Des fonctionnalités à coût maîtrisé ? Le modèle A-U appliqué à l'IA générative.

Robert Viseur[1]

[1] UMONS, FWEG, Service TIC

robert.viseur@umons.ac.be

Résumé

L'essor des IA génératives (IAG) donne lieu à des projections alarmistes quant à leur impact environnemental. Ce risque est-il correctement évalué ? L'innovation tend à suivre un cycle où les efforts se concentrent sur le produit puis son processus de production et enfin sa simplification (modèle A-U). Ce cycle s'applique-t-il également aux IAG ? L'analyse des premières IA basées sur le deep learning montre le potentiel lié aux optimisations matérielles et logicielles. Ce type d'optimisation est-il mis en œuvre dans le cas des IAG ? Les mesures d'optimisation peuvent porter sur le matériel mais aussi sur les données ou sur les modèles eux-mêmes. Compte tenu de la forte pression sur les coûts, la tendance actuelle pencherait davantage vers une croissance maîtrisée des coûts financiers et environnementaux, du fait notamment des économies d'échelle. La recherche met en avant l'importance de l'optimisation des coûts d'inférence pour l'atteinte de la rentabilité des grands chatbots internationaux. Une perspective se dégage par ailleurs en matière de développement collaboratif des IAG sur le principe des fondations open-sources.

Mots-clés

Intelligence artificielle générative, grand modèle de langage, sobriété numérique.

Abstract

The rise of generative AI (GAI) has led to alarmist projections regarding its environmental impact. Is this risk being properly assessed? Innovation tends to follow a cycle where efforts initially focus on the product, then on its production process, and finally on its simplification (A-U model). Does this cycle also apply to GAI? An analysis of early deep learning-based AI systems highlights the potential of hardware and software optimisations. Are such optimisations being implemented in the case of GAI? Optimisation measures can target hardware but also data or the models themselves. Given the strong cost pressures, the current trend seems to favour a controlled growth in financial and environmental costs, particularly due to economies of scale. Research highlights the importance of optimising inference costs to achieve profitability for major international chatbots. Another emerging perspective concerns the collaborative development of GAI, based on the principles of open-source foundations.

Keywords

Generative artificial intelligence, large language model, digital sobriety.

1 Contexte

Le développement de l'intelligence artificielle (IA) suscite des inquiétudes quant à son impact environnemental. Ainsi, Sundberg (2023) soutient que l'IA présente « *une empreinte carbone en croissance rapide* », liée à la consommation d'énergie due à son exploitation mais aussi à la fabrication du matériel. Ces alertes proviennent aussi de deux entités pourtant souvent antagonistes : les associations écologistes et les *bigtechs*. D'une part, dans un rapport publié en mars 2024, le CAAD (2024) s'alarmait d'un doublement possible du nombre de centres de données occasionnant « *une augmentation de 80 % des émissions globales de* CO_2 » pour ces infrastructures. D'autre part, la production d'énergie décarbonée, notamment nucléaire (Jeans, 2025), ressort comme une préoccupation importante des *bigtechs* (IEA, 2025a). Côté science, les inquiétudes quant aux impacts environnementaux ont conduit au développement de recherches sur les IA durables dont l'objectif est « *de développer des outils d'IA plus frugaux* » (Vuarin et al., 2023). Sur quels constats chiffrés ces inquiétudes sont-elles basées ?

Concrètement, l'entraînement du grand modèle de langage (LLM, *Large Langage Model*) GPT-3 aurait nécessité 1,3 GWh, soit la consommation annuelle moyenne de 120 foyers étasuniens, occasionnant l'émission de 522 tonnes de CO_2 (Sundberg, 2023). Par ailleurs, une seule requête ChatGPT générerait « *100 fois plus de carbone qu'une recherche Google classique* » (Sundberg, 2023). Quant à BLOOM, l'énergie consommée pour son entraînement est estimée à 433 MWh (Luccioni et al., 2023). Les performances à un instant donné d'une technologie peuvent-elles valablement être projetées pour en analyser l'évolution future ?

