L'IA générative, un nouveau Wikipédia ?

Illusions et réalités de la production du savoir en ligne

par Jeanne Vermeirsche & Eric Sanjuan

Les intelligences artificielles génératives, comme Wikipédia avant elles, prétendent ouvrir un accès élargi et immédiat au savoir. Mais au débat collectif et transparent de l'encyclopédie participative s'oppose l'opacité de production de textes plausibles mais peu précis, voire erronés.

Introduction

Parmi les dix sites les plus consultés au monde, aujourd'hui l'une des principales sources de données pour l'entraînement des larges modèles de langage (LLM) – tels que GPT, Gemini ou Mistral – Wikipédia occupe une place centrale dans nos usages quotidiens d'Internet. L'encyclopédie en ligne est devenue de fait un espace incontournable dans la fabrique contemporaine des savoirs dont le fonctionnement collaboratif permet, en théorie, à tout le monde d'écrire, de discuter et de modérer les contenus, indépendamment du statut économique, social ou culturel. Elle fonctionne en dehors des logiques marchandes qui gouvernent la plupart des autres plateformes en ligne et ne peut être achetée, même par un milliardaire comme

Elon Musk qui s'était pourtant ironiquement proposé d'en prendre le contrôle[1]. De fait, Wikipédia illustre un modèle original de gouvernance participative (Cardon et Levrel, 2009) où les contributeurs encadrent collectivement le contenu tout en se soumettant aux règles établies par la communauté.

Récemment, la Fondation Wikimédia a d'ailleurs réaffirmé la centralité de l'humain au cœur du projet, soulignant que le succès de Wikipédia repose avant tout sur l'engagement des wikipédiens bénévoles. Cette prise de position intervient dans un contexte de questionnements et de débats intenses puisque l'essor rapide des technologies d'IA génératives, conçues pour prédire et imiter le style humain, rendrait finalement possible la production quasi instantanée de textes qui sembleraient provenir de Wikipédia (Deckelmann, 2023) ou pourraient y être ajoutés presque tels quels. La communauté redoute de voir l'encyclopédie polluée par des contenus erronés et de mauvaise qualité, contraires aux articles créés, débattus et organisés par des humains, assurant un savoir fiable. Sur Wikipédia, il s'expérimente en effet des formes de délibération dans la co-construction des savoirs qui, aujourd'hui, doivent intégrer les défis posés par l'IA.

C'est dans le prolongement de ces questionnements actuels que nous proposons d'étudier Wikipédia et les sites d'IA générative accessibles aux particuliers, étudiants ou scolaires en tant que formes techniques et sociales de production du savoir, avec leurs atouts, leurs angles morts et leurs formes d'exclusion. Il ne s'agit pas ici d'en rejeter un plus que l'autre. D'ailleurs, tous deux ont été ou sont encore regardés avec suspicion par le monde universitaire. Les espaces numériques ne font pas disparaître la question de la légitimation du savoir, de son accès, mais la déplacent vers des formes parfois moins visibles, plus techniques, potentiellement pour certaines moins démocratiques, que nous nous proposons ici d'examiner.

Ce questionnement nous paraît particulièrement pertinent pour deux raisons principales. D'abord, Wikipédia nourrit les LLMs utilisés par l'IA générative puisqu'à ce jour, il constitue une de leur principale source de données d'apprentissage. Au-delà de l'entraînement initial, les sites d'IA conversationnelle les plus populaires tels ChatGPT continuent ensuite à puiser directement dans Wikipédia pour cadrer et contextualiser leurs réponses aux requêtes des utilisateurs. L'encyclopédie est donc à

[1] L'ironie est en effet double puisqu'Elon Musk ne peut justement pas acheter Wikipédia, dont le modèle collaboratif et non marchand repose sur un fonctionnement radicalement différent des logiques du capitalisme de plateforme fondé sur le rachat et l'accumulation. D'autre part, il s'est lui-même exprimé sur un ton ironique sur X, notamment lorsqu'il proposait de donner un milliard de dollars si l'encyclopédie changeait de nom pour « Dickipedia ».

la fois une source de données en amont pour l'entraînement et une source d'information en aval au moment de la requête. Ensuite, ces dispositifs d'IA générative, accessibles à tous aux mêmes conditions que les moteurs de recherche et les réseaux sociaux, produisent des contenus qui pourraient à leur tour être intégrés à Wikipédia plus ou moins directement. Cela alimente le débat sur le remplacement éventuel des contributeurs et soulève une série de questions fondamentales autour de la préservation de la qualité des contenus, la limitation des biais, et l'autorité chargée de déterminer ce qui est acceptable ou non dans l'encyclopédie. S'y ajoute le fait que les bénévoles wikipédiens produisent gratuitement un savoir qui alimente une IA privatisée[2] qui mobilise des capitaux sans limite, pour laquelle ils ne sont jamais crédités ou n'obtiennent aucune rétribution, même symbolique comme dans Wikipédia.

