sycomore am

JANVIER 2025

2025 SERA L'ANNÉE DE L'IA AVEC UN GRAND A

L'intelligence artificielle (IA) a beau alimenter les conversations depuis plus de deux ans, beaucoup d'investisseurs et d'observateurs se demandent encore quand cette technologie passera réellement à la vitesse supérieure. Les avancées successives que nous avons connues ces dernières années – notamment les « large language models » (LLMs) tels que GPT-4 – ont fait naître un engouement spectaculaire, mais aussi un certain scepticisme quant à la poursuite de cette dynamique à long terme.

Pourtant, à l'aube de 2025, un tournant majeur se profile : l'arrivée des « modèles de raisonnement » o1 et o3 d'OpenAI, annoncée fin 2024, devrait intensifier l'adoption de l'intelligence artificielle et enclencher une nouvelle phase d'investissements, faisant de 2025 une année décisive pour l'IA.

Par David Rainville, Gérant du fonds Sycomore Sustainable Tech

DE L'EFFERVESCENCE À LA CONSOLIDATION : POURQUOI 2025 ?

En 2023 et 2024, on a observé un engouement marqué pour l'IA, sans pour autant atteindre l'envergure d'autres cycles technologiques passés. De nombreux observateurs et investisseurs se sont penchés sur les « modèles de fondation » (ou « foundation models »), tels que GPT-4. Ils sont reconnus pour leur capacité à générer rapidement des réponses sur divers sujets ou à résumer rapidement des textes très longs et/ou complexes, permettant ainsi des gains de productivité, mais uniquement sur quelques cas d'usage.

L'adoption de masse des technologies LLMs a été assez limitée jusqu'à aujourd'hui. Les investisseurs comptaient sur le lancement de nouveaux modèles de fondation, comme GPT-5, pour augmenter le nombre de cas d'usage et conduire à une adoption rapide. Mais la date de sortie de GPT-5 n'est toujours pas connue. En cause ? Des progrès plus lents qu'attendus en raison d'un manque de nouvelles données de qualité pour entraîner des modèles plus intelligents.

L'emballement - y compris médiatique - des deux dernières années a laissé la place à de nouvelles interrogations : et si on avait épuisé toutes les données de qualité disponibles pour l'entraînement de nouveaux modèles ? Les modèles LLMs ont-ils atteint leur limite technique ? Les dépenses liées à l'IA vont-elles s'essouffler ?

D'après nous, en 2025, ces incertitudes vont se dissiper pour deux raisons principales :

1. L'arrivée des modèles de raisonnement o1 et o3, annoncés fin 2024, qui promettent d'apporter des capacités de réflexion et de résolution de problèmes bien supérieures à celles des modèles de fondation d'aujourd'hui. Cela devrait conduire à une forte adoption commerciale des technologies d'IA génératives.

« Les dépenses d'IA pourraient dépasser 250 milliards de dollars en 2025 et franchir la barre des 300 milliards de dollars dès 2026 »

2. La réaccélération de la demande en semi-conducteurs et en puces mémoires liée à ces nouveaux usages, et à des modèles de raisonnement qui se révèlent bien plus gourmands en puissance de calcul et en mémoire que les générations précédentes.

Ces deux facteurs combinés laissent entrevoir une croissance soutenue des dépenses d'IA, qui pourraient dépasser 250 milliards de dollars en 2025 et franchir la barre des 300 milliards de dollars dès 2026, d'après nos analyses.

QU'EST-CE QU'UN « MODÈLE DE RAISONNEMENT » ?

Pour comprendre en quoi les modèles de raisonnement constituent un point de bascule, il convient de distinguer ces nouveaux venus des modèles de fondation (LLMs) comme GPT-4.

Les modèles de fondation (LLMs) sont comparables à d'immenses bibliothèques de connaissances. En leur posant une question, l'utilisateur obtient une réponse quasi instantanée, puisée dans la masse d'informations déjà ingurgitée par le modèle lors de son entraînement. GPT-4 ou d'autres LLMs comme PaLM et LLaMA illustrent bien cette catégorie. Ces modèles sont très bons pour restituer et résumer l'information qu'on leur donne.