Cette comparaison entre ChatGPT et Google pourrait en effet erronément laisser penser que cette proportion est immuable. En réalité, l'innovation technologique suit traditionnellement un cycle de vie industriel. Décrit par William J. Abernathy et James M. Utterback, il a été baptisé « *modèle A-U* ». Les auteurs distinguent trois phases (Roth, 2016 ; Trott, 2021). La première phase voit

l'innovation stimulée par les besoins. Elle porte dès lors essentiellement sur le produit jusqu'à ce qu'une classe de produit obtienne la reconnaissance du marché. Ce « *design dominant* » va accélérer l'innovation de processus. L'objectif devient alors moins d'améliorer les performances de la technologie que de réduire les coûts de production ou d'exploitation. La maturité s'accompagne progressivement de sa simplification et, dès lors, d'une réduction importante des prix. Le produit devient petit à petit une commodité tandis que l'innovation de produit peut se relancer sur de nouvelles niches de marché. Peut-on appliquer ce cycle aux applications d'intelligence artificielle ?

2 Gains d'efficacité en IA

Patterson et ses co-auteurs (2022) critiquent en tout cas la surestimation des émissions dans les prévisions du fait, d'une part, de la propagation de données erronées fournies dans des études fréquemment citées, d'autre part, l'ignorance des améliorations en continu des systèmes d'apprentissage logiciel. Ainsi, l'optimisation de l'exploitation des modèles conduirait à des gains constatés de l'ordre de 100 pour la consommation et de 1000 pour les émissions (du fait de la décarbonatation de l'approvisionnement en énergie chez les gestionnaires d'infrastructures). Sont dès lors distingués quatre points sur lesquels agir pour réduire les émissions de CO_2 : (1) le modèle en lui-même, (2) le matériel permettant son exécution, (3) le centre de données centralisant le matériel et (4) l'utilisation d'énergie décarbonée par ce dernier (Patterson et al., 2022). Les auteurs pointent ainsi que le PUE[1] moyen des centres de données industriels est de 1,58 contre 1,1 environ pour les fournisseurs d'infrastructures publiques de *cloud computing*. Au final, la consommation des centres de données apparaît globalement sous contrôle (croissance modérée), excepté dans certains pays comme l'Irlande (IEA, 2025b), et ce, malgré un contexte d'explosion du trafic (IEA, 2025a). Ces estimations optimistes s'appliquent-elles également aux intelligences artificielles génératives et, plus particulièrement, aux grands modèles de langage (LLM) ?

Plusieurs recherches tendent à montrer que les gains observés sur la durée en apprentissage profond s'observent également avec les grands modèles de langage. Patterson et ses co-auteurs (2022) fournissent ainsi l'exemple de GLaM, un modèle apparu 18 mois après GPT-3, dont la consommation énergétique a été réduite par un facteur 3. Ho et ses co-auteurs (2024) ont réalisé une évaluation sur plus de 400 LLM. Les chercheurs ont montré que les modèles voyaient leurs besoins en ressources de calcul réduits d'un facteur 2 tous les 8 mois environ. IEA (2025a) fournit la répartition de la consommation d'énergie par les intelligences artificielles. Jusqu'à 10 % concerne le développement du modèle (expérimentation), entre 20 et 40 %, son entraînement ainsi que de 60 à 70 % pour l'inférence c'est-à-dire l'utilisation en production du modèle. L'IAG apparaît donc ici comme une industrie propice aux économies d'échelle. La consommation liée à l'entraînement reste inchangée quelle que soit la base d'utilisateurs tandis que la centralisation des infrastructures permet un abaissement du coût unitaire associé à chaque requête. Des optimisations permettent donc, d'une part, une réduction des coûts d'entraînement, d'autre part, une réduction des coûts d'inférence. Cependant, ne cachent-elles pas une évolution de la consommation globale des infrastructures artificielles génératives compte tenu d'un possible « *effet rebond* » ?