Notre analyse met ainsi en regard deux régimes de médiation, c'est-à-dire l'ensemble des règles, pratiques et outils qui encadrent la façon dont le savoir est produit, structuré et rendu accessible, ce qui conditionne aussi ce que l'on peut voir, discuter, ou corriger dans ce processus. Le premier, celui de Wikipédia, repose sur une médiation horizontale, délibérative et communautaire. Les contenus y sont entièrement traçables et discutables, grâce aux historiques de modifications et aux pages de discussion des articles, ouverts à tous. Le second, celui des LLMs, est invisible, algorithmique, non discutable. Ces dispositifs produisent du texte plausible, mais sans traçabilité ni débat contradictoire, et leur usage suppose des compétences spécifiques pour formuler des requêtes pertinentes, évaluer les réponses fournies, les trier et vérifier. Ces différences ne contredisent pas un point commun : Wikipédia et les LLMs entretiennent chacun une promesse d'égalité d'accès, pourtant illusoire. Ils sont tous deux gratuits et ouverts à tous, que ce soit pour contribuer à Wikipédia ou interroger un site d'IA conversationnelle, un assistant intégré tel que Gemini ou directement un LLM libre téléchargé, indépendamment de la position sociale ou économique. Dans les faits, leur usage effectif reste socialement et cognitivement situé. Contribuer à Wikipédia exige des compétences techniques et sociales, inégalement réparties, pour écrire selon les normes wikipédiennes, référencer et participer aux débats très codifiés. Certains profils y sont alors nettement sur-représentés, ce qui influence la manière dont les savoirs y sont construits et transmis : masculins, dotés d'un fort capital scolaire, politique et militant, particulièrement informés et engagés (Vermeirsche, 2025). Cette apparente égalité est renforcée pour la nature probabiliste

[2] Même si la publication en libre de LLMs de taille réduite pouvant être exécutés en local se développe via des sites collaboratifs tels que *huggingface.co* et *ollama.ai*.

des LLMs, puisqu'aucun mécanisme interne ne vient contredire l'utilisateur, contrairement aux débats collectifs publics sur Wikipédia, à propos de la création ou le retrait de pages et de règles, leur contenu et les corrections.

Expérimentation : générer des notices Wikipédia avec des LLMs

> « j'ai testé Storm, l'IA réputée la plus au point pour générer des articles *Wikipédia-like* selon un papier publié sur Wikimédia. Et c'est assez catastrophique, même en partant de bonnes sources. Cela peut faire effet quand on ne connaît pas le sujet traité, et c'est justement ce qui est dangereux, car au milieu d'un résultat *réaliste* et de forme impeccable se nichent de sacrées erreurs. »
>
> *Discussion sur le Bistro, octobre 2024*

La place de l'IA générative au sein de Wikipédia fait l'objet de débats au sein de la communauté, comme en témoignent de récentes discussions sur le Bistro, l'espace d'échanges informels entre les wikipédiens. Une partie d'entre eux s'opposent à toute utilisation de l'IA générative, estimant qu'elle va à l'encontre des fondements du projet, tel ce contributeur qui écrit : « *l'IA n'est pas soluble dans WP, et surtout touche des éléments des plus sensibles des articles, la neutralité et le sourçage, qui sont déficients avec l'utilisation l'IA.* ». D'autres adoptent une position plus nuancée, en distinguant « *"aider à faire" mais jamais pour "faire à la place"* » et rappellent également que l'utilisation finale reste humaine : « *C'est tout de même l'humain qui publie le texte rédigé par IA, donc lui qui peut veiller ou non à la qualité du texte.* » Leurs échanges révèlent combien, au sein même de la communauté, la question de l'utilisation de l'IA dans Wikipédia n'est pas tranchée, quoique toujours discutée avec la même préoccupation centrale, celle de la qualité, de la neutralité et de la traçabilité des contenus.