Les modèles de raisonnement sont construits sur la même fondation que les LLMs (les technologies de 'transformer'[1]) mais ils vont bien plus loin en apportant une "couche d'expertise" capable de raisonner. Concrètement, au lieu de se contenter de puiser dans leur base de connaissances, les modèles de raisonnement (tels que o1 et o3) décomposent la question en plusieurs étapes, élaborent une logique de résolution et l'exécutent pour obtenir une solution, même lorsqu'il s'agit d'un problème complexe ou potentiellement inédit.

Avec un quotient intellectuel estimé à 140 pour o1 Pro et de près de 160 pour o3, ces modèles s'approchent du seuil de ce que l'on appelle « AGI » (Artificial General Intelligence). Le modèle o3 se classerait même parmi les 200 meilleurs codeurs au monde, surpassant certains chercheurs chevronnés (y compris le chef de la recherche d'OpenAI). Ce saut qualitatif bouleverse la donne pour l'adoption de technologies d'IA générative, car il ouvre la porte à de nouveaux cas d'usage.

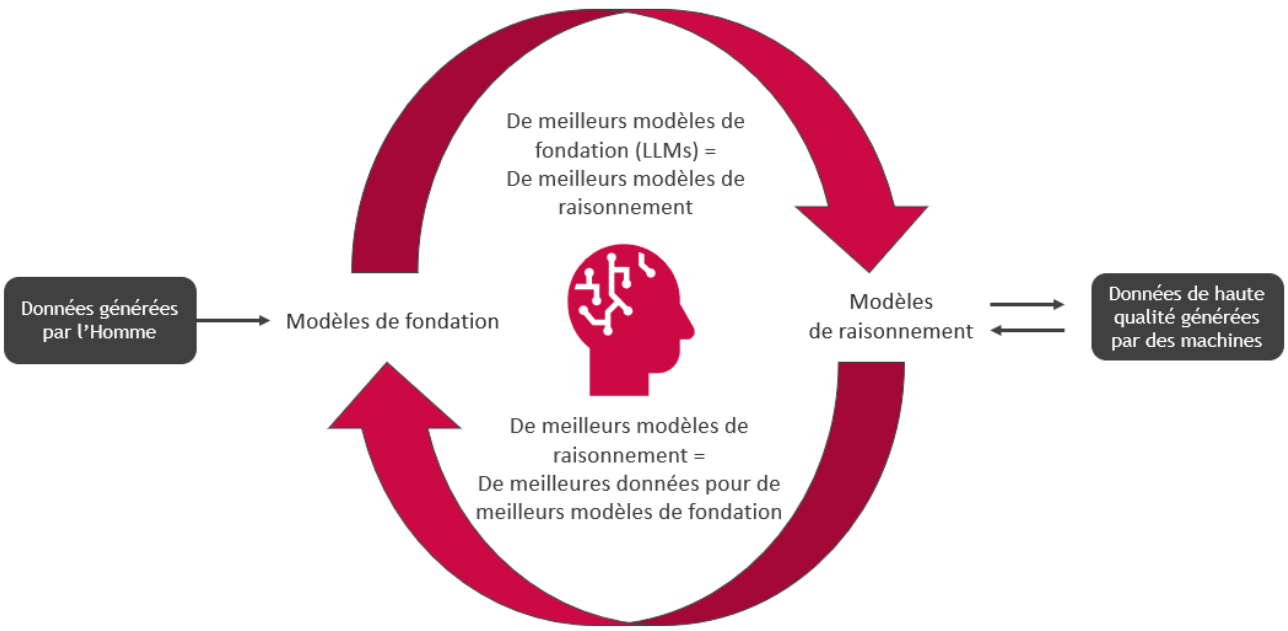
DES PERSPECTIVES D'INVESTISSEMENT RENOUVELÉES

1. Une nouvelle phase d'entraînement et de création de données

L'entraînement de modèles d'IA reposait jusqu'ici sur un volume colossal de données textuelles, visuelles ou audio provenant d'Internet et d'autres sources. Plusieurs laboratoires d'IA (de même que les investisseurs) craignaient une pénurie de « données de qualité », susceptible de freiner l'amélioration continue des LLMs.

Or, l'arrivée des modèles de raisonnement vient renverser cette équation : grâce à leur capacité à raisonner et à générer des contenus plus structurés, o1 et o3 sont à même de produire de la « donnée synthétique » qui servira à entraîner de futurs modèles de base. En d'autres termes, nous assistons à une boucle vertueuse où l'IA contribue elle-même à l'enrichissement des données nécessaires à sa propre évolution.