3 Question de l'effet rebond

L'effet rebond correspond au fait que « *l'accroissement des consommations de matières et d'énergie induit par l'utilisation généralisée des TIC efface largement les réductions de l'empreinte écologique obtenues par unité de produit* » (Flipo & Gossart, 2009). Dit autrement le développement des usages feraient plus que compenser les gains dus à l'optimisation.

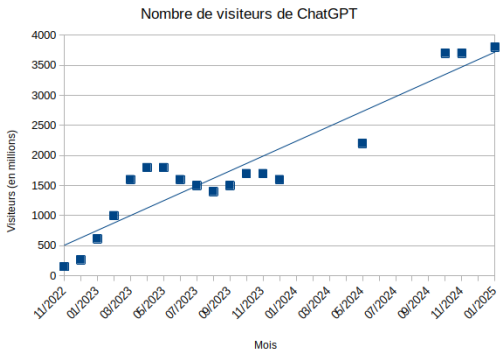

Figure 1. Nombre de visiteurs de ChatGPT (source : Similarweb).

Deux éléments sont susceptibles de contribuer à cet effet rebond : la croissance de la base d'utilisateurs et celle de la complexité des tâches déléguées à l'IA. Premièrement, la diffusion de la technologie d'intelligence artificielle générative conduit à un accroissement de la base d'utilisateurs. Cependant, force est de constater qu'après une croissance rapide en novembre 2022, pour atteindre 100 millions d'utilisateurs en deux mois (Hu, 2023), l'augmentation du nombre d'utilisateurs chez ChatGPT s'est ensuite déroulée de manière sensiblement plus lente (cf. Figure 1). Ce ralentissement[2] pourrait s'expliquer par plusieurs facteurs. D'une part, les agents conversationnels demeurent des outils pour adopteurs précoces (cf. Trott,

1 Le PUE, ou *Power Usage Effectiveness*, est une mesure d'efficacité définie comme le ratio entre l'énergie consommée par tout le centre de données et celle consommée par les seuls équipements informatiques (Patterson et al., 2022).

2 Ces statistiques, d'une part, ne couvrent pas l'usage des LLM au travers des API, d'autre part, négligent la croissance éventuellement captée, début 2025, par des *chatbots* concurrents (p. ex. Mistral et DeepSeek).

2021). En effet, ils sont majoritairement utilisés par les plus jeunes (Bianchi & Angulo, 2024 ; Ma et al., 2024). De plus, et malgré l'apparente convivialité des outils, les usages avancés supposent l'acquisition d'une expertise (Ma et al., 2024), notamment en promptologie (Yao et al., 2024). La diffusion à la majorité des utilisateurs potentiels est dès lors plus lente. D'autre part, le développement commercial de ces outils a été émaillé d'indisponibilités dues à la lourdeur des premiers modèles. Le manque de capacités de calcul a ainsi conduit à la mise en pause des inscriptions à la version payante ChatGPT Plus en novembre 2023[3]. Deuxièmement, les usages s'orientent progressivement vers des tâches complexes nécessitant des modèles capables d'exécuter des raisonnements (p. ex. OpenAI o1) ou d'automatiser le traitement de davantage de documents (p. ex. fonctions de recherche approfondie *i.e.* « Deep Search »). Samsi et ses co-auteurs (2023) montrent ainsi que l'utilisation de modèles de plus grande taille occasionne des coûts énergétiques accrus lors de l'inférence. Les modèles les plus simples, moins consommateurs, se trouveraient dès lors progressivement remplacés par des modèles aux capacités accrues, plus consommateurs. Qu'en est-il réellement ? Quelles motivations, quelles pratiques sous-tendraient des gains d'efficacité dans le cas des LLM ?