Pour tenter de comprendre concrètement ce que les modèles de langage produisent lorsque nous leur demandons d'écrire « à la manière » de Wikipédia, nous avons mené des tests en générant des notices avec différentes IA génératives. Les résultats permettent d'observer ce que ces productions révèlent sur les mécanismes propres aux LLMs [3] qui les propulsent et sur leurs limites dans le contexte

[3] Les LLMs génèrent du texte en se basant sur ce qu'ils ont appris à partir d'énormes quantités de données textuelles, dont Wikipédia. Durant la phase d'entraînement sur ce corpus, le LLM lit ces textes pour apprendre les relations statistiques entre les mots, c'est-à-dire quelles combinaisons sont les plus probables dans quel contexte (Ashkinaze et al., 2024). Après l'entraînement brut, il est affiné avec des données plus ciblées ou avec des corrections humaines. Il peut produire un texte très fluide

encyclopédique. Nous avons ainsi demandé à plusieurs dispositifs d'IA générative de rédiger deux notices Wikipédia – « Thèse du bouclier et de l'épée » et « Laurent Casanova » – en testant deux types de dispositifs :

- en ligne, accessibles depuis un navigateur aux mêmes conditions qu'un moteur de recherche, hébergés par une entreprise sur ses serveurs, par exemple OpenAI pour le site ChatGPT ;

- en local, installés sur un serveur personnel, exécutés hors-ligne avec un contrôle total sur le contexte exact fourni au LLM[4].

La distinction entre un modèle exécuté en ligne et en local a des conséquences directes sur les contenus produits. Les modèles en ligne sont exécutés sur de très larges infrastructures qui permettent d'intégrer de larges contextes. La requête de l'utilisateur est complétée par l'historique de ses précédentes interactions, le contenu d'une recherche sur le web public et les réseaux sociaux pour lesquels l'entreprise dispose d'accords (par exemple, X pour Grok). Ils bénéficient aussi de mises à jour grâce à l'apprentissage par renforcement en s'appuyant sur les retours des usagers. Enfin ces sites appliquent différents dispositifs de modération des textes générés. Si les infrastructures locales ne permettent pas de disposer de très larges contextes, elles permettent par contre de multiplier les requêtes de manière systématique pour approcher le cœur de ces modèles. Elles offrent une transparence et un contrôle plus grands sur les données et sur le processus de génération.

Les sites d'IA conversationnelle testés ici sont ChatGPT (OpenAI), Grok (xAI), Gemini (Google), LeChat (Mistral AI) et Perplexity (Perplexity AI, Inc.). Les LLMs testés directement en local, dont les architectures recouvrent celles des modèles qui propulsent les sites privés, sont les suivants : Gemma3, Mistral-small et gpt-oss[5]. La requête initiale adressée aux dispositifs était volontairement simple, et si nécessaire progressivement affinée[6] :

---

mais qui se trompe complètement (inventer des références, mélanger des information réelles et fausses, ou reproduire les biais présents dans ses données d'entraînement) et on parle dans ce cas d'hallucinations.

[4] Les LLMs testés directement en local sur un GPU unique limité à 30 Go de mémoire ont été téléchargés du site communautaire *ollama.ai* et testés via une interface locale OpenWebUI qui reproduit la même ergonomie que les sites d'IA conversationnelle.

[5] Parmi les dispositifs testés en ligne, seuls Grok et Perplexity n'ont pas de version pouvant être exécutée en local.

[6] Les tests ont été menés en août 2025.

> Bonjour, je te propose un exercice. Tu es un contributeur de Wikipédia et tu dois rédiger la page de la Thèse du bouclier et de l'épée / Laurent Casanova. Peux-tu écrire cette page ? Par toi-même. Qu'écrirais-tu ?

Les expérimentations réalisées sur ces deux pages permettent de tester à la fois un débat historiographique complexe et controversé et la biographie d'une figure politique moins connue du grand public. La page de la thèse du bouclier et de l'épée illustre bien les dynamiques de négociation sous contrainte au sein de l'encyclopédie. Peu consultée et faiblement surveillée pendant plusieurs années, elle a longtemps présenté une version ambiguë, historiquement discréditée, décrivant une collaboration simulée de Pétain pour mieux protéger la Résistance avec de Gaulle. Des contributeurs ont longuement participé à rendre fiable la page en respectant strictement les règles de Wikipédia. Ensuite, travailler sur la page de Laurent Casanova, un homme politique français, membre du parti communiste et résistant pendant la Seconde guerre mondiale, permet d'évaluer la capacité des modèles à restituer des trajectoires moins centrales, où les erreurs ou simplifications risquent d'être d'autant plus trompeuses pour un lecteur non spécialiste[7].

