Par Clément Dougé,
Responsable Marketing & Communication, Partenor Digital

GREEN AI

Rendre l'IA Générative plus vertueuse est-il possible ?

Dans un monde où la technologie avance à pas de géant et où l'IA, en particulier l'IA générative, occupe une place omniprésente, le concept de Green AI émerge de plus en plus.

La Green IA vise à rendre l'intelligence artificielle générative plus respectueuse de l'environnement, s'inscrivant ainsi dans une démarche de développement durable.

Compte tenu de l'impact environnemental significatif notamment lié à la consommation énergétique des data centers alimentant l'IA, l'importance de ce sujet ne peut être sous-estimée. L'importance de conjuguer l'IA Générative avec des mesures eco-responsables est désormais primordial.

Mieux comprendre l'impact environnemental de l'IA Générative

Depuis plusieurs mois, nous assistons à un développement de l'IA Générative à un rythme exponentiel. Qu'elle soit utilisée pour la création visuelle ou textuelle, son utilisation massive intensifie les préoccupations sur son impact environnemental.

Pour mieux comprendre cet impact, nous pouvons accorder une attention particulière à 2 aspects clés :
l'empreinte carbone des grands modèles de langage (LLM) et la consommation énergétique colossale des datacenters.

L'empreinte carbone des LLM

Les modèles de langage utilisés par les IA génératives nécessitent d'énormes capacités de calcul pour s'entraîner sur des milliards de données. Par exemple, l'entraînement de LLaMa 2-70B, un modèle relativement modeste, a émis 291,42 tonnes d'équivalent CO2. Pour LLaMa 3, cette empreinte atteint 2 290 tonnes, soit une hausse de 324,86 % par rapport à la version précédente.
D'autre part, l'inférence d'une IA, bien que moins énergivore que l'entraînement, peut rapidement s'avérer problématique lorsque les modèles sont interrogés des millions de fois par jour, et donc générer tout autant de CO2.

Pour illustrer cela, une conversation moyenne avec ChatGPT consomme environ 500 ml d'eau, un impact considérable compte tenu de ses milliards d'utilisateurs mensuels.

Pour rappel, les termes «inférence» et «entraînement» désignent deux phases distinctes du cycle de vie d'un modèle d'apprentissage. L'entraînement d'une IA est le processus par lequel un modèle d'IA apprend à partir des données. Pendant cette phase, le modèle ajuste ses paramètres internes (poids, biais, etc.) en fonction des exemples d'entrée et des sorties correspondantes.
De son côté, l'inférence est le processus d'utilisation d'un modèle préalablement entraîné pour faire des prédictions ou générer des données nouvelles à partir d'entrées inédites.

La consommation d'énergie des datacenters

Pour qu'une IA générative puisse s'entrainer sur des milliards de données, cela nécessite l'utilisation de serveurs puissants, situés dans des datacenters. Ces serveurs consomment de l'électricité, chauffent et doivent être refroidis avec des systèmes qui nécessitent à leur tour de l'énergie.

Selon l'Agence Internationale de l'Energie, d'ici 2026, la consommation énergétique des datacenters pourrait grimper à un tel point qu'elle équivaudrait à celle de tout le Japon, dépassant ainsi les 1 000 TWh.
Pour Sundar Pichai, le Directeur Général de Google, « le besoin de calcul informatique pour l'IA a été multiplié par un million en six ans et il décuple chaque année ».

A titre de comparaison, une requête faite à ChatGPT nécessite en moyenne 10 fois plus d'énergie qu'une simple requête sur Google selon plusieurs études.

Bien que des efforts soient déployés pour améliorer l'efficacité énergétique, l'empreinte environnementale de l'IA générative reste préoccupante. Des stratégies doivent être mises en œuvre pour réduire son impact et la rendre plus durable, notamment grâce à des innovations technologiques émergentes.

Quelles stratégies mettre en place pour une IA plus vertueuse ? Différentes pistes possibles.

Pour réduire la consommation énergétique de l'IA générative, plusieurs stratégies peuvent être mises en œuvre.

