SYNCICAP ASSET MANAGEMENT

MARKET INSIGHTS

26 JUILLET 2024

Un aperçu du monde de l'Intelligence Artificielle (IA)

Yixin GUO
Analyste actions
SYNCICAP AM

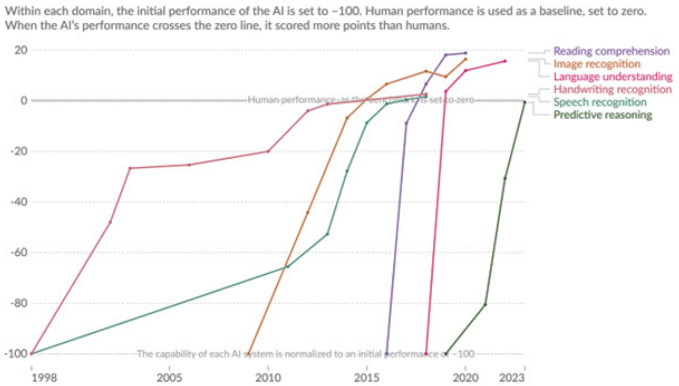
L'IA est-elle une bulle ? Voilà la question à mille milliards de dollars à laquelle il faudra répondre au milieu de l'année 2024. Mais où en sommes-nous dans ce cycle technologique axé sur l'IA ? Nous allons voir dans cet article comment l'IA va fondamentalement remodeler notre travail et notre vie, pourquoi nous sommes encore au début de ce mouvement technologique sur plusieurs décennies et pourquoi l'Asie est un élément clé de cette chaîne d'approvisionnement globale.

Tout d'abord, nous allons récapituler brièvement les principales étapes du développement de l'IA depuis le lancement de ChatGPT en novembre 2022. Nous tenterons ensuite de prédire certaines tendances futures du secteur. Enfin, nous présenterons nos valeurs préférées du secteur de l'IA sur les marchés asiatiques.

LOI DE MISE À L'ÉCHELLE ET LOI DE HUANG – AUGMENTATION DES CAPACITÉS DES MODÈLES PARALLÈLEMENT À LA RÉDUCTION DU COÛT DU MATÉRIEL

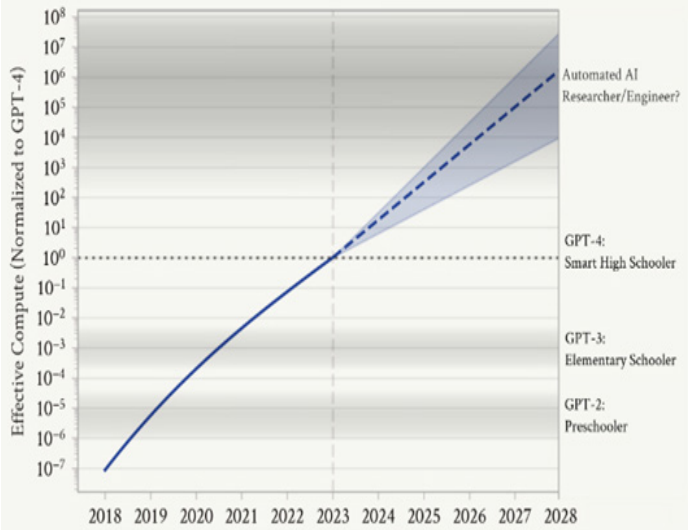
Dans un premier temps, nous avons progressé rapidement dans notre quête de l'AGI (Artificial General Intelligence). En 2019, le modèle le plus performant dont nous disposions était le GPT-2, qui pouvait à peine produire quelques phrases cohérentes. Seulement 5 ans plus tard, en 2024, notre meilleur modèle actuel, GPT-4, peut déjà produire de longues séquences de code et résoudre des problèmes mathématiques complexes. L'élément magique à l'origine de cet incroyable bond en avant des performances des modèles est **la loi de mise à l'échelle,** qui stipule que si nous utilisons plusieurs ordres de grandeur de données et de ressources informatiques supplémentaires pour entraîner notre modèle, nous pouvons nous attendre à une amélioration significative de ses capacités. Aschenbrenner, ancien chercheur de l'OpenAI, a prédit que les modèles seront capables de faire le travail d'un chercheur ou d'un ingénieur en IA d'ici à 2027. La poursuite de l'intelligence artificielle a incité les géants de la technologie disposant d'importantes réserves de liquidités, tels que Microsoft, Google et Meta, à construire des clusters d'entraînement d'une taille sans précédent, donnant lieu à une compétition effrénée pour s'approcher du Saint-Graal que représente l'AGI. Les GPU de Nvidia, qui sont au cœur de ces centres de données, offrent les meilleures performances et la meilleure efficacité énergétique de leur catégorie.

