IDÉE 6

# L'IA POURRAIT DEVENIR PLUS « LÉGÈRE » ET FONCTIONNER SUR DES APPAREILS COURANTS

Plutôt que quelques grands modèles d'IA fonctionnant sur des superordinateurs en nuage, les futurs modèles d'IA pourraient être diversifiés et personnalisés, certains d'entre eux fonctionnant sur de petits appareils demainlocaux tels que les téléphones intelligents. Cela pourrait compliquer la réglementation et le contrôle, et multiplier les risques de cybersécurité.

## AUJOURD'HUI

**Les améliorations apportées aux techniques d'entraînement et de compression de l'IA permettent aux modèles plus petits et moins gourmands en ressources de devenir plus performants.** La taille d'un modèle d'IA est souvent utilisée pour désigner sa puissance, sa capacité et sa qualité. Alors que les plus grands modèles sont souvent les plus puissants et les plus performants, les développeurs d'IA mettent sur le marché des versions plus petites et comprimées dérivées de modèles plus grands. Cela permet au modèle plus petit de conserver la plupart des performances du modèle plus grand tout en étant moins gourmand en énergie et fonctionnant sur du matériel moins puissant[1]. C'est pourquoi les modèles plus petits et plus récents sont plus performants que les modèles plus anciens et plus grands. Par exemple, Phi-3, qui a été lancé début 2024 et ne compte que 3,8 milliards de paramètres, a des performances comparables à GPT-3.5, qui a été lancé fin 2022 et compte 175 milliards de paramètres[2]. Des entreprises telles que Meta[3] et Mistral[4] ont publié des modèles d'IA en code source libre qui rivalisent avec les performances de ChatGPT, mais qui peuvent être exécutés sur un ordinateur portable. Les chercheurs dans le domaine du TinyML développent une IA plus petite et capable de fonctionner sur des appareils moins puissants pour permettre un Internet des objets (IdO) « intelligent ». Par exemple, le Raspberry Pi, un ordinateur de la taille d'une carte de crédit, très prisé des amateurs de programmation et d'ingénierie informatique, peut désormais exécuter une série de modèles d'IA, y compris la reconnaissance faciale[5].

### Taille du modèle

**La taille d'un modèle d'IA est déterminée par le nombre de ses paramètres.**

**Les paramètres sont des variables d'un système d'IA dont les valeurs sont ajustées au cours de l'entraînement. Les petits modèles peuvent avoir des millions de paramètres, voire moins, tandis que les grands modèles peuvent en avoir plus de 400 milliards.**

# AVENIRS

**Nous pourrions voir des milliers de modèles d'IA différents capables de fonctionner localement sur tous les types d'appareils numériques, des téléphones intelligents aux petits ordinateurs[6].** Ces modèles pourraient être développés par des amateurs, des jeunes poussespousses ou des criminels. Ils pourraient être basés sur des modèles à source ouverte et adaptés à différents objectifs par l'intermédiaire d'un entraînement sur des ensembles de données largement accessibles. Par exemple, Venice AI est un service d'IA basé sur le Web, construit à partir d'une poignée de modèles d'IA à source ouverte, qui permet aux utilisatrices et utilisateurs de générer du texte, du code ou des images avec peu ou pas de garde-fous et qui est vendu comme « privé et sans permission »[7]. Le déploiement de modèles d'IA de différentes tailles pourrait donner naissance à un écosystème de modèles d'IA présentant divers degrés d'interopérabilité. Les petits modèles pourraient interagir avec les grands modèles accessibles au public et basés sur l'informatique en nuage, en tirant parti de leur puissance pour effectuer des tâches ou apprendre (voir la figure 1). Ces petits modèles localisés peuvent manquer de mesures de sécurité et être déployés à grande échelle à l'insu de toute autorité.