Dans l'ensemble, tous les dispositifs testés parviennent à produire un texte reprenant grossièrement la structure d'un article Wikipédia dans sa forme la plus simple avec des titres, des paragraphes clairs, un ton globalement neutre et un enchaînement fluide des idées. Ils génèrent des textes plus ou moins longs, très cohérents en surface, mais avec des variations très importantes de précision et de pertinence.

Pour la page de la thèse du bouclier et de l'épée, seuls ChatGPT, Grok et Perplexity ne font pas de hors-sujet. Ils l'identifient d'emblée comme relevant de la Seconde Guerre mondiale et l'associent à Pétain et de Gaulle, avec des explications relativement correctes. Cependant, ChatGPT à cette date produit des références inventées, des erreurs historiques importantes, par exemple en citant le « général Paul Touvier ». Ainsi, là où tout article de Wikipédia rend accessibles l'historique des débats et les corrections successives depuis sa création, un modèle génératif peut produire une référence inexistante sans que l'on puisse en retracer l'origine. Grok livre un texte fluide, couvrant une large chronologie. Il est en effet le seul LLM à faire remonter le

[7] Cette page s'inscrit également dans un travail en cours au sein duquel nous sélectionnons 1000 notices Wikipédia correspondant à des requêtes, par exemple ici : « politique & (france | français | française | françaises) ». Nous demandons à des dispositifs en local de rédiger une notice Wikipédia pour chacun de ces titres. Les textes générés sont évalués selon plusieurs critères en comparant la production des LLMs à la version originale de Wikipédia. Les associations lexicales générées sont également analysées.

débat jusqu'à la période contemporaine avec une référence explicite à Eric Zemmour, présenté comme ayant « tenté de réhabiliter Pétain en 2018 en s'appuyant sur des arguments proches de cette thèse. » Il intègre des références réelles mais parfois approximatives, notamment quand il présente comme un ouvrage ce qui est en réalité un documentaire vidéo. Il accorde toutefois une place plus importante à la version apologétique de la thèse avant de la réfuter. Perplexity est le site d'IA conversationnelle qui a le plus investi dans un moteur de recherche avec un filtrage des sources selon des critères de fiabilité. Wikipédia en constitue le noyau initial. Il apparaît ici comme le plus précis et le plus référencé, et adopte un style très proche de l'encyclopédie, qui est ici la principale source du texte produit. Nous pouvons toutefois supposer que, si la notice n'existe pas encore ou n'est pas fiable, le problème se répercute immédiatement sur le texte généré.

Tous les autres dispositifs testés, en ligne ou en local, produisent, à la première requête, un texte hors-sujet malgré une cohérence formelle. Il faut systématiquement formuler une seconde requête, précisant que la thèse concerne la Seconde Guerre mondiale, pour qu'ils s'orientent vers le bon contexte historique, tout en continuant à produire un texte largement erroné. Ainsi, gpt-oss décrit « un courant de pensée politique et philosophique francophone qui considère la défense nationale comme composée de deux volets complémentaires : la protection passive (le « bouclier ») et la capacité offensive de dissuasion (la « épée ») [...] » et génère parmi les références de faux liens externes. LeChat interprète la thèse comme une « métaphore utilisée pour décrire une stratégie de défense et d'attaque dans divers contextes, notamment en politique, en économie et en stratégie militaire » tandis que Mistral-small se lance dans un récit médiéval, attribuant la paternité de la thèse, avec force détails, à un moine irlandais du VIIe siècle nommé Arculf. Même après avoir affiné la requête, Mistral-small tente artificiellement de relier ce texte médiéval à la Seconde Guerre mondiale, en citant de faux historiens et de fausses analyses, tout en précisant que cette thèse n'a pas de lien direct avec cette période historique. Gemini, qui dispose de l'accès le plus étendu au web et donc aussi le plus bruité, c'est-à-dire le plus exposé à des contenus contradictoires ou parasites, décrit pour sa part « une théorie controversée en ethnologie et en sociologie, proposée par l'anthropologue David Graeber [...] ». Même avec une requête affinée, le site conversationnel de Google ne parvient pas à produire une réponse moins hors-sujet que celle produite par Gemma3. Dans les deux cas, les textes fournis sont clairs et bien structurés mais assortis de faux auteurs et ouvrages indiquant la mention « (Hypothétique) ».

