

#Génération IA – Épisode 06 – Le coût caché de l'IA que personne n'ose révéler

David Di Pietro :

L'autre jour, il y a un de mes collègues qui m'a taquiné parce que j'ai demandé à l'IA de me trouver une recette de pain aux bananes au lieu d'aller chercher sur Google. Ce à quoi j'ai répondu : « Toi, tu as bien écouté une série sur Netflix l'autre jour ». Bref, toute cette anecdote, ça m'a fait réfléchir. Est-ce que c'est vraiment plus polluant d'utiliser ChatGPT? On entend souvent parler du coût environnemental de l'IA, mais, concrètement, ça veut dire quoi?

Bienvenue à un nouvel épisode de Hashtag Génération IA.

[musique]

Aujourd'hui, pour parler de tout ça, on a deux invitées avec nous : Tanya Grenier, qui est étudiante au DESS option personnalisée à HEC Montréal, et Camille Grange, qui est professeure agrégée au Département de technologie de l'information.

Tanya Grenier. Camille Grange. Bonjour. Merci d'être ici. Bonjour.

Camille Grange :

Bonjour, David. Bonjour Tanya.

Tanya Grenier :

Bonjour, David.

David Di Pietro :

Je me tourne tout de suite vers l'étudiante, Tanya Grenier, parce que j'ai une question pour toi. Quand on parle d'impact environnemental de l'IA, c'est quoi ton rapport avec ça?

Tanya Grenier :

En fait, c'est tout récent que j'ai commencé à utiliser l'intelligence artificielle. Pendant longtemps, je voyais ça comme le gros camion Hummer, qui est hyper polluant pour aller à l'épicerie du quartier, puis je ne suis pas très fan. Il y a aussi d'autres compréhensions, d'autres mythes et autres qui font en sorte que je suis moins fan de l'utiliser, mais, récemment, j'ai vu l'apport que ça pouvait avoir pour m'aider à m'organiser. Donc, je commence tranquillement à me mouiller les orteils.

David Di Pietro :

Parfait. On va profiter de la présence de Camille Grange. Je propose qu'on joue à déconstruire les mythes. Le premier mythe que tu évoquais Tanya, puis je pose la question à Camille. Premier mythe : l'IA c'est virtuel, donc ça ne pollue pas.

Camille Grange :

Le plus gros mythe du numérique. Tanya, ce que tu décris là, ton ambivalence, je pense qu'on est beaucoup à la ressentir. Moi, par ailleurs, ça fait quelques années que je la ressens, mais je suis en technologies de l'information depuis vingt ans, j'ai étudié à l'École et ce n'était jamais sous le radar des cours ou de la recherche, très peu. C'est récemment avec la croissance de l'IA que c'est venu plus nous chercher. Le plus gros mythe, le numérique, c'est virtuel, ça n'a aucun impact, aucune pollution, aucune empreinte environnementale.

Le numérique, c'est beaucoup de matérialité. On pense, par exemple, à tous les terminaux utilisateurs, les écrans, les ordinateurs, les téléphones, trente milliards d'équipements de terminaux utilisateurs sont en utilisation dans le monde, un milliard de téléphones intelligents sont livrés chaque année. Ça, c'est la partie que l'on connaît parce que ça supporte nos usages, les « devices », comme on dit, les terminaux utilisateurs. Mais il y a aussi une autre face cachée de la matérialité, ce sont les serveurs, les centres de données qui permettent d'héberger nos données et de traiter les données, de faire tout le travail comme dans un restaurant, la cuisine derrière, qui fournit tout. Et puis, la réseautique, c'est les câbles sous-marins, les antennes, les routeurs, tous les trucs qui permettent de faire passer les données des serveurs à nos « devices », à nos terminaux, pour qu'on puisse accéder aux services.

Ces trois composantes ont un impact. Ce sont des objets, il faut les fabriquer, il faut les assembler, les transporter. Il y a les usages, et puis après, malheureusement, il y a aussi beaucoup de déchets électroniques. Tout ce cycle de vie du numérique, avec tous ces objets qui est vraiment caché, mais qui, je pense, reflète bien que le numérique, c'est très, très matériel et il y a une vraie empreinte.

David Di Pietro :

Oui. Si on se donne une espèce d'ordre de grandeur, par rapport à tout ça, l'IA consomme plus que tout le reste. Voici une autre idée reçue peut-être à déconstruire. Ça consomme plus que Google, plus que Netflix par exemple. Donc, qu'est-ce que tu penses de ça?