4 Optimisation des LLM

Sur le plan des motivations, trois éléments poussent à l'amélioration de l'efficacité énergétique des LLM : les coûts de production (entraînement), les coûts d'exploitation (inférence) et les tensions d'approvisionnement des GPU (Oremus, 2023 ; Isaac & Griffith, 2024 ; Stokel-Walker, 2024). Les plans tarifaires, qu'il s'agisse de ChatGPT Plus, à 20 dollars par mois, ou ChatGPT Pro, à 200 dollars par mois, peinent d'ailleurs à être rentables (Quiroz-Gutierrez, 2025 ; Oremus, 2023). C'est ce qui explique, par exemple, que les premiers utilisateurs payants de GPT-4 « *pouvaient envoyer 25 requêtes seulement toutes les 3 heures car il était trop coûteux à exécuter* » (Oremus, 2023). Les producteurs d'IAG sont dès lors pris en tenaille entre, d'une part, l'importance des coûts et la confrontation à une limite physique (devoir faire avec un stock limité de puces dans un contexte de forte demande ; Bradshaw & Morris, 2024), et, d'autre part, la croissance limitée des revenus. Celle-ci est liée aux tarifs modestes des abonnements (*chatbots*) et du paiement à l'usage (API), maintenus sous pression par la concurrence (OpenAI, Gemini, Microsoft, Mistral, Claude, DeepSeek-AI…). De plus, la recherche de rentabilité encourage les innovations de processus. En quoi ces dernières consistent-elles ?

Sur le plan des pratiques, et pour analyser les opportunités d'optimisation des LLM, deux axes vont être retenus, d'une part, les quatre dimensions identifiées par Patterson et ses co-auteurs (2022), à savoir le modèle, le matériel, les centres de données et les sources d'énergie, et, d'autre part, les trois étapes identifiées par l'IEA, à savoir le développement, l'entraînement et l'inférence. Parmi les quatre dimensions précitées, nous allons négliger les centres de données et les sources d'énergie, car elles relèvent de politiques d'optimisation plus générales au secteur du numérique.

Sur le plan du matériel, l'entraînement reste tributaire de la disponibilité (Smith, 2024), et des progrès, des GPU (*Graphics Processing Unit*). Par contre, l'inférence s'accommode de puces spécifiques telles que, chez la société Groq, les LPU (*Language Processing Unit*) . Ces derniers se distinguent par un meilleur temps de réponse et une efficacité accrue (Ward-Foxton, 2023). Par ailleurs, les gestionnaires de grands centres de données (*hyperscalers*) tels que Amazon, Google, META et Microsoft travaillent sur leurs propres processeurs dédiés à l'IA (Smith, 2024).

Sur le plan des modèles, les optimisations sont à la fois organisationnelles et techniques. Premièrement, la collaboration permet le partage des coûts de développement puis, surtout, d'entraînement. Dans le premier cas, les jeux de données et les modèles peuvent être mis en commun au sein d'un consortium ou d'une communauté. Dans le second cas, le modèle des fondations, propres aux logiciels libres, pourrait servir d'exemple (Viseur, 2024). Deuxièmement, les méthodes d'entraînement peuvent être améliorées (p. ex. *early stopping*, *sparse training* et *gradient accumulation*). Troisièmement, les modèles en eux-mêmes peuvent être optimisés. Cela passe par la réduction du poids des nœuds au sein du réseau de neurones, le principe du MoE (*Mixture of Experts*) et la réduction de la taille des modèles (Gent, 2023). Eldan et Li (2023) ont ainsi montré qu'il était possible de développer un SLM (*Small Langage Model*), à l'image de TinyStories, capable de rivaliser, sur le plan des capacités de génération de texte, avec un LLM (GPT-2), grâce à un effort sur la conception du jeu de données. Des modèles moins lourds (poids réduits), exécutés partiellement (MoE), plus petits (SLM) nécessitent moins de ressources de calcul et voient donc leur empreinte environnementale réduite. Quelles sont les incitations économiques à mettre en œuvre ces différentes méthodes ?

