Stratégies open-sources : opportunités et limitations dans le domaine des *Large Language Models* (LLM)

Robert Viseur[1]

1. Service TIC, FWEG, UMONS
17 place Warocqué, B-7000 Mons, Belgique
robert.viseur@umons.ac.be

RÉSUMÉ.
L'avènement des modèles linguistiques de grande envergure (LLM), tels que GPT d'OpenAI, a marqué une avancée significative dans le domaine de l'intelligence artificielle. Ces développements ont été accompagnés par une réflexion sur l'importance des stratégies open-sources dans la recherche et le développement des LLM. Notre étude explore cette dynamique, mettant en lumière les bénéfices et les défis associés à l'adoption d'approches open-sources dans la création et l'utilisation de LLM. Nous examinons les différentes manières dont les données, les modèles et les applications sont partagées et développées de manière ouverte, contribuant à l'innovation et à l'amélioration continue dans le secteur. En dépit de la tendance à la privatisation et à la fermeture des modèles, ce papier argumente en faveur du potentiel des stratégies open-sources pour favoriser une intelligence artificielle éthique, transparente et accessible. En analysant les pratiques actuelles et en proposant une réflexion sur l'avenir de l'open-source dans le développement des LLM, nous soulignons comment ces stratégies peuvent répondre à divers besoins utilisateurs, de la personnalisation à la réduction des coûts, tout en stimulant l'innovation collaborative.

ABSTRACT.

The advent of Large Language Models (LLMs) such as OpenAI's GPT has marked a significant advancement in the field of artificial intelligence. These developments have been accompanied by a reflection on the importance of open-source strategies in the research and development of LLMs. Our study explores this dynamic, highlighting the benefits and challenges associated with adopting open-source approaches in the creation and use of LLMs. We examine the various ways in which data, models, and applications are shared and developed openly, contributing to innovation and continuous improvement in the sector. Despite the trend towards the privatisation and closing of models, this paper argues in favour of the potential of open-source strategies to foster an ethical, transparent, and accessible artificial intelligence. By analysing current practices and offering reflections on the future of open-source in the development of LLMs, we underline how these strategies can meet diverse user needs, from personalisation to cost reduction, whilst stimulating collaborative innovation.

Mots-clés : intelligence artificielle, LLM, FLOSS, éthique.

KEYWORDS: artificial intelligence, LLM, FLOSS, ethics.

1. Introduction

L'intelligence artificielle peut être vue comme « *un artefact informatique construit grâce à l'intervention humaine, qui pense ou agit comme les humains, ou comme nous nous attendons à ce que les humains pensent ou agissent* » (Dignum, 2019). Parmi les courants qui la traversent, l'apprentissage logiciel, ou « *machine learning* », a connu une popularité croissante suite à l'essor des réseaux de neurones profonds, ou « *deep learning* », et de leurs applications. Parmi celles-ci, citons les algorithmes de type Transformers, dont sont issus les *Large Language Models* (LLM). Parmi les LLM, le modèle GPT, développé par OpenAI, utilisé notamment dans l'agent conversationnel ChatGPT, est sans doute le plus populaire aujourd'hui. D'autres entreprises sont venues par la suite concurrencer OpenAI : Google (Bard, Gemini), Anthropic (Claude), Mistral (Mistral, Le Chat), META (Llama)… Parmi ces propositions, certaines se distinguent par leur caractère open-source. Cependant, force est de constater la tendance actuelle à la fermeture des modèles précédemment ouverts ou libres (OpenAI, Mistral...). Cette évolution traduit-elle un légitime désintérêt pour ce type d'approche collaborative ? Ce papier propose de faire le point sur l'intérêt des stratégies open-sources dans le domaine des LLM.

Notre article est décomposé en trois sections. Dans une première section, nous proposons un état de l'art relatif aux *Large Language Models* (LLM) et aux stratégies dites open-sources. Ensuite, nous inventorions les pratiques open-sources dans le domaine des IA génératives (*text-to-text*) et en analysons l'intérêt. Enfin, dans une troisième et dernière section, nous concluons.

2. Revue de littérature

Cette revue de littérature se focalise sur deux notions : les *Large Language Models* (LLM) et les stratégies dites open-sources.