Camille Grange :

Alors première chose, c'est très difficile d'être précis pour calculer, on dit estimer l'impact environnemental. Et quand je dis impact environnemental, ça peut être consommation d'énergie, génération de gaz à effet de serre, utilisation d'eau, utilisation de ressources. Il y a un paquet d'indicateurs. Ton exemple au départ de cake aux bananes, c'est typique du problème qu'on a. Quand on compare deux choses qui ne sont pas exactement comparables, où on n'a pas de précision sur ce qui est comparé. Comme tu dis, est-ce que je veux évaluer quel est l'impact de ma consommation d'une heure, de Netflix ou YouTube, de ma consommation d'électricité versus une requête, deux requêtes, un an de requêtes sur ChatGPT. Quel modèle? Le 4? Le 4.0?

Pour un ordre de grandeur, j'ai quelques chiffres pour vous, mais encore une fois, les études varient. Une recherche Google, c'est 0.3 watt-heure. Ce n'est pas grand-chose 0.3 wathheure.

Il y a presque dix milliards de requêtes Google chaque jour. Donc, c'est toujours ça qu'il faut voir quand on pense aux usages. Ce n'est pas grand-chose ce que je fais, mais c'est quand on le « scale », ça peut faire beaucoup. Une requête ChatGPT4 avec 400 « tokens ». « Tokens », c'est l'unité que l'on utilise pour évaluer la quantité de travail qu'on va demander au modèle de langage. Une requête ChatGPT4, c'est huit wattheures versus 0.3 pour Google. Un modèle beaucoup plus petit, ChatGPT-4o, moins de 1 wattheure. Quelques milliwattheures. Donc, ça dépend des modèles. Générer une image, ça c'est beaucoup plus coûteux, environ trois wattheures. Une heure de streaming YouTube ou Netflix, 77 wattheures. Donc, tu peux dire à ton ami, David, que son heure de consommation de YouTube nécessite beaucoup plus de ressources d'électricité que ta petite requête de pain aux bananes.

Une chose importante, quand on parle de la consommation d'électricité sur les usages, en fait des services numériques, c'est qu'il y a une dépendance importante de là où on est, parce que ça dépend du mix électrique dans lequel on se situe géographiquement. Au Québec, on a un mix électrique assez décarboné par rapport à d'autres pays. Trois heures de Netflix au Québec, c'est 0.7 gramme de CO₂. Trois heures de Netflix en Pologne, qui eux utilisent beaucoup de charbon pour produire de l'électricité, c'est 400g de CO₂. Là je parle juste de l'usage. Évidemment, on ne parle pas de toute la partie production des matériaux qui permettent de supporter le service, mais c'est important de connaître cette distinction sur le mix énergétique parce qu'on se dit, les serveurs d'IA, on les met où? Ils sont beaucoup aux États-Unis et en Chine, qui sont des pays où les mix électriques sont extrêmement carbonés encore.

Tanya Grenier :

Avec les chiffres que vous avez mentionnés. 0.3, un peu moins d'un, trois, huit, c'est sûr que je ne peux pas faire tous les calculs mentaux immédiatement, mais je me dis, avec ces informations, ça prend quand même plusieurs requêtes à l'intelligence artificielle avant d'équivaloir à une heure de « streaming ». C'est sûr que je peux passer une heure sur l'intelligence artificielle. Si, par exemple, j'essaie de générer un artefact. Des fois, la génération, sans rien changer dans le « prompt », juste la renvoyer pour que le résultat soit différent, comment est-ce que je peux limiter cet aller-retour de perte d'énergie quand le résultat que je reçois n'est pas exactement ce que je veux avoir?

Camille Grange :

Très intéressante, ta question, Tanya. Il y a beaucoup de facteurs qui rentrent en compte dans l'efficacité, la charge énergétique ou d'électricité, la pertinence ou la qualité de notre « prompting » de nos requêtes. Si on peut réussir à faire des « prompts », des requêtes qui sont bien formulées, claires, qu'on peut réutiliser, qu'on évite les va-et-vient, il y a un impact. Tu vas diminuer la charge que tu sollicites aux modèles d'IA. On a parlé du modèle, du type de modèle. Est-ce que c'est un modèle très, très sophistiqué ou un petit modèle? Est-ce qu'on fait tourner à l'externe sur un serveur ou est-ce qu'on l'installe sur notre ordinateur en local pour qu'il n'y ait pas besoin de faire d'aller-retour? Ça, peu de gens le font, j'imagine.

Tanya Grenier :

Je ne savais pas que c'était une option.

