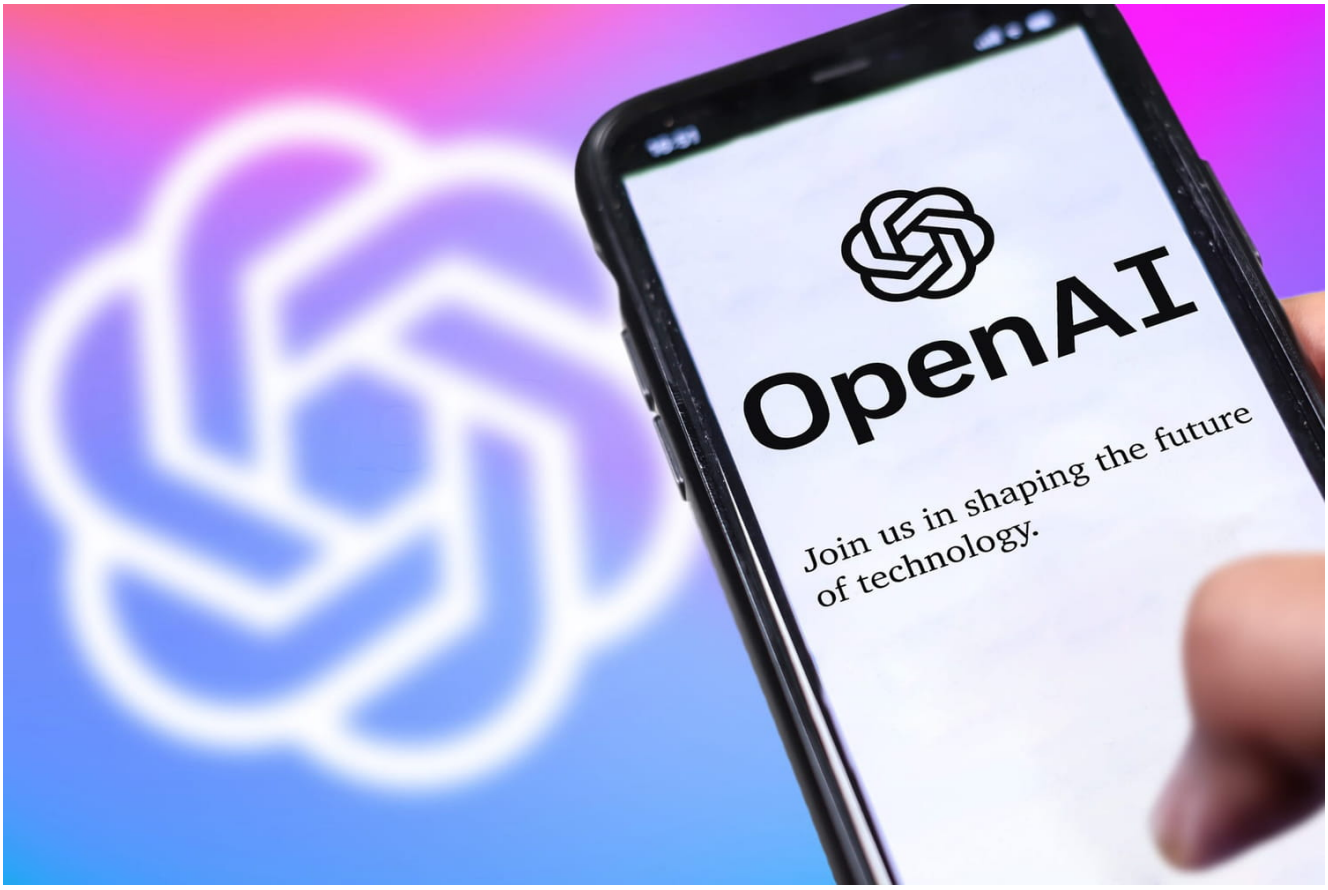
Les secrets dévoilés : L'envers fascinant d'une technologie colossale



Comment fonctionne l'IA générative d'OpenAI ? Quelles sont ses différentes briques ? Plongeon au cœur d'un projet qui a changé la face de l'intelligence artificielle.