La notice Wikipédia actuelle demeure donc nettement supérieure à toutes les productions obtenues, tant sur le fond que sur la forme, fruit de plusieurs années de contributions et de débats. Même à des périodes où l'article présentait encore des formulations biaisées, il restait plus fidèle au fond historique que la majorité des textes générés par les IAs. Pourtant, les productions des modèles, même lorsqu'elles sont erronées, restent fluides, plausibles et donc très convaincantes pour un lecteur non spécialiste. Si tous ces LLMs ont été entraînés sur Wikipédia, donc ont lu une version récente de cette page, l'encyclopédie est toutefois une ressource insuffisante pour entraîner une machine qui génère du texte syntaxiquement correct. De ce fait, la source est noyée dans un large corpus de textes aléatoires. Dans le cas de thèmes précis, peu fréquents sur le Web ou les réseaux sociaux, les LLMs produisent des réponses fausses, mais formulées avec la même assurance que si elles étaient correctes, souvent appelées hallucinations dans le vocabulaire de l'IA, et sont incapables de retrouver l'information exacte pourtant présente dans leur corpus d'apprentissage.

Les expérimentations menées pour la page de Laurent Casanova confirment les écarts observés. Les dispositifs en ligne fournissent immédiatement un récit centré sur le personnage communiste et résistant. ChatGPT et Grok font quelques approximations sur les dates ou les responsabilités exercées, et les sources nécessitent systématiquement une vérification. Perplexity se distingue par la variété et la vérifiabilité de ses sources par rapport aux autres modèles (Wikipédia, Assemblée nationale, plusieurs blogs…). Sa structure et le ton sont les plus proches d'une notice Wikipédia. Il faut aussi noter que, pour pouvoir ajouter autant de pages au contexte du prompt, ces dispositifs doivent mobiliser d'importantes ressources de calcul, *a minima* quatre fois plus que notre dispositif local et infiniment plus qu'une recherche Wikipédia[8]. LeChat, de la société française Mistral AI, qui développe une IA générative « souveraine » et ne dispose pas des mêmes ressources en calcul, génère bien une page concernant le personnage communiste mais contenant de très nombreuses erreurs et approximations, notamment sur les dates et références. Avec la recherche sur le web, il retrouve la page Wikipédia, ainsi que des informations sur le site de l'Assemblée nationale (avec des problèmes d'homonymie), mais celles-ci sont marginalisées par du contenu provenant du site *ruwiki*, une scission russe de Wikipédia. Gemini fait un rapprochement avec un footballeur homonyme existant, mais dont la notice générée est remplie d'erreurs.

[8] 11.5 secondes de calcul intensif sur processeurs graphiques pour Grok contre quelques millisecondes sur processeur à faible consommation.

Les dispositifs en local produisent des biographies inventées sur des homonymes fictifs. Mistral-small décrit « un auteur français », auteur d'un best-seller « traduit en plus de 30 langues » dont l'œuvre aurait « été adaptée au cinéma par l'acteur Gilles Lellouche ». Lorsque la requête est précisée, le modèle persiste dans son hallucination initiale l'enrichissant seulement de détails supplémentaires, notamment un supposé engagement militant. Gemma3 invente un économiste français connu pour ses travaux sur les nouvelles technologies. Ainsi, les LLMs sans dispositif de recherche de contexte dans des sources de référence montrent une forte difficulté à corriger leurs erreurs initiales, même après une clarification de la requête.

Ces exemples illustrent ce que l'on pourrait appeler une tendance au remplissage plausible, c'est-à-dire que les notices générées paraissent crédibles en première lecture, mais contiennent des approximations et des inventions qui pourraient d'autant plus facilement passer inaperçues que les sujets traités sont moins connus du grand public. Pour la thèse du bouclier et de l'épée, certaines erreurs (par exemple, la transformer en un traité médiéval) sont flagrantes pour un lecteur averti. D'autres, telles que la rapprocher des théories de relations internationales, sont moins manifestes. Dans tous les cas, pour un lecteur non spécialiste ou moins informé, ces récits peuvent paraître tout à fait recevables. Dans le cas de Laurent Casanova, les erreurs sont souvent plus subtiles. Elles concernent des dates (notamment de naissance et de mort) ou certains faits : par exemple, ChatGPT indique qu'il a été « déporté au camp de concentration de Mauthausen » ce qui est factuellement faux. Ces glissements, moins visibles, n'en sont pas moins dangereux puisqu'ils donnent l'illusion d'une information solide alors même que la notice est trompeuse, renforçant le risque d'adhésion par un lecteur non spécialiste. C'est ce que souligne un contributeur sur le Bistro : « *il y a un point sur lequel je considère l'IA (ChatGPT) comme définitivement non utilisable pour écrire des articles, c'est que lorsqu'elle ne sait pas, elle invente des réponses plausibles.* ». Ces constats rejoignent de précédentes analyses qui montrent combien les LLMs, dans le cas des contenus wikipédiens, ont tendance à générer des sources inventées, à produire un style non encyclopédique et à introduire des biais politiques (Brooks et al., 2024). Ils échouent aussi à appliquer le principe de neutralité de point de vue[9] pourtant central (Ashkinaze et al., 2024). Si leur capacité à produire beaucoup d'informations est forte, la précision et la justesse dans les faits

