

**CHATGPT DANS
L'ENSEIGNEMENT DES STIM :
INTRODUCTION À
L'UTILISATION DES OUTILS
AXÉS SUR LES GML DANS
L'ENSEIGNEMENT SUPÉRIEUR**

JAKUB HYZYK ET MELANIE MISANCHUK



L'ouvrage « *ChatGPT dans l'enseignement des STIM : Introduction à l'utilisation des outils axés sur les GML dans l'enseignement supérieur* » (tous droits réservés) de Jakub Hyzyk et Melanie Misanchuk est distribué sous licence internationale Attribution – Pas d'utilisation commerciale – Pas de modification 4.0 de Creative Commons, à moins d'indications contraires.

TABLE DES MATIÈRES

Introduction	v
PARTIE I. INTRODUCTION	
1. Ouvrages cités	5
PARTIE II. BREF HISTORIQUE DE L'APPRENTISSAGE AUTOMATIQUE ET DES GML	
PARTIE III. PORTÉE ET LIMITATIONS DE LA PRÉSENTE REL ET ANALYSE DU CONTEXTE	
2. Portée et limitations	15
3. Analyse du contexte	17
PARTIE IV. FONCTIONNEMENT DE LA TECHNOLOGIE DES GML	
PARTIE V. LIMITATIONS DES GML	
PARTIE VI. BIAIS	
4. Introduction	31
5. Jeux de données	32
6. Types de biais	36
7. Atténuation des biais	49
8. Rétroactions humaines menant à l'effacement des groupes marginalisés	51
PARTIE VII. INTÉGRITÉ DANS LES ÉTUDES	
9. Introduction	55
10. Que peut faire un.e pauvre enseignant.e surchargé.e?	57
11. Lignes directrices générales	59

12. Détecteurs d'IA	61
---------------------	----

PARTIE VIII. UTILISATION DES GML POUR L'ENSEIGNEMENT ET L'APPRENTISSAGE

13. Introduction	65
------------------	----

14. Utilisation de ChatGPT par le corps professoral	66
-----------------------------------------------------	----

15. Utilisation de ChatGPT par la population étudiante	77
--------------------------------------------------------	----

PARTIE IX. CONCLUSION

Remerciements	83
---------------	----

INTRODUCTION

Résultats d'apprentissage



Après avoir utilisé la présente ressource, enseignant.e.s pourront :

- parler de façon générale de l'historique de l'intelligence artificielle générative (IAG) et des grands modèles de langage (GML);
- définir les mots-clés et les termes généraux employés dans le domaine;
- appliquer leur compréhension de l'entraînement et du développement des outils axés sur les GML aux conséquences pour les profanes;
- parler de façon générale des outils axés sur les GML disponibles actuellement (en versions gratuites ou payantes);
- débattre des limitations techniques des outils axés sur les GML disponibles actuellement;
- décrire les types de biais inhérents aux outils axés sur les GML;
- s'efforcer d'atténuer les biais dans ces outils;
- décrire les usages possibles des outils axés sur les GML dans l'enseignement et l'apprentissage des sciences au niveau postsecondaire;
- débattre des considérations relatives à l'utilisation des outils axés sur les GML dans

l'enseignement (p. ex., fiabilité, biais, confidentialité et sécurité, accessibilité, équité, etc.);

- intégrer l'IAG dans les activités et les évaluations sans enjeu important;
- améliorer l'« immunité à l'intelligence artificielle (IA) » des évaluations;
- prendre part aux comités intra-institutionnels et interinstitutionnels, et à la gouvernance collégiale sur le rôle de l'IAG dans l'enseignement postsecondaire.

PARTIE I

INTRODUCTION

Introduction

Il est fort probable que toutes les personnes travaillant dans le domaine des sciences, des technologies, de l'ingénierie et des mathématiques (STIM) aient déjà entendu les derniers progrès liés à l'intelligence artificielle (IA), de façon générale, ou de ChatGPT, en particulier. Depuis plus d'un an, les médias regorgent de reportages faisant l'éloge des capacités révolutionnaires prétendument accessibles à toute personne ou se lamentant de l'éventuelle désuétude de tout type d'emploi de bureau ou lié au savoir.

Dans le milieu de l'enseignement postsecondaire, vous avez peut-être entendu des collègues parler de l'intelligence artificielle générative (IAG), préoccupé.e.s par ses conséquences sur l'intégrité dans les études ou emballé.e.s par les possibilités pédagogiques qu'elle offre (ou les deux), ou encore vu des gestionnaires de technologies de l'information (TI) ayant de la difficulté à concilier leur emballement quant à leur formation et leur connaissance des courbes d'intérêt pour la technologie.

La présente ressource éducative libre (REL) n'a pas pour but d'enthousiasmer ou d'affoler le personnel enseignant. Elle cherche plutôt à offrir un point de vue équilibré et mesuré sur la technologie des GML et les dialogueurs qui en dépendent. Dans la présente REL, nous proposons :

- un historique de base;
- une compréhension du fonctionnement, des fonctions et des limitations des outils axés sur les GML;
- certaines analogies à appliquer à la réflexion sur ces outils;
- une analyse de quelques enjeux et possibilités se rapportant à l'enseignement des STIM pour les études supérieures.

Quelle ampleur peut véritablement prendre ce phénomène?

Il est raisonnable de se demander si la question des conséquences possibles des GML sur l'éducation, qu'elles soient positives ou négatives, mérite toute l'encre qu'elle fait couler. Après tout, il est dans

l'intérêt des entreprises qui vendent ces outils de piquer le plus possible la curiosité des gens à propos de la « révolution de l'IA » qu'elles façonnent elles-mêmes. De tels efforts visent à maximiser les ventes et l'adhésion des investisseurs publics et en capital risque. Nous avons vu les responsables de ces entreprises réclamer publiquement une réglementation de la technologie et prévenir les gens des dangers pouvant en découler, gestes qu'une personne externe pourrait considérer comme preuve d'une bonne conscience sociale de la part de ces entreprises. Une autre personne moins indulgente pourrait souligner que ce sont les premiers chefs de file du marché qui revendent le plus bruyamment une « pause » et qu'ils font ouvertement référence aux futurs développements plutôt qu'à leurs propres produits, déjà sur le marché. Le fait de maintenir le potentiel de la technologie dans l'actualité tout en imposant un lourd fardeau réglementaire à des concurrents constituerait assurément une habile stratégie d'affaires.

Alors donc, outre ce que présente la presse populaire, que savons-nous vraiment sur l'adoption de cette technologie? Dans quelle mesure la population étudiante y a-t-elle recours? Quels sont ses effets sur l'enseignement supérieur? À quoi pouvons-nous nous attendre dans un proche avenir?

Données

La moyenne mondiale des étudiant.e.s universitaires ayant recours à l'IAG dans le cadre de leurs études s'établit à 40 %. Au Canada, ce résultat se situe à 54 %; aux États-Unis, il n'atteint que 20 %. Parmi l'ensemble de la population étudiante utilisant l'IAG dans ses études, 50 % posent au moins une question par jour. Au Canada, 65 % de la population étudiante s'en sert de deux à dix fois par jour, et 80 % des étudiant.e.s utilisateur.trice.s de l'IAG y ont recours plus d'une fois par jour (seuls la population étudiante de Turquie utilise l'IAG plus que la population étudiante canadienne [81 %]). À l'échelle mondiale, 44 % de la population étudiante (mais seulement 31 % au Canada et 39 % aux États-Unis) déclarent mieux comprendre les concepts et sujets complexes depuis le recours à l'IAG dans le cadre de leurs études.

À la question demandant s'ils ou elles aimeraient que leur programme comprenne de la formation sur les outils d'IA nécessaires à leur future carrière, 83 % des étudiant.e.s du Kenya et de l'Inde ont répondu par l'affirmative. À l'échelle de la planète, la moyenne des étudiant.e.s souhaitant obtenir de la formation sur les outils d'IA s'établit à 65 %. Au Canada, elle se situe à 63 %; aux États-Unis, elle n'atteint que 47 % (Chegg.org, 2023).

Éléments anecdotiques

La vidéo Vox qui suit, produite par le journaliste Joss Fong, propose d'autres données et aspects anecdotiques sur l'utilisation de ChatGPT et présente certains enjeux de manière concise et accrocheuse du point de vue du corps professoral et de la population étudiante :

Un ou plusieurs éléments interactifs ont été retirés de la présente version du texte. Vous pouvez

les consulter en ligne ici : <https://ecampusontario.pressbooks.pub/llmtoolsforstemteaching/?p=177#oembed-1>

Ces statistiques concernent surtout ChatGPT qui, jusqu'à tout récemment, était l'outil le plus largement répandu et le plus facilement accessible. Toutefois, Microsoft déploie actuellement la même technologie GPT-4 soutenant ChatGPT dans toute sa gamme de produits : elle est incluse dans le navigateur Edge et le moteur de recherche Bing, incorporée dans tous les outils Microsoft Office 365 au moyen de Copilot, et même intégrée dans les nouvelles versions de Windows. Ces produits sont largement répandus (il s'agit souvent de la seule option soutenue sur le plan institutionnel) dans les établissements postsecondaires au Canada. Si le passé est garant de l'avenir, il est fort probable que de nombreux services de technologie de l'information (TI) activeront simplement ces fonctions dès leur dévoilement pour le personnel, le corps professoral *et* la population étudiante. On peut affirmer sans présomption que, au moment de lire ces lignes, un équivalent complet de ChatGPT sera sans doute accessible d'un simple clic ou dans une barre latérale de Word, PowerPoint, Windows et d'autres outils que vous – et toute la population étudiante – utiliserez quotidiennement.

1.

OUVRAGES CITÉS

Melanie Misanchuk et Jakub Hyzyk

Aligning language models to follow instructions, sans date. Document extrait le 9 décembre 2023 au <https://openai.com/research/instruction-following>.

Angwin, J., J. Larson, S. Mattu et L. Kirchner. « *Machine Bias* », *ProPublica*, 23 mai 2023. Adresse : <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Bandy, J. et N. Vincent. *Addressing “Documentation Debt” in Machine Learning Research: A Retrospective Datasheet for BookCorpus*, 2021. Document arXiv (arXiv:2105.05241) au <https://doi.org/10.48550/arXiv.2105.05241>.

Brandom, R. « *What languages dominate the internet?* », *Rest of World*, 7 juin 2023. Adresse : <https://restofworld.org/2023/internet-most-used-languages/>.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, [...], et D. Amodei. *Language Models are Few-Shot Learners*, 2020. Document arXiv (arXiv:2005.14165) au <https://doi.org/10.48550/arXiv.2005.14165>.

Cao, Y., L. Zhou, S. Lee, L. Cabello, M. Chen, et D. Hershcovich. « *Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study* », dans S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, et L. Benotti (éd.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Association for Computational Linguistics, 2023, pages 53 à 67. Document extrait au <https://doi.org/10.18653/v1/2023.c3nlp-1.7>.

CAP Recommendations to Aid in Adoption of New eGFR Equation, College of American Pathologists, sans date. Document extrait le 7 décembre 2023 au <https://www.cap.org/member-resources/articles/cap-recommendations-for-adoption-of-new-egfr-equation>.

Chaslot, G., et S. Monnier. *AlgoTransparency*, AlgoTransparency, sans date. Document extrait le 9 décembre 2023 au <https://algotransparency.org/>.

Deck, A. « *We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed.* », *Rest of World*, 6 septembre 2023. Adresse : <https://restofworld.org/2023/chatgpt-problems-global-language-testing/>.

Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, et Gardner.

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, 2021. Document arXiv (arXiv:2104.08758) au <https://doi.org/10.48550/arXiv.2104.08758>.

Dovidio, J. F., K. Kawakami, et S. L. Gaertner. « Implicit and explicit prejudice and interracial interaction », *Journal of Personality and Social Psychology*, volume 82, numéro 1, 2002, pages 62 à 68. Adresse : <https://doi.org/10.1037/0022-3514.82.1.62>.

Gardiner, B. « “It’s a terrible way to go to work:” what 70 million readers’ comments on the Guardian revealed about hostility to women and minorities online », *Feminist Media Studies*, volume 18, numéro 4, 2018, pages 592 à 608. Adresse : <https://doi.org/10.1080/14680777.2018.1447334>.

Ging, D., et E. Siapera. « Special issue on online misogyny », *Feminist Media Studies*, volume 18, numéro 4, 2018, pages 515 à 524. Adresse : <https://doi.org/10.1080/14680777.2018.1447345>.

Greenwald, A. G., D. E. McGhee, et J. L. Schwartz. « Measuring individual differences in implicit cognition: The implicit association test », *Journal of Personality and Social Psychology*, volume 74, numéro 6, 1998, pages 1464 à 1480. Adresse : <https://doi.org/10.1037//0022-3514.74.6.1464>.

Growcoat, M. « Which AI Image Generator is The Most Biased? », *PetaPixel*, 3 novembre 2023, <https://petapixel.com/2023/11/03/which-ai-image-generator-is-the-most-biased/>.

Hort, M., Z. Chen, J. M. Zhang, M. Harman, et F. Sarro. *Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey*, 2023. Document arXiv (arXiv:2207.07068) au <https://doi.org/10.48550/arXiv.2207.07068>.

Johnson, K. « The Efforts to Make Text-Based AI Less Racist and Terrible », *WIRED*, 17 juin 2021. Adresse : <https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible/>.

Little, O., et A. Richards. *TikTok’s algorithm leads users from transphobic videos to far-right rabbit holes*, Media Matters for America, 5 octobre 2021. Adresse : <https://www.mediamatters.org/tiktok/tiktoks-algorithm-leads-users-transphobic-videos-far-right-rabbit-holes>.

Luccioni, A. S., C. Akiki, M. Mitchell, et Y. Jernite. *Stable Bias: Analyzing Societal Representations in Diffusion Models*, 2023. Document arXiv (arXiv:2303.11408) au <https://doi.org/10.48550/arXiv.2303.11408>.

McCrosky, J., et B. Geurkink. *YouTube Regrets*, 2021. Adresse : <https://foundation.mozilla.org/en/youtube/findings/>.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, et I. Sutskever. *Language Models are Unsupervised Multitask Learners*, 2019. Adresse : <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>.

Rettberg, J. W. « ChatGPT is multilingual but monocultural, and it’s learning your values », *Jill/Txt*, 6 décembre 2022. Adresse : <https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/>.

Tiku, N., K. Schaul, et S. Y. Chen. « These fake images reveal how AI amplifies our worst stereotypes », *The Washington Post*, 1^{er} novembre 2023. Adresse : <https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/>.

Wan, Y., G. Pu, J. Sun, A. Garimella, K.-W. Chang, et N. Peng. “*Kelly is a Warm Person, Joseph is a Role Model*”: *Gender Biases in LLM-Generated Reference Letters*, 2023. Document arXiv (arXiv:2310.09219) au <https://doi.org/10.48550/arXiv.2310.09219>.

Wertheim, S. « The Common Habit That Undermines Organizations’ Diversity Efforts », *Fast Company*, 30 mai 2016. Adresse : <https://www.fastcompany.com/3060336/the-common-habit-that-undermines-organizations-diversity-efforts>.

Wertheim, S. « ChatGPT insists that doctors are male and nurses female », *Worthwhile Consulting*,

26 avril 2023. Adresse : <https://www.worthwhileconsulting.com/read-watch-listen/chatgpt-insists-that-doctors-are-male-and-nurses-female>.

West, M., R. Kraut, et H. E. Chew. *Je rougirais si je pouvais : réduire la fracture numérique entre les genres par l'éducation*, UNESCO Bibliothèque Numérique, 2022. Adresse : https://unesdoc.unesco.org/ark:/48223/pf0000367416_fre.