Résultats des tests des systèmes d'IA sur différentes capacités par rapport aux performances humaines



Source : Our World in Data, fin 2023

Augmentation de la base de calcul effective



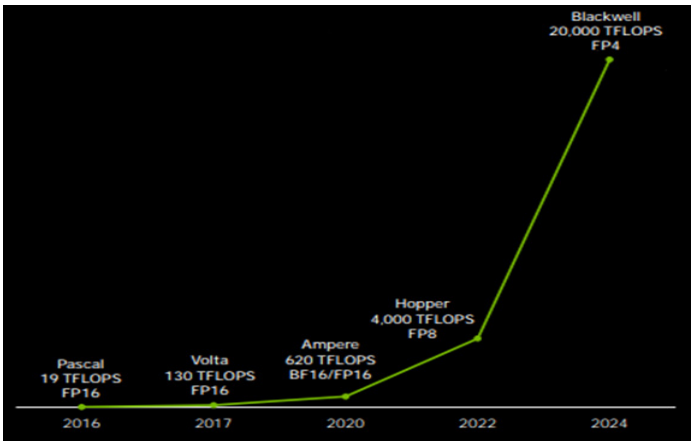
Source : Blog de Leopold Aschenbrenner, 2024

Le déploiement des systèmes d'IA se fait principalement en deux phases. Tout d'abord, une formation est nécessaire pour obtenir un modèle pouvant être utilisé à différentes fins. Ensuite, l'inférence est réalisée à chaque fois que l'on accède au modèle formé. Ainsi, chaque fois que nous faisons une demande à ChatGPT, le modèle effectue une tâche d'inférence.

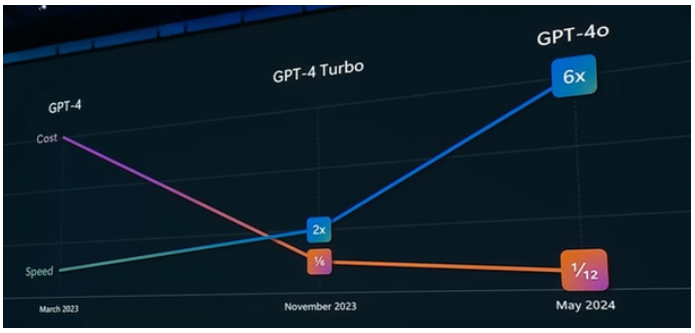
Le coût total de la **formation des modèles** de dernière génération ne cesse de progresser, sachant que la loi de mise à l'échelle implique une augmentation rapide en termes de quantités de paramètres et de données utilisées et que la formation des modèles de la génération suivante nécessitera une quantité croissante de ressources de calcul (GPU, réseau…). Ainsi, 8 000 GPU Hopper de Nvidia sont nécessaires pour entraîner GPT-4, le tout dernier modèle. En revanche, l'entraînement de GPT-5 nécessite environ 100 000 GPU Hopper, ce qui implique de multiplier par 10 le coût du matériel. Si l'infrastructure actuelle se limite principalement à l'adoption de la solution de Nvidia, à l'avenir, à mesure que d'autres fournisseurs de GPU ou d'ASIC (circuits intégrés spécifiques à une application) rivaliseront avec Nvidia, il sera possible d'améliorer le coût du matériel.

En termes d'inférence, le coût diminue rapidement. Pour le GPT-4, il ne représente aujourd'hui que le douzième du coût d'il y a un an, ce qui s'explique à la fois par l'amélioration des performances matérielles et par des optimisations logicielles. La **loi de Huang,** du nom de Jensen Huang, indique que la puissance de calcul de chaque puce GPU a doublé chaque année, ce qui signifie qu'elle a été multipliée par 1 000 en l'espace d'une décennie, l'élément déterminant étant l'utilisation de nombres de moindre précision pour effectuer les calculs. Ainsi, des techniques logicielles telles que la quantification (représentation des paramètres du modèle avec des chiffres de moindre précision) peuvent tirer parti des gains en termes de matériel pour réduire le coût de l'inférence. En conséquence, alors que le coût de formation du dernier modèle est en constante augmentation, le coût d'inférence va diminuer progressivement, ce qui ouvre la voie à une adoption plus large du LLM.

