KODRA

GUIDE PRATIQUE

Comprendre l'Intelligence Artificielle Générative

Notions essentielles et lexique
pour les professionnels et gestionnaires

Sébastien Bélisle — Kodra Conseil

Avril 2026 · kodra.ca

Table des matières

> *Ce document a pour objectif de démystifier l'intelligence artificielle générative en offrant des repères clairs, sans jargon inutile. Il s'adresse aux professionnels et gestionnaires qui souhaitent comprendre ces technologies pour prendre des décisions éclairées.*

1. Qu'est-ce que l'IA générative ?

L'intelligence artificielle générative (IAG) désigne une catégorie de systèmes d'IA capables de **créer du contenu nouveau** — texte, images, code, audio, vidéo — à partir de ce qu'ils ont appris durant leur entraînement. Contrairement aux systèmes classiques qui analysent ou classifient des données, l'IA générative produit quelque chose qui n'existait pas auparavant.

Le tournant majeur est survenu fin 2022 avec le lancement de ChatGPT par OpenAI, qui a rendu cette technologie accessible au grand public. Depuis, des entreprises comme Anthropic (créatrice de Claude), Google (Gemini) et Meta (Llama) ont développé leurs propres modèles, chacun avec des approches différentes en matière de sécurité et de capacités.

> *L'IA générative ne « comprend » pas le monde comme un humain. Elle identifie des patterns statistiques dans d'immenses quantités de texte et apprend à prédire ce qui devrait venir ensuite dans une séquence. C'est puissant, mais ce n'est pas de la pensée.*

Les grandes familles de modèles génératifs

Grands modèles de langage (LLM) — Ce sont les systèmes les plus connus (Claude, GPT, Gemini). Ils traitent et génèrent du texte, mais aussi de plus en plus d'images et de code. C'est le cœur de la plupart des assistants IA actuels.

Modèles de génération d'images — Des systèmes comme DALL-E, Midjourney ou Stable Diffusion créent des images à partir de descriptions textuelles (prompts).

Modèles multimodaux — La tendance actuelle : des modèles capables de traiter simultanément du texte, des images, de l'audio et de la vidéo. Claude, GPT-4 et Gemini sont des exemples de modèles multimodaux.

2. Comment fonctionne un modèle de langage ?

Pour bien utiliser l'IA générative, il est utile de comprendre ses mécanismes de base — même sans entrer dans les mathématiques.

L'entraînement

Un LLM est entraîné sur des quantités massives de texte provenant d'Internet : livres, articles, sites web, forums, documentation technique. Durant cet entraînement, le modèle ajuste des milliards de **paramètres** (des valeurs numériques) pour apprendre les relations statistiques entre les mots. C'est un processus coûteux qui nécessite des milliers de processeurs spécialisés (GPU) pendant des semaines.

La prédiction, mot par mot

Quand vous posez une question à Claude ou à ChatGPT, le modèle ne cherche pas la réponse dans une base de données. Il génère sa réponse **un mot à la fois** (techniquement, un « token » à la fois), en calculant à chaque étape quel est le mot le plus probable qui devrait suivre, compte tenu de tout ce qui précède.

> *Pensez à la saisie prédictive de votre téléphone, mais poussée à une échelle extraordinaire. Votre téléphone prédit le prochain mot ; un LLM prédit les prochains paragraphes avec une cohérence remarquable.*

La fenêtre de contexte

Chaque modèle a une **fenêtre de contexte** : la quantité d'information qu'il peut traiter en une seule conversation. Elle se mesure en tokens. Les modèles récents comme Claude offrent des fenêtres de 200 000 tokens, soit l'équivalent d'un livre de 500 pages. Cela signifie que vous pouvez lui soumettre de longs documents et avoir une conversation approfondie à leur sujet.

Le prompt : votre interface avec l'IA

Le **prompt** est l'instruction que vous donnez au modèle. C'est votre principal levier d'influence sur la qualité des résultats. Un prompt clair, précis et bien structuré produit des réponses nettement meilleures qu'une instruction vague. L'art de formuler de bons prompts s'appelle le « prompt engineering ».

3. Les usages concrets en milieu professionnel

L'IA générative n'est pas qu'un phénomène technologique — c'est un outil de travail qui transforme déjà le quotidien de nombreux professionnels. Voici les usages les plus courants et les plus porteurs de valeur.