La boucle vertueuse des modèles de raisonnement



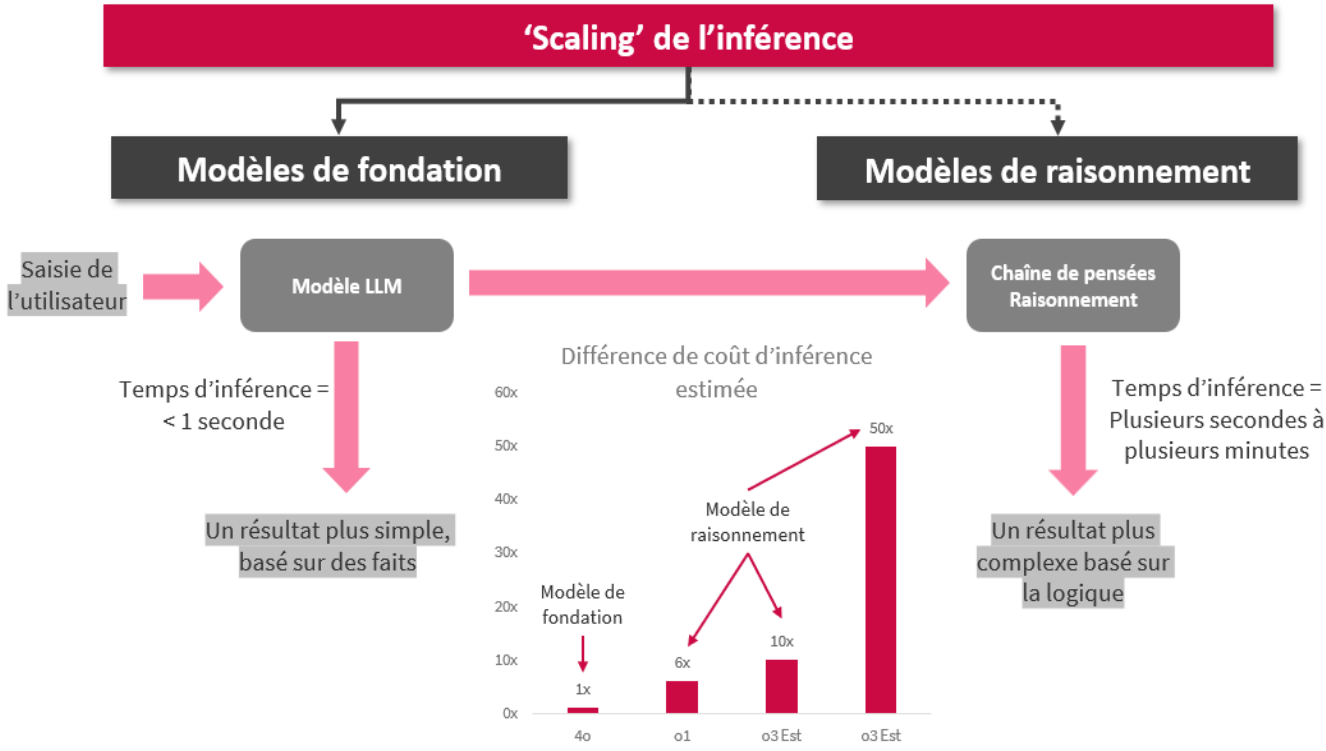
Source du graphique : Sycomore Asset Management.

[1] Le *transformer* est une architecture de réseau neuronal basée sur un mécanisme d'auto-attention et destinée au traitement du langage naturel.

2. Des coûts d'inférence (d'utilisation) en forte hausse

Le fonctionnement des modèles de raisonnement nécessite davantage de ressources que les modèles de base, aussi bien en calcul qu'en mémoire. Certaines estimations font état d'un coût 10 à 50 fois plus élevé que GPT-4. Cela s'explique par la durée de la phase de réflexion, ou d'« inférence » : au lieu de générer une réponse instantanée, un modèle comme o1 Pro peut « penser » pendant plusieurs minutes pour produire une solution détaillée et fiable.

