NOTES

LES LLMS, UN ANTIDOTE À LA POLARISATION ET À LA MÉSINFORMATION ?

Antoine Marie (Sciences Po),
Peter Barrett (enseignant à l'ESSEC),
Sacha Altay (Université de Zurich)

SPIRALES

RÉSUMÉ

Cette note examine le potentiel des grands modèles de langage (LLMs) pour réduire certaines dérives informationnelles et polarisantes des réseaux sociaux. En favorisant un accès plus factuel, nuancé et personnalisé à l'information, les LLMs pourraient contribuer à corriger certaines croyances erronées et à réduire la polarisation idéologique.

Des expériences récentes montrent des résultats encourageants, notamment sur les théories du complot et la qualité des échanges. Ce potentiel reste toutefois conditionnel : les LLMs peuvent produire des erreurs, refléter des biais ou être instrumentalisés à des fins de propagande.

Il apparaît donc essentiel de privilégier la transparence, la diversité des points de vue, l'écoute active et le maintien du jugement humain. Bien conçus et correctement encadrés, les LLMs pourraient ainsi devenir des outils utiles pour améliorer la qualité de l'information et du débat démocratique.

TEXTE INTÉGRAL

Depuis une vingtaine d'années, les réseaux sociaux ont bouleversé la manière dont nous nous informons et débattons collectivement. Ce bouleversement n'a pas été sans bénéfices : il a donné une caisse de résonance inédite aux lanceurs d'alerte, contribué à des mouvements comme #metoo, et cassé le monopole qu'exerçaient auparavant quelques rédactions sur la production de l'information.

Mais il a aussi eu un coût considérable pour la qualité du débat public. En dispersant les sources d'information entre une multitude de plateformes concurrentes, l'essor des réseaux sociaux (et avant lui le développement des chaînes partisanes) a fragmenté un peu plus le socle de faits partagés sur lequel devrait reposer la discussion politique.
Des camps opposés en sont venus parfois à habiter des réalités factuelles distinctes sur les sujets les plus contestés (changement climatique, vaccins, pandémie, immigration), chacun s'appuyant sur des sources qu'il tient pour fiables et que l'autre camp rejette comme suspectes. Or, sans accord minimal sur ce qui s'est réellement passé, débattre des valeurs et des choix collectifs devient très difficile.

À cela s'ajoute la logique propre aux algorithmes de recommandation, conçus pour maximiser le temps passé sur les plateformes en captant notre attention, ce qui les pousse

à privilégier les contenus divertissants, et parfois choquants, exagérés, peu nuancés et interpellant l'émotion plutôt que le raisonnement (Bail et al., 2018). Le résultat est bien documenté : une prime donnée aux théories du complot, aux lectures partisanes des enjeux, à la diabolisation des groupes rivaux et à l'outrage, souvent au détriment de l'analyse factuelle et prudente (Algan et al., 2025; Bor et al., 2026; Brady et al., 2021). Le débat politique s'est adapté à ce médium, gagnant en portée globale ce qu'il perd en nuance.

Une technologie qui inverse peut-être la tendance.

C'est dans ce contexte dégradé qu'émergent les grands modèles de langage (LLMs) — tels que ChatGPT, Gemini, Claude, ou Vibe — et le contraste avec les réseaux sociaux mérite qu'on s'y arrête. Le philosophe Dan Williams propose une manière utile de formuler ce contraste : là où les réseaux sociaux ont été une force de démocratisation de l'accès à la parole publique, arrachant le micro aux experts, aux journalistes et aux politiques pour le donner davantage aux masses et à leurs styles de communication, les LLM opèrent en sens inverse (Williams, 2025).