Capacités de calcul de l'IA x 1 000 en 8 ans



Source : Nvidia GTC, 2024



Source : Microsoft Build, 2024

NOUS SOMMES MOINS PRÉOCCUPÉS PAR LA SUFFISANCE DES DONNÉES

La taille des modèles augmentant rapidement, beaucoup commencent à se demander si nous disposons de suffisamment de données pour former les futurs modèles. Nous avons déjà utilisé à des fins de formation toutes les données textuelles publiques, qui ont constitué la plus importante source de données. Il existe cependant encore de nombreuses méthodes permettant d'**extraire davantage de données** pour répondre aux besoins de la loi de mise à l'échelle. Les entreprises conservent leurs **données propriétaires** dans leurs bases, de sorte que des entreprises modèles comme OpenAI négocient activement avec des entreprises privées pour utiliser leurs données. En outre, nous pouvons utiliser les anciens modèles (par exemple GPT-3.5) pour produire des données permettant d'entraîner les futurs modèles, appelées **données synthétiques.** Cela s'apparente à la manière dont AlphaZero, le modèle d'IA qui a atteint un niveau surhumain au jeu de Go, est devenu de manière itérative meilleur au jeu en jouant contre lui-même. Afin d'imiter les humains qui peuvent apprendre des concepts complexes avec seulement une poignée d'exemples, des techniques neuromorphiques telles que le **replay d'expérience** sont également employées, ce qui permet au modèle de développer plus d'idées à partir du même ensemble de données. Ainsi, même si la suffisance des données reste un sujet controversé, nous sommes moins préoccupés par le problème, du moins pour les 2 ou 3 prochaines années.

L'ABONDANCE DES APPLICATIONS DE L'AI REMODÈLE NOTRE TRAVAIL ET NOTRE VIE

Actuellement, l'IA générative est déjà en train de remodeler nos habitudes de travail et de vie. Les géants d'internet utilisent des modèles de recommandation basés sur l'apprentissage profond (DLRM) pour suggérer des contenus plus pertinents aux utilisateurs, tels que publicités ou posts sur les médias sociaux. Les entreprises utilisent les LLM pour aider à la réalisation de tâches fastidieuses telles que la traduction ou la synthèse de textes. Les programmeurs utilisent des copilotes de codage pour accélérer le développement d'applications, l'IA aidant à détecter les bugs et à suggérer des améliorations. À l'avenir, l'amélioration des performances de l'IA, associée à la réduction des coûts d'inférence, ouvrira sans aucun doute de nouveaux marchés dont nous n'avons encore aucune idée, de même que l'essor d'internet a donné naissance à des plateformes de commerce électronique. L'IA générative sera également l'un des principaux moteurs de la croissance économique. En effet, Goldman Sachs estime que la technologie sera à l'origine d'une augmentation de 0,4 % de la croissance du PIB aux États-Unis, de 0,2 % à 0,4 % dans les autres marchés développés et de 0,1 % à 0,2 % dans les marchés émergents avancés au cours de la prochaine décennie.

L'IA DÉVELOPPE LA COMPRÉHENSION ET LA GÉNÉRATION MULTIMODALES

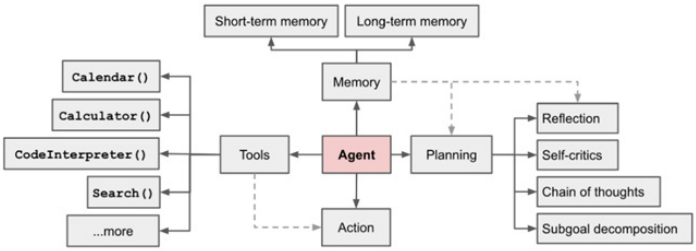
Comment l'IA va-t-elle poursuivre son évolution ? La première étape consistera à **doter l'IA d'un plus grand nombre d'organes sensoriels.** Nous n'entraînons pas seulement les modèles à partir de textes, mais aussi de diverses formes de données, telles que des images, des sons et des vidéos, afin de rendre le modèle plus performant. Comme l'a suggéré Andrew Parker, un zoologiste, l'émergence de la vision lors de l'explosion cambrienne a été cruciale pour les premiers animaux, non seulement pour trouver de la nourriture et éviter les prédateurs, mais aussi pour évoluer et s'améliorer. De même, le fait de permettre à l'IA d'appréhender des données au-delà du texte brut devrait favoriser de nouvelles avancées. Le psychologue suisse Jean Piaget a publié une théorie en 1952, selon laquelle les enfants développent leurs capacités cognitives à l'aide de la vue, de l'ouïe, du toucher, du goût et de l'odorat, et par le biais d'interactions avec le monde physique. Bien que les capacités de perception sensorielles des systèmes d'IA actuels soient encore limitées, le développement de robots humanoïdes alimentés par des LLM permettra probablement d'accroître et d'enrichir les interactions entre les modèles et le monde physique, ce qui nous rend optimistes sur le potentiel d'intelligence à long terme de l'IA.