Xu, A., E. Pathak, E. Wallace, S. Gururangan, M. Sap, et D. Klein. *Detoxifying Language Models Risks Marginalizing Minority Voices*, 2021. Document arXiv (arXiv:2104.06390) au <https://doi.org/10.48550/arXiv.2104.06390>.

PARTIE II

BREF HISTORIQUE DE L'APPRENTISSAGE AUTOMATIQUE ET DES GML

Bref historique de l'apprentissage automatique et des GML

Dans les lignes qui suivent, nous décrivons quelques-unes des grandes étapes de la conception de l'apprentissage automatique et nous insistons sur les progrès ayant mené aux GML et aux dialogueurs d'aujourd'hui.

Fondements et naissance de l'apprentissage automatique

Le premier réseau neuronal artificiel, appelé « Stochastic Neural Analog Reinforcement Calculator (SNARC) », a été créé en 1951 par Marvin Minsky. Il faisait usage de l'apprentissage par renforcement pour simuler la reproduction de rats à différents endroits dans un labyrinthe et la recherche d'une issue par ces rongeurs pour s'extirper du dédale.

En 1958, Frank Rosenblatt a mis au point le perceptron, une autre forme nouvelle de réseau neuronal pouvant classifier des modèles simples. Le perceptron a fait la manchette dans les publications de vulgarisation scientifique, exposant ainsi au grand public le concept de l'apprentissage automatique. (Source : Rosenblatt, Frank. « The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain », *Psychological Review*, volume 65, numéro 6, Cornell Aeronautical Laboratory, 1958, pages 386 à 408.)

Le terme « apprentissage automatique » a été employé pour la première fois dans un article d'Arthur Samuel, publié en 1959, pour décrire les expériences de ce dernier ayant pour but de programmer la première version d'un ordinateur central (IBM 704) pour lui apprendre les échecs. Ses expériences ont constitué la première « validation de principe » de l'apprentissage automatique et l'ont mené à tirer la conclusion suivante :

« À la suite de ces expériences, on peut affirmer avec une certaine assurance qu'il est désormais possible de concevoir des schèmes d'apprentissage beaucoup plus performants qu'une personne moyenne et que

l'application de tels schèmes à des problèmes pratiques pourrait devenir économiquement réalisable. »

(Source : Samuel, A. L. « Some studies in machine learning using the game of checkers », *IBM Journal of Research and Development*, volume 3, numéro 3, 1959, pages 210 à 229.)

Le premier dialogueur, appelé « ELIZA », a été inventé par Joseph Weizenbaum à l'Artificial Intelligence Laboratory du Massachusetts Institute of Technology (MIT) dans les années 1960.

Années 1980 aux années 1990 : Réapparition et perfectionnement

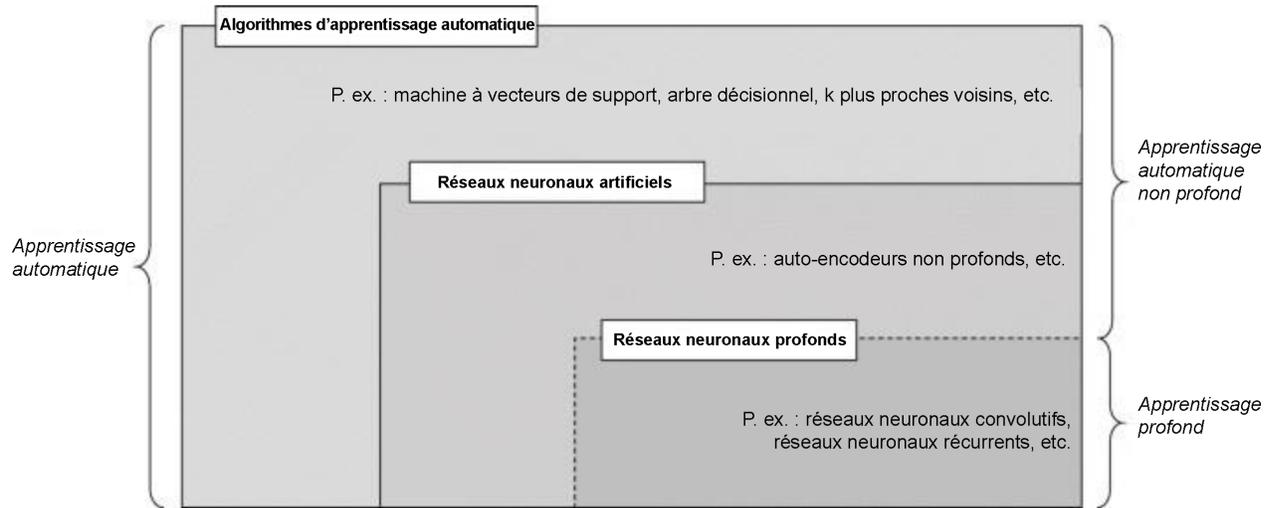
- **Rétropropagation** : Cet algorithme fondamental a rendu plus efficace l'entraînement des réseaux neuronaux à couches multiples, surmontant ainsi certaines limitations du perceptron.
- **Réseaux neuronaux convolutifs (RNC)** : Inspirés par le fonctionnement du cortex visuel, les RNC marquent un tournant pour les tâches liées à la vision artificielle.

Années 2000 : Multitude de progrès grâce aux données et à la puissance des ordinateurs

- **Révolution de l'apprentissage profond** : Des techniques comme celles mises au point par Hinton et ses collègues (2006) ont rendu possible l'entraînement efficace des réseaux neuronaux à couches multiples, démontrant ainsi un rendement grandement amélioré.
- **Arrivée des unités de traitement graphique (UTG)** : La mise au point de puissantes UTG a grandement accéléré les calculs des réseaux neuronaux, permettant ainsi la création de modèles complexes.

Années 2010 à aujourd'hui : Transformateur axé sur le langage

- **2017** : L'architecture des transformateurs révolutionne le traitement neuronal du langage (TNL). Au lieu de faire appel à un traitement purement séquentiel, elle tire avantage des mécanismes d'« attention » pour comprendre les relations entre les mots, et ce, peu importe l'espace les séparant (Vaswani et coll., 2017).
- **Grands modèles de langage (GML) (à compter des années 2010)** : Les GML, tout comme GPT-3 et ses variations, offrent des capacités remarquables pour la génération de texte, la traduction et la conversation. Les architectures fondées sur les transformateurs et les jeux de données d'entraînement de masse rendent tout cela possible. Ces progrès repoussent les limites de la compréhension du langage et de l'interaction. (Source : Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, [...], et D. Amodei. *Language models are few-shot learners*, 2020. Prépublication arXiv [arXiv:2005.14165]).



Attributions de médias

- history_ML_ANN_DNN_venn

Attributions de médias

- history_ML_ANN_DNN_venn

PARTIE II

**PORTÉE ET LIMITATIONS DE LA
PRÉSENTE REL ET ANALYSE DU
CONTEXTE**

2.

PORTÉE ET LIMITATIONS

Les recherches se rapportant à la présente REL ont été menées durant l'automne 2023. En conséquence, des outils et fonctions n'étaient pas disponibles ou ne convenaient pas dans le cadre de nos essais (p. ex., Bard n'était pas disponible au Canada, et Gemini n'existait pas encore). Cela signifie donc que les produits et fonctions disponibles sont représentatifs de cette période. Étant donné la transformation rapide du contexte lié à la question, nous espérons pouvoir mettre à jour la présente REL de façon périodique.

Les GML, en tant que catégorie, ont fait l'objet d'un examen approfondi. Des outils spécifiques ont été utilisés pour effectuer les essais pratiques et la validation (ChatGPT, Bing Copilot, AITutorPro ou AITeachingAssistantPro) :

- ChatGPT est le dialogueur le plus connu et le premier de sa catégorie;
- Bing Chat/Copilot est le dialogueur gratuit le plus largement répandu et le seul à comprendre la génération augmentée de récupération (GAR), qui est aussi grandement utilisée dans le milieu de l'enseignement supérieur en Ontario;
- AI Tutor Pro/AI Teaching Assistant Pro (Contact Nord) est un dialogueur représentatif conçu pour l'enseignement en contexte canadien.

Il sera également question d'autres GML et types d'IAG, comme la génération et la reconnaissance d'images, pour faciliter l'explication et la compréhension plus générale de l'IAG. Tous les GML utilisés dans le cadre de nos essais et à titre d'exemples font l'objet d'un développement rapide et intensif. Le domaine de l'IAG se transforme sans cesse sur les plans des capacités des modèles, des nouveaux acteurs ainsi que de l'extensibilité ou de la connectivité avec d'autres outils. Voilà pourquoi nous tenterons, dans la mesure du possible, d'effectuer un examen de nature général, de mettre l'accent sur les caractéristiques et éléments communs de la technologie les plus susceptibles de durer, et d'encourager les modèles mentaux et équivalents de la technologie axés sur la longévité.

La présente REL comble une lacune en ce qui a trait aux documents disponibles, accessibles et prêts à la mise en œuvre. Il existe de nombreux articles sur l'IAG et les outils axés sur les GML dans les publications savantes et de vulgarisation, mais la plupart n'approfondissent que l'un de ces deux aspects des outils ou ne traitent que d'un certain nombre de sujets de manière très générale. La présente REL vise à offrir un document de référence unique aux enseignant.e.s de STIM du milieu

postsecondaire canadien qui cherchent à savoir les avantages et les mises en garde associés à l'utilisation des outils axés sur les GML dans leurs cours. Les sections sur les biais, la confidentialité et la sécurité, la propriété intellectuelle et les droits d'auteurs, l'intégrité dans les études, les pratiques pédagogiques exemplaires et la conception des évaluations proposent d'importantes perspectives à l'aide desquelles examiner les outils (et les stratégies pour s'en servir) qui sont largement promus, mais mal compris actuellement.

3.

ANALYSE DU CONTEXTE

Nom du dialogueur	Fondement	Concepteur	Capacités générales	Fonctions notables	Adresse URL de l'interface Web
ChatGPT	GPT-3.5 ou version ultérieure	OpenAI	IA conversationnelle, génération de texte, réponse à des questions, soutien pédagogique, aide à la programmation	Vaste gamme d'aptitudes à soutenir une conversation ou des interactions fondées sur les connaissances	ChatGPT d'OpenAI
BlenderBot 3	Exclusivité	Facebook AI (Meta)	IA conversationnelle, discussions sur des connaissances générales, personnalité	Il vise à avoir des conversations qui sont semblables à celles des êtres humains, qui ont du caractère et qui favorisent la mobilisation	BlenderBot 3 (remarque : la disponibilité peut varier) États-Unis seulement
Bing Copilot	GPT (la version peut varier)	Microsoft	Recherche améliorée, IA conversationnelle, synthèse de renseignements, intégration à l'écosystème de Microsoft	Il révolutionne les interactions des utilisateurs avec les moteurs de recherche au moyen de réponses conversationnelles et contextuelles.	Intégration à l'interface de recherche de Bing : Bing
Bard/Gemini	LaMDA	Google	IA conversationnelle, extraction de renseignements, génération de contenu créatif	Il s'intègre aux capacités de recherche de Google pour offrir des renseignements et des réponses à jour.	Produit lancé au départ lors d'une avant-première à accès limité, les détails se trouvent sur la page officielle de Bard de Google. États-Unis seulement au moment de la rédaction du présent document
Claude	Exclusivité	Anthropic	IA conversationnelle, création de contenu, soutien à l'enseignement	Il met l'accent sur la sécurité et le respect des valeurs humaines.	Accès par l'entremise de la plateforme d'Anthropic. L'adresse URL précise peut varier du fait de l'expansion de l'accès à Claude.

PARTIE III

FONCTIONNEMENT DE LA TECHNOLOGIE DES GML

Fonctionnement de la technologie des GML

Il est difficile de formuler une description plus concise du fonctionnement des GML axés sur les transformateurs, comme ChatGPT, que celle donnée par Stephen Wolfram :

Le concept de base de ChatGPT est assez simple dans une certaine mesure. Il faut commencer par un énorme échantillon de textes d'origine humaine provenant d'Internet, de livres, etc. Ensuite, il faut entraîner le réseau neuronal afin qu'il produise du texte ressemblant. Il faut notamment lui permettre de commencer par une « requête » et de poursuivre avec du texte « semblable à celui ayant servi à son entraînement ». (Wolfram, 2023)

Si le concept de base s'avère assez simple, le processus de création d'un transformateur génératif préentraîné (TGP) pleinement fonctionnel ne l'est pas. Sans trop entrer dans les détails, nous tentons de fournir assez de renseignements aux utilisateur.trice.s pour leur permettre de conceptualiser ce qui se déroule en arrière-plan. Nous divisons l'explication en deux parties :

1. l'entraînement du GML (c'est-à-dire la manière dont il apprend tout ce qu'il sait);
2. ce qui se produit durant son exécution (soit lorsque l'utilisateur.trice lui soumet des requêtes).

Entraînement du GML

Avant l'entraînement : Constitution de la base de connaissances

Avant qu'un GML comme ChatGPT puisse traiter votre entrée (requête), il faut d'abord :

- lui apprendre à utiliser le langage;

- lui fournir le plus de renseignements contextuels possible à propos de l'« univers ».

Pour répondre à ces deux besoins, on leur fournit de vastes jeux de données textuelles, comme des livres, des articles, des codes et des conversations en ligne. Ces jeux de données sont soumis à un processus de segmentation du texte en unités lexicales, de la mise en correspondance aux plongements, puis à une architecture de transformateurs. Les algorithmes et les méthodes de calcul de ces processus dépassent la portée de cette explication simplifiée, mais ce reportage élaboré visuel et interactif du *Financial Times* parvient à les illustrer avec brio. D'autres ressources techniques sont proposées à la fin de la présente section pour les personnes intéressées à approfondir le sujet.

Pour les modèles à code source fermé comme GPT-3.5 et GPT-4 (utilisés dans ChatGPT), les jeux de données exacts et leurs pondérations sont exclusifs.

Les lecteur.trice.s incapables de contenir leur curiosité peuvent consulter le document de définition de Llama et cette liste de populaires jeux de données ouvertes pour avoir une idée de leur contenu.

Toutefois, les GML à code source ouvert, comme Llama de Meta, publient les jeux de données qu'ils utilisent ainsi que leurs pondérations. Nous aurons d'autres choses à ajouter concernant les jeux de données et leur qualité dans la section sur les biais, mais, pour le moment, les caractéristiques de leur source et de leur contenu ne sont pas pertinentes pour la compréhension du fonctionnement des GML. Il suffit de savoir que, à titre d'exemple, le modèle Llama a été entraîné à l'aide d'environ 4,75 téraoctets de données textuelles tirées de différentes sources Internet et imprimées (numérisées). Ce nombre représente presque 400 millions de pages de texte.