Leur assimilation encyclopédique des textes présents sur internet et leurs capacités de recoupement des sources et de production de synthèses structurées leur permet de faire remonter l'autorité de l'expertise dans les recherches d'information des citoyens,

et donc potentiellement dans l'opinion et le discours publique.
Un LLM fournit en un clic un résumé structuré de l'état des connaissances factuelles ou scientifiques sur le sujet en question. Il ne cherche pas à vous enrôler dans un camp. Il s'en tient aux faits établis et identifie les interprétations idéologiques qui peuvent en être faites, comme pour nous avertir. Les modèles les plus utilisés aujourd'hui, tels que ChatGPT, Claude ou Gemini, refusent, par exemple, de fournir des arguments à l'appui du négationnisme de la Shoah. Ils présentent la discussion, à raison, comme close par les historiens professionnels.

Mais ils donnent, en revanche, l'éventail des positions sur des questions controversées comme, par exemple, le changement climatique, l'efficacité des vaccins ou les effets de l'immigration sur l'emploi. Ce qui peut aider les citoyens à prendre conscience de la complexité des enjeux, et les encourager à l'humilité intellectuelle.

Incitations financières à produire des synthèses fiables

Cette bascule vers l'expertise n'a rien d'un effet de conception désintéressé : elle répond à une logique commerciale. Les grandes entreprises qui développent les LLMs se livrent une concurrence féroce pour construire les systèmes les plus utiles et les plus fiables possibles à un

public aussi large que possible et donc inévitablement hétérogène. Celui-ci inclut les nombreuses entreprises et professionnels dont l'activité dépend directement de la justesse des informations techniques, financières et légales obtenues. Produire un système idéologiquement orienté ou peu rigoureux sur le plan factuel irait rapidement à l'encontre de cet objectif : cela risquerait d'entamer la confiance qu'un public politiquement divers est prêt à accorder à l'outil, et donc sa capacité à convertir cette confiance en abonnements payants.

À cela s'ajoutent le risque juridique qui pèse sur une entreprise dont le système produirait des informations manifestement fausses, en matière médicale, financière ou électorale notamment. Autrement dit, la fiabilité factuelle des LLMs est encouragée par des forces de marché qui n'opèrent pas de la même manière sur les réseaux sociaux, dont le modèle repose sur la maximisation du temps d'attention plutôt que sur la qualité du contenu consommé.

Les expériences montrant un effet persuasif clair

L'hypothèse d'un effet de dépolarisation et de remontée de l'opinion experte dans les recherches d'information effectuées avec les LLMs n'est pas restée que de l'ordre de la spéculation philosophique. Elle a fait l'objet, ces deux dernières années, de tests expérimentaux, avec des résultats globalement encourageants.

TEXTE INTÉGRAL

L'étude la plus connue est celle de Costello, Pennycook et Rand, publiée dans Science en 2024 : plus de 2 000 Américains adhérant à une théorie du complot ont dialogué avec GPT-4, qui leur opposait des contre-arguments personnalisés et fondés sur des faits. L'adhésion moyenne à la théorie choisie par chaque participant a diminué d'environ 20 %, un effet qui s'est maintenu au moins deux mois plus tard (Costello et al., 2024).
Une réplication publiée en 2025 par Boissin, Costello, Rand et Pennycook a reproduit l'effet dépolarisant du dialogue argumenté avec une IA, y compris lorsque les participants croyaient s'adresser à un humain (Boissin et al., 2025). Ce qui suggère que l'effet tient bien à la qualité et à la personnalisation de l'argumentation, non à un quelconque prestige propre à la machine.

Deux études publiées simultanément fin 2025 dans Science et Nature élargissent le tableau du lien entre LLMs et dépolarisation. Dans Science, Hackenburg et ses collègues ont testé 19 modèles de langage sur 707 sujets politiques auprès de près de 77 000 participants (Hackenburg et al., 2025). Ils montrent que ce qui rend un LLM persuasif n'est pas tant sa taille ou la (micro)personnalisation du message, mais entraîner le modèle à la persuasion et la densité factuelle des informations.

TEXTE INTÉGRAL

Dans Nature, Lin et ses coauteurs ont répliqué cette capacité de persuasion fondée sur les faits lors de trois élections aux Etat-unis, Canada, et Pologne, avec des effets positifs modestes mais statistiquement significatifs sur les intentions de vote (choix d'un candidat et option dans un référendum) (Lin et al., 2025).

