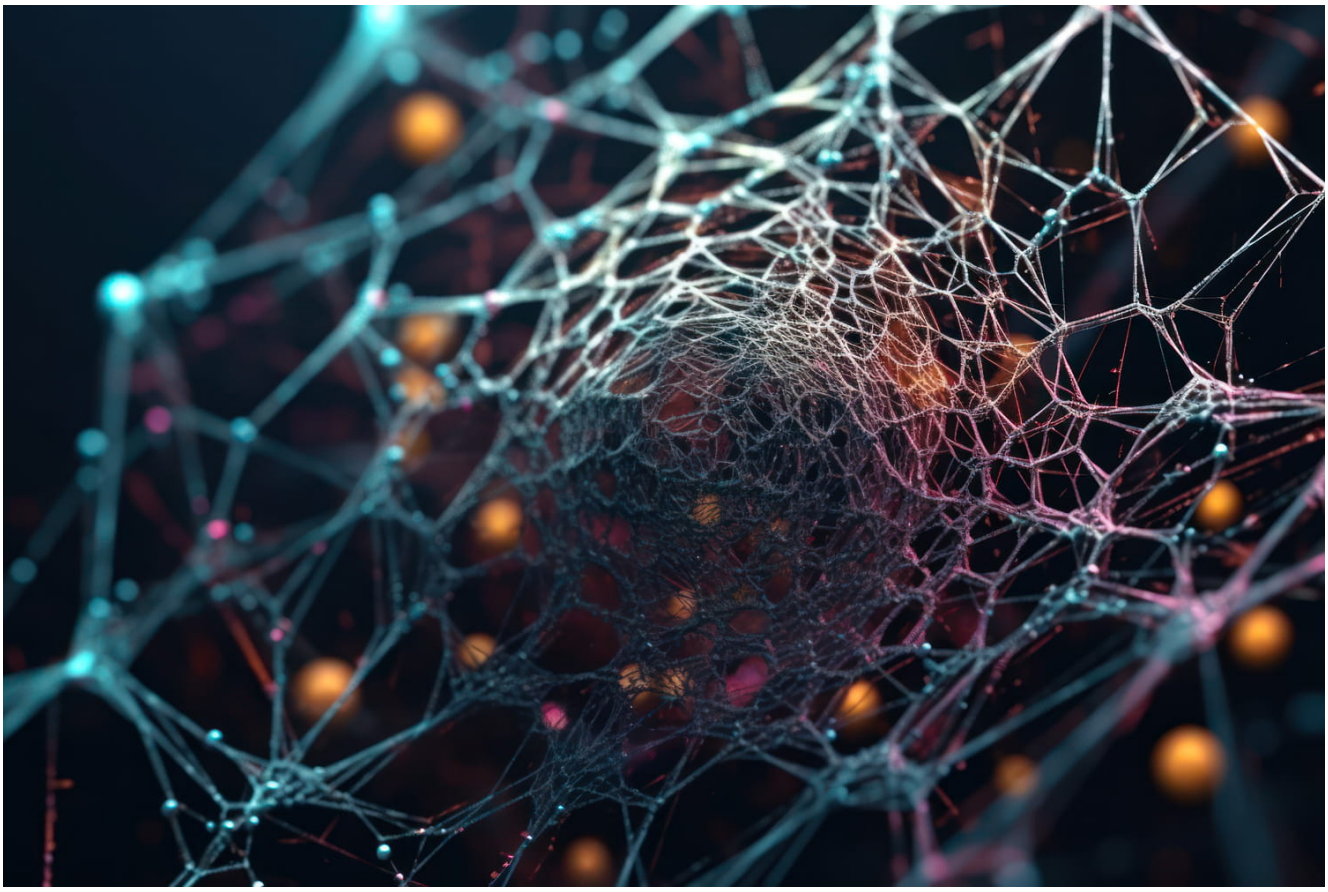
Démystifier le LLM en seulement 5 minutes : Comprendre son fonctionnement en profondeur !



GPT-4, Llama 2, Claude 2… Les large language models pullulent depuis 2022. Tour d'horizon de la technologie sous-jacente et de son principe de fonctionnement.