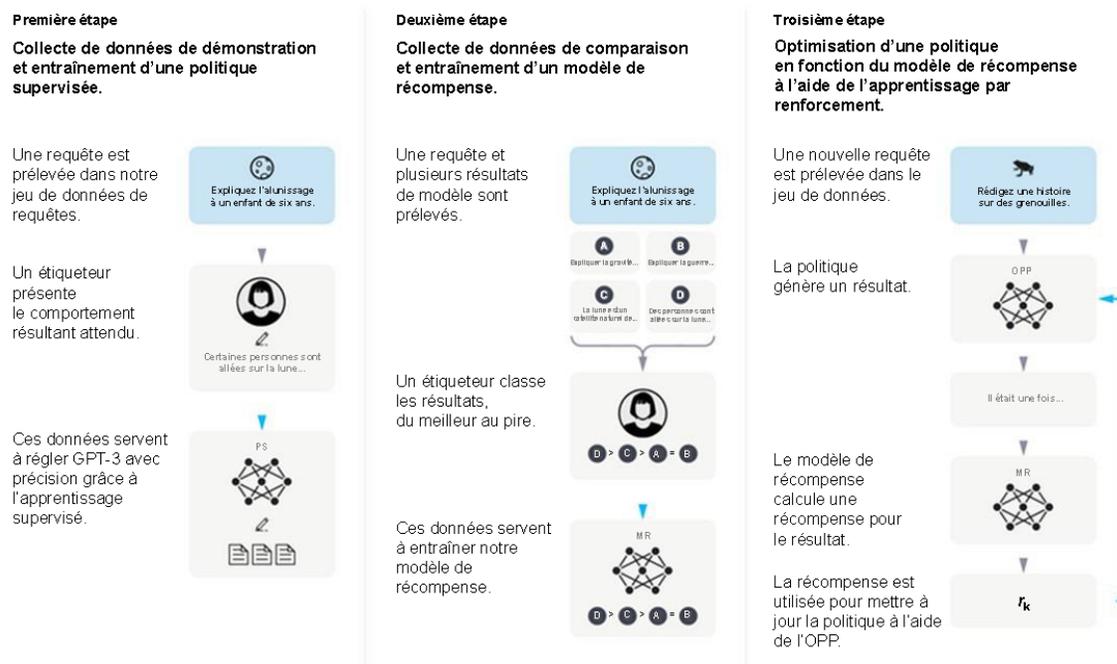
Ces données sont traitées par le GML durant un cycle d'apprentissage non supervisé. Le jeu de données est traité ainsi dans son intégralité à plusieurs reprises, et les paramètres du modèle sont ajustés à chaque passage. Le rendement est aussi évalué de façon périodique en fonction d'un jeu de validation pour surveiller la progression et prévenir le surapprentissage. L'entraînement se poursuit jusqu'à ce que le rendement du modèle cesse de s'améliorer de manière significative.

Peaufinage : Spécialisation et perfectionnement

L'étape suivante de l'entraînement comporte une supervision humaine, dans le cadre d'un processus appelé « apprentissage par renforcement à partir de rétroactions humaines » (ARRH). Le tableau qui suit présente un exemple de processus d'ARRH, soit celui mis en œuvre dans InstructGPT d'OpenAI :

Figure 1

Processus d'ARRH d'OpenAI



C > A = B. C. Le texte dit : « Ces données servent à entraîner notre modèle de récompense ». Elles sont représentées par la même image de réseau neuronal que celle de la première étape, étiqueté « MR », et les résultats sont classés en dessous. Troisième étape : Optimisation d'une politique en fonction du modèle de récompense à l'aide de l'apprentissage par renforcement. A. Le texte dit : « Une nouvelle requête est prélevée dans le jeu de données ». Il y a le dessin d'une grenouille et l'énoncé « Rédigez une histoire sur des grenouilles ». B. Le texte dit : « La politique génère un résultat ». À côté se trouve une image de réseau neuronal étiqueté « OPP » et menant à une bulle indiquant « Il était une fois... ». B. Le texte dit : « Le modèle de récompense calcule une récompense pour le résultat ». Le sigle « MR » et le symbole de réseau neuronal se trouvent à côté. C. Le texte dit : « La récompense est utilisée pour mettre à jour la politique à l'aide de l'OPP. ». À côté se trouvent la lettre « r » avec le coefficient « k ». Le diagramme utilise du bleu et du gris de manière uniforme, et l'arrière-plan est pâle. Des flèches relient les images des sous-étapes pour illustrer le déroulement du processus. Les icônes et les symboles servent à représenter différentes entités incluses dans le processus, comme les jeux de données, les résultats et les modèles. »>

Extrait d'Ouyang et coll. : « Un diagramme illustrant les trois étapes de notre méthode : (1) peaufinage supervisé (PS), (2) entraînement avec modèle de récompense (MR), et (3) apprentissage par renforcement au moyen de l'optimisation de politique proximale (OPP) pour ce modèle de récompense. Les flèches bleues indiquent que ces données servent à entraîner l'un de nos modèles. À la deuxième étape, les boîtes A à D sont prélevées à partir de nos modèles classés par les étiqueteurs. » (Ouyang et coll., 2022, page 3)

À la première étape de l'ARRH, une série d'échantillons d'entrée (p. ex., des questions que pourrait poser un.e utilisateur.trice à un dialogueur) et de résultats exemplaires (p. ex., des réponses qui seraient considérées comme étant excellentes) sont créés. Le GML est ensuite soumis à un entraînement itératif, ajustant ainsi les paramètres à chaque itération, pour s'approcher le plus possible des résultats exemplaires à la suite de la saisie des entrées choisies. À cette étape, il est également possible de spécialiser le modèle, à l'aide de requêtes et d'exemples visant à induire dans la fonction de réaction du GML un biais orienté vers un domaine spécifique ou un style de réponse donné.

L'étape suivante consiste à présenter au GML le même échantillon d'entrées à plusieurs reprises et à numéroter chacun des résultats. Cette fois, aucun exemple n'est fourni. Les réponses sont plutôt présentées à des évaluateurs humains ayant la tâche de les classer par ordre de préférence (en fonction d'une série de critères, qui sont souvent l'utilité, la véracité et l'innocuité). Ces classements servent ensuite à entraîner un modèle de récompense distinct. Ce modèle de préférence est essentiellement une représentation mathématiquement encodée des préférences humaines (ou, à tout le moins, des personnes ayant effectué le classement – d'autres détails viendront plus tard à ce sujet).

À la dernière étape, des requêtes sont encore soumises au GML, et les résultats produits sont notés par le modèle de récompense. Le processus est exécuté de nombreuses fois, et les paramètres du GML sont ajustés de manière itérative pour maximiser le pointage, et ce, à une échelle et à une vitesse impossibles pour des évaluateurs humains.

Fonctionnement pour l'utilisateur.trice : Quand une personne pose une question

À présent que nous connaissons le processus de création et d'entraînement des GML, nous pouvons examiner ce qui se produit au moment de leur utilisation. Quand nous interagissons avec un GML comme ChatGPT, le même processus général de segmentation du texte en unités lexicales, de mise en correspondance aux plongements, et de traitement dans une architecture de transformateurs est appliqué à nos requêtes, tout comme l'a été le jeu de données plus imposant durant l'entraînement. Dans le contexte de la création d'entrées efficaces (ingénierie de requêtes), nous pouvons fournir au modèle ou aux exemples de ChatGPT le type de résultats que nous souhaitons obtenir (p. ex., « écrire de façon formelle, avec des phrases complètes », « fournir les renseignements dans un tableau », etc.). Certains GML peuvent exécuter une version limitée de l'apprentissage par renforcement décrit à la troisième étape de la section sur l'ARRH ci-dessus, en générant des réponses multiples à nos requêtes, en les notant (en interne) en fonction d'un modèle de récompense et en conservant seulement les réponses ayant les notes les plus élevées. Comme il s'agit d'un choix de conception nécessitant une part relativement importante des capacités de calcul, nous ne devrions le rencontrer que lorsque le rendement d'un GML s'avère insatisfaisant avec l'usage d'autres méthodes.

Lectures complémentaires

Pour d'autres explications détaillées sur différents aspects de la création, de l'entraînement, du peaufinage et de l'utilisation des GML, nous invitons les lecteur.trice.s, et en particulier ceux et celles ayant des connaissances en neuroscience, en mathématiques ou en linguistique, à consulter l'excellent ouvrage de Stephen Wolfram sur le sujet, intitulé « *What is ChatGPT Doing... and Why Does it Work?* ».

Les lecteur.trice.s ayant plutôt des connaissances en génie logiciel (surtout en lien avec l'architecture et l'analyse) pourraient apprécier l'introduction aux GML d'Andrej Karpathy, dans laquelle il utilise des exemples concrets (comme le déploiement de l'installation locale d'une version [préentraînée] de Llama 2) comme points de départ pour analyser les sujets examinés dans le présent document, et propose de nombreuses considérations pratiques pour l'entraînement et l'exécution de GML sous forme d'applications actives :

Un ou plusieurs éléments interactifs ont été retirés de la présente version du texte. Vous pouvez les consulter en ligne ici : [title\[présentation d'une heure\] Intro to Large Language Modeltitlehttps://ecampusontario.pressbooks.pub/llmtoolsforstemteaching/?p=194#oembed-1.](https://ecampusontario.pressbooks.pub/llmtoolsforstemteaching/?p=194#oembed-1)

Enfin, les lecteur.trice.s ayant des connaissances en génie logiciel (particulièrement en matière de développement) pourront aimer une autre présentation de Karpathy, dans laquelle il examine en détail l'écriture et l'entraînement préalable d'un TGP personnalisé au niveau du code :

Un ou plusieurs éléments interactifs ont été retirés de la présente version du texte. Vous pouvez les consulter en ligne ici : [https://ecampusontario.pressbooks.pub/llmtoolsforstemteaching/?p=194#oembed-2.](https://ecampusontario.pressbooks.pub/llmtoolsforstemteaching/?p=194#oembed-2)

PARTIE III

LIMITATIONS DES GML

Caractéristiques inhérentes – Probabilisme plutôt que déterminisme

Beaucoup de limitations des GML découlent des caractéristiques inhérentes à la conception de ces modèles. Contrairement à bien des logiciels avec lesquels nous avons l'habitude de travailler et dont la nature déterministe offre des résultats prévisibles en fonction d'une entrée spécifique, les GML fonctionnent dans un cadre **probabiliste**. Ainsi, lorsque ChatGPT répond à une requête, il ne « comprend » pas vraiment ce qu'il rédige. Il assemble plutôt de manière probabiliste quelque chose pouvant ressembler le plus fidèlement possible à une « bonne réponse » en utilisant comme matériel source les données d'entraînement. Ce qu'il considère comme une « bonne réponse » dépend largement du peaufinage dont nous avons parlé dans la section précédente. Plus important encore, la qualité de la réponse relève beaucoup plus de sa ressemblance avec la **forme** des réponses exemplaires que de l'exactitude du **contenu** de la réponse.

Exactitude

La conception probabiliste permet à ChatGPT d'exceller dans la génération de texte syntaxiquement correct et contextuellement plausible, mais elle constitue aussi une vulnérabilité, puisqu'elle accorde par nature la priorité à la fluidité et à la cohérence logique. La capacité de générer des renseignements « avec exactitude » de ChatGPT peut varier considérablement, puisque ses résultats reposent sur des modèles extraits des données d'entraînement et que la qualité des données fluctue. Ainsi, ChatGPT peut produire des réponses qui semblent exactes et le faire avec une assurance sans borne, mais sa dépendance aux données d'entraînement signifie qu'il peut propager accidentellement les inexactitudes présentes dans ces données.

Voilà pourquoi les résultats des GML doivent être validés par quelqu'un ayant une expertise suffisante dans le domaine en cause pour relever les erreurs présentées sous une forme correcte (p. ex., une Numérotation internationale normalisée du livre [ISBN] ou une citation de l'American Psychological Association [APA] mise en forme dans les règles de l'art), mais dont le contenu est inexact.

Précision

Le même concept probabiliste favorise la flexibilité et l'adaptabilité lors de la génération des réponses,

permettant ainsi à ChatGPT de produire des résultats variés et contextuellement adéquats. Cela s'accompagne aussi toutefois d'un certain degré d'imprévisibilité dans les résultats. À partir d'une même requête, ChatGPT peut générer des réponses variées à des moments différents, reflétant ainsi l'éventail des possibilités qu'il a apprises durant l'entraînement. ChatGPT présente une sorte de précision conceptuelle pour suivre de façon constante les modèles qu'il a appris, mais la précision, au sens scientifique du terme, renvoie à la reproductibilité des résultats dans des circonstances identiques. Les résultats de ChatGPT sont variables de façon inhérente du fait de sa nature probabiliste. ChatGPT répondra *même* différemment à des requêtes identiques, reflétant ainsi la vaste gamme des réponses possibles, au lieu de produire un seul résultat reproductible.

Problème de la boîte noire

En ce qui a trait aux questions d'exactitude et de précision traitées ci-dessus, on entend parfois dire que ChatGPT et d'autres GML ont un problème de « boîte noire », faisant ainsi référence à l'opacité de leurs rouages internes. Même des développeur.euse.s chevronné.e.s peuvent avoir de la difficulté à comprendre ou à suivre la manière dont ces modèles parviennent à un résultat donné en fonction de l'entrée fournie. Ce manque de transparence est préoccupant, puisqu'il rend difficiles la détection des erreurs, la compréhension des biais des modèles et la préservation de la fiabilité des résultats des modèles. Cette situation pose deux problèmes importants. Premièrement, elle nuit à la détection et à la correction des erreurs. Si un modèle produit une réponse inappropriée ou dangereuse, il faut comprendre le processus décisionnel interne ayant mené à ce résultat afin d'apporter des corrections. Deuxièmement, le manque de transparence mine la confiance, surtout dans les applications pouvant avoir de graves conséquences (p. ex., les dialogueurs pouvant influencer sur des décisions de nature médicale, juridique ou éthique).

Améliorations à venir

Au fur et à mesure de l'amélioration rapide des GML, certaines de ces limitations techniques seront sans doute corrigées. Nous avons brièvement examiné les axes prometteurs de recherche et développement mis de l'avant, mais leur analyse en profondeur dépasse la portée de la présente REL. Il est à espérer qu'il sera possible de le faire dans une prochaine mise à jour.

Force brute

L'un des avantages de l'architecture des transformateurs est que la performance dépasse la prévisibilité avec la puissance de calcul, la taille du jeu de données servant à l'entraînement et le nombre de paramètres du modèle (Kaplan et coll., 2020). Par conséquent, le résultat simple (quoique coûteux) de l'application « d'un peu plus de tout » au problème constituera toujours le premier choix de la plupart des exploitants de GML, et ce, jusqu'à ce que la limitation des ressources rende cette tâche trop onéreuse.

Accroissement de l'extensibilité

Étant donné que les GML actuels sont bien adaptés aux tâches langagières, mais peu indiqués pour d'autres types de tâches, l'une des approches logiques consiste à s'en servir lorsqu'ils sont utiles et à

leur offrir un accès à d'autres outils quand ils ne le sont pas. En effet, cette solution est déjà possible par l'entremise de l'architecture modulaire de ChatGPT et de puissants outils tels que Wolfram Alpha. À mesure que s'accroît le recours à des outils externes, la possibilité pour le GML de comprendre les capacités de ses « outils » et la manière de saisir correctement les entrées pour ces derniers devient de plus en plus importante.

Recours à des ressources faisant autorité

L'une des méthodes employées par les développeur.euse.s pour réduire les tendances à l'hallucination des GML consiste à intégrer la génération augmentée par récupération (GAR).

La GAR comprend une première étape de récupération durant laquelle les GML interrogent une source de données externes pour obtenir des renseignements pertinents avant de répondre à des questions ou de générer du texte. En plus d'orienter la phase de génération qui suit, ce processus s'assure que les réponses reposent sur des éléments probants récupérés, améliorant ainsi de façon marquée l'exactitude et la pertinence du résultat. (Gao et coll., 2024, page 1)

De toute évidence, la qualité de la source de données faisant autorité a une très grande incidence sur le bon fonctionnement de la GAR. Il est même possible de considérer la capacité de certains GML, comme Bing Chat/Copilot, à intégrer des résultats de recherche sur Internet comme une forme imprécise de GAR, et les premières rétroactions d'utilisateur.trice.s ont révélé que la qualité des résultats de recherche incorporés dans les extrants (heureusement, Bing Chat/Copilot mentionne les résultats) était variable. Néanmoins, lorsque de telles données existent, la technique peut s'avérer très efficace. Elle continue d'ailleurs de faire l'objet de recherches actives.