Ce même travail de Hackenburg et ses collègues (Hackenburg et al., 2026) introduit toutefois une petite réserve. Les réponses les plus denses en informations, et donc les plus persuasives, même si elles avaient un haut niveau de factualité, contenaient aussi un peu plus d'affirmations factuellement inexactes, même si l'exactitude restait très haute dans l'ensemble (un peu plus d'inexactitudes étant certainement un sous-produit de la quantité d'information produite).

Des bénéfices contingents

Le potentiel persuasif des LLM dépend crucialement de la manière dont ces modèles sont entraînés, alignés et déployés. Si ces outils persuadent tout le monde dans le sens des arguments et des preuves experts disponibles, on peut espérer une contribution à la dépolarisation et à la lutte contre la mésinformation. Dans le même temps, il devient aussi malheureusement possible pour un acteur malveillant, ou un gouvernement autoritaire,

ou souhaitant désinformer ou influencer, d'orienter ces mêmes leviers de persuasion vers la propagande plutôt que vers la vérité.

Cependant, pour que tous ces effets de persuasion se traduisent dans le monde réel, il importe de souligner qu'il faut que les gens soient effectivement exposés à ces modèles et à leurs productions dans des chatbots, ce qui est loin d'être gagné, a fortiori le jour où se diffuse le soupçon qu'un modèle a été biaisé stratégiquement par un Etat. Qui, en Europe, dans le contexte actuel, voudrait utiliser un modèle russe ou manipulé par la Russie ?

Des outils de persuasion dépassionnée et personnalisée

Lorsqu'un expert corrige un citoyen sceptique sur les vaccins ou le réchauffement climatique, la réponse émotionnelle souvent déclenchée, comme le sentiment d'être pris de haut ou d'être traité d'ignorant, freine l'assimilation du message. De ce point de vue, les LLMs offrent plusieurs avantages. Ils délivrent l'information d'une manière non condescendante, et même diplomatique. Réglés pour adopter un ton encourageant, ils ne donnent pas l'impression de juger.

Des recherches en psychologie sociale suggèrent que ce ton conciliant ou "sycophantique" facilite l'acceptation de messages allant à l'encontre des croyances préexistantes, parce qu'il ne déclenche pas les mécanismes de "réactance" souvent associés aux efforts de persuasion (Coppock, 2023).

TEXTE INTÉGRAL

Autre avantage essentiel, un LLM peut expliquer sans jamais se fatiguer pourquoi les études sur le lien entre vaccins et autisme ont été invalidées, ou contredire des idées complotistes, en s'adaptant à chaque objection de l'utilisateur (Costello et al., 2024). Ce que même le meilleur vulgarisateur scientifique a du mal à faire lors d'un débat public devant des centaines de personnes.
La persuasion peut donc changer d'échelle, en s'adaptant aux doutes de chacun.

Convergence des faits, pas nécessairement des coeurs
Un point mérite d'être souligné : réduire les divergences de croyances factuelles au sein de la population—sur l'efficacité des vaccins, par ex., ou le nombre d'immigrés arrivant dans votre département chaque année—n'implique pas de faire converger les citoyens sur leurs préférences morales, ou de réduire l'hostilité et l'intolérance entre camps politiques, ou "polarisation affective".

Un rapport de décembre 2025 ayant fait dialoguer des participants avec un LLM instruit à défendre une position centriste sur le soutien à l'Ukraine a bien constaté une dépolarisation idéologique de 10-20% sur cette question, mesurable un mois plus tard, mais pas d'effet sur la tolérance ressentie envers les gens soutenant la position opposée (Walter, 2025).

À l'inverse, d’autres travaux montrent que ce qui réduit le plus la polarisation affective n'est pas tant le contenu factuel de l'échange que son style conversationnel (Hruschka & Appel, 2026). Des chatbots entraînés à faire preuve d'écoute active et de réceptivité conversationnelle (reformuler, valider l'expérience de l'interlocuteur avant de le contredire) produisent une dépolarisation affective plus forte, et augmentent l'humilité intellectuelle des participants, davantage que des chatbots focalisés sur la correction factuelle.