5 Rationalité économique

Afin d'illustrer notre réflexion, nous traitons le cas suivant. En prenant des informations publiées sur le coût d'entraînement et le coût d'inférence par mille *tokens* de GPT-4[4], ainsi que d'autres hypothèses incluant un nombre mensuel de requêtes par utilisateur de 100[5], une taille

3 Cf. https://x.com/sama/status/1724626002595471740.

4 Cf. https://patmcguinness.substack.com/p/gpt-4-details-revealed.

5 Ce nombre de 100 *prompts* par utilisateur en moyenne est estimé sur base des chiffres évoqués fin 2024 par Sam Altman, et donnant le nombre quotidien de messages et le nombre hebdomadaire d'utilisateurs

moyenne de requête de 250 *tokens*[6] et un taux de conversion de 1 % au modèle *freemium*[7], nous calculons, pour un nombre croissant de *prompts*, le nombre d'utilisateurs, une estimation du prix de revient au *prompt* et le surplus (chiffre d'affaires moins coûts d'entraînement moins coûts d'inférence). Cette évaluation très simplifiée (cf. Tableau 1) permet de dresser quelques conclusions intéressantes quant à la rationalité de certains choix.

Coût total d'entraînement :	$63.000.000
Durée de vie d'un modèle (mois) :	12
Coût d'inférence par 1000 *tokens* :	$0,002
Taille moyenne d'une requête (*tokens*) :	250
Coût d'inférence par *prompt* :	$0,0010
Prompts mensuels par utilisateur :	100
Taux de conversion (Plus) :	5 %
Revenus par utilisateur (Plus) :	$20

Prompts	**Utilisateurs**	**Prix de revient**	**Surplus**
1.000.000	10.000	$5,2510	-$5.241.000
100.000.000	1.000.000	$0,0535	-$4.350.000
583.333.333	5.833.333	$0,1090	$0
5.000.000.000	50.000.000	$0,0021	$39.750.000
50.000.000.000	500.000.000	$0,0011	$444.750.000
500.000.000.000	5.000.000.000	$0,0010	$4.494.750.000

Tableau 1. Simulation de prix de revient.

Premièrement, nous constatons que le coût total d'inférence devient rapidement, pour les prestataires internationaux devant satisfaire plusieurs centaines de millions d'utilisateurs, d'un ordre de grandeur comparable au coût d'entraînement (ramené à un coût fixe mensuel compte tenu de la durée de vie moyenne d'un modèle). Le modèle des fondations propre au logiciel libre peut donc se révéler pertinent pour le partage des coûts d'entraînement, en particulier pour les organisations de taille moyenne, devant gérer mensuellement quelques millions de *prompts* au maximum. Deuxièmement, la rentabilité des agents conversationnels n'est possible que pour des prestataires de grande taille. De fait, un surplus n'est dégagé qu'au-delà des 6 millions d'utilisateurs mensuels. Troisièmement, nous constatons qu'un prestataire dominant tel qu'OpenAI présente un problème plus global de modèle d'affaires. En effet, ce surplus reste éloigné d'un bénéfice d'exploitation. D'une part, il n'intègre pas d'autres coûts incluant par exemple les salaires. D'autre part, il est calculé en négligeant les coûts supplémentaires, liés aux usages avancés (p. ex. chaînes de raisonnement, recherche approfondie, multimodalité et agents) ainsi que pour l'entraînement et l'inférence d'autres modèles inclus dans l'abonnement. Cela concerne notamment DALL-E et Sora, respectivement pour les images et les vidéos. Luccioni et ses co-auteurs (2024) montrent ainsi que le coût d'inférence pour une génération d'images est en moyenne 61 fois plus important que pour du texte. Notre simulation permet donc de tirer des conclusions utiles à l'atteinte de la rentabilité. Tout d'abord, compte tenu de l'importance des coûts fixes (p. ex. coûts d'entraînement), des bases installées importantes sont nécessaires. Au-delà des 50 millions d'utilisateurs, le coût d'entraînement (mensualisé) devient en effet inférieur au coût d'inférence. De plus, les opportunités d'économies d'échelle poussent à la consolidation du secteur. Ensuite, le modèle *freemium* montre ses limites. La concurrence empêche de réduire les prestations de la version gratuite (ce qui met les taux de conversion sous pression) et d'augmenter les coûts d'abonnement. Augmenter les revenus se révèle difficile sauf à faire évoluer le modèle d'affaires. C'est notamment ce que fait Microsoft, dont l'IA est désormais financée au travers des abonnements aux logiciels de productivité (Crider, 2025). Sur ce plan, OpenAI pourrait par exemple développer une offre de liens sponsorisés adaptée à l'usage de ChatGPT comme moteur de recherche. Enfin, l'abaissement des coûts unitaires d'inférence impacte directement les coûts compte tenu du volume de requêtes exécuté sur les agents conversationnels internationaux. Quant à celui des coûts d'entraînement (fixes), il conduit à une réduction mécanique du prix de revient d'une requête. L'optimisation, tant de l'entraînement que de l'inférence, se révèle donc une nécessité, voire un impératif de survie, pouvant s'appuyer, d'une part, sur des innovations matérielles (p. ex. nouvelles puces), d'autre part, sur des innovations logicielles (p. ex. nouvelles stratégies d'entraînement et réduction du poids des modèles utilisés).

