Policy Horizons Canada Horizons de politiques Canada

Canada

INSIGHT 6

AI COULD BECOME "LIGHTER" AND RUN ON COMMONLY HELD DEVICES

Rather than a few large AI models running on cloud-based supercomputers, future AI models could be diverse and customized, some of them running on small, local devices such as smartphones. This could make regulation and control more complicated and multiply cybersecurity risks.

TODAY

Improvements in AI training and compression techniques are allowing smaller, less resource-intensive AI models to become more capable. The size of an AI model is often used as a shorthand for its power, capability, and quality. While the largest models are often the most powerful and capable, AI developers are releasing smaller, compressed versions derived from larger models. This allows the smaller model to retain most of the performance of the larger model while also allowing it to be much smaller, less energy demanding, and run on less powerful hardware.[1] This has led smaller, newer models to outperform older and larger models. For example, Phi-3, which was released in early 2024 and has only 3.8 billion parameters, has comparable performance to GPT-3.5, which was released in late 2022 with 175 billion parameters.[2] Companies including Meta[3] and Mistral[4] have released open-source AI models that rival ChatGPT's performance but can run on a laptop. Researchers in the field of TinyML are developing AI that is smaller and can run on less powerful devices to enable the "smart" Internet of Things (IoT). For example, the Raspberry Pi, a credit card-sized computer popular with programming and computer engineering enthusiasts, can now run a suite of AI models including facial recognition.[5]

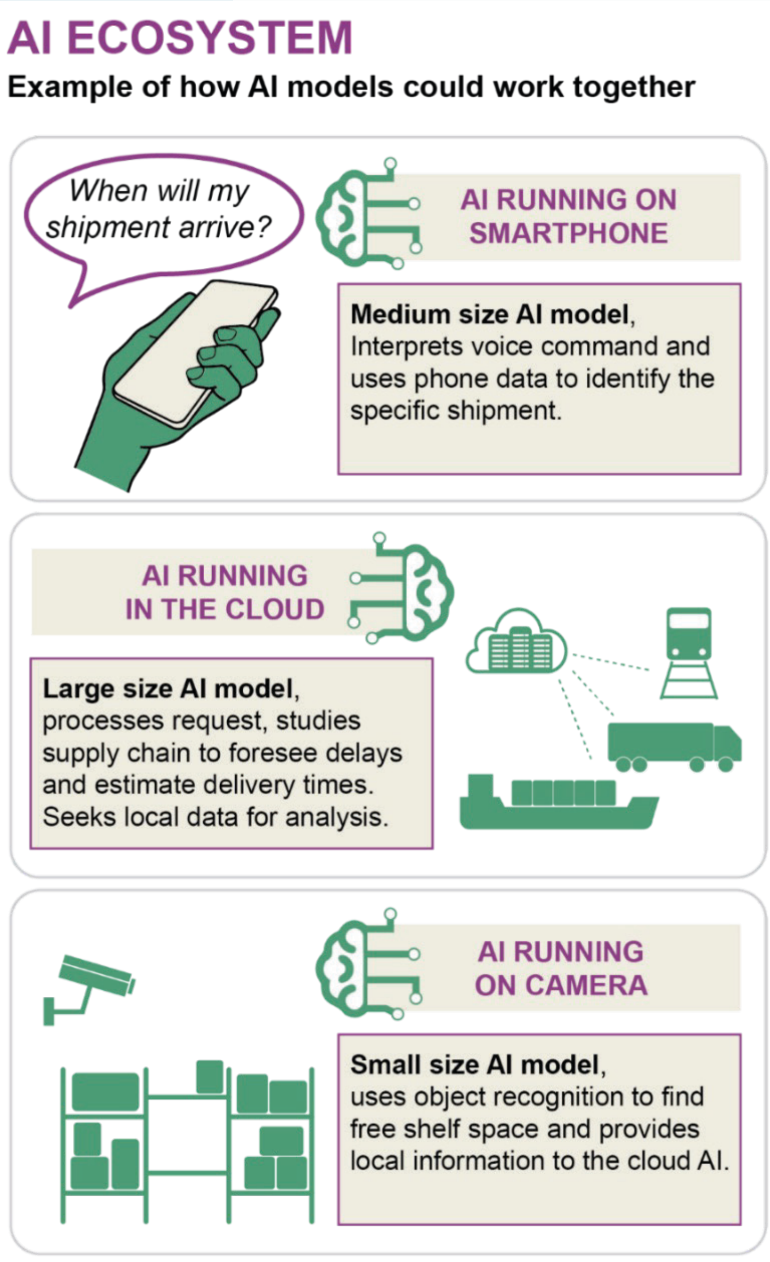
Model size

The size of an AI model is determined by how many parameters it has.

Parameters are variables in an AI system whose values are adjusted during training. Smaller models can have parameters numbering in the millions or fewer, while larger models can have more than 400 billion.

FUTURES

We may see thousands of different AI models capable of running locally on every type of digital device, from smartphones to tiny computers.[6] These models could be developed by amateurs, startups, or criminals. They could be based on open-source models and customised for different purposes through training on widely accessible datasets. For example, Venice AI is a web-based AI service, built from a handful of open-source AI models, that allows users to generate text, code, or images with little to no guardrails and is sold as 'private and permissionless.'[7] As AI models of different sizes become more widely deployed, this may give rise to an ecosystem of AI models with various degrees of interoperability. Small models could interact with large, cloud-based, publicly accessible models, leveraging their power to perform tasks or learn (see Figure 1). Such small, localized models may lack safety measures and be deployed broadly without the knowledge of any authority.



Medium size AI model, Interprets voice command and uses phone data to identify the specific shipment.

Large size AI model, processes request, studies supply chain to foresee delays and estimate delivery times. Seeks local data for analysis.

Small size AI model, uses object recognition to find free shelf space and provides local information to the cloud AI.

Figure 1. AI Ecosystem, an example of how AI models could work together

IMPLICATIONS

- Regulations focused only on large AI models **may not be effective**[8]
- **Open-source AI could allow the circulation of models which are problematic,** whether because they incorporate bias, lack safety measures, or facilitate illegal activities[9]
- **It could be hard to track bad actors** training or running small but powerful AI models
- By analysing data locally, on-device AI models could help individuals **protect their data and privacy**
- **Small businesses could customise their own AI tools** to better meet their needs[10]
- Compatibility between AI-enabled devices could **provide users with more options** but also create cybersecurity vulnerabilities[11]

Endnotes

1 Hou, Jilei. Quantization: What It Is & How It Impacts AI. Qualcomm (blog), 11 March 2019.

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

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

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

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

6 Mearian, Lucas. GenAI Is Moving to Your Smartphone, PC and Car — Here's Why. Computerworld, 30 January 2024.

7 Irwin, Kate. Venice's Privacy-Focused AI Chatbot Won't Store Your Data, Judge Your Questions. PCMAG. 9 August 2024.

8 The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 October 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 June 2021.

11 Apple. Building a Trusted Ecosystem for Millions of Apps: The Important Role of App Store Protections, June 2021.

PDF: PH4-216/2025E-PDF
ISBN: 978-0-660-76891-5

Aussi disponible en français sous le titre : L'IA pourrait devenir plus « légère » et fonctionner sur des appareils courants.

DISCLAIMER
Policy Horizons Canada (Policy Horizons) is the Government of Canada's centre of excellence in foresight. Our mandate is to empower the Government of Canada with a future-oriented mindset and outlook to strengthen decision making. The content of this document does not necessarily represent the views of the Government of Canada, or participating departments and agencies.