D'autres expériences publiées dans PNAS (Argyle et al., 2023) montrent que des interventions conversationnelles pilotées par IA améliorent la qualité perçue des échanges politiques en ligne — le sentiment d'être compris, la réciprocité démocratique — sans nécessairement faire bouger les opinions politiques elles-mêmes.
Autrement dit : les LLMs semblent capables d'agir à la fois sur ce que nous croyons vrai et sur la civilité avec laquelle nous nous parlons, mais ce sont, en partie, deux leviers distincts, qui ne s'actionnent pas nécessairement ensemble.

Une promesse conditionnelle

Espérons que cette force technocratisante des consommations d’information politique que représentent les LLMs restera stable dans le temps.
Elle dépend des choix commerciaux des entreprises qui

développent ces modèles, de la qualité et de la diversité des données sur lesquelles ils sont entraînés, et se trouverait annulée, dans les régimes autoritaires notamment, par des modifications conduisant à des usages propagandistes plutôt qu'informatifs.

Mais dans un écosystème informationnel dégradé et sous pression, la conjonction de ces résultats dessine une hypothèse solide : celle d'une technologie qui, bien conçue et bien surveillée, pourrait contribuer à inverser certaines des dynamiques les plus délétères des dernières décennies numériques.

Recommandations

À la lumière des recherches récentes que nous avons passées en revue sur les effets persuasifs des LLMs, nous pensons raisonnable de formuler les recommandations suivantes. Elles visent à permettre aux LLMs de servir d'outils d'accès à une information de qualité et de renforcement de la démocratie, plutôt que de fragmentation idéologique et sociale.

TEXTE INTÉGRAL

1. Optimisme mais communication sur les limites

Présenter les LLMs comme des outils puissants, mais faillibles. Les risques d'information approximative demeurent. Le langage et les positions par défaut des principaux LLMs penchent, selon les audits disponibles, du côté du centre gauche (Motoki et al., 2024; Rozado, 2024), un penchant qui tient aux données d'entraînement et à l'alignement plus qu'à une conception délibérée, et qui varie selon les questions, la langue et la formulation des requêtes. Au cours du temps, il est possible que certains LLMs s'adaptent progressivement à une audience partisane d'utilisateurs, et deviennent à leur tour plus orientés politiquement qu'ils ne le sont actuellement (été 2026). Et comme avec les médias partisans, les utilisateurs pourraient s'auto-sélectionner vers des LLMs alignés avec leurs opinions. Grok se positionne déjà comme une alternative « anti-woke » à ChatGPT. Cette tendance pourrait reproduire, dans une certaine mesure, la logique d'exposition sélective qui a fragmenté l'audience des chaînes d'information. Ceci étant dit, début 2026, même Grok proposait des fact-checks non partisans d'évènements dont l'interprétation était polarisée, comme la mort d'Alex Pretti à Minneapolis, aux Etats-Unis, et refusait de produire des arguments niant l'existence de l'Holocauste (Williams, 2025).

TEXTE INTÉGRAL

2. Privilégier des modèles à "transparence délibérative"

Au lieu de promouvoir une neutralité artificielle souvent perçue comme illégitime ou illusoire, les concepteurs de modèles pourraient envisager la transparence sur l'espace des positions légitimes. Plutôt que de proposer une réponse unique, les systèmes pourraient exposer explicitement l'éventail des arbitrages défendables sur une question donnée, permettant à l'utilisateur de se situer dans la diversité des opinions fondées.

3. Sanctuariser le jugement humain dans les processus décisionnels

Il est impératif de délimiter des domaines (décisions politiques, arbitrages moraux) où l'intervention de l'IA doit être limitée au rôle de facilitateur. L'IA doit compenser les limites humaines (mémoire de travail, accès aux informations, écoute active) sans se substituer à la délibération et à la prise de décision humaine, qui seule garantit l'apprentissage civique et l'acceptabilité des propositions.