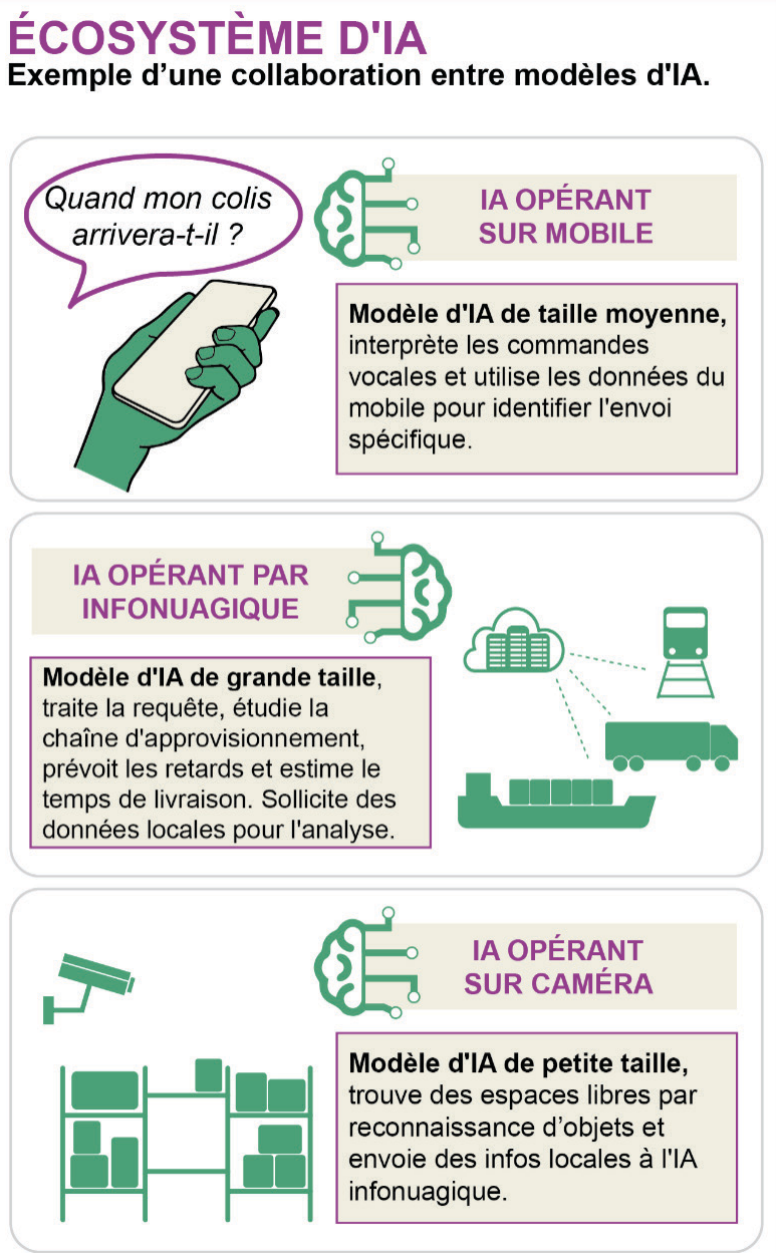


Figure 1. Écosystème d'IA, un exemple d'une collaboration entre modèles d'IA.

# IMPLICATIONS

- Les réglementations axées uniquement sur les grands modèles d'IA **risquent d'être inefficaces**[8]
- **L'IA en libre accès pourrait permettre la circulation de modèles qui posent problème**, que ce soit parce qu'ils comportent des biais, manquent de mesures de sécurité ou facilitent des activités illégales[9]
- **Il pourrait être difficile de repérer les mauvais acteurs** qui forment ou utilisent des modèles d'IA petits, mais puissants
- En analysant les données localement, les modèles d'IA qui opèrent directement sur appareil pourraient aider les personnes **à protéger leurs données et leur vie privée**
- **Les petites entreprises pourraient personnaliser leurs propres outils d'IA** pour mieux répondre à leurs besoins[10]
- La compatibilité entre les appareils dotés d'IA pourrait **offrir plus d'options aux utilisatrices et utilisateurs**, mais aussi créer des vulnérabilités en matière de cybersécurité[11]

## Notes de fin de texte


1 Hou, Jilei. « Quantization: What It Is & How It Impacts AI ». *Qualcomm* (blogue), 11 mars 2019.

2 Edwards, Benj. « Microsoft's Phi-3 Shows the Surprising Power of Small, Locally Run AI Language Models ». Ars Technica, 23 avril 2024.

3 Ramlochan, Sunil. « How Does Llama-2 Compare to GPT-4/3.5 and Other AI Language Models ». Prompt Engineering Institute, 1er septembre 2023.

4 Ali Awan, Abid. « Running Mixtral 8x7b On Google Colab For Free ». KDnuggets, 12 janvier 2024.

5 Dunn, Caroline. « How to Train Your Raspberry Pi for Facial Recognition ». Tom's Hardware, 17 septembre 2022.

6 Mearian, Lucas. « GenAI Is Moving to Your Smartphone, PC and Car - Here's Why ». Computerworld, 30 janvier 2024.

7 Irwin, Kate. « Venice's Privacy-Focused AI Chatbot Won't Store Your Data, Judge Your Questions ». PC Magazine, 9 août 2024.

8 La Maison-Blanche. « Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence », 30 octobre 2023.

9 Thiel, David. « Identifying and Eliminating CSAM in Generative ML Training Data and Models », 2023.

10 Pham, Nguyen. « Open Source Tools as an Opportunity for SMEs to Use AI? » foojay, 2 juin 2021.

11 Apple. « Bâtir un écosystème fiable pour des millions d'apps », juin 2021.



© Sa Majesté le Roi du chef du Canada, 2025

Pour obtenir des informations sur les droits de reproduction : https://horizons.gc.ca/fr/contactez-nous/



PDF : PH4-216/2025F-PDF
ISBN : 978-0-660-76892-2

Also available in English under the title: AI could become "lighter" and run on commonly held devices.



**AVERTISSEMENT**
Horizons de politiques Canada (Horizons de politiques) est le centre d'excellence en prospective du gouvernement du Canada. Notre mandat est de doter le gouvernement du Canada d'une perspective et d'un état d'esprit tournés vers l'avenir afin de renforcer la prise de décisions. Le contenu de ce document ne représente pas nécessairement le point de vue du gouvernement du Canada ou des agences et ministères participants.