GPT est « *un modèle linguistique autorégressif de troisième génération qui utilise l'apprentissage profond pour produire des textes semblables à ceux des humains* » (Floridi & Chiriatti, 2020). Ce modèle linguistique est formé par entraînement « *sur un ensemble de données non étiquetées composé de textes, tels que Wikipédia et de nombreux autres sites, principalement en anglais, mais aussi dans d'autres langues* » (Floridi & Chiriatti, 2020). OpenAI exploite un ensemble diversifié de jeux de données incluant, pour GPT-3, le Common Crawl (60 % des données d'entraînement), WebText (22%), Books1/Books2 (16%) et Wikipédia (3%) (Brown et al., 2020). Le modèle produit est utilisable au travers de ChatGPT (version gratuite ou payante : ChatGPT Plus) ou au travers d'une API payante.

L'éthique des intelligences artificielles génératives (IAG) est associée à des critères d'équité (« *fairness* »), de transparence (« *transparency* ») et de responsabilité (« *accountability* ») (Ferrara, 2023). Elle porte tant sur la conception que sur l'utilisation des IA. Les IA génératives font en particulier l'objet d'efforts importants en matière de lutte contre les biais, définis comme « *la présence de déformations systématiques, d'erreurs d'attribution ou de distorsions factuelles qui favorisent certains groupes ou idées, perpétuent des stéréotypes ou font des suppositions incorrectes basées sur des schémas appris* », ce qu'une ouverture des données, des méthodologies et des outils facilite (Ferrara, 2023).

Les logiciels libres et open-sources recouvrent un modèle d'innovation construit par et pour les utilisateurs-innovateurs (Jullien et al., 2022). Plutôt que de privatiser le logiciel par le recours à la propriété intellectuelle, il s'appuie sur cette dernière pour « *organiser l'évolution continue de la demande et de l'innovation* ». S'ils sont pensés par et pour les utilisateurs, ces logiciels permettent cependant le développement de stratégies commerciales, dites stratégies open-sources. La proposition de valeur associée au modèle d'affaires recouvre des prestations incluant le développement sur mesure, l'édition logicielle, la production de modules spécialisés ou encore l'hébergement des applications (Jullien & Viseur, 2021). Ce modèle de développement a fait l'objet d'une extension, qualifiée d'« *open source innovation* », au-delà du seul logiciel (Pénin, 2011). Des stratégies open-sources sont dès lors déployables aussi pour des données (open-data) ou du matériel (open-hardware).

Jullien et Viseur (2021) analysent les modèles d'affaires open-sources sous l'angle de la segmentation des besoins des utilisateurs, et du coût total de possession de la solution (Shaikh & Cornford, 2011). L'intérêt du FLOSS (*Free Libre Open Source Software*) dépend en effet des besoins des clients, lesquels conditionnent les modèles d'affaires praticables. Quatre types de besoins sont identifiés par les auteurs : « *Contrôle* » (nécessité d'un haut degré de personnalisation impliquant du développement sur mesure), « *Lock-out* » (peu ou prou de besoins de personnalisation mais évitement du « *vendor lock-in* »), « *Lean* » (personnalisation de masse) et « *Prix* » (recherche d'un produit standardisé à prix modique). Le FLOSS est d'autant plus intéressant que le besoin de maîtrise de la solution (Contrôle, Lock-out, voire Lean) est important.

3. Stratégies open-sources appliquées aux LLM

Lors du développement et de l'exploitation d'une intelligence artificielle générative (IAG), les stratégies open-sources peuvent porter sur trois artefacts : les données, les modèles et les applications. Elles s'ajoutent à des stratégies d'innovation ouverte plus classiques, notamment en matière d'optimisation des infrastructures (p. ex. Open Compute Project).

Données :

Le jeu de données partagé de référence est le Common Crawl (CC). Alimenté par la Common Crawl Foundation, il est couvert par les Terms of Use de cette dernière. Il n'est pas strictement open-source puisque, d'une part, il contient des données issues du Web (donc couverte par le droit d'auteur de leurs créateurs), d'autre part, il n'est pas couvert par une licence libre. Le fichier publié fin décembre 2023 contient 3,35 milliards de documents pour un total d'environ 125 TiB après compression. Le partage des jeux de données permet dès lors d'éviter que tous les réutilisateurs doivent mettre en place une coûteuse infrastructure de collecte de données. Le robot d'exploration associé au CC, issus d'un *fork* du moteur de recherche open-source Nutch est cependant publié sous licence libre[1].

