GUIDE 2 FONDATIONS DE L'IA

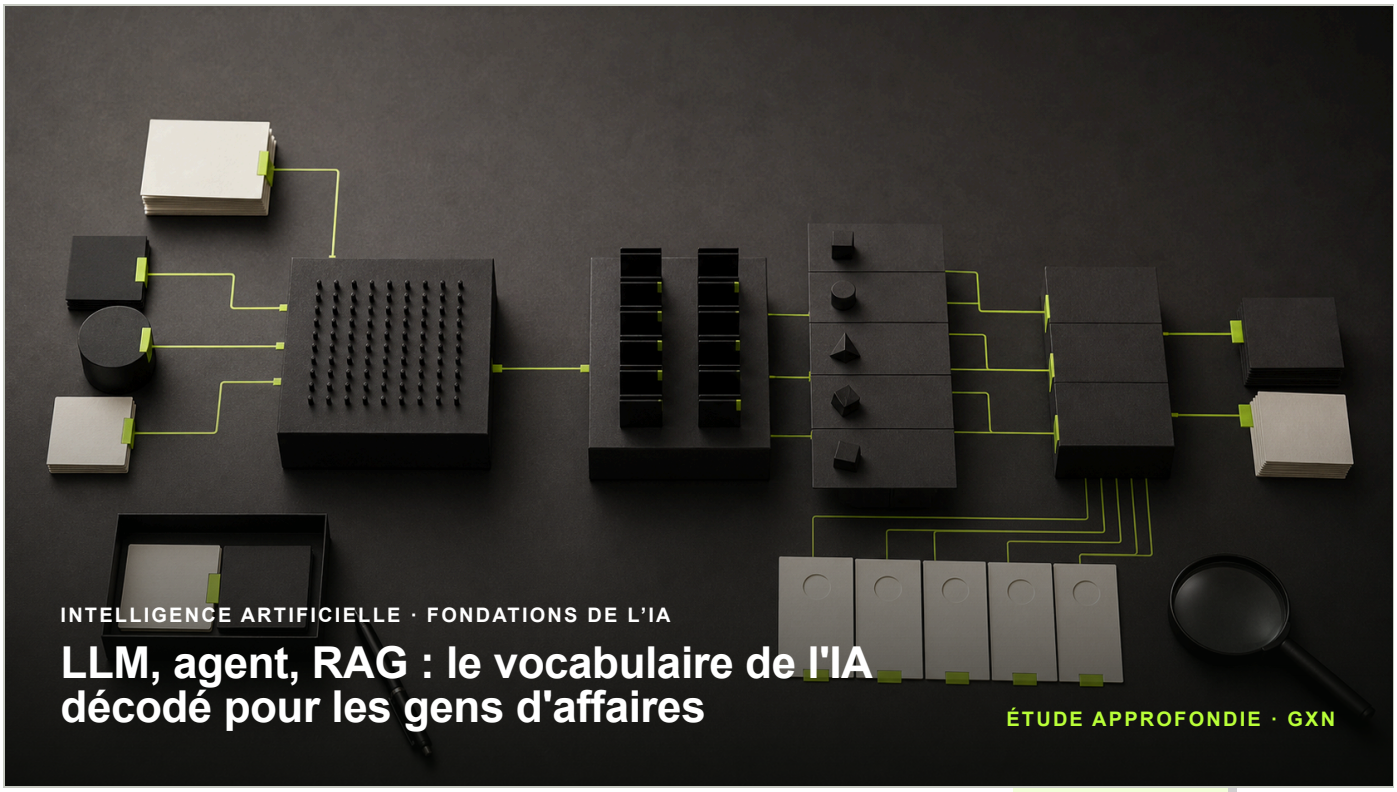
LLM, agent, RAG : le vocabulaire de l'IA décodé pour les gens d'affaires

M

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 · FONDATIONS DE L’IA

LLM, agent, RAG : le vocabulaire de l'IA décodé pour les gens d'affaires

ÉTUDE APPROFONDIE · GXN

TL;DR

La réponse en une minute

- Un **LLM** est un moteur qui prédit du texte. Tout le reste est construit par dessus.
- Un **agent** est un LLM à qui on a donné des outils et le droit d'enchaîner des étapes seul.
- Le **RAG** permet au modèle de répondre à partir de vos documents plutôt que de sa mémoire générale. C'est ce qui rend l'IA utile en entreprise.
- Une **hallucination** est une invention livrée avec assurance. Ce n'est pas un défaut temporaire, c'est une caractéristique du fonctionnement.
- Un **jeton** est l'unité de facturation. Vos coûts se calculent en jetons, pas en questions posées.
- Le **réglage fin** est presque toujours la mauvaise réponse pour une PME. Le RAG règle le même problème pour beaucoup moins cher.