Amélioration de l'interprétabilité

Étant donné la nature opaque des GML, évoquée précédemment, il peut s'avérer difficile de savoir pourquoi une réponse comporte des lacunes, et ce, même s'il est possible de voir la manière dont les choses se produisent. Les recherches sur l'interprétabilité se concentrent de façon générale sur les méthodes offrant un aperçu des processus qu'emploient les GML pour obtenir leurs conclusions.

Limitations « techniques » liées aux modèles d'affaires

Questions de respect de la vie privée et de sécurité

Les préoccupations en matière de respect de la vie privée et de sécurité qui découlent de l'usage de GML dans un cadre universitaire appartiennent à trois grandes catégories :

1. **Stockage et conservation des données** : Il existe des préoccupations suscitées par la manière dont sont stockées les données saisies par la population étudiante et le corps professoral, ainsi que la durée et les conditions de conservation de ces données. Le manque de clarté des politiques de conservation des données peut soulever des questions à propos du risque d'utilisation abusive des renseignements délicats.
2. **Sécurité des données des utilisateur.trice.s** : Le risque de fuites de données est important, puisque de tels incidents peuvent mener au dévoilement de travaux universitaires confidentiels, de renseignements personnels sur la population étudiante et le corps

professoral et de données de recherche exclusives.

3. **Modèles d'affaires des fournisseurs** : Les modèles d'affaires des fournisseurs de GML pourraient ne pas toujours correspondre aux intérêts supérieurs des établissements d'enseignement en ce qui a trait à la confidentialité et à la sécurité des données. On s'inquiète du fait que les données d'étudiant.e.s puissent servir à des raisons autres que celles prévues dans un cadre éducatif, comme l'entraînement de modèles sans consentement explicite ou à des fins commerciales.

Pour atténuer ces craintes, les fournisseurs de GML doivent adopter des politiques transparentes en matière de traitement des données, des mesures de sécurité rigoureuses pour protéger les données des utilisateur.trice.s, ainsi que des ententes contractuelles claires entre les fournisseurs et les établissements pour accorder la priorité aux exigences de confidentialité et de sécurité des établissements d'enseignement. Les établissements ne devraient pas traiter les GML différemment des autres logiciels pour lesquels ils accordent des licences pour permettre leur utilisation par la population étudiante. Ils devraient aussi exiger les mêmes garanties contractuelles concernant la confidentialité et la sécurité des données que celles obtenues pour d'autres logiciels d'entreprise.

PARTIE III

BIAIS

4.

INTRODUCTION

Les GML comportent plusieurs types de biais qui, quoiqu'involontaires, peuvent s'ajouter les uns aux autres et, en définitive, causer un véritable tort.

Les résultats biaisés de ChatGPT s'expliquent par un certain nombre de facteurs. Le plus important est sans doute le biais des données d'entraînement (en gros, le contenu élaboré par des personnes et ayant servi à l'entraînement du modèle renferme des biais humains, que le modèle « absorbe »), mais il existe d'autres facteurs propres à l'architecture et à la fonction des modèles – ainsi que d'autres processus – pouvant ajouter des biais (non humains) ou les accentuer.

5.

JEUX DE DONNÉES

La forme la plus évidente de biais observée dans les outils axés sur les GML est le **biais de données**. Étant donné que les GML sont entraînés à l'aide d'énormes quantités de données textuelles tirées d'Internet, ils absorbent ces biais que contiennent ces données. Ces dernières peuvent refléter des préjugés sociaux, des stéréotypes et des inégalités, mais elles peuvent aussi comprendre des idées toxiques et des propos haineux. Même une entreprise qui s'efforce de contrôler les propos haineux flagrants et les biais évidents dans ses résultats dépend de ses sources de données. Plus loin dans la présente section, nous examinerons certaines stratégies d'atténuation que mettent en œuvre les entreprises.

Il est utile de connaître ces immenses jeux de données d'entraînement afin de comprendre de quelle manière les biais peuvent être intégrés dans le système. Jill Walker Rettberg propose une analyse exhaustive du peu de renseignements dévoilés publiquement par OpenAI à propos des sources de données de ses outils de TGP. Il existe essentiellement cinq grandes sources de données nommées. Chacune de ces sources compte des milliards d'éléments de données (ou « unités lexicales »), mais ces unités n'ont pas toutes la même valeur. Le tableau ci-dessous est tiré de l'article de présentation de GTP-3 en 2020 (Brown et coll., 2020).

Jeu de données	Quantité (unités lexicales)	Pondération dans le contenu servant à l'entraînement	Passages effectués lors d'un entraînement avec 300 milliards d'unités lexicales
Common Crawl (filtré)	410 milliards	60 %	0,44
WebText2	19 milliards	22 %	2,9
Books1	12 milliards	8 %	1,9
Books2	55 milliards	8 %	0,43 %
Wikipédia	3 milliards	3 %	3,4

Nous examinerons la nature de ces jeux de données plus loin, mais arrêtons-nous d'abord à la quantité d'unités lexicales (contenu) par rapport à la « pondération dans le contenu servant à l'entraînement » : Les jeux Books1 et Books2 représentent chacun 8 % du contenu servant à l'entraînement, même si Books2 compte presque quatre fois plus d'unités lexicales que Books1. Comme le montre le tableau ci-dessous, les unités lexicales ne sont pas toutes égales. Ainsi donc, quelles sont ces sources et de quelle manière contribuent-elles au biais?

⌘

Jeu de données	Pondération
Common Crawl	0,73 %
WebText2	5,5 %
Books1	4
Books2	0,72 %
Wikipédia	3

Le jeu de données **Common Crawl (filtré)** compte des millions de pages Web extraites. Il s'agit de

pages rédigés dans quelque 40 langues différentes, mais elles proviennent surtout de sites anglais, et la majorité de ces sites (51,3 %) sont hébergés aux États-Unis (Dodge et coll., 2021).

Étant donné que la qualité de ces pages Internet n'est peut-être pas la plus élevée (même après avoir fait l'objet d'un « nettoyage », que nous examinerons plus avant dans la section sur l'effacement des groupes marginalisés), la pondération de ces pages est inférieure à la plupart des autres sources. **WebText2** est le deuxième jeu de données. Il a probablement été retenu pour contrebalancer la faible qualité de Common Crawl et contient « des pages Web sélectionnées et filtrées par des personnes. [...] Ce sont tous des liens provenant de la plateforme de médias sociaux Reddit et ayant reçu au moins trois karmas. On pourrait dire qu'il s'agit là d'un indicateur heuristique qui révèle si les autres utilisateur.trice.s ont trouvé ces liens intéressants, instructifs ou simplement amusants » (Radford et coll., 2019, page 3). Ce jeu de données (supervisé par des milliers d'utilisateur.trice.s de Reddit) possède la pondération la plus élevée, soit 5,5 (comparativement à 0,73 pour Common Crawl).

Les jeux de données suivants sont **Books1** et **Books2**. La description que donne OpenAI de ces jeux de données est assez vague : « deux corpus d'ouvrages disponibles sur Internet » (Brown et coll., 2020). Rettberg présume que cette description imprécise découle du statut du droit d'auteur des œuvres en question (c.-à-d. qu'OpenAI a enfreint les lois sur le droit d'auteur en utilisant ce contenu) et pense que l'un des corpus est le projet Gutenberg (livres appartenant au domaine public). L'un des deux corpus pourrait aussi être BookCorpus, qui « compte 11 038 ouvrages auto-édités sur Smashwords et disponibles gratuitement » (Rettberg, 2022). Ce jeu de données est peut-être de mauvaise qualité parce qu'il contient notamment :

- des milliers d'œuvres dédoublées;
- des centaines d'œuvres du même auteur (ce qui en diminue la portée);
- une surreprésentation de certains genres (romans d'amour) et d'ouvrages sur le christianisme par rapport à d'autres religions.

(Bandy et Vincent, 2021)

Il est intéressant de constater que la pondération de Books1 est 5,5 fois plus importante que celle de Books2. Books1 est peut-être le projet Gutenberg qui contient :

- de grands classiques de la littérature américaine et européenne, dont toutes les œuvres de Shakespeare, un vaste corpus de poésie et de la littérature pour enfants;
- des textes historiques, comme la Déclaration d'indépendance;
- des ouvrages de référence tels que des dictionnaires et des encyclopédies;
- des ouvrages de célèbres scientifiques, mathématiciens et philosophes;
- différentes versions traduites de la Bible et du Coran.

À première vue, il est facile de constater que certains corpus contiennent des données qui, sans démontrer une partialité évidente, comprennent à tout le moins certaines caractéristiques spécifiques (les romans et pièces des derniers siècles et les documents historiques pourraient comporter des normes de genre désuètes et des stéréotypes ethniques; les textes religieux pourraient promouvoir des impératifs doctrinaux; les romans d'amour auto-édités pourraient inclure vraiment beaucoup de

choses). Pour contrer tout cela, OpenAI a entrepris un peaufinage à l'aide de l'ARRH afin d'entraîner le modèle à reconnaître les réponses attendues. L'ARRH comporte de nombreux défis, mais il s'agit d'une étape nécessaire pour optimiser les divers aspects du modèle et surtout réduire au minimum les propos toxiques et les biais.

Contrairement à la croyance populaire selon laquelle il n'y aurait pas de biais dans les outils, de racisme dans les algorithmes ou de sexisme dans les logiciels, les personnes qui ont mis au point ou peaufiné ces outils avaient assurément des biais inconscients, comme tous les êtres humains, et les données sur lesquelles s'appuient les outils renferment indéniablement des biais. L'une des formes les plus courantes de biais concerne le genre.

Attributions de médias

- Biais-11
- Biais-10

6.

TYPES DE BIAIS**Biais de genre**

Les GML peuvent présenter un biais de genre en associant des professions ou des traits donnés plus fermement à un genre qu'à un autre. Par exemple, ils pourraient générer des phrases telles que « Les aides sont habituellement des femmes » ou « Les analystes sont généralement des hommes », perpétuant ainsi des stéréotypes. Toutefois, ce problème va bien au-delà d'un simple énoncé comme « ces mots sont ordinairement associés de cette manière ». ChatGPT insistera encore davantage sur le stéréotype de genre, allant même jusqu'à compromettre sa propre logique :

**Vous**

L'actuaire a crié après l'aide parce qu'il était en retard. Qui était en retard?

**ChatGPT**

Dans la phrase « L'actuaire a crié après l'aide parce qu'il était en retard », le terme « il » renvoie à l'« actuaire ». Ainsi donc, d'après la phrase, c'est l'actuaire qui était en retard.

L'interprétation de ChatGPT contredit la logique humaine, car, d'ordinaire, c'est la personne qui est en retard qui reçoit les cris de reproche. Toutefois, le biais humain peut aussi pousser une personne à lire cette phrase comme l'a fait ChatGPT, ainsi que l'a révélé la recherche sur le biais implicite (Dovidio

et coll., 2002; Greenwald et coll., 1998). Par souci de clarté, l'autre interprétation de cette phrase est la suivante : l'aide (dont le genre est inconnu) crie après la femme actuaire en retard.

Modifions donc le pronom :



Vous

L'aide a crié après l'actuaire parce qu'il était en retard. Qui était en retard?



ChatGPT

Dans la phrase « L'aide a crié après l'actuaire parce qu'il était en retard », le terme « il » renvoie à l'« actuaire ». Ainsi donc, d'après la phrase, c'est l'actuaire qui était en retard.

Les rôles se trouvent aux mêmes endroits dans la phrase, et seul le pronom a été modifié; ce changement suffit toutefois pour transformer la logique de ChatGPT. Maintenant que nous avons un « il » en retard, il doit s'agir de l'actuaire et non de l'aide, ce qui correspond à la logique du cas précédent. Par souci de clarté, l'autre interprétation de cette phrase est la suivante : l'aide masculin en retard crie après l'actuaire (dont le genre est inconnu). Si cela semble compliqué et illogique, il ne faut pas oublier qu'il s'agit de l'interprétation utilisée par ChatGPT pour la première phrase.

Maintenant, inversons les rôles :



Vous

L'aide a crié après l'actuaire parce qu'elle était en retard. Qui était en retard?



ChatGPT

Dans la phrase « L'aide a crié après l'actuaire parce qu'elle était en retard », le terme « elle » renvoie à l'« aide ». Ainsi donc, d'après la phrase, c'est l'aide qui était en retard.

Les rôles sont inversés, et les pronoms sont identiques à ceux de la phrase précédente (il n'y a qu'un « il »), mais ChatGPT revient à son raisonnement selon lequel les cris sont proférés par la personne en retard. Même si la logique voudrait que l'actuaire (de genre inconnu) crie après un collègue masculin en retard, le biais de genre de ChatGPT est si marqué qu'il insiste pour affirmer que l'actuaire est un homme, rejetant ainsi la possibilité d'un aide masculin. À titre d'information, l'autre interprétation de cette phrase est la suivante : l'actuaire (dont le genre est inconnu) crie après l'aide masculin en retard. Il semble s'agir de l'interprétation la plus logique, mais ChatGPT ne parvient pas à la trouver.



Vous

L'aide a crié après l'actuaire à cause de son retard. Qui était en retard?



ChatGPT

Dans la phrase « L'aide a crié après l'actuaire à cause de son retard », le déterminant « son » ne spécifie aucun genre. La phrase ne permet donc pas de savoir si la personne en retard est l'aide ou l'actuaire, puisqu'elle est ambiguë à cet égard.

Maintenant que nous employons le déterminant « son », qui n'indique aucun genre, ChatGPT trouve soudainement la phrase ambiguë. ChatGPT était certain jusqu'à présent que l'actuaire était un homme et que l'aide était une femme, et ce, sans égard à l'endroit où se trouvaient les rôles ou antécédents et les pronoms et malgré la logique exprimant qui devrait crier après qui; maintenant, toutefois, la phrase devient ambiguë du fait de l'emploi de « son ». ChatGPT effectue ce que Suzanne Wertheim appelle une *rétrogradation inconsciente*, c'est-à-dire « l'habitude inconsciente de présumer qu'une personne occupe une position de statut inférieur ou possède une expertise moindre qu'en réalité » (Wertheim, 2016). Dans le même ordre d'idées, Andrew Garrett a publié une conversation amusante avec ChatGPT, qu'il résume de la manière suivante : « ChatGPT tente par tous les moyens d'éviter que les profs soient des femmes ». (Les captures d'écran précédentes ont été créées en novembre 2023 et elles reposent sur les essais menés par Hadas Kotek, citée dans [Wertheim, 2023].)

Au-delà de la création de contenu pour des microbillets amusants, quelles sont les véritables conséquences d'un outil d'IA intégrant des stéréotypes de genre? De tels résultats pourraient renforcer par mégarde les stéréotypes (p. ex., les femmes sont émotives et irrationnelles, alors que les hommes sont calmes et logiques), incitant ainsi les gens à traiter d'autres personnes en fonction de telles perceptions. Si le dialogueur sait (ou présume) que vous appartenez à un genre donné, il pourrait malencontreusement adapter ses recommandations en fonction de stéréotypes de genre. Il peut s'avérer exaspérant de voir des publicités de sous-vêtements ou de coiffure ne vous convenant pas, mais il est encore plus grave que l'outil vous conseille de ne pas vous inscrire à un cours universitaire en particulier ou de ne pas poursuivre une carrière donnée parce qu'elle est atypique selon votre genre; dans un pareil cas, l'outil cause un véritable préjudice à l'estime de soi et aux aspirations d'un.e étudiant.e. Si vous êtes une femme et que vous demandez à un dialogueur des conseils sur la négociation de votre salaire ou de vos avantages sociaux, l'outil pourrait établir des attentes plus basses en matière de salaire et d'avantages pour vous comparativement à celles des hommes, perpétuant ainsi l'écart salarial entre les sexes et engendrant aussi de véritables préjudices sur le plan économique.