Le développement des LLM s'est accompagné de la publication de nombreux jeux de données (Liang et al., 2014). Le Common Crawl est en effet composé de données de qualité très variable. Deux stratégies sont dès lors déployées. La première stratégie consiste à filtrer le Common Crawl. C'est notamment ce qui est

[1] Cf. https://github.com/commoncrawl/cc-nutch-example.

appliqué par Google avec le C4 Colossal Clean Crawled Corpus (Dodge et al., 2021). Y sont par exemple filtrés les documents comportant des mots issus d'une liste de mots bannis. Au final, ce jeu de données favorise les sources de qualité comme les sites de presse (NYTimes, LATimes...) ou les plateformes de contenus scientifiques (PLoS, Springer...). La seconde stratégie consiste à produire de nouveaux jeux de données porteurs de qualités spécifiques. L'OpenWebText reprend ainsi l'esprit du WebText utilisé par OpenAI. Il est alimenté par des informations plébiscitées par les utilisateurs de Reddit (Liang et al., 2014).

L'ouverture des données en facilite l'audit et simplifie l'identification des biais induits (Ferrara, 2023). Les dispositifs de filtrage sont-ils proportionnées ou conduisent-ils à invisibiliser certaines communautés (Dodge et al., 2021) ? Les jeux de données incluent-ils des contenus toxiques (Liang et al., 2024) ? La publication des règles et des logiciels de filtrage permet ainsi une amélioration continue du processus, et d'aller vers des intelligences artificielles plus éthiques.

Modèles :

Plusieurs organisations proposent des LLM sous licence libre et open-source. C'est notamment le cas de Google (T5) et, jusqu'à récemment, de Mistral[2]. Ces modèles tendent à être diffusés sous des licences permissives telles que la licence Apache ou la licence MIT. Cela autorise les utilisateurs à intégrer le modèle dans leurs applications, tel quel ou après une étape de spécialisation (*fine tuning*), avec un accès à l'architecture, à la stratégie d'entraînement et aux poids. Notons l'usage de licences « *partly open* » (West, 2003) par certains de ces acteurs, à l'image de META[3] (Llama), notamment justifié par des considérations éthiques (voir par exemple la Responsible AI Licenses ; Contractor et al., 2022).

Plusieurs bénéfices peuvent être associés à ces LLM open-sources. Le premier bénéfice est un gain de visibilité pour l'entreprise qui publie ce modèle, grâce à la plus grande diffusion de la marque associée à l'éditeur open-source (Jullien & Viseur, 2021). Le second bénéfice découle de l'accélération de la diffusion du modèle, par téléchargement ou inclusion au sein de plateformes d'exécution (Amazon Sagemaker, NVIDIA NeMo…). D'une part, la disponibilité accrue de ces modèles stimule les retours des utilisateurs quant à leurs performances. Par exemple, les LLM font l'objet d'évaluations quant à leurs capacités ou leurs limitations (p. ex. hallucinations[4]). D'autre part, cette disponibilité des modèles permet d'accélérer le rythme d'innovation. Les améliorations portent par exemple sur la réduction de la taille des modèles (Eldan & Li, 2023). Le troisième bénéfice a trait au partage des coûts entre partenaires (si le modèle open-source est construit collaborativement). Les intelligences artificielles génératives suscitent en effet des inquiétudes en matière d'impact environnemental (Sundberg, 2024). Dès lors que l'entraînement des intelligences artificielles compterait actuellement pour 20 à 40 % de leur consommation (IEA, 2024), la création de modèles en consortium permettrait d'en réduire l'impact environnemental.

Applications :

Les besoins d'intégration des LLM entraîne la création de logiciels capables de les exploiter, soit sous la forme de progiciels (p. ex. agents conversationnels :

---

2 Cf. https://github.com/eugeneyan/open-llms.
3 Cf. https://opensource.org/blog/metas-llama-2-license-is-not-open-source.
4 Cf. https://github.com/vectara/hallucination-leaderboard.

GPT4All), soit sous la forme de plateformes d'exécution multi-modèles (p. ex. Ollama, Hugging Face et LangChain). Ces logiciels suivent alors des logiques open-sources plus classiques.

4. Conclusion

Les besoins clients identifiés par Jullien et Viseur (2021) permettent de discuter l'intérêt des stratégies open-sources pour les IAG (cf. Tableau 1). Les clients à logique « Contrôle » sont davantage sensibles aux questions de personnalisation, de confidentialité et de sécurité. Ils ont un intérêt à pouvoir accéder à des jeux de données spécifiques, incluant des données internes à l'organisation, permettant la spécialisation de modèles génériques. Les clients à logique « Lock-Out » sont peu ou prou intéressés par la personnalisation des IA génératives. Le caractère open-source des modèles et des applications leur garantit par contre une relative indépendance vis-à-vis des prestataires informatiques. Les clients à logique « Lean » peuvent être intéressés par des modèles spécialisés, utilisés en combinaison, pour satisfaire leur exigence de personnalisation à coût maîtrisé. Par ailleurs, les modèles open-sources peuvent leur permettre de bénéficier de technologies en évolution rapide et d'optimiser leurs coûts d'usage. Les clients à logique « Prix » seront peu ou prou intéressés par le caractère open-source des modèles et applications. Ils privilégieront plutôt les API ou les applications web leur permettant un déploiement rapide (SaaS) et à moindre coût (tarification au *pay-per-use*).