01 · GUIDE DE FOND

Pourquoi ce guide

Vous êtes en rencontre avec un fournisseur. Il enchaîne : agent, orchestration, RAG, fenêtre de contexte, réglage fin, garde fous. Vous hochez la tête. Vous ne pouvez pas juger ce qu'il propose.

Ce guide règle ce problème. Chaque terme reçoit une définition simple, une image concrète, et surtout la seule chose qui vous intéresse vraiment : ce que ça change pour votre entreprise et votre facture.

Lisez le une fois. Gardez le sous la main pour votre prochaine rencontre.

02 · GUIDE DE FOND

Les termes de base

Intelligence artificielle

Terme parapluie qui recouvre à peu près tout, ce qui le rend presque inutile en réunion. Quand un fournisseur dit « notre solution utilise l'IA », il n'a rien dit. Demandez lui quel modèle, pour faire quoi, sur quelles données.

Ce que ça change pour vous : rien tant que la question suivante n'est pas posée.

Apprentissage automatique

Un système qui améliore ses résultats à partir d'exemples plutôt qu'à partir de règles écrites par un humain. C'est la famille technique. Votre filtre antipourriel en fait partie depuis vingt ans.

Ce que ça change pour vous : ça vous rappelle que l'IA n'est pas apparue en 2023 dans votre entreprise. Elle y était déjà, discrètement.

Grand modèle de langage, ou LLM

Le moteur derrière ChatGPT, Claude, Gemini et compagnie. Il a lu une quantité immense de texte et il est devenu très bon pour prédire quelle suite de mots répond bien à une demande.

L'image utile : imaginez un stagiaire brillant qui a tout lu, qui écrit vite, qui ne se souvient de rien d'une rencontre à l'autre, et qui n'admet jamais qu'il ignore une réponse.

Ce que ça change pour vous : c'est excellent pour lire, écrire, classer et extraire. Ce n'est pas une base de données, pas une calculatrice, et pas une source de vérité sur votre entreprise.

Modèle

Un LLM précis, avec sa version et ses caractéristiques. Il en existe des dizaines, du très gros et cher au petit et rapide. Le choix du modèle est une décision d'ingénierie, pas une préférence de marque.

Ce que ça change pour vous : un fournisseur qui n'utilise qu'un seul modèle pour tout, quel que soit le problème, travaille par habitude et non par analyse. Un modèle plus petit et bien encadré bat souvent un gros modèle mal intégré, pour une fraction du prix.

Jeton

L'unité de découpage du texte. Environ trois quarts d'un mot en français. Tout ce que vous envoyez au modèle et tout ce qu'il répond se compte en jetons, et c'est exactement ce qui est facturé.

Ce que ça change pour vous : c'est la clé de vos coûts d'usage. Un système qui envoie 30 pages de contexte à chaque question coûte cent fois plus cher qu'un système qui envoie le paragraphe pertinent. Deux solutions qui font la même chose peuvent avoir des factures mensuelles complètement différentes. Posez la question.

Fenêtre de contexte

La quantité de texte que le modèle peut avoir sous les yeux en même temps. Au delà, il ne voit plus le début.

L'image : un bureau de travail. Vous pouvez y étaler un certain nombre de documents. Le reste reste dans le classeur.

Ce que ça change pour vous : cette limite explique pourquoi on ne peut pas simplement « donner tous nos documents au modèle ». C'est justement le problème que le RAG résout.

Invite, ou prompt

Les instructions données au modèle. Dans un système d'entreprise, l'invite n'est pas une question tapée par un utilisateur, c'est un texte construit, testé, versionné, qui contient les règles, le ton, le format attendu et les interdictions.

Ce que ça change pour vous : c'est une partie sérieuse du travail d'implantation. Un fournisseur qui traite l'invite comme un détail livrera un système instable.

Hallucination

Quand le modèle produit une information fausse, formulée avec la même assurance qu'une information vraie. Une date inventée, un article de loi qui n'existe pas, un montant plausible mais faux.

Point important, et souvent mal compris : ce n'est pas un bogue en voie d'être corrigé. Ça découle de la façon même dont ces systèmes fonctionnent. Le taux baisse avec les meilleurs modèles et une bonne architecture. Il ne tombe pas à zéro.

Ce que ça change pour vous : on ne règle pas ça par la confiance, on le règle par la conception. Vérification humaine là où l'erreur coûte cher, citation des sources, contraintes de format, et jamais de décision automatique irréversible.