6 Cas de bonnes pratiques

Deux modèles se sont distingués début 2025 pour leurs pratiques propices à davantage de frugalité : LUCIE et DeepSeek-R1.

Le LLM LUCIE, développé par OpenLLM France et l'ESN Linagora, met en œuvre certaines des bonnes pratiques organisationnelles identifiées. Premièrement, le projet s'appuie sur une infrastructure partagée, le supercalculateur Jean Zay. Lucie-7B a ainsi été entraîné sur 512 GPU NVIDIA H100 pour un total d'environ 550.000 heures de calcul[8]. Cette infrastructure permet d'atteindre une taille critique minimale, d'en accroître

respectivement à 1 milliard et à 300 millions ; cf. https://www.linkedin.com/posts/rowancheung_sam-altman-just-dropped-some-new-chatgpt-activity-7270172267223904257-lAfL/.

6 Cette valeur inclut les *tokens* donnés en entrée (*prompt*) et ceux inclus dans la réponse.

7 Cette valeur de 5 % a été retenue sur base du nombre de clients (soit environ 5 millions), extrapolables du chiffre d'affaires (2023) en rythme annuel (1,3 milliards), rapporté au nombre d'utilisateurs hebdomadaires (environ 100 millions), pris comme base d'utilisateurs actifs.

8 Cf. https://huggingface.co/OpenLLM-France/Lucie-7B.

l'efficacité et d'en augmenter le taux d'utilisation. De plus, LUCIE-7B adopte une stratégie open-source complète. En effet sont publiés les données d'entraînement (cf. Lucie-Training-Dataset), les scripts d'entraînement (cf. Lucie-Training, un *fork* de Megatron-DeepSpeed) et les modèles eux-mêmes (cf. LUCIE-7B sur Hugging Face). Cette approche permet une optimisation des jeux de données, par ailleurs réutilisables sans nécessiter une collecte massive, et des méthodes d'entraînement, ainsi qu'une réutilisation plus aisée, sous sa forme originale ou après spécialisation.

Le LLM DeepSeek-R1, développé par la société chinoise DeepSeek-AI, a été publié dans le contexte de tensions et de concurrence croissante entre les États-Unis et la Chine. Le modèle se distingue par un entraînement partiel sur des GPU domestiques, sans que ne soit ici claire la frontière entre gains réels et propagande visant à démontrer l'inefficacité des restrictions étasuniennes à l'exportation des puces les plus puissantes. Par ailleurs, le modèle se distingue par, d'une part, l'adoption d'une stratégie open-source (publication du modèle et de la méthode d'entraînement), d'autre part, une architecture MoE permettant un abaissement des coûts d'inférence[9]. DeepSeek-R1 revendique des performances comparables aux meilleurs modèles étasuniens en matière de raisonnement (Guo et al., 2025). Malgré ses performances, DeepSeek-R1 relève davantage de l'innovation de processus du fait des efforts entrepris pour abaisser les coûts d'entraînement et d'inférence.