Rédaction et communication

Ébauches de courriels, rapports, résumés exécutifs, comptes rendus de réunions, publications sur les réseaux sociaux. L'IA accélère le premier jet ; le jugement humain assure la qualité finale.

Analyse et synthèse de documents

Résumer un rapport de 50 pages, extraire les points clés d'un contrat, comparer des propositions. Les modèles avec de grandes fenêtres de contexte excellent dans cette tâche.

Recherche et veille

Explorer un sujet, identifier des tendances, préparer un dossier de recherche. L'IA offre un point de départ solide, mais les informations doivent toujours être vérifiées.

Programmation et automatisation

Écrire du code, créer des formules Excel, automatiser des tâches répétitives. Même sans être développeur, on peut utiliser l'IA pour créer des outils simples.

Formation et apprentissage

Expliquer un concept complexe, créer des exercices, simuler des scénarios. L'IA est un tuteur patient et disponible qui s'adapte au niveau de l'apprenant.

4. Limites et précautions

L'IA générative est un outil puissant, mais elle comporte des limites importantes qu'il faut connaître pour l'utiliser de manière responsable.

Les hallucinations

Les modèles peuvent générer des informations fausses avec une grande assurance. Ce phénomène, appelé « hallucination », se produit parce que le modèle optimise la *vraisemblance* de sa réponse, pas sa *véracité*. Il est donc essentiel de vérifier les faits, les chiffres et les citations produits par l'IA.

Les biais

Les modèles reflètent les biais présents dans leurs données d'entraînement. Ils peuvent reproduire des stéréotypes ou présenter des perspectives déséquilibrées sur certains sujets. Une lecture critique reste indispensable.

La confidentialité

Les informations que vous partagez avec un modèle d'IA peuvent, selon la plateforme et les paramètres utilisés, être conservées ou utilisées pour améliorer le modèle. Il est crucial de ne jamais partager d'informations confidentielles, personnelles ou sensibles sans avoir vérifié les politiques de confidentialité du service utilisé.

La date de coupure

Chaque modèle a une **date de coupure des connaissances** (knowledge cutoff) : il ne connaît pas les événements survenus après cette date. Certains modèles compensent avec un accès à Internet en temps réel, mais ce n'est pas systématique.

> *Règle d'or : traitez l'IA comme un collègue brillant mais parfois trop sûr de lui. Ses contributions sont précieuses, mais elles méritent toujours une relecture humaine attentive.*

5. Lexique de A à Z

Les termes essentiels pour naviguer dans l'univers de l'IA générative, expliqués en langage clair.

Agent IA (AI Agent)

Un système d'IA capable d'accomplir des tâches complexes de manière autonome en enchaînant plusieurs actions : rechercher de l'information, exécuter du code, interagir avec des outils externes. C'est l'évolution naturelle des assistants conversationnels vers des assistants d'action.

> *Exemple : Claude Code est un agent IA qui peut écrire, tester et déployer du code de manière autonome.*

Alignement (Alignment)

L'ensemble des techniques visant à s'assurer qu'un modèle d'IA se comporte conformément aux intentions et aux valeurs humaines. C'est l'un des grands défis de la recherche actuelle en IA.

Apprentissage automatique (Machine Learning)

Branche de l'IA où les systèmes apprennent à partir de données plutôt que d'être explicitement programmés. L'IA générative est une sous-catégorie de l'apprentissage automatique.

Biais algorithmique (Algorithmic Bias)

Tendance d'un modèle à produire des résultats déséquilibrés ou discriminatoires, souvent héritée des données d'entraînement. Par exemple, un modèle entraîné principalement sur des textes en anglais sera moins performant en français.

Chatbot (Chatbot)

Interface conversationnelle permettant d'interagir avec un modèle d'IA par le biais de messages texte. Claude, ChatGPT et Gemini sont des chatbots propulsés par des LLM.

Date de coupure (Knowledge Cutoff)

Date au-delà de laquelle le modèle n'a pas d'information fiable. Les événements postérieurs à cette date lui sont inconnus, sauf s'il dispose d'un accès à Internet.

> *Exemple : Un modèle avec une coupure en mars 2025 ne connaîtra pas les résultats d'une élection tenue en juin 2025.*

Embedding (Embedding)

Représentation numérique d'un mot, d'une phrase ou d'un document sous forme de vecteur (liste de nombres). Les embeddings permettent au modèle de mesurer la proximité sémantique entre des concepts.