Camille Grange :

Ce n'est pas facile. C'est encore un petit peu technique quelque part, mais tu vas voir, par exemple, un logiciel qui s'appelle LM Studio. Tu installes LM Studio, et puis, de LM Studio, tu peux aller télécharger des modèles ouverts, qui sont performants, qui sont, je pense, téléchargés sur la plateforme Huggingface. Tu dis, « je vais vouloir le modèle Phi de Microsoft ». En fait, tu as comme une librairie de modèles qui sont plus ou moins lourds, donc il faut que tu voies par rapport à ta machine ce qu'elle est capable de supporter. Il y a de petits modèles et puis tu peux les tester pour faire des résumés de texte, pour faire de l'écriture de courriel. Pense aux tâches pour lesquelles tu aimes utiliser de l'IA, si c'est du traitement d'images, tu prendras un modèle plus spécialisé sur la génération d'images, sachant que ça va être plus lourd pour ta machine probablement. Ça c'est mieux parce que tu utilises tes ressources en local de ta machine plutôt que de faire des appels à un serveur.

Tanya Grenier :

Et par rapport à ça, c'est quoi les plus grandes différences? J'imagine que l'impact environnemental doit être drastiquement plus bas. Mais à quel point les performances sont affectées, à quel point c'est similaire en termes d'expérience?

Camille Grange :

Tout à l'heure, tu mentionnais le fait de prendre le Hummer pour aller à l'épicerie, qui est à cinq minutes de chez moi. Tu as cette dimension de ce que tu veux faire. Quel est le besoin que tu veux satisfaire, puis quel est l'outil qui va permettre de mieux satisfaire ce besoin? Et je pense qu'on n'a pas encore l'habitude, parce qu'on ne connaît pas trop ça, on ne connaît pas trop les outils. Le chat GPT-5.1, le quatre, le trois, qu'est-ce que ça peut faire de vraiment différent? Il faut un petit peu expérimenter, tester. Par exemple, regarde ce petit modèle, il va bien pour mes tâches à l'école. Pour mes tâches de professionnel, de design, de site web, d'image, là j'ai besoin d'un truc plus particulier. Quel est l'outil plus spécifique qui va me permettre de faire ça? Mais il y a encore une méconnaissance du bon outil pour la bonne tâche. Il y a un petit peu d'exploration et d'expérimentation à faire pour chacun.

David Di Pietro :

C'est hyper pratico-pratique comme question, mais pourquoi les médias générés d'images, vidéos, voix synthétique consomment plus?

Camille Grange :

Alors c'est important de distinguer textes, images et vidéos.

Pour le texte avec les modèles de langage, on parle en « token ». C'est comme une quantité. C'est le volume qui est requis, la charge de travail qui est requis pour les GPU, les puces qui traitent l'information.

Pour les images, on ne génère pas du texte, on génère des pixels. Ce n'est pas des « tokens » mais on peut avoir des « token » qui sont générés parce qu'il y a de la description qui est envoyée, puis le modèle traite du texte pour aller décrire l'image. Mais ce qui est généré, ce sont des pixels. Alors, certains modèles donnent des équivalents. Par exemple, une image, c'est combien de « token »? Je vais vous donner des ordres de grandeur. Et une vidéo, c'est X images par vidéo.

Tanya Grenier :

Donc c'est exponentiellement plus gros.

Camille Grange :

Oui, puis, si tu veux une image avec beaucoup de résolution, une vidéo avec une très fine résolution, tu augmentes la charge, la demande.

Tanya Grenier :

Est-ce qu'il y a déjà des outils qui existent pour comparer les différents modèles? Pour pas que j'aie à faire un million de tests et augmenter l'empreinte, justement. Le temps de « figure it out »?

Camille Grange :

Eh bien, oui. Il y a des modèles, deux sites en fait, que je peux te recommander pour aller évaluer le coût énergétique, le coût environnemental de différents modèles. Il y a un site ou un outil qui s'appelle EcoLogits. Sur ce site, tu as leur « Calculator ». Je crois que ce sont des Français en plus [rire], mais en bon français « EcoLogits Calculator ». Ce site te permet d'aller comparer. Tu te dis : « j'aimerais bien comparer le modèle ChatGPT-3.0 avec le modèle Claude X, Y, Z sur une tâche particulière ». Exemples : Échange texte prompt 400, créer une image, écrire un rapport de 5 000 mots.