Pour les entreprises, cet investissement supplémentaire peut se justifier lorsqu'il s'agit de résoudre des problèmes complexes où la précision prime sur la rapidité. Et pour les fournisseurs d'infrastructures (fabricants de semi-conducteurs, de puces graphiques ou de mémoires HBM [*High Bandwidth Memory*]), cette hausse des coûts d'inférence se traduit par une demande continue, voire accrue, de la part de leurs clients.



Sources du graphique : Sycomore Asset Management ; Semianalysis.

LA MÉMOIRE, UN ENJEU CLÉ

En 2025, l'un des points de blocage les plus critiques sera la capacité de mémoire, notamment la HBM. Contrairement à la mémoire traditionnelle, la HBM empile plusieurs puces pour offrir une bande passante et une densité supérieures. Essentielle pour supporter les longs temps de calcul des modèles de raisonnement, elle présente aussi un coût nettement plus élevé – de l'ordre de 3 à 5 fois celui de la mémoire classique.

En 2024, les investisseurs ont découvert que les composants de base du réseau étaient essentiels au bon fonctionnement des data centers. Cette prise de conscience a largement soutenu la performance et les niveaux de valorisation des entreprises concernées. On s'attend à un « moment de vérité » similaire pour le secteur de la mémoire en 2025.

Le marché de la HBM est un marché très profond. En 2024, il a déjà atteint 16 milliards de dollars ; il pourrait franchir les 30 milliards de dollars dès 2025 et grimper jusqu'à 100 milliards de dollars en 2030. Ce montant équivaut aujourd'hui à l'intégralité du marché mondial de la mémoire. Une progression aussi rapide témoigne de l'enjeu stratégique que représente la HBM pour l'IA de nouvelle génération.

DES VALEURS À SURVEILLER ?

Lorsqu'on parle d'investissements dans le domaine de l'IA, le nom de **Nvidia** vient souvent à l'esprit : cette entreprise a fortement bénéficié de la vague initiale d'adoption des LLMs. Cependant, pour 2025, d'autres acteurs méritent toute l'attention des investisseurs.

TSMC : Leader incontesté de la fonderie de puces, TSMC bénéficie d'une position de quasi-monopole pour les composants IA de grandes entreprises comme Nvidia, Google, AMD ou Meta. Avec une capitalisation boursière d'environ 800 milliards de dollars, TSMC se négocie à des multiples de valorisation raisonnables (autour de 15 fois les bénéfices) selon nous. D'après nos estimations, plus de 25 % de son chiffre d'affaires proviendra directement de l'IA en 2025. Comme ASML avant lui (qui avait connu son « moment de reconnaissance » en tant que monopole essentiel de la chaîne de valeur des semi-conducteurs avec sa technologie d'EUV), TSMC pourrait voir sa valorisation s'envoler à mesure que les investisseurs prendront conscience de son rôle croissant en tant que monopole de la fabrication de puces les plus avancées technologiquement.

Micron : Avec une capitalisation boursière d'environ 100 milliards de dollars, Micron est l'un des principaux fabricants de mémoire HBM, technologie-clé de la nouvelle génération de l'IA. L'entreprise pourrait capter près de 30 % de ce marché du HBM, qui se dirigerait vers 100 milliards de dollars d'ici 2030. Cela ferait plus que doubler son chiffre d'affaires issu de ce segment. Par ailleurs, le lancement de la nouvelle gamme GB300 de Nvidia au second semestre 2025 devrait permettre à Micron de gagner de nouvelles parts de marché.

LE RAISONNEMENT CONDUIRA À L'ADOPTION MASSIVE DE L'IA

Si les prouesses de l'intelligence artificielle ont déjà fait couler beaucoup d'encre depuis 2023, l'année 2025 s'annonce comme le véritable moment de bascule pour le début de son adoption à grande échelle. Les incertitudes qui pesaient sur la pérennité de la demande s'estompent grâce à l'apparition de modèles de raisonnement tels que o1 et o3, beaucoup plus gourmands en ressources et surtout capables de générer de la donnée synthétique utile à l'entraînement de futures IA.

Cette dynamique, couplée à des innovations technologiques (comme la mémoire HBM) et à des opportunités d'investissement majeures (TSMC, Micron, Nvidia, etc.), promet une année décisive au cours de laquelle l'IA passera d'une mode encore immature à un outil incontournable dans de nombreux domaines de l'économie. C'est en cela que 2025 sera l'année de l'IA avec un grand A.