4. **Déployer des protocoles de détection contre la manipulation**

Face au risque de création de consensus artificiels par des essaims d'agents autonomes, les plateformes et régulateurs doivent développer des protocoles spécifiques de détection des bots coordonnés. La lutte contre la fabrication de "fausse majorités" (astroturfing) est importante pour éviter que l'IA ne devienne un outil de désinformation à grande échelle.

5. **Adopter des benchmarks de "performance démocratique"**

La mesure de la performance des modèles doit évoluer. Au-delà de la précision technique, il serait bénéfique d’intégrer des benchmarks évaluant la capacité à maintenir une acceptabilité politique mutuelle. Il est recommandé d'utiliser des métriques d'approbation croisée et de "bridging", testant si une réponse est jugée utile par des groupes idéologiquement opposés. Ce changement de paradigme incite les développeurs à optimiser l'IA pour l'utilité d’un public politiquement divers.

6. Valoriser l'écoute active et la réceptivité conversationnelle

Pour réduire la polarisation affective, les interfaces de dialogue doivent être entraînées à des comportements de modération active (reformulation, validation de l'expérience de l'interlocuteur avant contradiction). La priorité ne doit pas être la correction factuelle seule, mais aussi la qualité de l'échange, favorisant l'humilité intellectuelle et le sentiment d'écoute et de réciprocité.

7. Encadrer le déploiement des outils de vérification automatique

Le fact-checking par IA déployé « à l'état sauvage » présente des risques. Des erreurs de classification peuvent discréditer des sources fiables et renforcer involontairement des théories du complot. Ces outils gagnent donc à garder les humains dans la boucle, comme le fait Community Notes sur X. Depuis 2025, des agents IA y rédigent des propositions de notes d'évaluation des posts, ce qui accélère la production, mais rien n'est publié sans avoir été jugé utile par des contributeurs de bords politiques opposés. L'IA augmente ainsi la capacité du dispositif sans se substituer au jugement collectif qui fonde sa légitimité. Ce modèle est préférable à l'oracle unique du « @Grok, c'est vrai ? », en moyenne assez fiable (Renault et al. 2025), mais dont les erreurs sont publiées avec la même autorité que les réponses justes, faute d'évaluation collective.

Références

Algan, Y., Renault, T., & Subtil, H. (2025). La Fièvre parlementaire : Ce monde où l'on catche ! CEFREMAP.

Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. Proceedings of the National Academy of Sciences, 120(41), e2311627120. https://doi.org/10.1073/pnas.2311627120

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences of the United States of America, 115(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. PNAS Nexus, 4(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

Bor, A., Marie, A., Pradella, L., & Petersen, M. B. (2026). Social media users experience more political hostility in less economically equal and less democratic societies. Nature Human Behaviour. https://doi.org/10.1038/s41562-026-02432-5
Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. Science Advances, 7(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Coppock, A. (2023). Persuasion in Parallel: How Information Changes Minds about Politics. University of Chicago Press.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. Science, 385(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hackenburg, K., Wagner, C., Hewitt, L., Tappin, B. M., Saunders, E., Kirk, H. R., Margetts, H., & Summerfield, C. (2026). AI systems out-persuade expert humans (arXiv:2606.16475). arXiv. https://doi.org/10.48550/arXiv.2606.16475

Hruschka, T. M. J., & Appel, M. (2026). Reducing political polarization through conversations with artificial intelligence. Journal of Computer-Mediated Communication, 31(2), zmag003. https://doi.org/10.1093/jcmc/zmag003

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. Nature, 648(8093), 394–401. https://doi.org/10.1038/s41586-025-09771-9

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. Public Choice, 198(1), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Rozado, D. (2024). The political preferences of LLMs. PLOS ONE, 19(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

Walter, J. (2025). Using AI Persuasion to Reduce Political Polarization. Working Paper. https://johanneswalter.github.io/

Williams, D. (2025, March 3). How AI Will Reshape Public Opinion. Substack. https://substack.com/@conspicuouscognition/p-188535644