L'IA DÉVELOPPE LES CAPACITÉS DE RAISONNEMENT PAS À PAS DU SYSTÈME 2

Nous explorons également les manières de permettre aux modèles de penser davantage comme des humains. Ainsi, les chercheurs intègrent des méthodes telles que le « chain of thought prompting », qui suggère une séquence de pensées afin de permettre aux modèles de penser étape par étape lorsqu'ils résolvent un problème complexe. Cette méthode est analogue à la pensée du système 2, qui se réfère à un processus cognitif lent et délibéré par opposition à la pensée du système 1, qui est une pensée rapide, automatique et intuitive (davantage similaire à la façon dont les LLM fonctionnent aujourd'hui). Les chercheurs ont constaté que l'utilisation de chaînes de pensées permet d'améliorer les performances des modèles en matière de raisonnement arithmétique, symbolique et de bon sens.

L'AGENT AI COMME VÉHICULE PROMETTEUR POUR LA POURSUITE DE L'AGI

La baisse du coût de l'informatique, les progrès en matière de compréhension et de génération multimodales, les architectures de modèles plus efficaces et l'amélioration des capacités de raisonnement et de planification contribueront tous au développement **d'agents d'IA** avancés. L'agent IA est considéré comme un véhicule prometteur pour la recherche d'une intelligence artificielle de niveau humain. Lilian Weng, d'Open AI, décrit un agent d'IA doté de LLM comme un système qui utilise ces LLM comme cerveau, disposant d'une mémoire à court terme et d'une mémoire à long terme, capable de planifier en plusieurs étapes et d'utiliser des outils tels que les appels d'API. Certaines démonstrations préliminaires, telles que Devin, montrent le potentiel de l'agent IA, qui est conçu pour résoudre individuellement des tâches de codage complexes. Les chercheurs développent également des systèmes multi-agents, car ils ont découvert qu'un agent spécialisé dans un domaine de tâches étroit est plus performant que les agents généralistes, ce qui est analogue au concept de « division du travail » évoqué pour la première fois par Adam Smith. À l'avenir, nous pensons être en mesure d'envisager une coexistence harmonieuse entre les systèmes multi-agents et les humains.



Source : Blog de Lilian Weng 2023

L'IA, MOTEUR DE LA TRANSFORMATION DE LA ROBOTIQUE

L'IA est aussi à l'origine d'un changement transformationnel dans le domaine de la robotique, qui débloquera des ressources illimitées et des gains de productivité, créant ainsi un changement radical dans l'économie mondiale. Ainsi, les chercheurs ont doté les robots de capacités de vision, de planification et d'interaction avec les humains, par le biais des LLM. Une équipe de Google a développé un projet dans lequel un robot peut agir en raisonnant à partir des paramètres de son environnement présent et selon ses capacités. Lorsqu'on lui demande « J'ai renversé mon coca, peux-tu m'aider ? », le robot évalue les outils disponibles autour de lui qui peuvent être utiles et les actions qu'il peut éventuellement entreprendre, et peut par exemple choisir une éponge plutôt qu'un objet non pertinent.

UN MONDE HÉTÉROGÈNE OÙ COEXISTENT DE GRANDS ET DE PETITS MODÈLES

Les modèles doivent être commercialisés et transformés en produits. Certains fabriquent **des modèles linguistiques spécialisés de petite taille,** adaptés à des tâches ou à des flux de travail spécifiques. Ainsi, Apple utilise un modèle doté d'environ 3 milliards de paramètres, spécialisé pour des tâches spécifiques telles que la synthèse, la relecture d'épreuve ou la production de réponses à des e-mails. Les techniques permettant de transformer les grands modèles en modèles de plus petite taille, telles que la distillation ou l'élagage, ont beaucoup évolué. Nous voyons l'avenir comme un monde hétérogène où les chercheurs continueront de développer des modèles plus compétents basés sur le Cloud avec des comportements davantage « émergents », tandis que des modèles plus petits basés sur l'Edge AI (intelligence artificielle en périphérie du réseau), ayant une meilleure sécurité et une latence plus faible, seront déployés sur nos ordinateurs et nos appareils mobiles.