Tableau 1. Utilisation de stratégies d'ouverture de type open-source en fonction des besoins des utilisateurs (grisé : technologies closed-source).

	Contrôle	**Lock-Out**	**Lean**	**Prix**
Données	Jeux de données partagés (Common Crawl, C4...)	-	-	-
Modèles	LLM open-source (LLama 2, Mistral 7B…) avec *fine tuning*	LLM open-source (LLama 2, Mistral 7B…)	LLM (open-source ou non) spécialisés et combinés	API LLM (Mistral, GPT...)
Applications	Agents conversationnels open-source (GPT4All…) Applications spécifiques (Ollama, LangChain...)	Agents conversationnels open-source (GPT4All...)	Applications spécifiques (Ollama, LangChain...)	ChatGPT (OpenAI, Microsoft Azure), Copilot

Reste le cas des utilisateurs-innovateurs, assimilés par Jullien et Viseur (2021) à des utilisateurs de pointe au sens de von Hippel, développant de nouveaux LLM pour leurs besoins propres. Il recouvre des situations comme celle de META. Le développement de Llama est actuellement internalisé par l'entreprise. Cependant, il sera intéressant de suivre dans quelle mesure un modèle de fondation (Riehle, 2010), typique des FLOSS (Apache, Eclipse, Linux…), émergera en vue de partager les coûts de recherche et de création des LLM ou d'autres modèles d'intelligence artificielle nécessitant d'importante ressources informatiques (à l'image de PyTorch, transféré par META vers Linux Foundation).

5. Références

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

Contractor, D., McDuff, D., Haines, J. K., ... & Li, H. (2022). Behavioral use licensing for responsible AI. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 778-788). https://doi.org/10.1145/3531146.3533143.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. arXiv preprint arXiv:2104.08758. https://doi.org/10.48550/arXiv.2104.08758.

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

Liang, P., Hashimoto, T., Ré, C., Bommasani, R., Xie, S.M. (2024). Data. CS324 - Large Language Models. https://stanford-cs324.github.io/winter2022/lectures/data/ (consulté le 06/03/2024).

Dignum, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 2156). Cham: Springer. ISBN : 978-3-030-30373-0.

Ferrara, E. (2023). Should chatgpt be biased? Challenges and risks of bias in large langage models. arXiv preprint arXiv:2304.03738. https://doi.org/10.5210/fm.v28i11.13346.

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

IEA (2024). Data Centres and Data Transmission Networks. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks (consulté le 12/02/2024).

Jullien, N., Viseur, R., Zimmermann, J.-B. (2022). Gouvernance d'un projet libre : contrôler un flux d'innovation. Enjeux numériques, juin 2022, n°18. https://www.annales.org/enjeux-numeriques/2022/en-2022-06/2022-06-13.pdf.

Jullien, N., & Viseur, R. (2021). Les stratégies open-sources selon le paradigme des modèles économiques. Systèmes d'Information et Management, 26(3), 67-103. https://doi.org/10.3917/sim.213.0067.

Pénin, J. (2011). Open source innovation: Towards a generalization of the open source model beyond software. Revue d'économie industrielle, (136), 65-88. https://doi.org/10.4000/rei.5184.

Riehle, D. (2010). The economic case for open source foundations. Computer, 43(01), 86-90. https://doi.ieeecomputersociety.org/10.1109/MC.2010.24.

Shaikh, M., & Cornford, T. (2011). Total cost of ownership of open source software: a report for the UK Cabinet Office supported by OpenForum Europe. https://eprints.lse.ac.uk/39826/1/Total_cost_of_ownership_of_open_source_software_(LSERO).pdf.

Sundberg, N. (2024). Tackling AI's Climate Change Problem. MIT Sloan Management Review, 65(2), 38-41. https://sloanreview.mit.edu/article/tackling-ais-climate-change-problem/.

West, J. (2003). How open is open enough?: Melding proprietary and open source platform strategies. Research policy, 32(7), 1259-1285. https://doi.org/10.1016/S0048-7333(03)00052-0.