UMONS
Université de Mons

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

Dr Ir Robert Viseur
Associate Professor

Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle
(APIA 2025 ; événement affilié à PFIA 2025)

30 juin - 1er juillet 2025 (Dijon, France)

POLYTECH MONS

Problématique et question de recherche

Le développement des IA génératives suscite des inquiétudes quant à leur impact environnemental (Sunberg, 2023).

Ces inquiétudes s'inscrivent dans une réflexion plus générale, non seulement sur l'impact environnemental des IA, mais aussi du numérique (Viseur, 2025).

Ces inquiétudes sont relayées par, d'une part, des associations militantes (p. ex. The Shift Project et CAAD), d'autre part, les *bigtechs* elles-mêmes (p. ex. Sam Altman).

D'autres acteurs se révèlent davantage rassurants et pointent la tendance à la réduction du poids des modèles d'IA au fil du temps (p. ex. Gent, 2023).

Le développement de l'IA générative, et en particulier des LLM, va-t-il s'accompagner d'une explosion de l'impact environnemental du numérique, ou les processus d'innovation intègrent-ils une tendance naturelle à la frugalité ?

Évaluation de l'impact environnemental #1

Sur quels constats chiffrés ces inquiétudes sont-elles basées ?

La transparence sur l'impact environnemental de l'IA est faible (Luccioni, 2025), ce qui occasionne un manque de statistiques fiables.

Les chiffrages de l'impact environnemental tournent souvent autour des modèles GPT-3 et BLOOM.

L'entraînement de GPT-3 aurait nécessité 1,3 GWh contre 433 MWh pour BLOOM (Sundberg, 2023 ; Luccioni *et al.*, 2023).

Par comparaison à d'autres outils, une seule requête ChatGPT générerait « *100 fois plus de carbone qu'une recherche Google classique* » (Sundberg, 2023).

Évaluation de l'impact environnemental #2

Les performances à un instant donné d'une technologie peuvent-elles valablement être projetées pour en analyser l'impact futur ?

Modèle A-U

Le modèle A-U, proposé par Abernathy et Utterback, distingue trois phases dans l'innovation d'un produit (Roth, 2016 ; Trott, 2021) :

- la phase fluide (innovation de produit),

  L'innovation est stimulée par les besoins, et se porte sur les caractéristiques du produit, avec une moindre attention sur les coûts.
- la phase transitionnelle (innovation de processus),

  Dès lors que le marché plébiscite un certain design, dit « design dominant », l'innovation se porte progressivement vers la réduction des coûts.
- la phase spécifique (commoditisation).

  L'innovation se concentre sur la simplification du produit (moins de fonctionnalités, moindre coûts).

Peut-on appliquer le modèle A-U au processus d'innovation applicable aux technologies d'intelligence artificielle ?

Optimisation des IA

Patterson et ses coauteurs (2022) dénoncent une surestimation de émissions attribuables aux IA du fait, d'une part, de la propagation de données erronées, d'autre part, de la méconnaissance des mécanismes d'amélioration continue.

Ces améliorations portent sur :

- les modèles,
- le matériel,
- les centres de données,
- les sources d'énergie.

L'histoire récente du *deep learning* comporte des exemples documentés de réduction du poids des modèles (Viseur, 2025)… souvent à performance comparable.

Ces estimations optimistes s'appliquent-elles également aux IA génératives et, plus particulièrement, aux grands modèles de langage (LLM) ?

Optimisation des IAG

La littérature pointe des signes encourageants :

- la réduction par un facteur 3, comparativement à GPT-3, du modèle GlaM (Patterson *et al.*, 2022),
- la réduction des ressources de calcul nécessaires à 400 LLM évalués par un facteur 2 tous les 8 mois environ (Ho *et al.*, 2024).

La consommation d'énergie par une IA se répartit comme suit (IEA, 2025a) :

- 10 % pour le développement du modèle,
- 20 à 40 % pour son entraînement,
- 60 à 70 % pour l'inférence.

L'optimisation de l'entraînement sera rentable quels que soient les usages tandis que celle de l'inférence sera cruciale en cas de base d'utilisateurs importante.

Ces optimisations ne cachent-elles pas une évolution de la consommation globale des IA génératives compte tenu d'un possible effet rebond ?

Question de l'effet rebond

Deux éléments peuvent contribuer à un effet rebond, d'une part, la croissance de la base d'utilisateurs, d'autre part, la complexité des tâches déléguées à l'IA.

La croissance de la base d'utilisateurs a été rapide au lancement de ChatGPT (Hu, 2023), avant de se ralentir, peut-être parce que l'agent conversationnel reste un outil complexe à l'usage (p. ex. prompts), plutôt destiné à des adopteurs précoces.

La croissance des usages a été perturbée par des problèmes de capacité de serveurs gérés par OpenAI pour ses outils d'IA génératives.

Les usages s'orientent partiellement vers la délégation de tâches complexes, avec les LRM (*Large Reasoning Models*) et les fonctionnalités type « *Deep Search* ».

Les modèles anciens vont-ils être remplacés par des modèles plus puissants et plus lourds ?