03 · GUIDE DE FOND

Les termes d'architecture

RAG, ou génération augmentée par récupération

La méthode qui permet au modèle de répondre à partir de vos documents à vous. Le système cherche d'abord les extraits pertinents dans votre documentation, puis les donne au modèle en lui demandant de répondre uniquement avec ça.

L'image : au lieu de demander à votre stagiaire ce dont il se souvient, vous lui déposez les trois bons dossiers sur le bureau avant de poser la question.

Ce que ça change pour vous : c'est la technologie derrière la recherche interne d'entreprise et la majorité des assistants utiles. C'est aussi ce qui réduit le plus les hallucinations, parce que le modèle peut citer d'où vient sa réponse. Pour une PME, c'est presque toujours la bonne approche, et presque toujours moins chère que les alternatives.

Vectorisation et plongement

La technique qui permet de retrouver un document par le sens plutôt que par les mots exacts. Une recherche sur « congé parental » ramène un document qui parle de « absence à la naissance d'un enfant ».

Ce que ça change pour vous : c'est la mécanique interne du RAG. Vous n'avez pas à la comprendre en détail, mais vous devez savoir qu'elle explique pourquoi une recherche assistée par IA trouve des choses que votre moteur de recherche actuel manque.

Agent

Un LLM à qui on a donné des outils, ainsi que le droit d'enchaîner plusieurs étapes sans supervision entre chacune. Un assistant classique répond. Un agent agit : il consulte votre système, calcule, écrit dans une base de données, envoie un courriel.

Ce que ça change pour vous : c'est là que se trouve le potentiel réel, et aussi le risque réel. Un agent qui se trompe ne produit pas juste une mauvaise phrase, il pose un mauvais geste. Toute conversation sur les agents doit commencer par la liste de ce qu'ils ont le droit de faire, et surtout de ce qu'ils n'ont pas le droit de faire seuls.

En 2026, les agents fonctionnent bien sur des tâches délimitées et bien définies. Les démonstrations d'agents autonomes qui gèrent des processus complets sont encore, dans la majorité des cas, des démonstrations.

Outil, ou appel de fonction

Ce qu'on branche à un modèle pour qu'il puisse faire autre chose qu'écrire : consulter votre inventaire, créer une fiche client, calculer une taxe.

Ce que ça change pour vous : c'est exactement ici que se situe le travail d'intégration, et donc une bonne part du coût d'un projet. Le modèle est la partie facile. Le brancher correctement à vos systèmes est la vraie job.

Orchestration

La coordination de plusieurs étapes, modèles ou outils dans un même processus. Extraire, valider, classer, acheminer, alerter.

Ce que ça change pour vous : un processus réel demande presque toujours de l'orchestration. C'est aussi ce qui distingue un projet d'entreprise d'un gadget branché sur une seule requête.

Humain dans la boucle

Un point d'arrêt obligatoire où une personne valide avant que le système continue.

Ce que ça change pour vous : c'est votre principal levier de contrôle du risque. Ma règle : validation humaine partout où une erreur touche un client, engage l'entreprise, ou modifie une donnée financière. Automatisation complète là où l'erreur est mineure, détectable et corrigeable. Cette frontière se décide avec vous, explicitement, avant la conception.

Garde fous

Les contraintes qui empêchent le système de sortir de son couloir : formats imposés, sujets interdits, limites de montants, vérifications automatiques des réponses.

Ce que ça change pour vous : demandez à voir les garde fous prévus. Un fournisseur qui n'y a pas pensé n'a pas conçu pour la production, il a conçu pour la démonstration.

INFOGRAPHIE 01

Du moteur au geste contrôlé

01

LLM

Prédit le texte

↓

02

RAG

Fournit vos sources

↓

03

Agent

Utilise des outils

↓

04

Humain

Assume la décision

Les termes qui coûtent cher

Réglage fin, ou fine tuning

Réentraîner un modèle sur vos propres exemples pour qu'il adopte un comportement ou un style particulier.

Ce que ça change pour vous : c'est souvent proposé, rarement nécessaire. Neuf fois sur dix, ce qu'une PME veut réellement, c'est que le modèle réponde à partir de ses documents, et c'est le RAG qui fait ça, pour beaucoup moins cher et sans devoir tout recommencer quand vos documents changent. Si un fournisseur propose du réglage fin d'entrée de jeu, demandez lui pourquoi le RAG ne suffirait pas. Sa réponse vous en dira long.

Modèle ouvert et modèle fermé

Un modèle ouvert peut être téléchargé et hébergé sur votre infrastructure. Un modèle fermé s'utilise à distance, chez son fournisseur.

