23 SEPTEMBRE 2026

IA : EXPLOSION DE LA PUISSANCE DE CALCUL, POURQUOI PARLE-T-ON D'UN RALENTISSEMENT DES MODÈLES ?

Temps de lecture – 8 minutes.

La semaine dernière, le débat sur un possible ralentissement du développement des modèles d'IA est revenu sur le devant de la scène. Alors que la puissance de calcul disponible continue d'augmenter fortement et que les capacités des modèles progressent, plusieurs dirigeants de premier plan ont appelé à davantage de prudence. OpenAI a mobilisé plus de **100 000 GPU pour entraîner Astra**, son nouveau modèle de pointe. Ce chiffre donne une idée de l'échelle atteinte par la course à l'intelligence artificielle : pour continuer à améliorer leurs modèles, les laboratoires construisent désormais de véritables usines de calcul. *(source : Jensen Huang, Nvidia)*

La prochaine génération de Nvidia, Vera Rubin, doit encore repousser cette frontière. Nvidia estime que sa plateforme pourra entraîner certains grands modèles de type Mixture-of-Experts avec quatre fois moins de GPU que Blackwell pour un même objectif d'entraînement. *(source : Nvidia)*

Surtout, **la montée en puissance industrielle change** elle aussi **d'échelle**. Andrew Bell, responsable de l'ingénierie hardware chez Nvidia, a indiqué que la douzaine de partenaires chargés d'assembler les racks Rubin pourraient, à terme, disposer d'une capacité allant jusqu'à **1 000 racks NVL72 par jour**. Avec 72 GPU par rack, cela représente théoriquement 72 000 GPU par jour. *(source : Nvidia, cité par The Information)*

Pour faciliter la lecture, retrouvez notre lexique de termes techniques en fin de note.

À cette cadence maximale, l'équivalent en nombre de GPU du cluster ayant servi à entraîner Astra pourrait être produit en moins de deux jours.

Il ne s'agit évidemment pas d'une prévision de livraisons quotidiennes. Et surtout, **fabriquer davantage de GPU ne suffit pas**. Entre les centaines de milliards de dollars de capex annoncés et la puissance de calcul réellement disponible, il faut encore construire les centres de données, les raccorder au réseau, sécuriser l'alimentation électrique, installer le refroidissement, la mémoire et les équipements de réseau.

Ces goulots d'étranglement physiques peuvent eux-mêmes ralentir la montée en puissance de l'IA. Pour les fournisseurs d'infrastructures, c'est un point essentiel : les ventes dépendront moins des annonces de capex que du rythme réel auquel les projets seront construits, équipés et mis en service.

L'acceptabilité locale peut également devenir un facteur limitant. La construction de nouveaux centres de données concentre d'importants besoins en électricité, parfois en eau, et peut provoquer des résistances autour du coût des réseaux ou de leur impact environnemental.

Autrement dit, le ralentissement pourrait aussi être imposé par le monde physique : la puissance de calcul ne peut croître qu'aussi vite que les infrastructures capables de l'accueillir.

Et toute cette puissance ne servira pas uniquement à entraîner des LLM plus gros.

Après les LLM, le monde agentique

La première vague de l'IA générative était essentiellement réactive : un utilisateur posait une question, le modèle produisait une réponse, puis s'arrêtait.

L'étape actuelle est celle des agents. Un agent peut poursuivre un objectif, utiliser des logiciels, rechercher des informations, écrire et exécuter du code, vérifier le résultat puis recommencer. Une seule tâche peut ainsi nécessiter des dizaines, voire des centaines d'appels successifs au modèle.

Cela change profondément l'économie de la puissance de calcul. Selon des données OpenRouter citées par Nvidia, **une requête agentique consomme en moyenne environ 15 fois plus de tokens** (les unités de texte traitées par le modèle) qu'une simple conversation avec un chatbot. *(sources : Nvidia, Montpensier Arbevel)*

Un chatbot mobilise des GPU pendant quelques secondes pour produire une réponse. Un agent peut, lui, travailler pendant plusieurs minutes ou plusieurs heures, utiliser différents outils et solliciter plusieurs fois le modèle.

Si des millions d'agents fonctionnent simultanément, faire fonctionner les modèles au quotidien peut devenir un immense consommateur de puissance de calcul, en plus de celle nécessaire pour les entraîner.