Porté par OpenAI et Microsoft, ChatGPT a démocratisé les LLM. Dans le sillage de sa sortie fin 2022, la concurrence s'est mise en ordre de bataille. Google a lancé Bard. Meta a sorti Llama 2 en open source. Amazon Web Service a déployé Bedrock. Le pure player Anthropic a mis en ligne Claude 2. Le français Mistral a dévoilé Mistral 7B. Désormais, les LLM font partie du paysage numérique. Mais comment fonctionne-t-il ? Le point.

Qu’est-ce qu’un LLM ?

Un LLM pour large language model, ou modèle massif de langage en français, est une architecture de réseau de neurones artificiel reposant sur l’infrastructure dite des transformers. Taillée pour le traitement automatique des langues (TAL) ou natural language processing (NLP), cette technologie de deep learning a été initialement développée par Google, qui l’a publiée en open source en 2017.

Comment fonctionne un transformer ?

Comme un réseau de neurones récurrents (RNN), un transformer est taillé pour ingérer des données séquentielles. Schématiquement, il permet à la machine d’apprendre des séquences informatiques de manière automatique, sans avoir été programmé spécifiquement à cet effet. Le transformer est par conséquent bien adapté au traitement de suites de mots, et donc des langues.

A la différence d’un RNN, un transformer n’implique pas cependant de traiter les informations sous forme de flux continu, en respectant par exemple l’ordre des mots dans une phrase. Partant de là, un modèle de ce type peut paralléliser les calculs de la phase d’entrainement. Ce qui lui permet d’ingérer des volumes massifs de données d’apprentissage en un temps réduit.

Partant de là, quels sont les cas d’usage des LLM ?

Traitant les données de manière séquentielle, les LLM sont utilisés historiquement pour la traduction et la synthèse de texte. Ils sont par exemple utilisés par les traducteurs en ligne pour traiter de manière automatique du langage naturel. Contrairement aux anciens traducteurs en ligne utilisant des

RNN, les traducteurs modernes basés sur des transformers ont la capacité de lier les mots entre eux (notion d'interdépendance). Cela leur permet notamment d'obtenir des tournures de phrase bien plus proches du langage écrit ou parlé, et de donner le bon sens à un mot qui peut en avoir plusieurs.

Les transformeurs peuvent être utilisés dans d'autres domaines comme le traitement d'images. Ils peuvent également combiner plusieurs types de média. ChatGPT est un excellent exemple de transformer multimodal. Il permet d'intégrer à ses invites de commande à la fois du texte, de l'image et du son.

Quelle différence entre ChatGPT et le LLM GPT ?

GPT n'est que la couche de LLM de ChatGPT. Ce dernier se découpe au total en cinq couches. En amont, il fait appel à une couche d'embendding non-supervisé pour vectoriser les mots. Ensuite seulement vient l'apprentissage auto-supervisé pour le traitement du langage. C'est là que se situe la technologie des transformers et le LLM en tant que tel.

A ces deux premières couches s'ajoute encore un mode d'entrainement supervisé qui permet d'apprendre au bot à répondre aux questions sur la base de grands ensembles de données labélisées. L'objectif ? Permettre non seulement d'aligner des mots qui ont un sens (ce qui est la mission des transformers), mais aussi de gérer des scénarios de plus haut niveau : répondre à une question, converser en mode chatbot, résumer un texte… En aval, l'apprentissage par renforcement entre dans la danse. Il consiste à soumettre les réponses fournies par ChatGPT à des experts humains qui leur attribuent une note. Sur la base de cette notation, le modèle affine la pertinence de ses résultats.

Comment se définit la performance d’un LLM ?

Historiquement, la performance d’un LLM se définit au regard de son volume de paramètres. Ces derniers sont représentés par les connexions entre les différentes couches du réseau de neurones, et par les poids attribués par l’algorithme à ces derniers. Dans sa dernière itération, le LLM de ChatGPT (GPT-4) compte 1,7 trillion de paramètres. C’est l’un des plus vastes LLM jamais créé à ce jour. Revers de la médaille : en fonction de leur poids, les LLM peuvent se révéler complexes à appréhender et coûteux à réentraîner sur des données spécifiques. Il est donc essentiel de tenir compte de la taille et de la capacité d’un LLM afin de décider de la meilleure façon de l’utiliser.

Mais le nombre de paramètres n’est pas le seul critère de performance d’un LLM. L’architecture globale du LLM a aussi un rôle à jouer dans ce domaine. Pour preuve : le large language model Claude 2 édité par Anthropic (et principal concurrent de
Source : journaldunet.com