Ce que ça change pour vous : le modèle ouvert donne un contrôle maximal sur les données, avec un coût d'infrastructure et d'entretien réel. Le modèle fermé est plus simple et souvent plus performant, mais vos données transitent ailleurs. Le choix se fait sur vos contraintes de confidentialité, pas sur une préférence idéologique. Un guide complet couvre cette décision.

Inférence

Le fait de faire tourner le modèle pour obtenir une réponse. Chaque inférence a un coût.

Ce que ça change pour vous : c'est votre coût variable. Il grandit avec votre usage. Exigez une estimation sur douze mois, au volume réel prévu.

Température

Un réglage qui contrôle le degré de créativité des réponses. Basse pour de l'extraction de données, plus élevée pour de la rédaction.

Ce que ça change pour vous : un détail technique, mais un révélateur. Si un système d'extraction de factures vous donne des résultats différents pour le même document, ce réglage fait partie des choses mal ajustées.

05 · GUIDE DE FOND

Le tableau de traduction en une ligne chacun

- Un LLM prédit du texte.
- Un jeton est votre unité de facture.
- Le RAG lui donne vos documents.
- Un agent lui donne des mains.
- Les garde fous lui donnent des limites.
- L'humain dans la boucle lui donne un patron.
- Une hallucination est le risque permanent qui justifie tout ce qui précède.

TRADUCTION OPÉRATIONNELLE DES PRINCIPAUX TERMES

Terme	Rôle	Question d'affaires
LLM	Prédit du texte	Quelle tâche de texte doit-il accomplir?
RAG	Donne vos documents	D'où vient chaque réponse?
Agent	Donne des mains	Quels gestes peut-il poser seul?
Humain	Valide	Où l'entreprise reprend-elle la décision?

Ce que je vois sur le terrain

Le mot « agent » est en train de devenir ce que le mot « infonuagique » a été en 2012. Il se colle sur à peu près n'importe quoi.

Mon test en rencontre, et il fonctionne toujours : demandez au fournisseur ce que son agent a le droit de faire sans qu'un humain valide. Si la réponse est floue ou enthousiaste plutôt que précise, ce n'est pas un agent. C'est un formulaire avec du vocabulaire moderne.

Deuxième test : demandez d'où vient la réponse quand le système répond. Un bon système d'entreprise cite ses sources. Un système qui répond sans jamais montrer sur quoi il s'appuie vous demande une confiance que rien ne justifie.

07 · GUIDE DE FOND

Quand ce vocabulaire ne suffit pas

Ce guide vous donne de quoi juger une proposition. Il ne fait pas de vous un architecte technique, et ce n'est pas le but.

Trois situations demandent un vrai regard technique : quand des données sensibles ou réglementées sont en jeu, quand le système doit écrire dans vos systèmes de production, et quand un fournisseur vous demande un engagement pluriannuel. Dans ces cas, faites regarder la proposition par quelqu'un d'indépendant. Une heure de second regard coûte moins cher qu'un mauvais contrat.

QUESTIONS ET RÉPONSES

Questions fréquentes

01 **Quelle est la différence entre un robot conversationnel et un agent.**

Un robot conversationnel répond avec du texte. Un agent utilise des outils et pose des gestes dans vos systèmes. La différence n'est pas cosmétique, elle est de nature. Elle change complètement le niveau de risque et le travail d'intégration.

02 **Le RAG est il mieux que le réglage fin.**

Pour la grande majorité des besoins en PME, oui. Le RAG répond à partir de vos documents actuels, se met à jour quand vos documents changent, et coûte beaucoup moins cher. Le réglage fin sert surtout à modifier un comportement ou un style, pas à ajouter des connaissances.

03

Est ce qu'un modèle apprend de nos conversations.

Ça dépend du fournisseur et du contrat. Sur les offres destinées aux entreprises, l'entraînement sur vos données est généralement exclu, mais c'est une clause à vérifier, pas à supposer. Le guide sur la confidentialité couvre les questions exactes à poser.

04

Pourquoi les réponses changent d'une fois à l'autre.

Ces systèmes sont probabilistes par nature. Pour les tâches qui exigent de la constance, comme l'extraction de données, on réduit fortement cette variabilité par les réglages et par des vérifications automatiques.

05

Combien de ces termes dois je vraiment connaître.

Cinq suffisent pour être un bon acheteur : LLM, jeton, RAG, agent, humain dans la boucle. Le reste est du confort.

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.

M

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 ↗

PROCHAINE ÉTAPE

Vous avez une proposition d'IA sur votre bureau et vous voulez un second regard.

Le diagnostic IA examine la faisabilité, les coûts réels et les risques, y compris ceux qui ne sont pas dans la soumission. De 750 $ à 1 500 $, crédité sur votre projet lorsque pertinent.

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