7 Conclusion

Le cycle décrit par le modèle A-U se retrouve dans le développement actuel de l'IAG. Premièrement, des modèles permettant des tâches simples ont été mis sur le marché (p. ex. GPT-3.5). Ces modèles ont démontré leur utilité, ainsi que le plébiscite pour le design de l'agent conversationnel, mais aussi leurs limites pour des tâches plus complexes impliquant par exemple des chaînes de raisonnement. Deux pressions à l'innovation en ont découlé. D'une part, les limitations ont encouragé la mise sur le marché de modèles plus performants et plus lourds (p. ex. OpenAI o1). D'autre part, l'existence d'une base d'utilisateurs pour des modèles simples a encouragé leur optimisation. La continuation de l'innovation de produit, sur de nouveaux modèles, s'est dès lors complétée d'une innovation de processus visant à réduire les coûts de production (nouvelles stratégies d'entraînement, publication de modèles open-sources…) et d'exploitation (optimisation des centres de données ; agents, puces spécialisées...), à performance inchangée. La *commoditisation* s'observe également avec la publication de modèles sensiblement plus petits, plus frugaux, mais néanmoins adaptés à la réalisation de tâches spécifiques (p. ex. RAG). Si des signaux de croissance de la consommation de ressources énergétiques et matérielles existent, ils s'accompagnent d'une amélioration continue des infrastructures et des modèles, propice à une perpétuation du découplage actuellement constaté entre les usages et les ressources consommées. L'IEA prévoit cependant une croissance régulière de la consommation électrique des centres de données, autour de 15 % par an (ce qui aboutirait à un doublement de la consommation après 5 ans), avec de fortes disparités locales ou régionales (IEA, 2025b). Par ailleurs, dresser un bilan complet nécessiterait une mise en balance avec les économies éventuellement permises par l'automatisation à l'aide de modèles génératifs (Tomlinson et al., 2024). Enfin, cette recherche identifie un ensemble de thématiques sur lesquelles des activités de recherche ciblées permettraient de contribuer à la réduction du poids environnemental de ces systèmes techniques.

9 Cf. https://api-docs.deepseek.com/news/news250120.

Références

[1] Bianchi, T. & Angulo, F. (2024), Online search after ChatGPT: the impact of generative AI. *Semrush & Statista*. https://static.semrush.com/file/docs/evolution-of-online-after-ai/Online_Search_After_ChatGPT.pdf.

[2] Bradshaw, T. & Morris, S. (2024). Microsoft acquires twice as many Nvidia AI chips as tech rivals. *Financial Times*, 17 décembre 2024. https://www.ft.com/content/e85e43d1-5ce4-4531-94f1-9e9c1c5b4ff1.

[3] CAAD (2024). Artificial Intelligence Threats to Climate Change. *Climate Action Against Disinformation*, 7 mars 2024. https://foe.org/wp-content/uploads/2024/03/AI_Climate_Disinfo_v6_031224.pdf.

[4] Crider, M. (2025). En intégrant Copilot, Microsoft365 voit ses tarifs augmenter pour les particuliers. *Le Monde Informatique*, 21 janvier 2025. https://www.lemondeinformatique.fr/actualites/lire-ia-et-cybersecurite-priorites-des-ssii-et-editeurs-francais-en-2024-96067.html.

[5] Eldan, R., & Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English?. *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.07759.

[6] Flipo, F., & Gossart, C. (2009). Infrastructure numérique et environnement. L'impossible domestication de l'effet rebond. *Terminal. Technologie de l'information, culture & société*, (103-104). https://doi.org/10.4000/terminal.3093.

[7] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681-694. https://www.doi.org/10.1007/s11023-020-09548-1.

[8] Gent, E. (2023). When AI's Large Language Models Shrink. *IEEE Spectrum*, 31 mars 2023. https://spectrum.ieee.org/large-language-models-size.

[9] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*.

https://doi.org/10.48550/arXiv.2501.12948.

[10] Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2403.05812.