Et puis ils ont fait des évaluations, des estimations parce que, parfois, avec certains modèles, il n'y a pas tous les détails. Ce sont des estimés sur un paquet d'hypothèses qui sont, de manière transparente, partagée. Tu peux aussi avoir la méthodologie sur le site. Tu vas avoir cette tâche avec ces deux modèles et voilà ce que ça requiert en eau, en énergie, l'impact CO₂ et tu as même des équivalences qui sont intéressantes pour les ordres de grandeur. Parce que si je te dis neuf wattheures, ça ne nous parle pas trop. Mais, si je te dis que c'est équivalent à cinq kilomètres en voiture, c'est l'équivalent de courir une heure et demie. Tu vois? Il y a des équivalences sur des choses qui nous parlent plus. Donc, il y aurait ça, le site EcoLogits.

Et puis le site Compar:IA, que vous connaissez peut-être et qui est un projet pilote mené par le ministère de la Culture en France. Comparia.gov, je crois. Là aussi, tu vas pouvoir comparer

des modèles. Alors, tu vas te dire : « je vais comparer le modèle Goliath, le super modèle, ton Hummer du départ, avec un petit modèle léger » et tu vas écrire un « prompt » et tu vas voir les résultats. Là, tu vas évaluer la qualité des résultats. Tu vas dire, ça c'est utile, ça c'est concis, ça c'est bien pour les deux. Après, tu appuies sur le bouton et tu dis : « révèle-moi les résultats », parce que tu ne les connais pas les modèles. Et là, on va dire : « ça c'était le modèle A, ça c'était le B ». Donc, c'est un petit peu un jeu. Eux, ça leur permet de collecter les données sur comment les utilisateurs vont évaluer les modèles. Il y a des indicateurs environnementaux aussi qui sont fournis. Le but, je crois, de ce projet, c'était plus de travailler sur les biais culturels de différents modèles, projet porté par le ministère de la Culture, mais tu vas avoir des indicateurs environnementaux aussi.

Tanya Grenier :

Alors, avec un de ces outils, je serais capable de voir lequel répond le mieux à mon besoin. Puis, avec l'autre, je peux voir, avec une plus grande portée, l'impact que ça va avoir d'utiliser ce modèle-là versus un autre.

Camille Grange :

Oui, c'est ça. Puis, il y a sûrement d'autres outils. Ce sont des outils intéressants ces deux-là, parce que ce sont des outils grand public. Gardez en tête que c'est tous des estimés, mais qu'on est de plus en plus capables de faire des estimés corrects de l'empreinte environnementale de ces modèles d'IA.

Tanya Grenier :

Ça permet de faire des choix pas mal plus informés.

Camille Grange :

Tu sais, il n'y a pas juste l'aspect électricité et l'impact climatique. D'ailleurs, une étude récente que j'aimerais te recommander « The Green IT », des Français aussi, je crois, publiée en octobre 2025, qui s'appelle l'impact ou l'empreinte environnementale et sanitaire de l'IA. Leur estimation, c'est qu'à peu près 30 % de l'impact environnemental, c'est du réchauffement climatique, c'est l'émission de CO₂, de gaz à effet de serre. Les autres 70 %, c'est la diminution de la qualité de l'eau, l'eutrophisation, c'est l'eau qui devient trop riche en nutriments, en azote, etc., qui est très néfaste pour les écosystèmes, et donc, pour nous. Diminution de biodiversité et tout cela. L'épuisement des ressources. Tous les minerais dont on a besoin pour aller extraire les métaux dont on a besoin pour construire les serveurs, etc. Tout ça plus l'émission de particules, donc la pollution de l'air. Ces trois éléments, c'est 60 % de l'impact.

Il n'y a pas juste les besoins en électricité et l'impact sur le réchauffement climatique, qui est déjà en soi dramatique, alors que tous les secteurs essaient évidemment de réduire et de revoir leur modèle économique pour adresser la crise climatique.

David Di Pietro :

Tanya, je te vois osciller entre l'espoir et le doute [rire].

Tanya Grenier :

Et le désespoir peut-être [rire]. Mais oui, en fait, la crise climatique est quand même le plus gros enjeu mondial qu'on a en ce moment. Et de voir à quel point l'IA vient taper dans toutes les catégories, c'est pas mal plus stressant que juste prendre son auto pour aller à l'épicerie.

Camille Grange :

Tu sais, Tanya, il faut garder en tête un chiffre. Les émissions de gaz à effet de serre liées au numérique, c'est à peu près 4 %. C'est des estimés.

Tanya Grenier :

Ça reste encore des avions.