Optimisation des LLM

Plusieurs facteurs encouragent l'optimisation des LLM :

- l'importance des coûts,
  (coûts d'entraînement, coûts d'inférence)
- les tensions d'approvisionnement sur les puces spécialisées,
  (en particulier : GPU)
- la croissance limitée des revenus (abonnement et paiement à l'usage).
  (causes possibles : concurrence vs modèle freemium)

Les optimisations portent sur :

- la conception de puces spécialisées (Smith, 2024),
- la réduction du poids des modèles (Gent, 2023 ; Eldan & Li, 2023).

Quelles sont les incitations économiques à mettre en œuvre ces optimisations ?

Rationalité économique

La simulation de la marge brute de GPT-4, utilisé au travers de ChatGPT, permet de discuter la rationalité économique des choix d'optimisation.

Premièrement, si pour les *bigtechs* les coûts d'inférence deviennent rapidement prépondérants, il n'en va pas de même pour les acteurs plus modestes, pour qui les modèles des fondations (logiciel libre) pourraient se révéler intéressants.

Deuxièmement, la rentabilité (et la survie) est surtout envisageable pour des acteurs de très grande taille, capables de dégager des économies d'échelle substantielles.

Troisièmement, l'atteinte de la rentabilité est contrariée par la mise à disposition de fonctionnalités (p. ex. « Deep Search ») et de modèles (p. ex. DALL-E) plus lourds.

Quatrièmement, le modèle d'affaires type « freemium » montre ses limite. Se révèlent nécessaires des modèles de revenus alternatifs ou complémentaires comme les abonnements (cf. Microsoft) ou la publicité ciblée (cf. Google).

Cas de bonnes pratiques

Cas de DeepSeek :

DeepSeek-R1 est un LLM avancé capable de rivaliser avec les meilleurs LRM, malgré des coûts réduits d'entraînement et d'inférence (innovation de processus).

La stratégie est open-source (publication du modèle, de la méthode d'entraînement…), dans un contexte de luttes géopolitiques autour de l'IAG (Jolie, 2025).

Cas de LUCIE :

LUCIE-7B est un « petit » LLM développé en open-source. Sont publiés :

- les données d'entraînement,
- les scripts d'entraînement,
- les modèles eux-mêmes.

L'entraînement a été réalisé sur l'infrastructure partagée Jean Zay.

Conclusion #1

Le marché a rapidement validé un design dominant, celui de l'agent conversationnel génératif (p. ex. ChatGPT). Sur les modèles, les trois formes d'innovation (produit, processus, commoditisation) sont visibles.

Le marché des LLM reste marqué par une « course à l'armement » sur les modèles capables de raisonnement (p. ex. GPT-o3) ou de recherches approfondies. Cependant, leur poids financier, donc aussi environnemental, doit être abaissé.

Chez les *bigtechs*, les économies passent par l'optimisation des modèles et la commercialisation de modèles allégés adaptés aux tâches plus simples, dans le *chatbot* (p. ex ; GPT-4o), mais aussi, avec une incitation tarifaire, dans les API GPT.

L'innovation de processus et la commoditisation s'inscrit notamment dans le cadre de stratégies open-sources (p. ex. DeepSeek et LUCIE).

Dès lors, la croissance de la consommation devrait être régulière, soit 5 % par an sur 10 ans selon l'IEA (2025b), ce qui correspond à un découplage modéré (x3).

Conclusion #2

Limitations :

- Les usages de LLM par des organisations via les API monétisées par OpenAI n'ont pas été discutés dans cette recherche exploratoire.
- Les externalités positives liées à la substitution de tâches physiques par les IAG mériteraient une évaluation soigneuse (voir par exemple Tomlinson *et al.*, 2024).
- L'impact des centres de données (consommation d'eau, d'électricité) nous semble mériter une analyse davantage contextuelle (p. ex. cas général vs Irlande).

Perspectives de recherche :

- L'évolution sur le MT/LT du tarif des API nous semble un *proxy* intéressant du poids des modèles (car tarification proposée au million de *token*).
- Le modèle économique des modèles open-sources (cf. Viseur, 2024) reste actuellement peu étudié (définitions, incitations, acteurs, revenus…).

Références

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.

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

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.

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/.

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

Jolie, P. (2025). La stratégie chinoise en matière d'IA. L'intelligence artificielle au quotidien. Réalités industrielles. Annales des Mines, mai 2025.

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.

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.

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.

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.

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/.

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.

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

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.

Viseur, R. (2025). Viseur, R. (2025). Sobriété numérique et traitement des données massives: vers des stratégies d'innovation plus durables. Terminal. Technologie de l'information, culture & société, (140), 61-83. https://doi.org/10.4000/13vhv.

Ce support de présentation est diffusé sous licence CC-BY-ND.

UMONS
Université de Mons

Université de Mons
Faculté Warocqué d'économie
et de gestion - Service TIC
Place Warocqué, 17
B-7000 Mons

Tél. : +32.65.373.201

www.umons.ac.be
info.warocque@umons.ac.be

Plus d'information...

Dr Ir Robert VISEUR
Chargé de cours

Tél. : +32.65.374.054
robert.viseur@umons.ac.be