La cybersécurité donne déjà un aperçu de ce monde. Astra est le premier modèle qu'OpenAI classe au niveau « critique » dans son propre cadre d'**évaluation des risques cyber**. Avec les outils et les accès appropriés, il peut découvrir des vulnérabilités jusque-là inconnues et développer des moyens de les exploiter dans des systèmes bien protégés, sans qu'un humain guide chaque étape.

La même technologie peut évidemment **servir à la défense**. Nvidia et CrowdStrike ont par exemple présenté SafeMind, un système dans lequel un agent offensif recherche des chemins d'attaque tandis qu'un agent défensif cherche à les détecter et à les bloquer.

La cybersécurité illustre donc bien le passage du LLM à l'agent : **l'IA ne produit plus seulement de l'information, elle agit dans un environnement numérique**.

L'IA agit dans le numérique comme dans le monde physique

Le même mouvement commence à gagner le monde physique.

Pour conduire une voiture ou piloter un robot, il ne suffit plus de maîtriser le langage. Il faut interpréter la vidéo, comprendre le mouvement et l'espace, puis anticiper les conséquences d'une action.

C'est l'objectif des **« world models »** : permettre à une IA de représenter son environnement et d'en simuler l'évolution. Nvidia développe notamment Cosmos pour l'IA physique, GR00T pour la robotique et Alpamayo pour la conduite autonome.

L'enjeu est aussi celui des données. Une voiture autonome ne peut pas parcourir physiquement des milliards de kilomètres pour rencontrer toutes les situations rares. Un robot ne peut pas répéter des milliards de gestes dans le monde réel. Dans un environnement simulé, ces expériences peuvent en revanche être générées et répétées à très grande échelle.

Le prochain cycle de puissance de calcul pourrait donc alimenter simultanément les LLM, les agents, la cybersécurité, la robotique, la voiture autonome, la recherche scientifique ou encore la défense.

Quand l'IA commence à améliorer l'IA

Il existe enfin une autre boucle, potentiellement plus puissante encore : **l'IA commence à contribuer à son propre développement.**

Chez Anthropic, plus de 80 % du code intégré à la base de code de l'entreprise était écrit par Claude en mai 2026. Les modèles servent désormais à programmer, conduire des expériences, analyser des résultats et accélérer le travail des chercheurs. *(source : Anthropic)*

La boucle devient alors simple : **plus de puissance de calcul ➜ meilleure IA ➜ recherche en IA plus rapide ➜ meilleure IA.**

C'est ce que l'on appelle **l'auto-amélioration récursive**. Nous sommes encore loin d'une IA capable de concevoir seule son successeur, ce qu'Anthropic reconnaît explicitement. Mais le début de la boucle est déjà visible : plus les modèles deviennent performants en programmation et en raisonnement, plus ils peuvent accélérer le travail des équipes qui développent la génération suivante.

Alors, pourquoi parler de ralentir ?

C'est dans ce contexte que Dario Amodei, le dirigeant d'Anthropic, a appelé à ralentir le rythme de progression des modèles de frontière afin de laisser davantage de temps à leur évaluation et à leur contrôle.

En parallèle, certains chercheurs de l'industrie ont adopté un ton beaucoup plus alarmiste. Jacob Coxon, ancien chercheur d'OpenAI puis d'Anthropic, a démissionné en accusant les deux groupes de courir vers une « superintelligence auto-améliorante » et de « jouer avec nos vies ». Il affirme que certains de ceux qui développent ces technologies pensent sincèrement qu'elles pourraient tuer l'humanité d'ici la fin de la décennie.

Evan Hubinger, responsable des recherches sur l'alignement chez Anthropic, a de son côté estimé personnellement à plus de 10 % la probabilité que l'IA puisse provoquer l'extinction humaine au cours de la prochaine décennie.

Ces **scénarios** restent évidemment très **spéculatifs** et **sont loin de faire consensus**. Mais ils montrent à quel point le débat interne à l'industrie a changé. Il ne porte plus seulement sur les biais d'un chatbot ou la désinformation, mais aussi sur la capacité à garder le contrôle de systèmes de plus en plus autonomes.

Il est également légitime de s'interroger sur la dimension communication de ce nouveau discours.