Exemple : « Roi » et « reine » auront des embeddings proches dans l'espace vectoriel.

Fenêtre de contexte (Context Window)

Quantité maximale de texte qu'un modèle peut traiter en une seule interaction. Mesurée en tokens. Plus la fenêtre est grande, plus le modèle peut intégrer d'information dans ses réponses.

Exemple : Claude offre une fenêtre de 200 000 tokens, soit environ 500 pages de texte.

Fine-tuning (Fine-tuning)

Processus d'ajustement d'un modèle pré-entraîné sur un ensemble de données spécifique pour le spécialiser dans un domaine ou une tâche particulière. C'est comme offrir une formation continue à un employé déjà compétent.

Hallucination (Hallucination)

Génération d'informations fausses, inventées ou incohérentes par un modèle d'IA, présentées avec la même assurance que des faits vérifiés. C'est l'une des limites les plus importantes à connaître.

Exemple : Un modèle peut inventer une citation attribuée à une personne réelle, ou référencer un article scientifique qui n'existe pas.

IAG (Generative AI)

Intelligence artificielle générative. Désigne les systèmes d'IA capables de créer du contenu original (texte, image, code, audio, vidéo) à partir de patterns appris durant l'entraînement.

Inférence (Inference)

Le processus par lequel un modèle entraîné génère une réponse à partir d'une entrée. Chaque fois que vous posez une question à Claude, vous déclenchez une inférence. C'est cette étape qui consomme des ressources de calcul et détermine le coût d'utilisation.

LLM (Large Language Model)

Grand modèle de langage. Réseau de neurones entraîné sur de vastes corpus de texte, composé de milliards de paramètres, capable de comprendre et de générer du langage naturel. Claude, GPT-4 et Gemini sont des LLM.

Modèle multimodal (Multimodal Model)

Modèle capable de traiter et de générer différents types de données : texte, images, audio, vidéo. La tendance actuelle va vers des modèles de plus en plus multimodaux.

Exemple : Claude peut analyser une image et en décrire le contenu en texte.

Paramètres (Parameters)

Les valeurs numériques internes d'un modèle, ajustées durant l'entraînement. Le nombre de paramètres (souvent en milliards) est un indicateur de la taille et de la complexité du modèle. Plus n'est pas toujours synonyme de meilleur.

Prompt (Prompt)

L'instruction ou la question que vous soumettez à un modèle d'IA. La qualité du prompt influence directement la qualité de la réponse. Un bon prompt est clair, précis et fournit suffisamment de contexte.

Prompt engineering (Prompt Engineering)

L'art et la science de formuler des instructions efficaces pour obtenir les meilleurs résultats d'un modèle d'IA. Cela inclut des techniques comme le « chain-of-thought » (demander au modèle de raisonner étape par étape) ou le « few-shot » (donner des exemples).

RAG (Retrieval-Augmented Generation)

Technique qui combine un LLM avec une base de données externe. Le système recherche d'abord de l'information pertinente dans la base, puis la fournit au modèle pour enrichir sa réponse. Cela réduit les hallucinations et permet au modèle d'accéder à des données à jour.

Température (Temperature)

Paramètre qui contrôle le degré de créativité (ou d'aléatoire) dans les réponses du modèle. Une température basse (0) produit des réponses prévisibles et factuelles ; une température élevée (1) génère des réponses plus créatives et variées.

Token (Token)

Unité de base que le modèle traite. Un token correspond approximativement à 3/4 d'un mot en anglais (un peu moins en français). Les tokens servent à mesurer la taille des entrées et des sorties, et déterminent le coût d'utilisation des API.

Exemple : Le mot « intelligence » est généralement découpé en 3 tokens.

Transformer (Transformer)

Architecture de réseau de neurones inventée en 2017 par des chercheurs de Google, qui est à la base de tous les LLM modernes. Son innovation clé — le mécanisme d'attention — permet au modèle de comprendre les relations entre tous les mots d'un texte simultanément.

L'IA ne remplace pas les humains —
elle amplifie ce qu'ils font de mieux.

Sa valeur réside dans la manière dont nous choisissons de l'utiliser — avec discernement, curiosité et responsabilité. Se concentrer sur ce qui compte.

Sébastien Bélisle — Kodra Conseil
seb@kodra.ca · kodra.ca · Avril 2026