Optimisation des modèles et rationalisation des cas usages

Bien que les modèles de langage offrent des performances impressionnantes, ils nécessitent des ressources énergétiques importantes lors de leur entraînement et de leur utilisation. Pour pallier ce problème, l'optimisation des modèles existants est cruciale. Cela peut inclure la compression des modèles pour réduire leur taille sans compromettre les performances, ou l'amélioration de l'efficacité algorithmique pour accélérer le traitement et diminuer la charge sur les serveurs.

En parallèle, la rationalisation des cas d'usages implique une sélection plus judicieuse des tâches pour lesquelles l'IA est employée. Cela signifie privilégier des modèles mini ou micro-modèles qui, bien que moins puissants que les grands modèles de langage, sont souvent suffisants pour des tâches spécifiques. Ces modèles réduits consomment moins d'énergie et sont plus rapides à entraîner, ce qui les rend idéaux pour des applications ciblées.

Concrètement :

1. Évaluer rigoureusement la nécessité d'utiliser un LLM pour une tâche donnée.
2. Envisager des alternatives plus légères ou conventionnelles lorsque cela est possible.
3. Réserver l'utilisation des LLM aux cas où ils apportent une réelle valeur ajoutée.

Adopter des modèles plus spécialisés

Au lieu d'utiliser des modèles généralistes pour toutes les tâches, l'adoption de modèles spécialisés peut être une stratégie à envisager pour réduire l'empreinte carbone de l'IA générative. Ces modèles sont conçus pour des applications précises et peuvent donc être plus petits et plus économes en énergie. Par exemple, un modèle dédié uniquement à la traduction n'a pas besoin de la même capacité qu'un modèle destiné à la génération de texte.

Utiliser des techniques comme le RAG (Retrieval Augmented Generation) pour réduire la sollicitation des LLM

Le RAG est une technique qui combine la recherche d'informations et la génération de texte. En récupérant des informations pertinentes avant la génération de contenu, elle permet de réduire la charge sur les LLM (car le modèle peut s'appuyer sur des données existantes plutôt que de générer du contenu à partir de zéro). Cela peut diminuer la quantité de calcul nécessaire et, par conséquent, l'énergie consommée.

Gérer et utiliser les données plus efficacement

Une utilisation efficace des données implique d'optimiser les processus de collecte et de traitement des données. Cela peut être illustré par l'utilisation de techniques de réduction de dimensionnalité et de nettoyage des données, afin de ne conserver que les informations pertinentes pour l'entraînement des modèles d'IA. De plus, l'apprentissage incrémental, où le modèle apprend continuellement de nouvelles données au lieu d'être réentraîné à partir de zéro, peut également contribuer à une utilisation plus rationnelle des ressources.

Alimenter les datacenters en énergie renouvelable

Une autre piste plus difficile à envisager pourrait être l'utilisation d'énergies renouvelables pour les datacenters. Cela peut inclure l'installation de panneaux solaires, l'utilisation d'éoliennes, ou encore la connexion à des réseaux fournissant de l'énergie hydroélectrique. En s'approvisionnant en énergies propres, les datacenters peuvent fonctionner de manière plus durable et réduire leur empreinte carbone.

Éduquer les entreprises et les utilisateurs sur les bons et les mauvais usages des LLM

La formation et l'éducation des entreprises et des utilisateurs finaux sur les pratiques responsables dans l'utilisation des outils d'IA Générative sont essentielles. Cela peut par exemple inclure la sensibilisation aux impacts environnementaux, la promotion de l'utilisation de modèles plus petits lorsque cela est possible, et l'encouragement à l'adoption de pratiques d'IA éthiques et durables.

Bien que l'IA Générative puisse débloquer de nombreuses opportunités pour nos métiers et sera probablement indispensable dans les années à venir, elle navigue pour le moment à contre-courant, dans son utilisation, avec les grands principes de durabilité. L'évolution d'une technologie doit être accompagnée de mesures concrètes pour réduire son empreinte carbone.

En vue des implications considérables de l'IA générative sur notre environnement, il apparaît indispensable de persévérer dans la recherche et l'application d'approches responsables et innovantes. Les pistes mentionnées plus haut font sont des axes à envisager, mais il est avant tout important d'appeler à une prise de conscience et un engagement collectif pour la mise en place de solutions eco-efficientes.

starclay

www.starclay.com

hello@starclay.com