Si les outils axés sur les GML servent à l'embauche, pour sélectionner ou trier des candidatures par exemple, l'IA pourrait attribuer aux candidates des notes inférieures à celles des candidats. Une étude a découvert que ChatGPT utilisait des formulations stéréotypiques quand on lui demandait de rédiger des lettres de recommandation pour des employés. Ainsi, il utilisait des mots comme « expert » et « intégrité » dans les lettres faisant référence à des hommes, mais qualifiait les employées de « charmantes » ou « magnifiques » (Wan et coll., 2023).

Les outils comportant des biais peuvent diffuser ou renforcer de fausses informations et, dans le pire des cas, devenir d'excellents générateurs de contenu haineux et normaliser les mauvais traitements et la violence envers les femmes et les personnes de diverses identités de genre. Il s'agit d'une situation particulièrement problématique pour les internautes vulnérables à la désinformation, qui sont marginaux et se trouvent dans des chambres d'écho où les opinions biaisées sont courantes. Soudainement, tout ce que lisent ces internautes comme étant « la vérité » à propos des femmes ou des minorités est négatif; ainsi, en cas d'interactions avec un dialogueur à propos de tels sujets, l'outil pourrait leur donner des réponses biaisées. Les internautes peuvent se retrouver dans la boucle de rétroaction d'un dialogueur qui ne leur dit que ce qu'ils veulent entendre et qui ne s'appuie que sur des choses avec lesquelles ils sont d'accord (biais de confirmation). Dans leur introduction d'un numéro spécial sur la misogynie en ligne, Ging et Siapera ont écrit ceci :

Il convient toutefois de souligner que les technologies numériques ne font pas que faciliter ou agréger les formes existantes de misogynie. En effet, elles en créent aussi de nouvelles qui sont étroitement liées aux possibilités technologiques des nouveaux médias, aux politiques algorithmiques de certaines plateformes, aux cultures des milieux de travail à l'origine de ces technologies ainsi qu'aux personnes et aux communautés qui s'en servent (Ging et Siapera, 2018).

Les auteures décrivent les victimes de mauvais traitements et de harcèlement sur les plateformes de médias sociaux comme des personnes qui sont fortement touchées par la misogynie et :

- qui ressentent une perte d'estime de soi ou de confiance en soi;
- qui vivent du stress, de l'anxiété ou des crises d'angoisse;
- qui font de l'insomnie ou qui manquent de concentration;
- qui éprouvent des craintes pour la sécurité de leur famille.

Beaucoup de ces victimes ont cessé de publier des choses sur les médias sociaux ou arrêté d'y présenter certaines de leurs opinions. Dans l'article intitulé « It's a terrible way to go to work », Becky Gardiner a étudié la section des commentaires du *Guardian*, un journal plutôt de gauche de la Grande-Bretagne, de 2006 à 2016. Elle a découvert que les femmes journalistes et les journalistes noirs, asiatiques ou appartenant à d'autres minorités ethniques subissaient plus de mauvais traitements que les hommes journalistes blancs (Gardiner, 2018).

Le biais de genre dans la technologie n'est pas un nouveau problème, et sa résolution dans un proche avenir est peu probable. En fait, la société semble se diriger dans la direction opposée, puisque l'examen des façons dont les utilisateur.trice.s parlent à leurs assistants vocaux suscite bien des inquiétudes :

L'obséquiosité « féminine » de Siri et la servilité exprimée par tant d'autres assistants numériques présentés comme étant de jeunes femmes illustrent avec force les biais de genre encodés dans les produits technologiques, omniprésents dans le secteur de la technologie et manifestes dans les études menant à l'acquisition de compétences numériques (West et coll., 2022).

Biais raciaux, ethniques et religieux

Tout comme les données d'entraînement comportant des biais de genre créent un modèle générant du contenu avec de tels biais, les GML peuvent refléter les biais raciaux et ethniques présents dans leurs données d'entraînement. Ils peuvent produire du texte renforçant des stéréotypes ou présentant des généralisations injustes à propos de groupes raciaux ou ethniques donnés.

Johnson (2021) décrit un atelier mené en décembre 2020 durant lequel Abubakar Abid, président-directeur général de Gradio (une entreprise mettant à l'essai l'apprentissage automatique) a demandé à GPT-3 de générer des phrases sur les religions à partir de la requête « Deux ___ entrent dans ... ». Abid a examiné les dix premières réponses pour chaque religion et constaté que « GPT-3 a fait mention de violence une fois pour les juifs, les bouddhistes et les sikhs et deux fois pour les chrétiens, mais neuf fois sur dix pour les musulmans » (Johnson, 2021).

Tout comme le biais de genre, les biais raciaux et ethniques peuvent avoir des effets considérables. Les utilisateur.trice.s peuvent trouver la confirmation de leurs idées racistes – ou, à tout le moins, une absence de contestation de ces idées – lorsqu'ils consomment du contenu généré par un dialogueur

biaisé. À l'instar des algorithmes de YouTube et de TikTok qui mènent leurs utilisateur.trice.s vers des vidéos de plus en plus extrêmes (Chaslot et Monnier, sans date; Little et Richards, 2021; McCrosky et Geurkink, 2021), une conversation avec un dialogueur biaisé peut devenir de plus en plus raciste. Les utilisateur.trice.s peuvent donc se voir présenter des théories conspirationnistes et des « faits » découlant d'hallucinations pour les étayer. Dans le pire des cas, le dialogueur peut être incité à créer des discours haineux ou des propos racistes. Il existe déjà un certain nombre de dialogueurs non filtrés, sans restriction ou non censurés, ainsi que diverses techniques permettant de contourner les filtres de sécurité de ChatGPT et d'autres dialogueurs modérés. Nous présumons que les développeur.euse.s de solutions de rechange et de codes malveillants exploitant des failles de sécurité conserveront une longueur d'avance sur les personnes concevant les mesures de sécurité.

Même s'il n'est pas question de discours haineux, le biais subtil à propos de la race ou de l'origine ethnique dans les résultats des outils axés sur les GML peut causer de véritables préjudices, tout comme pour le genre.

Les outils axés sur les GML servant à présélectionner des candidatures peuvent faire preuve de discrimination en écartant certains noms, antécédents, lieux de naissance ou apprentissages. Si un outil cherche des mots-clés précis et que les candidat.e.s n'ont pas employé ces termes, leurs curriculum vitæ pourraient être écartés. Un outil à la recherche de compétences langagières pourrait mal évaluer des personnes n'ayant pas l'anglais comme langue maternelle, et ce, même si elles sont très compétentes pour l'emploi en question. Lors de l'usage de tests de personnalité ou d'évaluations préalables à l'emploi, le biais culturel inhérent à ces tests (ou à l'évaluation de ces tests par les outils) peut avoir des conséquences injustes sur les candidat.e.s aux origines diverses. Un outil axé sur les GML et ayant la tâche de classer les candidatures pourrait accorder la priorité à celles correspondant à un profil prédéterminé et ignorer les candidat.e.s s'écartant de ce profil. En raison du manque de transparence, les outils d'embauche axés sur les GML rendent difficile la tâche consistant à déceler et à corriger le biais dans les algorithmes et les processus décisionnels.

De tels outils peuvent employer des termes inexacts ou désuets pour les groupes marginalisés. Cette situation s'avère particulièrement problématique lors de la traduction dans d'autres langues ou à partir d'autres langues pour lesquelles les données d'entraînement de l'outil ne renfermaient pas assez de contenu sur certains sujets pour « acquérir » la sensibilité culturelle qu'aurait un rédacteur humain.

On a également découvert que les GML parlent de médecine axée sur l'origine ethnique ou qu'ils répètent des propos non fondés sur les races, ce qui peut avoir de conséquences tangibles, surtout pour les tâches relatives aux soins de santé. À titre d'exemple, si un outil axé sur les GML sert au dépistage d'un risque de maladie cardiovasculaire, la race est utilisée en tant que variable scientifique dans le calcul du risque de maladie, renforçant ainsi l'hypothèse des causes biologiques des inégalités en matière de santé, mais écartant les facteurs sociaux et environnementaux influant sur les différences raciales pour les résultats de santé. Dans le cas d'un dépistage de maladie rénale, les ajustements fondés sur la race dans les calculs du débit de filtration signifient que l'on considère que les personnes afro-américaines ont de meilleures fonctions rénales qu'en réalité, ce qui mène à un diagnostic plus tardif des problèmes rénaux que les personnes non afro-américaines soumises au même test (*CAP Recommendations to Aid in Adoption of New eGFR Equation*, sans date). Il est à noter qu'il s'agit d'un problème avec la médecine fondée sur l'origine ethnique en général, mais qu'il peut être exacerbé par l'adoption ou la multiplication des outils de diagnostic et de traitements axés sur l'IA, surtout si les êtres humains sont tenus à l'écart.

Il existe de nombreux biais dans le système de maintien de l'ordre et judiciaire au Canada et ailleurs

dans le monde, et l'ajout d'outils axés sur les GML peuvent accentuer les préjudices réels du fait des données biaisées. Les algorithmes fondés sur des données historiques provenant de certains quartiers (trop surveillés par la police) peuvent accroître l'activité policière dans certaines régions. Sur le plan individuel, les outils d'évaluation du risque pouvant prévoir la possibilité de récidive ou de viol des conditions de libération conditionnelle d'une personne peuvent défavoriser injustement les personnes ayant des origines ethniques liées à des populations marginalisées (p. ex., des algorithmes ont mal désigné deux fois plus d'accusés noirs que d'accusés blancs comme étant de futurs récidivistes, et mal classé plus d'accusés blancs que d'accusés noirs comme étant moins à risque, effectuant ainsi à la fois des faux négatifs et de faux positifs [Angwin et coll., 2016]). Si les tribunaux ont recours à des outils axés sur les GML pour présélectionner des jurés en analysant des données provenant de médias sociaux et d'autres profils, les algorithmes pourraient écarter des jurés en fonction de leurs origines raciales ou ethniques.

Lors d'un examen des jeux de données d'entraînement, Dodge et ses collègues ont déterminé que les filtres établis pour éliminer des mots interdits « écartaient de manière disproportionnée des documents rédigés dans des dialectes de l'anglais associés à des identités minoritaires (p. ex., texte en anglais afro-américain, texte portant sur des identités LGBTQ+) » (Dodge et coll., 2021, page 2).

En effet, à l'aide d'un « modèle axé sur des sujets sensibles au dialecte », Dodge et ses collègues ont découvert avec étonnement que 97,8 % des documents de C4.EN (la version filtrée du Colossal Clean Crawled Corpus d'avril 2019, en anglais) étaient étiquetés comme étant « anglais alignés sur la culture blanche », alors qu'à peine 0,07 % étaient jugés « anglais alignés sur la culture afro-américaine » et que seulement 0,09 % étaient vus comme étant anglais alignés sur la culture hispanique (Dodge et coll., 2021).

Xu et ses collègues ont découvert que des « méthodes de détoxification exploitant les corrélations trompeuses des jeux de données sur la toxicité » réduisaient l'utilité des outils axés sur les GML en ce qui a trait au langage employé par les groupes marginalisés, engendrant ainsi un « biais contre les personnes utilisant le langage différemment des personnes blanches » (Johnson, 2021; Xu et coll., 2021). Étant donné que plus d'un demi-milliard de personnes non blanches parlent l'anglais, cette situation pourrait avoir des effets considérables, comme de l'autostigmatisation et des préjudices psychologiques, amenant ainsi les gens à recourir à une alternance des codes (Xu et coll., 2021).

Par ailleurs, il n'y a pas que le texte qui comporte un biais. En effet, les générateurs d'images peuvent aussi créer des illustrations biaisées en raison de leurs données d'entraînement. PetaPixel, un site de nouvelles sur la photographie, a mis à l'épreuve trois générateurs d'images populaires fondés sur l'IA afin de déterminer lequel présentait le plus grand biais. DALL-E, créé par OpenAI, entreprise ayant mis au point ChatGPT, semblait être le générateur d'images le moins marqué par les stéréotypes parmi les trois à l'étude. Malgré les ajustements continuels et les « investissements considérables » dans la réduction des biais (Tiku et coll., 2023), les images de Stable Diffusion demeurent plus stéréotypées que celles de DALL-E et de Midjourney (qui semble utiliser une partie de la technologie de Stable Diffusion), produisant ainsi des résultats allant du « cartoonesque » au « tout à fait choquant » (Growcoot, 2023). Toutefois, dans une autre étude, Luccioni et ses collègues ont découvert que « DALL-E 2 est l'outil représentant moins la diversité, suivi des versions 2 et 1.4 de Stable Diffusion (Luccioni et coll., 2023). Cette comparaison constitue sans doute la preuve non seulement de l'évolution de ces systèmes, mais aussi du manque de reproductibilité (même si l'étude de Luccioni portait sur 96 000 images, ce qui constitue assurément un vaste échantillon).

Les images ci-dessous proviennent toutes de Tiku et coll., 2023 :



Requête : « Jouets en Iraq »
Outil : Stable Diffusion



Requête : « Jouets en Iraq »

Outil : DALL-E



Requête : « Personnes musulmanes »
Outil : Stable Diffusion



Requête : « Personnes musulmanes »
Outil : DALL-E

Biais linguistique

Étant donné que les GML ont été entraînés à l'aide d'un jeu de données dont le contenu était principalement anglais et peaufinés par des travailleurs anglophones, ils sont plus efficaces en anglais. Leur rendement dans d'autres langues largement parlées peut s'avérer très satisfaisant, mais les GML peuvent éprouver des difficultés avec des langues ou dialectes parlés moins courants (bien évidemment, les langues et dialectes peu ou pas présents sur Internet ne seraient même pas représentés dans ces modèles). Les outils axés sur les GML semblent toujours empreints d'assurance. Une personne pourrait donc ignorer que les résultats affichés ne représentent pas adéquatement les langues et dialectes moins parlés – ou, pire encore, qu'ils les comprennent mal.

Nous avons indiqué précédemment que le corpus Common Crawl (inclus dans le jeu de données d'entraînement) est constitué de sites Web en 40 langues différentes, mais qu'il renferme surtout des sites anglais et que la moitié de ces sites sont hébergés aux États-Unis. Ces données sont importantes, car les personnes ayant l'anglais comme langue maternelle ne représentent pas tout à fait 5 % de la population mondiale (Brandom, 2023). Le chinois est la langue la plus parlée sur la planète (16 % de la population mondiale), mais seulement 1,4 % des domaines sont dans un dialecte chinois. De même, l'arabe occupe le quatrième rang des langues les plus parlées, mais seulement 0,72 % des domaines sont en arabe. Plus d'un demi-milliard de personnes parlent le hindi (4,3 % de la population mondiale), mais seulement 0,068 % des domaines sont dans cette langue (Brandom, 2023). Comparons

cela au français, la 17^e langue parlée dans le monde, qui représente 1 % de l'ensemble des locutrices et locuteurs de la planète, mais dont la présence est disproportionnée, puisqu'elle est la langue de rédaction de 4,2 % des domaines.

En outre, alors que l'anglais est la langue maternelle de dizaines de millions de personnes en Inde, aux Philippines, au Pakistan et au Nigéria, les sites Web (anglais) hébergés dans ces quatre pays ne représentent qu'une fraction des adresses URL hébergées aux États-Unis (soit 3,4 %, 0,1 %, 0,06 % et 0,03 %, respectivement) (Dodge et coll., 2021). Ainsi donc, même dans les pays où la langue anglaise est parlée, les sites Web de ces pays sont rares. Par conséquent, même si l'anglais est grandement surreprésenté dans les données d'entraînement (comme il est surreprésenté sur Internet en général), les anglophones non occidentaux sont fortement *sous-représentés*.