NOUS SOMMES OPTIMISTES AU SUJET DE L'IA GÉNÉRATIVE

De nombreux chercheurs estiment que les réseaux neuronaux numériques sont conceptuellement analogues aux cerveaux biologiques. Les paramètres des modèles sont analogues aux synapses du cerveau, c'est-à-dire à la connexion entre les neurones. Un cerveau humain compte 100 000 milliards de synapses. Plus les synapses sont nombreuses, plus l'intelligence est élevée. Aujourd'hui, nous estimons que le GPT-4 compte 1,8 trillion de paramètres, sachant que nous nous orientons vers un objectif de 100 trillions pour tester la limite de la loi de mise à l'échelle.

En conclusion, même si le scepticisme persiste quant à la durabilité de la loi de mise à l'échelle et au calendrier de commercialisation de l'IA générative, nous restons optimistes quant à la tendance générale. Le coût du matériel continuera à diminuer en vertu de la loi de Huang, ce qui permettra une adoption plus large de l'IA générative. En outre, des avancées logicielles telles que la formation avec davantage de modalités de données ou la possibilité pour le modèle de s'autoformer permettront de débloquer davantage de capacités. Les progrès de l'IA physique pourraient être le chaînon manquant dans la réalisation de l'AGI, les robots s'intégrant de manière transparente dans notre société pour résoudre des tâches plus complexes.

TAÏWAN, CORÉE ET JAPON SONT AU CŒUR DE LA CHAÎNE D'APPROVISIONNEMENT

Comment participer à ce voyage dans l'IA ? De nombreux investisseurs pourraient penser qu'ils ne peuvent investir que dans des entreprises américaines pour en bénéficier, mais il existe de nombreuses entreprises asiatiques qui contribuent au développement de l'IA. **Taïwan,** la **Corée** et le **Japon** sont au cœur de la chaîne d'approvisionnement. Voici quelques exemples.

TSMC (2330-TW) détient une position de monopole de fait dans la fabrication des puces GPU de Nvidia, des puces en silicium sur mesure destinées aux fournisseurs américains de Cloud, et des puces de mise en réseau. En outre, la société permet d'élargir le thème de l'IA en intégrant ces puces aux téléphones mobiles et aux PC, où les applications de l'IA seront disponibles.

SK Hynix (000660-KR) est le plus grand fournisseur au monde de puces mémoires à large bande passante (HBM), spécialisées pour répondre aux exigences de bande passante élevée de la formation et de l'inférence pour la réalisation des modèles d'IA.

Disco (6146-JP) est le leader mondial des outils de broyage et de découpage utilisés dans la fabrication des puces à semi-conducteurs. Les machines de Disco jouent un rôle essentiel dans la fabrication de ces HBM hautement sophistiqués, qui nécessitent le broyage des puces à un niveau ultra-mince. Seules les machines capables d'assurer une grande précision et une grande planéité sont en mesure de le faire. Disco, avec ses 50 ans d'histoire dans l'équipement de précision pour semi-conducteurs, est un acteur dominant dans la fabrication de HBM.

SYNCICAP ASSET MANAGEMENT

Syncicap AM est une société de gestion détenue par le groupe Ofi Invest (66 %) et Degroof Petercam Asset Management (34 %), agréée le 4 octobre 2021 par la Securities and Futures Commission de Hong Kong. Cette société, spécialisée dans les pays émergents, permet d'établir une présence en Asie, depuis Hong Kong. Elle gère également une gamme de fonds émergents proposés aux investisseurs européens par Ofi Invest Asset Management.

Les chiffres des performances citées ont trait aux années écoulées. Les performances passées ne sont pas un indicateur fiable des performances futures.
Ces placements permettent de profiter du potentiel de performance des marchés financiers en contrepartie d'une certaine prise de risque. Le capital investi et les performances ne sont pas garantis et il existe un risque de perte en capital. Source des indices cités : www.bloomberg.com

www.syncicap.com
2801-02, 28/F, Man Yee Building, 68 Des Voeux Road Central, Central, Hong Kong • Tél.: +852 3842 8204

SYNCICAP ASSET MANAGEMENT | Ofi invest