Les grands laboratoires doivent désormais convaincre non seulement les investisseurs, mais aussi les gouvernements et une opinion publique plus attentive aux conséquences de l'IA. Anthropic prépare actuellement son introduction en Bourse, potentiellement à une valorisation extrêmement élevée.

Dans ce contexte, afficher une position volontariste sur la sécurité peut aussi contribuer à rassurer et à démontrer que les acteurs les plus avancés sont capables de contrôler la technologie qu'ils développent.

Surtout, un **véritable ralentissement** semble très difficile à organiser.

La course à l'IA est désormais aussi une compétition technologique entre les États-Unis et la Chine. Si les laboratoires américains ralentissent seuls, leurs concurrents disposent d'une incitation évidente à continuer. La Chine développe par ailleurs une industrie très compétitive autour d'acteurs comme DeepSeek, Z.AI ou Alibaba, dont plusieurs diffusent leurs modèles avec des poids ouverts.

C'est le deuxième obstacle : les modèles open-weight. Une fois leurs paramètres publiés, ils peuvent être téléchargés, exécutés et adaptés indépendamment de leur créateur. La frontière technologique devient alors beaucoup plus difficile à contrôler depuis quelques laboratoires américains.

C'est tout le paradoxe : ceux qui sont aujourd'hui en tête commencent à demander davantage de prudence, mais chacun sait que ralentir seul peut permettre à un concurrent de prendre l'avantage.

Et même si un ralentissement volontaire intervenait, il s'ajouterait à un autre facteur, beaucoup plus concret : la **capacité réelle à déployer les infrastructures**. Le rythme du progrès dépendra donc autant des choix des laboratoires que de la vitesse à laquelle l'industrie pourra transformer les capex annoncés en mégawatts raccordés, en centres de données opérationnels et en GPU effectivement utilisés.

Le scénario le plus probable n'est donc peut-être pas l'arrêt de la course à la puissance de calcul, mais un **déplacement de son usage**.

Jusqu'ici, l'industrie s'est surtout concentrée sur produire de l'intelligence : entraîner des **modèles toujours plus performants et toujours plus coûteux**.

Le prochain cycle pourrait être celui d'utiliser cette intelligence : faire travailler des agents pendant des heures, automatiser la cybersécurité, simuler le monde physique, piloter des robots ou accélérer la recherche scientifique.

Même si les laboratoires ralentissaient le rythme des nouveaux entraînements de pointe, toute cette puissance de calcul ne resterait donc probablement pas inutilisée. Elle pourrait simplement être réallouée.

Après avoir passé plusieurs années à produire toujours plus d'intelligence, l'industrie entre peut-être dans une nouvelle phase : celle où il s'agit désormais d'utiliser cette intelligence, partout et en permanence.

Glossaire :

- **GPU :** puce spécialisée dans les calculs massivement parallèles, devenue centrale pour entraîner et faire fonctionner les modèles d'IA.
- **LLM :** Large Language Model, grand modèle de langage capable de comprendre et générer du texte, du code ou d'autres contenus.
- **Mixture-of-Experts (MoE) :** architecture où seule une partie du modèle est activée pour chaque requête, ce qui permet d'augmenter la taille du modèle sans faire exploser proportionnellement le coût de calcul.
- **Token :** petite unité de texte traitée par un modèle ; un mot peut être découpé en plusieurs tokens.
- **Agent IA :** système capable de poursuivre un objectif, d'utiliser des outils et d'enchaîner plusieurs actions de manière relativement autonome.
- **World model / modèle du monde :** modèle qui apprend à représenter un environnement et à anticiper son évolution ou les conséquences d'une action.
- **Open-weight :** modèle dont les paramètres entraînés sont rendus accessibles, ce qui permet de le télécharger, l'exécuter et l'adapter.
- **Alignement :** ensemble des méthodes visant à faire en sorte qu'un système d'IA poursuive des objectifs compatibles avec les intentions humaines et se comporte de manière contrôlable.

Valentin François
Gérant Actions & Thématiques

Sébastien Lalevée
Directeur Général

RETROUVEZ-NOUS SUR NOTRE SITE INTERNET
montpensier-arbevel.com
LinkedIn X.com
MONTPENSIER ARBEVEL
58 avenue Marceau, 75008 Paris, France - Tél. : +33 (0)1 45 05 55 55