Assistant informationnel, génération de texte, traduction, résumé, analyse de données, codage d'application… Les fonctionnalités de ChatGPT sont nombreuses et ne cessent de surprendre par leur acuité. Des milliers d'articles ont été publiés sur le sujet. Reste à savoir sur quelle architecture repose le service d'OpenAI et quelles sont les briques d'IA mises en œuvre pour la faire tourner.

D'abord, ChatGPT s'adosse globalement à une architecture de

deep learning ou apprentissage profond. Elle s'articule autour d'un réseau de neurones artificiels de toute dernière génération. Initialement lancé sur la base de GPT-3.5 en novembre dernier, ChatGPT est désormais basé sur GPT-4, le dernier-né des large language models d'OpenAI. Les principaux apports de GPT-4 : il analyse également les images (ce qui le rend multimodal) tout en étant capable de glaner des réponses sur le web. Ses capacités vont donc bien au-delà de la précédente version dont la base d'apprentissage ne dépassait pas mi-2021.

300 milliards de mots ingérés

Comment ChatGPT a-t-il été entraîné ? Dès sa première itération, l'assistant d'OpenAI a ingéré un corpus de 300 milliards de mots. Un corpus composé de Wikipedia, de nombreux livres, de contenus issus des réseaux sociaux, et d'autres sources accessibles publiquement. Comme le traduit l'acronyme GPT (pour generative pre-training transformer), GPT est modèle génératif de langage basé sur l'architecture des transformers, un type de réseau de neurones profonds conçu pour ingérer des données d'apprentissage séquentielles en utilisant des mécanismes d'attention.

Les transformers ont été conceptualisés à l'origine par Google en 2017. Schématiquement, ils permettent à la machine d'apprendre des séquences informatiques de manière automatique, sans avoir été programmés spécifiquement à cet effet. Ils sont par conséquent bien adaptés au traitement de suites de mots, et donc des langues. D'où le choix d'OpenAI de partir sur cette architecture.

Mais ChatGPT ne s'adosse pas seulement aux transformers. 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. 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, mais aussi de gérer des scénarios de plus haut niveau : répondre à une question, converser en mode chatbot, résumé un texte…

> *"C'est principalement pour ses capacités de calcul haute performance ultra-haut débit qu'OpenAI a opté pour le cloud de Microsoft"*

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.

Pour entrainer ChatGPT, OpenAI capitalise sur son partenariat avec Microsoft. L'ensemble du processus d'apprentissage de l'assistant repose sur le cloud Azure du groupe de Satya Nadella. Le supercalculateur Azure d'OpenAI est gigantesque. Il comprend 285 000 cœurs de CPU AMD, et 10 000 processeurs GPU Nvidia V100 Tensor. Chacune des deux infrastructures repose sur un réseau ultra haut débit. "Les liaisons InfiniBand fournissent 400 gigabits par GPU, soit un total de 3,2 téraoctets par serveur. (…) C'est principalement pour ses capacités de calcul haute performance ultra-haut débit qu'OpenAI a opté pour le cloud de Microsoft", précise Mark Russinovich, CTO d'Azure (lire le détail dans ce post de l'éditeur).

700 000 dollars par jour

OpenAI s'adosse à l'infrastructure de machine learning distribué DeepSpeed de Microsoft pour paralléliser les calculs, et ainsi en optimiser le temps d'exécution. De très

nombreuses instances du modèle sont entrainées sur de petits lots de data d'apprentissage en les répartissant sur de très nombreux GPU. "C'est pourquoi vous avez besoin d'un système aussi grands", résume Mark Russinovich. "Issu de la collaboration entre l'ingénierie Microsoft, Microsoft Research, l'organisation OpenAI et Nvidia, ce système que nous avons construit en 2020 était à l'époque le cinquième plus grand supercalculateur au monde et le plus grand supercalculateur jamais construit dans le cloud public."

Pour fonctionner et répondre aux requêtes des utilisateurs, ChatGPT requiert 28 936 GPU. Ce qui correspond à un coût de près de 700 000 dollars par jour, selon une estimation du cabinet d'études SemiAnalysis (lire le post).