[9] NPOV, pour « Neutral Point of View », une règle fondamentale qui guide la rédaction des articles. Les contributeurs doivent s'efforcer de présenter les informations de manière impartiale, sans privilégier un point de vue plutôt qu'un autre. Ils doivent ainsi éviter les opinions personnelles ou les prises de position partisanes, et, s'il existe plusieurs interprétations d'un sujet, les présenter toutes de manière équilibrée.

rapportés restent faibles, ce qui entraîne en réalité un surcoût de vérification et de modération pour l'utilisateur. Il semble ainsi qu'un LLM peut imiter la forme d'un article wikipédien mais non, par nature, en comprendre l'esprit ni en respecter les exigences de vérifiabilité et de neutralité.

Les LLMs échouent ainsi à reproduire fidèlement un savoir déjà existant. Sans Wikipédia comme point d'appui, ils se perdent et même lorsqu'ils y puisent directement, ils produisent encore de nombreuses erreurs. L'enjeu semble d'autant plus inquiétant pour les notices qui n'existent pas encore alors même que l'une des grandes promesses est précisément de combler ces vides. Par ailleurs, les résultats sont non reproductibles puisqu'une même requête adressée plusieurs fois peut conduire à des réponses différentes. À cette variabilité s'ajoute l'appui sur des sources hétérogènes, allant de sites établis (Wikipédia, *Le Monde*…) à des plateformes moins transparentes comme *franco.wiki* ou *ruwiki.press*. Ces sites, se présentant comme encyclopédiques, ont récemment émergé et se sont rapidement multipliés en parallèle de l'irruption des IA génératives.

Les enjeux de la qualité et des biais

De fait, que se passe-t-il si l'on s'en tient uniquement aux résultats produits par les LLMs ? Ces dispositifs ne génèrent pas à proprement parler des connaissances, mais des textes donnant l'illusion du savoir sans en avoir la consistance. Censés libérer l'accès au savoir, les IA génératives risquent en réalité de devenir un facteur d'exclusion supplémentaire. Les compétences, qu'elles soient rédactionnelles, techniques ou critiques, étant inégalement distribuées, le risque est de se retrouver à consommer du texte généré plutôt qu'à construire ou comprendre réellement les connaissances. Nous pourrions bien entendu préciser les requêtes mais, si pour obtenir une réponse pertinente, je dois être capable de préciser ma demande avec exactitude, alors cela suppose déjà de disposer d'un savoir préalable, et cela contredit la promesse initiale de son accès simple et immédiat.

Ainsi, le savoir encyclopédique ne peut se réduire à une accumulation ou un résumé d'informations, mais résulte d'un processus collectif de débats, parfois de conflits. Là où un contributeur wikipédien peut corriger, contester ou nuancer, les modèles génératifs tendent à conforter l'utilisateur, produisant des textes sans discussion ni contradiction. Leurs contenus relèvent de boîtes noires où les sources, les

choix et les biais qu'ils peuvent contenir restent invisibles. Or, ces biais, loin d'être seulement techniques, reflètent aussi les déséquilibres déjà présents dans les corpus d'apprentissage, comme la surreprésentation de certaines langues et points de vue dominants, qui affectent particulièrement les groupes marginalisés ou sous-représentés. L'affaire récente autour de Grok révèle aussi des biais idéologiques assumés, le modèle étant volontairement paramétré pour tordre certains faits et les présenter selon une lecture politique orientée[10].