[11] Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*, 2 février 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[12] IEA (2025b), Energy and AI, *International Energy Agency*. https://www.iea.org/reports/energy-and-ai.

[13] IEA (2025a). Data Centres and Data Transmission Networks, *International Energy Agency*. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

[14] Isaac, M., & Griffith, E. (2024). OpenAI Is Growing Fast and Burning Through Piles of Money. *The New York Times*, 27 septembre 2024. https://www.nytimes.com/2024/09/27/technology/openai-chatgpt-investors-funding.html.

[15] Jeans, D. (2025). Sam Altman's Fusion Power Startup Is Eyeing Trump's $500 Billion AI Play. *Forbes*, 5 février 2025. https://www.forbes.com/sites/davidjeans/2025/02/05/stargate-sam-altman-fusion-helion/.

[16] Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 85-99). https://doi.org/10.1145/3630106.3658542.

[17] Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253), 1-15. https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf.

[18] Ma, L., Xu, X., He, Y., & Tan, Y. (2024). Learning to Adopt Generative AI. *arXiv preprint*. https://doi.org/10.48550/arXiv.2410.19806.

[19] Oremus, W. (2023). AI chatbots lose money every time you use them. That is a problem. *The Washington Post*, 5 juin 2023. https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/.

[20] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7), 18-28. https://doi.org/10.1109/MC.2022.3148714.

[21] Quiroz-Gutierrez, M. (2025). Sam Altman says he's losing money on OpenAI's $200-per-month subscriptions: 'People use it much more than we expected'. *Fortune*, 7 janvier 2025. https://fortune.com/2025/01/07/sam-altman-openai-chatgpt-pro-subscription-losing-money-tech/.

[22] Roth, F. (2016). V. William J. Abernathy et James M. Utterback. *Le cycle des innovations technologiques*. In Les Grands Auteurs en Management de l'innovation et de la créativité (pp. 103-120). EMS Editions. ISBN : 978-2-84769-812-1.

[23] Samsi, S., Zhao, D., ... & Gadepally, V. (2023). From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1-9). IEEE. https://doi.org/10.1109/HPEC58863.2023.10363447.

[24] Smith, M. S. (2024). Challengers are Coming for NVidia's Crown: In AI's Game of Thrones, Don't Count Out the Upstarts. *IEEE Spectrum*, 61(10), 40-44. https://doi.org/10.1109/MSPEC.2024.10705376.

[25] Stokel-Walker, C. (2024). AI chatbots are improving at an even faster rate than computer chips. *New Scientists*, 27 mars 2024. https://www.newscientist.com/article/2424179-ai-chatbots-are-improving-at-an-even-faster-rate-than-computer-chips/.

[26] Sundberg, N. (2023). Tackling AI's Climate Change Problem. *MIT Sloan Management Review*, 65(2), 38-41. https://sloanreview.mit.edu/article/tackling-ais-climate-change-problem/.

[27] Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports (Sci Rep)*, 14(1), 3732. https://doi.org/10.17605/OSF.IO/YHTMQ.

[28] Trott, P. (2021). *Innovation Management and New Product Development – Seventh Edition*. Pearson. ISBN : 978-1-2922-5152-3.

[29] van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3), 213-218. https://www.doi.org/10.1007/s43681-021-00043-6.

[30] Viseur, R. (2024). Stratégies open-sources : opportunités et limitations dans le domaine des Large Language Models (LLM). *Inforsid*, Nancy (France), 31 mai 2024. https://hdl.handle.net/20.500.12907/50980.

[31] Vuarin, L., Lopes, P. G., & Massé, D. (2023). L'intelligence artificielle peut-elle être une innovation responsable? *Innovations*, 72(3), 103-147. https://doi.org/10.3917/inno.pr2.0153.

[32] Ward-Foxton, Sally (2023). Groq Demonstrates Fast LLMs on 4-Year-Old Silicon. *EETimes*, 12 septembre 2023. https://www.eetimes.com/groq-demos-fast-llms-on-4-year-old-silicon/.

[33] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.