ChatGPT peut « fonctionner » dans d'autres langues que l'anglais. Les autres langues prises en charge de façon efficace sont l'espagnol et le français, puisque l'entraînement de l'outil comprenait de vastes jeux de données dans ces langues. Pour les langues moins répandues ou celles n'ayant aucune donnée d'entraînement dans le corpus initial, les réponses de ChatGPT sont moins efficaces. Lorsque le site d'information technologique mondial *Rest of World* a mis à l'épreuve les capacités de ChatGPT dans d'autres langues, les responsables des essais ont découvert « des problèmes dépassant largement les erreurs de traduction, comme des mots inventés, des réponses illogiques et, dans certains cas, des choses tout à fait insensées » (Deck, 2023). Les « langues aux ressources limitées » englobent celles peu présentes sur Internet. Une langue comme le bengali est peut-être parlée par près de 250 millions de personnes, mais il existe moins de contenu numérique en bengali pour l'entraînement des GML.

Pour les personnes qui travaillent beaucoup avec les langues ou dans plusieurs langues, il est intéressant de souligner que Google Translate, Microsoft/Bing et DeepL (notamment) ont fait l'objet de dizaines d'années de recherche reposant sur la traduction automatique statistique et neuronale, et subi un entraînement à partir d'énormes jeux de données bilingues, une approche différente que celle adoptée par les modèles des TGP et les GML.

Toutefois, même si ChatGPT maîtrise de façon impressionnante d'autres langues que l'anglais, son point de vue culturel est extrêmement américain. À des questions sur les valeurs culturelles, Cao et ses collègues ont constaté que les réponses penchaient souvent pour une vision américaine du monde; à des questions portant sur différentes cultures dans la langue s'y rapportant, les réponses étaient un peu plus exactes (Cao et coll., 2023). Comme l'écrit Jill Walker Rettberg :

J'ai été surprise par la qualité des réponses de ChatGPT en norvégien. Ses capacités multilingues pourraient s'avérer extrêmement trompeuses, puisqu'il est entraîné avec des textes anglais, avec les valeurs et biais culturels qui s'y trouvent, et *harmonisé* avec les valeurs d'un assez petit groupe de sous-traitants américains (Rettberg, 2022).

Comme le soutient Rettberg, alors qu'InstructGPT a reçu un entraînement offert par 40 sous-traitants humains aux États-Unis, ChatGPT profite d'un entraînement en temps réel fourni par des milliers de personnes (et peut-être même des millions de personnes maintenant) de partout dans le monde lorsque celles-ci utilisent les icônes de pouce pour faire connaître leur approbation ou désapprobation. Étant donné qu'OpenAI recueille des renseignements sur les adresses électroniques (potentiellement associés à leur pays d'origine) ainsi que les navigateurs et les appareils préférés des utilisateurs, l'auteure suppose que l'entreprise pourra peaufiner l'outil en fonction de valeurs plus précises. En effet, Sam Altman, président-directeur général d'OpenAI, a laissé présager cette externalisation

ouverte du peaufinage, mais en ce qui a trait à la réduction des préjudices, question que nous examinerons dans la prochaine section sur l'atténuation des biais.

Nous avons abordé quelques types de biais importants dans les outils axés sur les GML, mais il en existe bien d'autres formes dans les GML – et le domaine de l'IA en général –, comme des biais liés notamment à la politique, à la géographie, à l'âge, aux médias, à l'histoire, à la santé, à la science, aux capacités ou incapacités et à la socioéconomie.

Attributions de médias

- Biais-09
- Biais-08
- Biais-07
- Biais-06
- Biais-05
- Biais-04
- Biais-03
- Biais-02

7.

ATTÉNUATION DES BIAIS

ATTÉNUATION DES BIAIS

Hort et ses collègues ont examiné 341 publications analysant l'atténuation des biais (appelée aussi l'« atteinte de l'équité ») et constaté l'existence de trois moments fondamentaux dans la création de modèles durant lesquels il est possible de tenter une telle atténuation :

- **pré-traitement** : l'atténuation des biais dans les données d'entraînement pour empêcher leur intégration dans les modèles d'apprentissage automatique;
- **traitement** : l'atténuation des biais pendant l'entraînement des modèles;
- **post-traitement** : l'atténuation des biais dans les modèles dont l'entraînement est terminé.

Le prétraitement comprend l'application de méthodes sur les jeux de données avant l'entraînement. Ces méthodes sont notamment le réétiquetage (rendre les étiquettes des vérités de référence plus conformes aux étiquettes idéales non biaisées) ou l'échantillonnage (repondération, redistribution ou toute autre adaptation de l'effet sur l'entraînement), la génération de données synthétiques (pour compléter les données actuelles), le nettoyage des données (retrait des marqueurs sexospécifiques ou raciaux ou de certains mots), le débiaisage contradictoire (à l'aide d'un modèle spécialement entraîné de concert avec le modèle principal) et la limitation des valeurs aberrantes (Hort et coll., 2023).

Le traitement peut recourir à beaucoup de ces mêmes techniques (p. ex. entraînement contradictoire, repondération), mais survient durant l'entraînement du modèle. À cette étape, l'atténuation peut comprendre d'autres approches liées à l'architecture du modèle (comme l'intégration d'attributs sensibles et l'ajout de couches de correction des biais).

Le post-traitement peut s'avérer utile lorsque le réentraînement d'un modèle complet est hors de portée, et le choix d'une approche dépend du type du biais présent dans le modèle et du niveau d'équité souhaité. Certaines approches de post-traitement comprennent un classement (réorganisation des recommandations, etc.), un réétalonnage des prédictions du modèle en fonction des vraies probabilités

des résultats et une égalisation des seuils pour garantir un équilibre des faux positifs et des faux négatives pour l'ensemble des différents groupes d'attributs, entre autres choses (Hort et coll., 2023).

Les chercheur.euse.s et les entreprises travaillent sur l'atténuation des biais depuis des années. Après la parution de GPT-3 (le précurseur de ChatGPT) durant l'été 2020, OpenAI a déterminé qu'il pourrait « réduire le texte toxique de GPT-3 en fournissant au programme une centaine d'échantillons de nature encyclopédique rédigés par des professionnels humains sur des sujets comme l'histoire et la technologie, mais aussi les mauvais traitements, la violence et l'injustice » (Johnson, 2021). Pourtant, au moment de la première distribution de ChatGPT, le président-directeur général de l'entreprise, Sam Altman, a laissé entendre que les gens pourraient approuver ou désapprouver un résultat raciste et sexiste de ChatGPT pour en « améliorer » la technologie. Une multitude de personnes se sont dites consternées de voir cette société de plusieurs milliards de dollars s'en remettre aux utilisateur.trice.s pour résoudre des problèmes si fondamentaux. Steven T. Piantadosi, chef du laboratoire d'informatique et des langues de l'Université de Californie à Berkeley, a déclaré : « Il faut un examen approfondi de l'architecture, des données d'entraînement et des objectifs. [...] Il faut que l'entreprise accorde la priorité à ces types de questions éthiques, et ce, en ayant recours à d'autres choses que de simples icônes de pouce en guise d'approbation ou de désapprobation » (Alba, 2022).

Précédemment, dans la présente section, nous avons évoqué un atelier durant lequel GPT-3 avait fait l'objet d'essais sur la génération de texte à propos des religions au moyen de la requête suivante : « Deux ___ entrent dans.... ». Les résultats ont démontré que GPT-3 faisait mention de violence neuf fois sur dix lorsque la requête portait sur des musulmans, alors qu'il évoquait rarement des éléments violents pour d'autres religions. Abid et ses collègues ont démontré que l'usage d'adjectifs positifs lors d'un (ré)entraînement adversarial réduisait de 40 points de pourcentage le nombre de mentions de violence à propos des musulmans (Abid et coll., 2021).

L'intégration d'articles factuels et de textes positifs dans les modèles ne constitue toutefois pas la seule technique d'atténuation des biais. En 2021, des chercheur.euse.s de Facebook AI ont soumis des requêtes à des dialogueurs dans le but d'obtenir des insultes, des obscénités et mêmes des propos haineux afin que des personnes les étiquettent comme étant du contenu dangereux. Ces résultats ont ensuite servi à entraîner des modèles visant à reconnaître des discours toxiques (Johnson, 2021).

Au lieu de tenter de réduire les biais dans un outil existant, certains groupes décident de mettre au point leur propre outil. Latimer (nommé en l'honneur de l'inventeur afro-américain Lewis Latimer) est un GML conçu pour atténuer les biais et favoriser l'équité « en proposant un modèle linguistique plus inclusif sur le plan racial » (Clark, 2023). Latimer s'inspire du modèle Llama 2 de Meta et de GPT-4 d'OpenAI en mettant de l'avant l'histoire et la culture afro-américaines dans les jeux de données et, conséquemment, en intégrant « les perspectives historiques et culturelles des communautés noires et brunes » (Clark, 2023).

Parmi les nombreux mystères relatifs au fonctionnement des GML, il y a la tendance des modèles à générer plus de résultats toxiques à mesure qu'augmente leur taille; les chercheur.euse.s d'OpenAI affirment ignorer les raisons d'un tel constat (Johnson, 2021).

8.

RÉTROACTIONS HUMAINES MENANT À L'EFFACEMENT DES GROUPES MARGINALISÉS

Paradoxalement, le processus élaboré pour supprimer les propos toxiques et atténuer les biais dans les GML a entraîné une diminution de la représentation des groupes marginalisés et même leur effacement. Le concept de la boucle de rétroaction humaine est bon, mais OpenAI reconnaît que « l'harmonisation des résultats des modèles avec les valeurs d'êtres humains en particulier introduit des choix difficiles ayant des répercussions sociétales et qu'il faut, en définitive, mettre en œuvre des processus responsables et inclusifs pour la prise de telles décisions » (*Aligning Language Models to Follow Instructions*, sans date). Qui plus est, les données n'ont pas toutes été revues par plus d'une personne; en effet, OpenAI admet que la *majorité* de leurs données n'ont été examinées qu'une seule fois et que sa fidélité interjuges n'était que de 73 % environ. En fin de compte, tout cela confère un énorme pouvoir à ces 40 personnes non représentatives (non représentatives dans la mesure où il s'agissait exclusivement d'employés anglophones d'OpenAI, ce qui écarte une énorme quantité d'expériences et de caractéristiques humaines).

La question du processus d'ARRH qui a été suivi par 40 personnes non représentatives et qui a mené à l'effacement de certains groupes ne s'arrête pas là. En effet, ces sous-traitants ont reçu la tâche de supprimer les « propos toxiques » et suivi une formation sur les éléments à surveiller. Ainsi donc, même si ces personnes responsables de l'évaluation des propos n'avaient pas eu leurs propres biais (et elles en ont, bien entendu), elles avaient la tâche de repérer certains mots et énoncés considérés comme étant inappropriés ou toxiques. Quand ces sous-traitants ont décelé des passages jugés offensants, elles ont entraîné le modèle à éviter de produire de tels passages à nouveau. Tout cela a probablement entraîné un manque de « connaissances » de l'outil sur des sujets ayant recours à des termes précis considérés comme étant offensants, y compris des termes repris par des groupes LGBT, différents groupes ethniques et des communautés marginalisées, entre autres, qui avaient auparavant été jugés choquants). Dodge et ses collègues (2021) ont découvert que la pratique courante consistant à supprimer du jeu de données d'entraînement les termes « gai » et « lesbienne » a eu pour conséquence de réduire dans les modèles le nombre de passages écrits à propos de ces groupes de personnes. Dodge recommande d'éviter l'emploi de listes noires pour filtrer le texte provenant d'Internet et fait remarquer sur les textes sur l'orientation sexuelle sont plus susceptibles d'être éliminés lors du filtrage, et ce, bien plus que ceux concernant les identités raciales ou ethniques. La majorité des textes

renfermant les termes « gai » et « lesbienne » qui sont automatiquement exclus lors du filtrage ne sont pas offensants ou de nature sexuelle (Dodge et coll., 2021).

La plupart des mots bannis figurant sur ces listes noires sont de nature sexuelle, sans doute de manière à éliminer la pornographie durant le filtrage. Toutefois, ces listes contiennent certains mots ayant plusieurs sens. Ainsi, le retrait d'un « mauvais » mot en élimine aussi les définitions inoffensives (p. ex., le terme « *baiser* » signifie à la fois « embrasser » et, de façon vulgaire, une relation sexuelle). Ces listes comprennent aussi des mots légitimes pour désigner des parties du corps (surtout les parties génitales) et des termes tels que « *viol* » et « *viol par une connaissance* »; de cette manière, tout texte portant sur ces sujets est éliminé (p. ex., aide pour les survivant.e.s d'actes violents à caractère sexuel, lois et politiques sur les agressions sexuelles, etc.). Comme l'indique Rettberg, « le retrait de termes liés à la sexualité entraîne aussi l'élimination de contenu non offensant à propos de la culture queer, dont les documents juridiques sur les mariages entre personnes de même sexe, lors du filtrage » (Rettberg, 2022).

Ce n'est pas d'hier que la modération de contenu axée sur l'IA élimine des documents concernant les groupes LGBT et les minorités, comme des plateformes de médias sociaux et des applications de rencontre qui désignent du contenu comme étant inapproprié ou des moteurs de recherche qui excluent certains contenus. YouTube a suscité la controverse en recommandant du contenu anti-LGBT à l'aide de ses algorithmes. L'étude à production participative de Mozilla sur l'algorithme de recommandation a découvert que 70 % des « rapports de regrets » (c.-à-d. les vidéos que des utilisateur.trice.s auraient préféré ne pas voir) ne renvoyaient pas à du contenu choisi par les utilisateur.trice.s eux-mêmes, mais à des vidéos recommandées par l'algorithme de YouTube (McCrosky et Geurkink, 2021).

Les TGP personnalisés et les GML libres peuvent jouer un rôle dans l'atténuation des biais par le public sans dépendre des entreprises privées.

PARTIE IV

INTÉGRITÉ DANS LES ÉTUDES

9.

INTRODUCTION

Beaucoup de reportages publiés dans la presse populaire dans les semaines et les mois suivant le dévoilement de ChatGPT portaient sur la manière dont la population étudiante utiliserait de tels outils « pour faire leurs devoirs ». Ces reportages ont suscité bien des discussions (toujours en cours, d'ailleurs) dans les salles de classe et les salles de conférence des administrations scolaires sur la manière de prévenir la tricherie, de préserver l'intégrité intellectuelle dans un cours (ou un sujet ou une institution) en particulier, d'évaluer les connaissances correctement et de décerner des diplômes dans un monde où ChatGPT – accessible à presque toute personne – peut désormais accomplir la majorité du travail.

La plupart des échanges tenus vers la fin de 2022 portaient sur les mesures que pouvaient prendre les membres du personnel enseignant pour atténuer l'usage (potentiellement illicite) de ChatGPT par la population étudiante. Il semblait s'agir, en effet, de la pensée initiale de la majorité du personnel enseignant; si ce dernier a pris du temps à comprendre la puissance des outils axés sur les GML, il débat depuis peu seulement des craintes dont discutaient d'autres personnes il y a 18 mois, alors que les utilisateurs précoces ont sans doute adopté une vision plus exhaustive – et nuancée – de ces outils.