Un point crucial est déjà que les LLMs n'ont pas accès à l'ensemble du savoir disponible, notamment celui qui n'est pas accessible en ligne, et donc « ne peuvent pas utiliser des informations qui ne figurent pas dans leurs données d'entraînement » (Deckelmann, 2023). Tous les ouvrages non disponibles en ligne ou les contenus produits dans des langues moins représentées échappent à leur champ de « connaissance ». Un contributeur souligne ce point : *« les IA exploitent les données qui sont sur Internet. Or, même si les portails en Open [access, ndlr] sont de plus en plus nombreux, volumineux et efficaces, une bonne partie de la production intellectuelle reste accessible seulement en version papier, y compris des ouvrages qui paraissent actuellement »* Ce constat amène à une autre conséquence majeure qui est celle de la dépendance des LLMs aux contenus produits par les humains. Pour rester performants, ils ont en effet besoin de textes originaux, argumentés et référencés. Les LLMs entraînés sur leurs propres productions deviennent moins précis, oublient même certaines informations qu'ils « connaissaient » auparavant, et finissent par s'effondrer dans un phénomène appelé *« model collapse »* (Shumailov et al., 2024). Les productions humaines de Wikipédia sont donc des ressources indispensables au maintien de la qualité et de la diversité des modèles. Également, si des textes générés par l'IA venaient à être intégrés à l'encyclopédie sans relecture critique, puis réutilisés dans l'entraînement, cela mènerait vers une boucle où des erreurs inventées par l'IA se valideraient et se propageraient elles-mêmes. L'illusion d'autorité produite par ces contenus amplifierait alors un appauvrissement progressif du savoir disponible en ligne. Or, les LLMs tendent à devenir des canaux d'information primaires, les internautes préférant les réponses générées instantanément par IA. Dans ce contexte, chaque biais ou chaque omission pourrait influencer de plus en plus la perception des utilisateurs, parfois sans qu'ils en aient conscience, soulevant la question, à terme, d'une nouvelle forme de

[10]https://techcrunch.com/2025/07/23/trumps-anti-woke-ai-order-could-reshape-how-us-tech-companies-train-their-models/
https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/

« propagande codée », d'autant plus insidieuse qu'elle se déploie sous l'apparence d'une neutralité technique et d'une fluidité discursive.

Wikipédia, laboratoire d'une gouvernance hybride ?

Pour autant, il faut rappeler que les outils d'IA sont déjà présents dans l'espace wikipédien. Depuis longtemps, des bots automatisent certaines tâches fastidieuses, comme l'annulation de vandalismes. Leur fonctionnement repose sur des scripts transparents, paramétrés et validés collectivement. Ils viennent assister les contributeurs, mais ne se substituent jamais à eux. Ils sont pleinement intégrés dans la culture de Wikipédia. Les LLMs pourraient s'inscrire dans cette même logique d'assistance comme le souligne la stratégie en matière d'IA de la Wikimedia Foundation, pour suggérer des sources, corriger des tournures, améliorer la fluidité d'un texte ou traduire. Ils offriraient aux contributeurs un gain de temps pour se concentrer sur le cœur du travail encyclopédique à savoir chercher, évaluer et insérer des sources, (ré)écrire des contenus, débattre et argumenter. L'encyclopédie constituerait ainsi un espace d'expérimentation hybride, où des outils d'IA viendraient soutenir le travail collaboratif sans en remplacer les fondements. Mais les LLMs ne sont pas de simples bots. Ils sont capables de générer des articles entiers structurés et apparemment crédibles. Ce gain apparent de temps et de fluidité pourrait se payer au prix fort, celui de la disparition du processus délibératif et de la traçabilité du travail collectif. De plus, ces LLMs, faciles à déployer mais incapables de justifier leurs actions, pourraient progressivement remplacer les bots, plus complexes à programmer mais fondés sur des règles vérifiables.

Sur le Bistro, si une partie des contributeurs reconnaît l'intérêt potentiel des IA pour des tâches techniques, beaucoup expriment une forte réserve. En 2024, une proposition a d'ailleurs été formulée pour lancer un projet consacré spécifiquement aux usages de l'IA générative pour encadrer ces pratiques « *dans le respect des principes de l'encyclopédie particulièrement sur les questions de fiabilité et d'intégration au caractère collaboratif de Wikipédia, avec sa communauté et ses attentes.* » Il ne s'agit pas, comme le rappelle un contributeur dans la discussion, d'« *être le théâtre d'un remplacement de l'humain par des IA pour la rédaction d'articles.* » L'insistance sur la valeur irremplaçable de l'intervention humaine rejoint un constat largement partagé qui est qu'au-delà de la simple production de texte, c'est le processus même de négociation et de relecture mutuelle qui fonde la fiabilité de Wikipédia. Un contributeur rappelle d'ailleurs que

