

Fonctionnement d'outils d'IAg comme ChatGPT

Les outils d'IA générative (IAg) comme les chatbots ChatGPT, Vibe, Claude, etc. se sont largement installés dans les usages personnels et professionnels. Il apparaît important de faire un focus sur leur fonctionnement global afin de mieux les utiliser.

Les Large Language Models

◆ Qu'est-ce que c'est :

- > Les chatbots conversationnels comme ChatGPT reposent sur des LLM (Large Language Models ou grands modèles de langage). Les LLM sont des modèles d'apprentissage profond (*deep learning*) entraînés sur d'immenses quantités de données dont il est difficile de retracer l'origine.

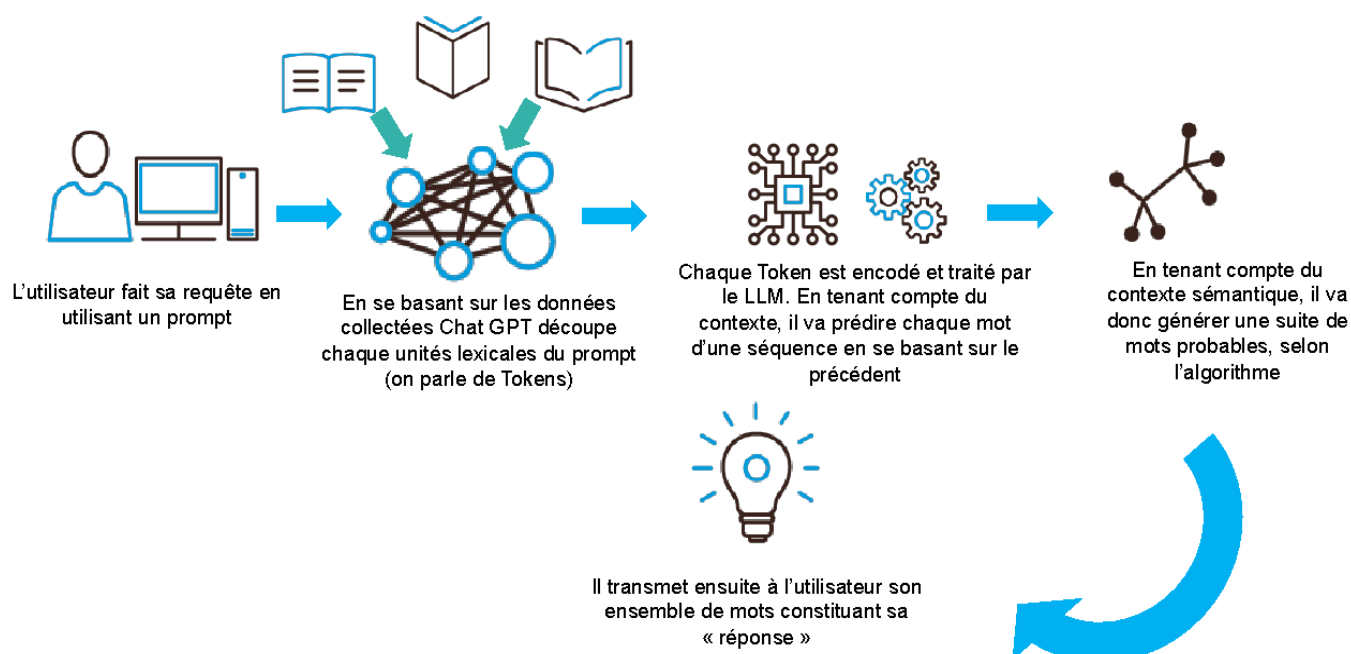
Si vous souhaitez en savoir plus sur le Machine learning et le Deep learning, vous pouvez vous référer à la fiche [Qu'est-ce que l'IA.](#)

- > Ainsi ChatGPT, avec GPT qui signifie « **Generative Pre-trained Transformer** » (« transformeur génératif pré-entraîné »), dispose de milliards de paramètres (réglages internes). On peut imaginer ces paramètres comme des quantités exactes d'ingrédients qu'un chef ajuste pour composer un plat parfait. Ces paramètres permettent au modèle, grâce à son entraînement, de prédire la suite de bouts de mots (*tokens*) la plus crédible en réponse à une requête. Pour mieux comprendre comment cette génération de texte fonctionne, le site [Vittascience](#) montre pour chaque mot généré, le score de probabilité exact qui a déterminé ce choix.

ChatGPT ne « comprend » pas le monde comme un humain. Il peut traiter le sens et le contexte d'une situation et produire une réponse adaptée. Mais il ne ressent pas, ne désire pas, et ne perçoit pas directement les contraintes physiques du monde.

Une réponse générée peut être plausible sans être vraie. Le modèle peut produire des informations erronées ou inventées : on parle alors d'**hallucination**. Ce qui arrive parfois quand les informations nécessaires pour répondre à la requête ne sont pas présentes ou sont sous-représentées (ex : dans les données d'entraînement du modèle).

◆ Le processus :



Pour aller plus loin : [ChatGPT ou la percée des modèles d'IA conversationnels](#)



Il est important de rappeler qu'il est quasiment impossible d'identifier la source exacte des éléments composant la réponse d'un LLM. En revanche, lorsqu'un modèle récupère des informations dans des sources externes (ex : documents, sites internet) avant de générer une réponse (**approche RAG** pour *Retrieval-Augmented Generation*), la source des informations est en général plus transparente. Par exemple, le processus RAG peut être enclenché lorsque l'utilisateur téléverse un fichier ou active la recherche web / recherche approfondie dans ChatGPT. Le LLM peut alors récupérer des informations, soit citer dans sa réponse des extraits de passage de documents sources, soit produire une synthèse avec des références cliquables. Cette synthèse s'appuie à la fois sur les informations récupérées et sur le modèle de langage lui-même.

Par conséquent, **la réponse de ChatGPT dépend de sa connaissance statique (apprise), des informations dynamiques qui lui sont fournies éventuellement via le RAG, et de la qualité du prompt**. Si vous souhaitez parfaire vos prompts vous pouvez vous référer à la fiche [L'art du prompt](#). Pour connaître les bonnes pratiques dans l'utilisation de ChatGPT, vous pouvez vous référer à la fiche [Les bonnes pratiques](#).

Source :

- Amazon Web Service. *Qu'est-ce que la génération à enrichissement contextuel (RAG) ?* <https://aws.amazon.com/fr/what-is/retrieval-augmented-generation/>

Dernière MAJ septembre 2026



université
de BORDEAUX