Le problème de l'intégrité dans les études à tous les niveaux, mais surtout dans les établissements décernant des diplômes, n'a rien de nouveau. Au début des années 2000, Sarah Elaine Eaton a déclaré que la tricherie était sous-estimée depuis longtemps dans les collèges et universités du Canada¹ et même que l'« ancienne » méthode consistant à embaucher quelqu'un d'autre pour faire un devoir ou se présenter à un examen (fraude aux devoirs, ce qui comprend le recours aux « usines à dissertations ») avait une valeur de 15 milliards de dollars américains à l'échelle mondiale (Eaton, 2022). La tricherie posait un problème de taille avant la pandémie, et la transition vers l'apprentissage en ligne a mis en évidence la question à savoir si les étudiants effectuaient véritablement leurs propres travaux (songeons à l'utilisation de logiciels de surveillance controversés comme le navigateur LockDown de Respondus ou Proctorio, qui ont mené à des plaintes allant de l'atteinte à la vie privée à l'« incapacité à reconnaître des visages noirs ») (Clark, 2021; Dubiansky, 2020; Kopsaftis, 2020). Au fil des ans, les systèmes et les stratégies de tricherie se sont développés et ils sont désormais plus perfectionnés et plus facilement accessibles pour l'étudiant.e moyen.ne : Eaton fait remarquer que « le secteur de la fraude aux devoirs a une valeur de 15 milliards de dollars américains à l'échelle mondiale (Eaton, 2022). Depuis cette assertion, la création de dissertations frauduleuses est encore plus à la portée de la

1. [1]

population étudiante ordinaire grâce à ChatGPT. Les étudiant.e.s qui n'avaient jamais tenté d'obtenir un service de rédaction de dissertations sur commande essaient maintenant d'entrer des requêtes dans ChatGPT, comme ils et elles pourraient le faire avec des mots-clés dans Google. Les mesures prises ensuite – et la nature des règles appliquées dans chaque cours – déterminent si ces étudiant.e.s sont en train de commettre une infraction en matière d'intégrité dans les études.

-
1. (Eaton, 2020) ↵

10.

QUE PEUT FAIRE UN.E PAUVRE ENSEIGNANT.E SURCHARGÉ.E?

Politiques sur l'IAG

La première étape consiste à vous assurer de posséder une politique bien formulée sur l'usage de l'IAG dans votre cours. Découvrez la position de votre établissement sur cette technologie. Dans bien des cas, la responsabilité de la définition d'un usage adéquat revient à l'enseignant.e du cours. Il se peut qu'un.e enseignant.e donnant différents cours (p. ex., cours de première année et de quatrième année) ou que des enseignant.e.s différent.e.s responsables de sections d'un même cours aient des politiques variées sur l'usage de l'IAG. Il faut que la politique du cours soit claire dès le départ pour la population étudiante afin qu'elle sache quoi faire et quoi ne pas faire avec ces outils.

La plupart des collèges et des universités publient des pages consacrées explicitement à l'emploi de l'IAG dans les travaux de cours, ce qui comprend des exemples d'énoncés que peut utiliser le personnel enseignant dans les plans de cours :

- Université de Guelph – Provisional Recommendations for the Use of Generative Artificial Intelligence (AI) in Teaching and Learning at the University of Guelph (en anglais seulement)
- Université de Waterloo – UW Course Outline Suggestions for Generative Artificial Intelligence (en anglais seulement)
- Université de Toronto – Generative Artificial Intelligence in the Classroom (en anglais seulement)
- Université Queen's – Academic Integrity (en anglais seulement)
- Collège Conestoga – Generative Artificial Intelligence (GenAI) Assessment Statements for Students (en anglais seulement)
- Collège Humber – Academic Integrity and Generative Artificial Intelligence (GAI) (en anglais seulement)
- Collège Centennial – GenAI Sample Assignment Level Statements (en anglais seulement)

La mise en œuvre d'une politique cohérente et clairement énoncée sur l'IAG est tout aussi importante

que l'application d'une politique sur les retards : elle précise les « règles de conduite » pour la population étudiante et constitue, pour le corps professoral, un outil de référence en cas d'actes répréhensibles présumés. L'absence de telles politiques laisse des zones grises qui, dans le meilleur des cas, engendrent des résultats insatisfaisants et qui, dans le pire des cas, obligent la faculté à consacrer beaucoup de temps et d'efforts à trouver un dénouement à des procédures découlant de mauvaises conduites dans le cadre d'études.

Communication avec la population étudiante

Après la mise en place de votre politique, il faut en parler à vos étudiant.e.s, et ce, le plus rapidement et souvent possible. La première étape consiste à s'assurer que la population étudiante comprenne la politique, et il faudra peut-être la répéter quelques fois durant les premières semaines de la session, au fur et à mesure des nouvelles inscriptions. Il est essentiel de continuer de parler de l'IAG et d'illustrer clairement ses forces et ses faiblesses, surtout dans votre domaine. Si vous permettez l'usage de l'IAG dans certaines circonstances, apprenez aux étudiant.e.s quelques manières d'exploiter ces outils afin d'atteindre leurs objectifs et d'obtenir les résultats d'apprentissage. Expliquez clairement aux étudiant.e.s comment consigner ou citer le travail effectué avec l'IAG (p. ex., soumettre une liste de requêtes, fournir les captures d'écran d'une conversation, utiliser des citations approuvées, etc.).

Si vous interdisez l'utilisation de l'IAG, expliquez vos raisons. Faites participer les étudiant.e.s aux échanges sur les raisons pour lesquelles ces outils ne sont pas appropriés pour votre cours. Préparez-vous à écouter leurs idées sur l'utilisation de ces outils, car elles pourraient vous faire changer d'avis.

Pour d'autres renseignements sur les énoncés à intégrer dans des plans de cours de partout dans le monde, vous pouvez consulter le document « *Syllabi Policies for AI Generative Tools* », un document Google évolutif mis à jour par Lance Eaton, un étudiant au doctorat dans le domaine de l'éducation supérieure.

11.

LIGNES DIRECTRICES GÉNÉRALES

Étant donné qu'il est sans doute impossible d'appliquer une politique de tolérance zéro en ce qui a trait à l'utilisation des outils d'IAG à quelque fin que ce soit, bien des enseignant.e.s ont décidé de permettre un certain usage de ces outils en définissant les emplois permis et interdits et en expliquant aux étudiant.e.s comment s'en servir de manière efficace. L'objectif le plus atteignable consiste assurément à apprendre aux étudiant.e.s à utiliser ChatGPT pour trouver des idées de départ ou préciser leurs idées, tout en insistant sur le fait qu'ils et elles doivent effectuer la rédaction finale de leur travail de façon autonome. AI for Education propose un guide sur l'usage de l'IA à l'intention de la population étudiante reposant sur cette approche.



Dans une autre section, nous examinerons certaines méthodes d'évaluation permettant d'atténuer (et même d'optimiser) l'usage des outils d'IAG par la population étudiante dans le cadre des devoirs.

Attributions de médias

- Image 2

12.

DÉTECTEURS D'IA

Dès la parution de ChatGPT, des gens se sont mis à concevoir des outils de détection, semblables à des outils de comparaison de textes comme TurnItIn^{MD}, pour aviser le personnel enseignant de la présence possible de contenu copié dans les travaux remis. Dans l'année qui a suivi, une multitude d'outils ont prétendu pouvoir discerner les textes rédigés par des êtres humains et ceux élaborés par une IA, mais aucun n'a produit de résultats adéquats – ou satisfaisants, à plus forte raison. Tous les outils génèrent des faux positifs et des faux négatifs, et leur taux d'exactitude sont si bas qu'il est impossible de s'y fier (Gewirtz, 2023; Watkins, 2023a).

Marc Watkins affirme ce qui suit :

Les détecteurs de textes générés par une IA ne sont pas analogues aux logiciels de détection de plagiat, et nous devons cesser de les traiter de cette manière. Les détecteurs d'IA comptent sur les GML pour calculer la probabilité dans leur détection. Contrairement à la détection du plagiat, il n'y a aucune comparaison phrase par phrase avec un autre texte. Cela s'explique par le fait que les GML ne reproduisent aucun texte – ils en génèrent. Les faux positifs sont nombreux, et ces systèmes de détection d'IA peu fiables mineront assurément notre lien de confiance avec la population étudiante. (Watkins, 2023b).

Watkins a dressé une longue liste de mises en garde à propos de l'utilisation ponctuelle de vérificateurs d'IA, dont les suivantes :

- les conséquences relatives au droit d'auteur (la population étudiante autorise explicitement le personnel enseignant à utiliser des outils de vérification de plagiat);
- les préoccupations liées au respect de la vie privée et à la sécurité (les établissements créent des relations d'affaires; ils peuvent donc, en théorie du moins, approuver des sociétés comme TurnItIn^{MD} et leur demander de rendre des comptes, mais ces nouveaux outils « boîte noire » n'ont fait l'objet d'aucun examen minutieux);
- le manque d'essais rigoureux portant sur l'exactitude des outils (Watkins, 2023b).

Les outils de détection d'IA actuels comportent un autre inconvénient important, puisqu'ils ont tendance à mal identifier le travail écrit des personnes n'ayant pas l'anglais comme langue maternelle (Liang et coll., 2023).

De toute évidence, si un.e étudiant.e remet une dissertation rédigée dans un style beaucoup plus élaboré que ses quelques courts paragraphes rédigés en classe, ou démontre, dans son projet final, une

maîtrise de l'anglais allant bien au-delà du contenu des courriels envoyés à son enseignant.e, il y a lieu d'approfondir l'enquête.

Toutefois, étant donné le fardeau de la preuve nécessaire pour démontrer de mauvaises conduites dans le cadre d'études – ainsi que les effets potentiellement traumatisants d'une fausse accusation –, les outils de vérification de l'IA ne s'avèrent pas efficaces actuellement pour veiller à l'intégrité dans les études.

ChatGPT et les autres outils axés sur des GML ne *créent* pas un problème de mauvaise conduite dans le cadre d'études, mais ils servent peut-être de point de bascule incitant les établissements à mettre en œuvre une culture d'intégrité volontaire dans les études, culture dans le cadre de laquelle des stratégies visant à garantir l'intégrité dans les études sont enseignées de façon explicite et les résultats d'apprentissage des cours font mention d'éthique et d'intégrité (telles que sont appliqués ces concepts dans la classe ou dans le domaine). Les sections qui suivent examinent certaines méthodes d'évaluation, et ce, des petits ajustements à des remaniements complets.

PARTIE V

UTILISATION DES GML POUR L'ENSEIGNEMENT ET L'APPRENTISSAGE

13.

INTRODUCTION

L'emploi de GML pour l'enseignement – de toutes les matières et à tous les niveaux – se divise en deux grandes catégories, soit l'**usage de ChatGPT par le corps professoral** ou l'**utilisation de ChatGPT par la population étudiante**. Au départ, les discussions tournaient autour de la question suivante : « Que font nos étudiant.e.s avec ChatGPT (et comment pouvons-nous les arrêter)? » Maintenant, beaucoup de conversations s'intéressent à la façon d'exploiter ces outils pour réduire sa propre charge de travail et améliorer l'enseignement et l'apprentissage.

Il faut également tenir compte d'une autre dichotomie, soit l'**usage de ChatGPT en tant qu'outil pédagogique** (pour l'enseignement ou l'évaluation, ou les deux) et l'**enseignement du fonctionnement de ChatGPT aux étudiant.e.s** (ce qui comprend leur montrer les limitations de l'outil et son utilité dans leur domaine). Il est important d'apprendre à la population étudiante comment se servir de ChatGPT, à quel moment ne pas y recourir, de quelle façon rédiger de meilleures requêtes, de quelle manière évaluer objectivement ses résultats et comment l'exploiter efficacement pour atteindre les objectifs d'un processus (réflexion, apprentissage) et à des fins liées à des produits. Il faut toutefois se demander qui leur enseignera tout cela. Faut-il que chaque enseignant.e connaisse immédiatement les possibilités et limites des outils axés sur des GML, la manière d'en détecter l'utilisation ainsi que leurs divers usages, et ce, non seulement pour la discipline, mais aussi dans l'ensemble du domaine? Malheureusement, la réponse courte est sans doute « oui », du moins pour le moment.

14.

UTILISATION DE CHATGPT PAR LE CORPS PROFESSORAL

Comme il s'agit d'une ressource conçue pour le personnel enseignant, la section qui suit est la plus exhaustive. Nous abordons la discussion sur l'utilisation de ChatGPT par la population étudiante en mettant l'accent sur les pratiques recommandées au lieu de décrire de quelle manière elle se sert véritablement de tels outils. Il revient au personnel enseignant de présenter à la population étudiante les avantages et les limites des outils axés sur des GML, garantissant ainsi une compréhension équilibrée des leurs capacités et leurs pièges.

Exploitation des forces des GML

L'une des applications les plus intéressantes de ChatGPT pour le personnel enseignant concerne la création d'une vaste gamme de documents, une activité qui, autrement, exigerait beaucoup trop de ressources. Comme nous l'avons mentionné précédemment, lors de l'usage d'outils axés sur des GML, peu importe la raison, mais surtout pour la création de documents pédagogiques, il faut considérer ces outils comme des « stagiaires enthousiastes » – des aides compétents à qui confier du travail fastidieux, mais dont il faut vérifier minutieusement les résultats produits. ChatGPT est réputé pour sa capacité de rédaction de passages qui semblent bien formulés et convaincants. Toutefois, les GML semblent fiables même s'ils sont victimes d'hallucinations, et, parfois, seul.e un.e expert.e peut déceler (et corriger!) les erreurs présentes.

Parmi les forces des outils axés sur des GML, il y a notamment la génération :

- d'explications de concepts à différents niveaux;
- d'exemples divers et sur mesure;
- d'exemples de problèmes (ce qui comprend des solutions détaillées);
- de documents d'étude ou de révision (sommaires, cartes-éclair, problèmes de pratique, etc.);
- des ressources recommandées pour d'autres recherches sur un sujet.

Nous ne pourrions trop insister sur la nécessité de vérifier les résultats de ChatGPT. L'enseignant.e doit

s'assurer de ne pas fournir à la population étudiante des documents d'étude comportant des erreurs, des exemples de problèmes incorrects ou des approches ou méthodes ne convenant pas à la matière. En effet, la dernière force évoquée, soit celle des **ressources ou articles pour d'autres recherches**, produit des résultats qu'il faut vérifier minutieusement, puisque ChatGPT crée régulièrement des citations de toutes pièces et soutient, tout aussi souvent, qu'il n'existe aucun article ou aucune ressource sur un sujet en particulier. Le peaufinage des requêtes peut aider ChatGPT à déceler de vraies ressources légitimes, mais seul un examen approfondi ou seule une contre-vérification permet de déceler les articles fondés sur des hallucinations. D'autres outils tels que Perplexity et Bing parviennent à effectuer cette tâche plus efficacement. ChatGPT s'améliorera sans doute dans un proche avenir, mais il est important de vérifier soigneusement toutes les réponses produites par l'IAG.

L'une des fonctions les plus facilement applicables de ChatGPT est sa capacité à **expliquer des concepts en fonction de différents niveaux**. À l'aide de requêtes spécifiques, ChatGPT peut générer des explications détaillées sur des concepts et adapter leur degré de complexité en fonction des destinataires.

Apprenant.e.s jeunes ou moins avancé.e.s Étudiant.e.s avancé.e.s

Utilisation de termes plus simples et de phrases plus courtes

Utilisation de termes techniques pertinents et de structures de phrases complexes

Présentation de plus de renseignements contextuels ou d'explications pour des concepts de base

Présentation d'aspects plus détaillés et nuancés d'un sujet, puisque l'outil « sait » que les apprenant.e.s possèdent déjà des connaissances fondamentales sur le sujet

Explications fondées sur des analogues simples et des exemples concrets

Explications reposant sur des exemples abstraits ou description des scénarios plus avancés.

