Fonctionnement de ChatGPT

ChatGPT s'installe de plus en plus dans les usages du quotidien professionnel. Il apparaît important de faire un focus sur son fonctionnement global, afin de vous aider à mieux appréhender cet outil.

Les Large Language Models

◆ Qu'est-ce que c'est :

> Tous les chatbots conversationnels reposent sur les LLM (Large Language Models ou grands modèles de langage). Les LLM sont des réseaux de neurones artificiels profonds entraînés sur d'immenses corpus de textes dont il est difficile de retracer l'origine.

> ChatGPT, développé par la société OpenAI, utilise le Deep Learning (apprentissage profond). Il s'agit d'un type de machine learning qui utilise des algorithmes fonctionnant un peu à la manière d'un cerveau humain. Le Deep Learning utilise un réseau de neurones artificiels capables d'apprendre automatiquement à partir de données.

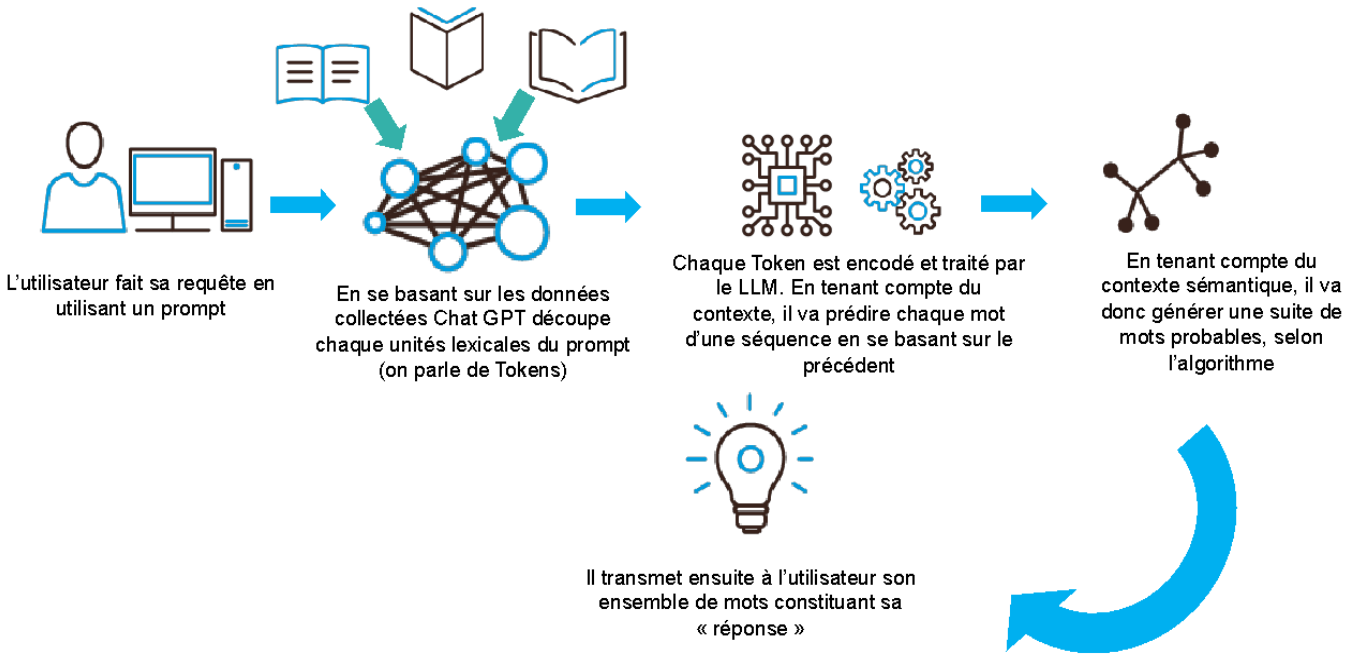
Si vous souhaitez en savoir plus sur le Machine learning et le Deep learning, vous pouvez vous référer à la fiche **Qu'est-ce que l'IA.**

> Ainsi ChatGPT, qui signifie « **Generative Pre-trained Transformer** » (« transformeur génératif pré-entraîné »), dispose de milliards de paramètres. 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 mots la plus crédible en réponse à une requête. Pour mieux comprendre comment cette IA générative de texte fonctionne, le site Vittascience montre pour chaque mot généré, le score de probabilité exact qui a déterminé ce choix.

ChatGPT et les LLMs ne comprennent ni le sens des questions qui leur sont posées ni le sens de leurs « réponses ».

Si les informations nécessaires pour répondre à la requête ne sont pas présentes (ou sont sous-représentées) dans leurs données d'entraînement, ils peuvent inventer des éléments de réponse. On parle alors d'**hallucination**.

◆ 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 modèle de langage classique. En revanche, lorsqu'un modèle utilise une **approche RAG** (Retrieval-Augmented Generation), il peut consulter des documents ou des sources externes pour améliorer ses réponses, rendant la source des informations plus transparente. Par exemple, le processus RAG est enclenché lorsque l'utilisateur téléverse un fichier ou active la recherche web / recherche approfondie dans ChatGPT. Le LLM peut alors soit citer dans sa réponse des extraits de passage de documents sources, soit produire une synthèse. Cette synthèse s'appuie à la fois sur les informations récupérées et sur le modèle de langage lui-même, ce qui peut compliquer le traçage exact de chaque mot jusqu'à la source.

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/