> « de toute façon, qu'un contenu wikipédien ait été créé par intelligence artificielle ou non [...] ce sont toujours nos règles, recommandations et principes fondateurs qui doivent servir à l'évaluer [...] au même titre que n'importe quel contenu que nous produisons ici. À mon sens, il est donc inutile de redouter d'avance les menaces que l'utilisation d'IA pourrait induire sur WP : nous sommes en réalité déjà armés pour y faire face, et ce depuis longtemps. Au pire, ce que l'on peut craindre, c'est l'explosion de contenus pauvres générés à la chaîne, dont il sera aisé de bloquer les éventuels instigateurs. »

Conclusion. La reconfiguration de nos rapports au savoir

Nous pourrions ainsi être tentés de considérer que l'usage le plus pertinent des LLMs consisterait à les cantonner à une de leur fonction première, c'est-à-dire réécrire du texte. Mais même dans ce cadre apparemment limité, ces modèles ne sont pas neutres. Ils véhiculent inévitablement des biais, liés aux rapprochements qu'ils opèrent de manière statistique entre des fragments de textes, qui reflètent la composition et les déséquilibres du web – principalement anglophone – des dix dernières années, base principale de leur entraînement. Là où Wikipédia vise à corriger ces déséquilibres par la confrontation et la traçabilité des sources, les modèles génératifs tendent au contraire à les amplifier en les habillant d'un style fluide et crédible.

La surreprésentation de certaines sources dans les corpus d'entraînement fait ainsi que certains récits sont renforcés, tandis que des thématiques moins couvertes sont marginalisées. C'est ce que l'on observe dans le cas de la thèse du bouclier et de l'épée ou de Laurent Casanova, sous-documentés dans les résultats générés par les modèles. À l'inverse, des thématiques fortement présentes sur le web bénéficient d'une abondance d'informations que les modèles restituent avec beaucoup plus de justesse. Ces déséquilibres illustrent combien il est aussi crucial de réfléchir à ce que signifie désormais « former aux savoirs » dans les espaces numériques. Il ne s'agit pas seulement d'apprendre à chercher une information ou à formuler une requête, mais aussi à interpréter, recouper, vérifier, et surtout débattre dans un environnement technique de plus en plus empreint d'automatismes. Former aux savoirs, aujourd'hui, ce serait donc aussi former à la critique des outils qui prétendent les délivrer.

Pour aller plus loin :

- Ashkinaze, J. et al. (2024) « Seeing Like an AI: How LLMs Apply (and Misapply) Wikipedia Neutrality Norms », arXiv.org
- Brooks, C., Eggert, S et Peskoff, D. (2024). « The Rise of AI-Generated Content in Wikipedia », arXiv.org
- Broudoux, E. (2015) « Wikipédia, objet de recherches : entre observations, expérimentations et co-constructions ». In *Wikipédia, objet scientifique non identifié*, édité par Lionel Barbe, Louise Merzeau, et Valérie Schafer, 55-73. Presses universitaires de Paris Nanterre, 2015.
- Cardon, D., et Levrel, J. (2009) « La vigilance participative. Une interprétation de la gouvernance de Wikipédia ». *Réseaux* 154, no 2 : 51-89.
- Deckelmann, S. (2023). « Wikipedia's value in the age of generative AI », Wikimedia Foundation.
- Shumailov, I., Shumaylov, Z., Zhao, Y. et al. (2024) « AI models collapse when trained on recursively generated data. » Nature 631, 755–759 (2024).
- Vermeirsche, J. (2025). Sur les traces d'un Wikipédia politique. Écrire le politique dans un espace encyclopédique : discours, contributeurs et discussions en ligne à l'épreuve d'une démocratie numérique. Thèse de doctorat en science politique, Avignon Université.
- Zhou, Y. et al. (2024). « Trustworthiness in Retrieval-Augmented Generation Systems: A Survey », arXiv.org.
- Rapport d'information déposé en application de l'article 145 du règlement, par la commission de la défense nationale et des forces armées, en conclusion des travaux d'une mission d'information sur le thème de « l'opérationnalisation de la nouvelle fonction stratégique influence» (Mme Natalia Pouzyreff et Mme Marie Récalde), n° 1661.

Publié dans laviedesidees.fr, le 16 septembre 2025.