real agent. It’s just a form with fancy words.

Second test: Ask where the response comes from when the system responds. A good business system cites its sources. A system that responds without ever showing what it relies on requires a level of trust that nothing justifies.

When this vocabulary is not enough

This guide gives you what you need to judge a proposal. It doesn't make you a technical architect, and that's not the goal.

Three situations require a real technical look: when sensitive or regulated data are involved, when the system has to write in your production systems, and when a supplier asks you for a multi-year commitment. In these cases, have someone look at the proposal. An hour of second thought costs less than a bad contract.

Frequently asked questions

01 **What is the difference between a conversational robot and an agent?**

A conversational robot responds with text. An agent uses tools and gestures in your systems. The difference is not cosmetic, it is of nature. It completely changes the level of risk and integration work.

02 **Is RAG better than fine adjustment?**

For the vast majority of SMEs, yes. The RAG responds from your current documents, updates when your documents change, and costs a lot less. Fine tuning is mainly used to change behaviour or style, not to add knowledge.

03 **Does a model learn from our conversations.**

Depends on the supplier and the contract. On offers to companies, training on your data is usually excluded, but it is a clause to check, not to assume. The privacy guide covers the exact questions to ask.

04 **Why the answers change from time to time.**

These systems are probabilistic in nature. For tasks that require consistency, such as data extraction, this variability is greatly reduced through adjustments and automatic checks.

05 **How many of these terms do I really need to know?**

Five are enough to be a good buyer: LLM, token, RAG, agent, human in the loop. The rest is comfort.

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](#) [↗](#)

NEXT STEP

You have an AI proposal on your desk and you want a second look.

The AI diagnosis examines the feasibility, actual costs and risks, including those not in the quote.

No call required. No sales sequence.