Camille Grange :

Par rapport, comme tu dis, aux transports, à l'agriculture, aux bâtiments, à l'industrie, c'est moins. Mais tous ces autres secteurs sont en train de réfléchir à des solutions pour réduire drastiquement, pour revoir leur modèle d'affaires. Je l'ai vécu cette dissonance cognitive. Elle ne me plaît pas, j'essaie de mieux comprendre et de sensibiliser via les fresques du numérique notamment. Je ne sais pas si tu connais. Il faut voir de manière aussi globale et penser qu'il n'y a pas juste le numérique, il y a d'autres domaines qui ont des empreintes plus importantes, évidemment, et qui pourraient avoir un impact très fort si on arrive à changer les modèles et à passer sur des modèles qui n'ont pas besoin de ressources.

Tanya Grenier :

Ça n'empêche pas le besoin de responsabilisation pour les IAG qui sont en train d'exploser.

David Di Pietro :

Tu as donné des trucs pour réduire, un tant soit peu son empreinte carbone, tout à l'heure. Notamment d'avoir une approche qui est réfléchié minimalement. Est-ce qu'il y a d'autres gestes qu'on peut adopter justement pour essayer de réduire cette empreinte numérique.

Camille Grange :

Oui, on a parlé beaucoup des usages et du choix de l'outil ou du modèle. La première chose qu'on peut faire tous, c'est de réduire les usages d'une certaine manière, de réduire l'achat de nos équipements. Prendre soin de nos appareils, aller les faire réparer quand on peut. Il y a une super organisation qui s'appelle Insertech à Montréal, aux Shops Angus, qui font des

ateliers pour nous aider à réparer ou qui peuvent réparer nos terminaux. Arrêter d'acheter cinq milliards d'objets connectés qui ne servent pas à grand-chose et qui ne finissent souvent pas recyclés. Puis, même quand on recycle, il n'y a pas grand-chose qui peut être extrait. Et tout ça, c'est invisible pour nous, toute cette partie déchets électroniques. Et quand on s'y penche un petit peu, c'est assez dramatique. Toute la partie extraction, fabrication et déchets, c'est la super face cachée pour nous. C'est une des premières choses à faire.

Après, sur l'utilisation, de petites choses simples. Utiliser le wifi plutôt que la 5G pour faire vos requêtes. Éteindre votre ordinateur en fin de journée. Essayez de fermer les onglets ouverts dans le navigateur et, je sais, moi aussi, je pense qu'on est tous coupables. Ce sont de petites choses, mais en fait, c'est d'essayer de nous habituer à être réflexifs sur nos usages.

David Di Pietro :

Parce que, je dis coupable, mais c'est plus d'être réfléchi finalement. Comme toujours, face aux grands enjeux, je pense qu'il faut y aller, un pas à la fois. Puis, malheureusement, c'est là-dessus qu'on va terminer notre échange qui était vraiment très intéressant. Mais, en terminant, parlant d'un pas à la fois, on termine toujours en demandant à la personne invitée de nous lancer le défi de la semaine. Est-ce que tu as un défi pour nous, Camille? Pour faire un prochain petit pas, peut-être.

Camille Grange :

Un défi qui pourrait être un défi de mesure. Une première chose hyper facile que vous pouvez faire, c'est aller télécharger, par exemple, un petit « plug-in » de Chrome ou de Firefox, ou de votre navigateur qui s'appelle « Carbon Analyzer ». J'imagine des Français. Et vous partez une session de « Carbon Analyzer » qui est connectée à votre navigateur, puis là, vous allez faire des choses, vous allez naviguer le web, faire vos recherches Google, allez sur YouTube, etc. Et puis, quand tu fermes la session, tu peux voir où est-ce que ton empreinte a été la plus forte : tes besoins, ta consommation, tes émissions de gaz à effet de serre... Tu indiques où tu es, ou bien il te détecte. Et tu peux voir, si tu étais en Chine, ce que cette utilisation aurait coûté. Et pour les plus aventuriers, essayez d'aller réfléchir à l'utilisation de modèles locaux pour des tâches de tous les jours. Ça pourrait être un bon challenge.

David Di Pietro :

Merci beaucoup. Alors défi accepté.

Tanya Grenier :

Tout à fait [rire]!

Camille Grange :

Super!

David Di Pietro :

Alors, Tanya Grenier, Camille Grange, merci beaucoup pour votre présence avec nous.

Tanya Grenier :

Merci bien.

Camille Grange :

Merci. C'était très plaisant.

David Di Pietro :

Merci d'avoir été à l'écoute de Hashtag Génération IA. À très bientôt pour un prochain épisode!

[musique]

Hashtag Génération IA est un balado soutenu par les cellules de travail en intelligence artificielle générative à HEC Montréal.