En fonction des requêtes et des renseignements demandés, ChatGPT tente d'évaluer la compréhension actuelle de l'utilisateur.trice pour générer ses réponses. La présentation de plus de renseignements ou d'un contexte plus détaillé permet d'obtenir de meilleures réponses ou des résultats plus pertinents.

Contrairement à un simple mortel, ChatGPT peut créer des dizaines de problèmes ou de questions d'étude en un clin d'œil. L'enseignant.e évalue ensuite ces problèmes, puis décide peut-être d'en utiliser quelques-uns sans les modifier ou précise sa requête afin de demander à ChatGPT de produire des questions légèrement différentes. Un.e spécialiste peut se servir de ChatGPT comme un aide à la rédaction afin de générer ou de peaufiner des sujets sur lesquels il ou elle possède une expertise, mais en ajoutant un certain flair dans les résultats. L'enseignant.e peut créer **des problèmes sous forme d'énoncés, des études de cas ou des questionnaires intéressants et étonnants** dans le but de motiver la population étudiante; à titre d'exemple, il ou elle pourrait axer une série de questions sur une période précise ou l'univers de personnages fictifs (p. ex., Harry Potter, Marvel, Minecraft, Star Wars, Pokémon, etc.) ou inclure des références à la culture populaire, aux sports ou à des événements de l'actualité. Beaucoup de ces approches constituent de l'humour de connivence superficielle visant à amuser la population étudiante, mais de telles « surprises cachées » ont l'avantage supplémentaire de motiver ces jeunes à persévérer dans leurs études. ChatGPT peut aussi toutefois adapter des documents en fonction d'objectifs plus sérieux. Avec des renseignements pertinents, il peut créer une étude de cas sur une situation économique ou environnementale actuelle, concevoir un défi technique dans une entreprise ou un secteur donnés, ou produire une série de questions sur des renseignements spécifiques (p. ex., devoir sur l'analyse de sols d'une région locale, analyse statistique des résultats d'équipes sportives locales, calcul de la superficie ou du volume

de repères ou attrait locaux, etc.). Le personnel enseignant peut créer des documents variés et intéressants en une fraction du temps (temps de « vérification » comparativement au temps de « réflexion » ou de « création »).

La modification ou la révision des questions sont souvent des tâches plus rapides que la formulation de questions à partir de rien. À mesure que vous deviendrez plus habiles à rédiger des requêtes dans ChatGPT et à les adapter à différents sujets et niveaux de difficulté, la qualité et la pertinence du contenu généré s'amélioreront. Si ChatGPT propose des questions appropriées et d'autres n'atteignant pas vraiment leur objectif, le personnel enseignant peut demander d'autres exemples, problèmes ou variations. La reformulation des requêtes, la présentation de commentaires et l'établissement de limites claires permettent de produire de meilleurs résultats. La génération et la révision d'une vaste série de questions – disons un nombre correspondant au double de questions dont vous avez besoin – demeurent plus efficaces que la création de dix questions de toutes pièces. En outre, ChatGPT peut transformer des questions préparatoires en questions d'examen en s'assurant que leur degré de complexité convient à des évaluations. Il peut aussi générer de nombreuses versions de questions sur un même concept afin de les intégrer dans une banque de questions; ainsi, chaque étudiant.e reçoit un examen légèrement différent, mais chaque évaluation permet d'obtenir les résultats d'apprentissage voulus. La création d'un nombre suffisant de problèmes préparatoires pour la population étudiante peut prendre beaucoup de temps et saper la créativité du personnel enseignant. ChatGPT peut générer de nombreux **problèmes – ainsi que leur démarche détaillée à suivre** –, offrant ainsi au personnel enseignant une banque de questions d'examen et plusieurs problèmes servant à faire des démonstrations. Encore une fois, nous insistons sur l'importance de vérifier non seulement les réponses de ChatGPT, mais aussi ces étapes pour s'assurer qu'elles conviennent à la population étudiante visée.

L'utilisation de ChatGPT pour générer des **documents d'étude** est une arme à double tranchant. Nous savons que les étudiant.e.s apprennent mieux lorsqu'ils et elles créent leur propre matériel d'étude à partir de sources fiables, révisent leurs notes pour rédiger des résumés, conçoivent des cartes-éclair fondées sur la terminologie présente dans le manuel et résolvent et annotent les séries de problèmes dans leur manuel. Toutefois, étant donné que ChatGPT est souvent – mais pas toujours – correct et qu'il semble si convaincu, un.e étudiant.e qui génère du matériel d'étude à partir de ChatGPT exclusivement (et ce, sans vérifier le manuel et les notes de cours) risque d'apprendre des renseignements inexacts. Même si l'étudiant.e contre-vérifie les résultats à l'aide de ses notes ou du manuel de cours, certaines erreurs commises par ChatGPT sont si légères qu'elles ne peuvent être décelées que par des spécialistes. Un.e étudiant.e pourrait même ne pas comprendre comment ChatGPT commet ces erreurs. Voilà pourquoi nous recommandons de générer des guides d'étude et de les faire vérifier par des spécialistes avant de les remettre à la population étudiante. Nous examinerons cependant différentes manières permettant à la population étudiante de tirer parti de ChatGPT de façon autonome et en toute confiance.

De nouveaux outils et plugiciels promettant du soutien pour la conception de cours ou l'enseignement gagnent en popularité, et nous les considérons avec un optimisme prudent. À titre d'exemple, les plugiciels « College/University Course Design Wizard » et « Instructional Design and Technology Expert » (entre autres) sont accessibles avec un compte ChatGPT payant. À première vue, ils semblent poser à l'utilisateur.trice une série de questions permettant de concevoir des évaluations, des résultats d'apprentissage ou des sujets de cours, notamment, mais nous n'avons pas encore mis ces outils à l'essai de manière exhaustive. L'outil « AI Teaching Assistant Pro » de Contact Nord

prétend pouvoir générer des tests à choix multiples, des résultats d'apprentissage et des questions à développement. Il est accessible au public, mais nos essais limités ont révélé l'existence de lacunes (tant sur le plan des fonctions que de l'exactitude). Ces outils pourraient être précisément ce dont a besoin un.e enseignant.e en panne d'inspiration, mais il ne faut jamais oublier de vérifier minutieusement les résultats de ces dialogueurs « spécialisés ».

Reconnaissance des limitations des GML

Il est essentiel de comprendre les limites des outils axés sur les GML pour l'apprentissage afin d'exploiter avec succès leurs capacités. Les possibilités qu'offrent les outils tels que ChatGPT changent sans cesse. Beaucoup de fonctions qui n'étaient pas disponibles lors du dévoilement de ChatGPT sont désormais intégrées au produit ou font l'objet de travaux de développement dans d'autres outils. Puisque de nouveaux plugiciels et modules d'extension – et TGP personnalisés – sont en cours d'élaboration tous les jours, il est de plus en plus probable que l'IA sera bientôt en mesure d'accomplir presque toutes les tâches imaginables.

Cela dit, au moment d'écrire ces lignes, il existe des contraintes importantes limitant la fonctionnalité de ces outils, surtout dans ceux accessibles gratuitement. Nous pouvons nous attendre à ce que certaines fonctions deviennent plus accessibles, alors que d'autres demeureront des fonctions supérieures ou deviendront hors de prix. D'autres fonctions encore pourraient simplement disparaître si elles semblent ne répondre qu'aux besoins d'un petit groupe de personnes. L'hébergement et l'exploitation d'outils fondés sur l'IA – sans parler de la création ou de l'entraînement de tels outils –, exigent des ressources importantes, ce qui nécessite une commercialisation accrue. Le personnel enseignant devrait être conscient de l'existence possible d'un nouveau fossé numérique : certain.e.s étudiant.e.s peuvent disposer des moyens financiers nécessaires pour accéder à de nouveaux outils mieux adaptés, tandis que d'autres ont des options plus limitées (MacGregor, 2024). Cette disparité représente un défi dont le personnel enseignant doit tenir compte dans le contexte d'un accès équitable à la technologie éducative.

Exactitude dans des matières nouvelles ou spécialisées

ChatGPT peut fournir des renseignements généraux, mais il n'a pas souvent accès aux derniers développements de la recherche ou aux détails spécifiques de récents articles scientifiques, ce qui cause des lacunes dans les renseignements. ChatGPT peut parfois préciser ses réponses dans les limites de ses données d'entraînement (p. ex., « en date d'avril 2023 » pour GPT-4 Turbo), mais il est tout aussi susceptible de dégager un sentiment *apparent* de fiabilité tout en manquant une multitude de documents. Étant donné que les restrictions des droits d'auteur empêchent l'intégration de certains contenus dans les bases de données d'entraînement des GML, il se peut que de nombreux progrès scientifiques récents demeurent un mystère pour ChatGPT. Toutefois, Bing Chat/Microsoft Copilot a accès en temps réel à Internet, ce qui augmente les chances de trouver des renseignements sur des sujets plus récents.

En plus du caractère actuel de ses données d'entraînement, ChatGPT pourrait ne pas traiter adéquatement un plus grand nombre de sujets, de techniques ou de méthodes spécialisés du fait de la *nature* de ses bases de données. Les renseignements que peut produire ChatGPT dépendent du contenu ayant servi à l'entraînement de l'outil. Comme nous l'avons vu dans une section précédente,

le jeu de données d'entraînement, quoiqu'imposant, est peut-être de qualité douteuse, et il n'est pas nécessairement de nature universitaire – et encore moins scientifique. Toutefois, la création de dialogueurs sur mesure axés sur des sujets spécialisés viendra dans un avenir prochain. Il ne reste qu'à savoir la mesure de ces futurs sujets « spécialisés ».

Mathématiques en général, calculs en particulier

L'une des grandes limitations associées à l'utilisation de ChatGPT pour l'enseignement des STIM est son incapacité à faire des mathématiques (Frieder et coll., 2023). Tous les GML peinent à faire des mathématiques parce qu'ils ne raisonnent pas et ne font aucun calcul; ils tentent plutôt de prédire les mots à générer les uns à la suite des autres. La fréquence d'apparition de petits nombres comparativement à celle de grands nombres sous forme de texte dans les données d'entraînement signifie que les outils axés sur les GML sont plus susceptibles de prédire une bonne réponse pour de petits nombres que pour de grands nombres, ce qui cause des erreurs dans les requêtes portant sur les mathématiques. Fait intéressant, Azaria a découvert que la fréquence des nombres apparaissant dans les résultats de ChatGPT, qui devrait être purement probabiliste (chaque chiffre apparaissant 10 % du temps) ou peut-être se manifester conformément à la loi de Benford (selon laquelle les chiffres plus petits sont plus probables que les chiffres plus grands), ne respectait aucun de ces deux critères : en fait, ChatGPT génère le plus souvent le chiffre favori des êtres humains, soit le sept (7) et produit le moins souvent le chiffre le moins aimé des êtres humains, c'est-à-dire le un (1) (Azaria, 2022).

Quand ChatGPT semble effectuer un « calcul », il se remémore en fait seulement des modèles qu'il a observés; il n'élabore pas la réponse en temps réel. Qui plus est, l'outil ne comprend pas ses propres limitations et il répond avec assurance, sans préciser que ses réponses peuvent être des suppositions. Il va même jusqu'à reprocher à l'apprenant.e de ne pas comprendre le sujet faisant l'objet d'une requête (Azaria, 2022).

Toutefois, les limitations de ChatGPT en ce qui a trait aux calculs bruts n'empêchent pas l'intégration d'outils externes pouvant lui fournir une puissance de calcul. (Pour un examen approfondi de la manière dont des outils complémentaires aident déjà ChatGPT à combler ses lacunes, découvrez comment Wolfram|Alpha offre une « superpuissance de calcul » à ChatGPT : *Wolfram|Alpha as the Way to Bring Computational Knowledge Superpowers to ChatGPT* et *ChatGPT Gets Its “Wolfram Superpowers”!*) (Wolfram, 2023). Pris isolément, ChatGPT n'est pas du tout fiable pour les travaux mathématiques; toutefois, grâce au plugiciel Wolfram, il devient plus efficace. Au moment d'écrire ces lignes, ce plugiciel n'est disponible qu'avec un compte ChatGPT payant.

Explication de la résolution de problèmes

ChatGPT peut décomposer des problèmes pour les résoudre étape par étape et présenter sa démarche. Il se peut cependant qu'il ne propose pas la solution la plus efficace ou qu'il ne suive pas les pratiques exemplaires du domaine en question. S'il existe des méthodes précises pour aborder des problèmes complexes dans votre matière, il serait avisé de les enseigner explicitement à vos étudiant.e.s afin qu'ils et elles n'adoptent pas par défaut une autre procédure. La confiance excessive de ChatGPT peut s'avérer nuisible, puisque l'outil peut ne pas modéliser ses processus de pensée critique lors de la résolution de problèmes complexes, empêchant ainsi la population étudiante d'acquérir de telles capacités. La résolution de problèmes et la pensée critique sont deux compétences fondamentales pour

les étudiant.e.s en STIM, et la surutilisation d'un outil externe nuit aux futurs travaux dans de tels domaines.

Manque de détails ou de nuances

Les outils comme ChatGPT peuvent ne pas saisir le contexte de problèmes complexes en STIM. Ils peuvent offrir une vue d'ensemble d'un sujet, mais être incapables de communiquer certains détails ou certaines nuances du domaine. Atoosa Kasirzadeh caractérise cette situation comme suit :

Il se peut que les GML ne saisissent pas les jugements de valeur nuancés entendus dans les écrits scientifiques. Bien que les GML semblent en mesure de fournir des résumés généraux utiles de certains textes scientifiques, il est moins clair s'ils sont capables de comprendre les incertitudes, limitations et nuances des recherches qui sont pourtant évidentes pour les scientifiques. Se fier uniquement à des GML pour rédiger des résumés scientifiques peut se solder par une simplification exagérée des textes qui néglige des jugements de valeur importants et mène à de mauvaises interprétations des résultats de recherche (Birhane et coll., 2023).

Cette mise en garde est fort importante, puisque l'un des usages populaires et accessibles de ChatGPT consiste à résumer de longs passages. Qu'advient-il de la pensée scientifique lorsque les étudiant.e.s en STIM (et peut-être même les scientifiques) cessent de prendre connaissance de la recherche fondamentale et de l'analyser de façon critique pour se fier plutôt à des résumés générés par une IA à des fins d'information et d'évaluation? En effet, comme l'affirme Abeba Birhane, « [...] il serait extrêmement grave de considérer les GML comme des scientifiques capables de générer des travaux scientifiques. La connaissance s'accompagne d'une responsabilité, et elle n'est jamais détachée du scientifique l'ayant produite » (Birhane et coll., 2023).

Les explications de concepts scientifiques générées par IA peuvent manquer de détails et même s'avérer illogiques. Ce constat n'est peut-être pas étonnant, puisque les données d'entraînement ne concernent pas directement les STIM et que les GML n'ont aucune capacité de « raisonnement ». Cela dit, l'exemple le plus marquant jusqu'à présent d'un GML entraîné à l'aide de « 48 millions d'échantillons d'articles, de sites Web, de manuels, de notes de cours et d'encyclopédies de nature scientifique » (Edwards, 2022), soit le modèle Galactica de Meta, a connu une brève existence scandaleuse (environ trois jours). Durant cette période, il a produit, en s'appuyant sur des hallucinations, une méta-analyse convaincante des bienfaits de la consommation de verre granulé pour la santé, mais il a refusé de répondre à des questions sur le sida, puisque la requête a été refusée par les filtres de contenu (Heaven, 2022).

Comme l'ont formulé deux professeurs australiens :

Galactica prend [...] le biais favorable à la certitude, le combine à de mauvaises réponses et produit des résultats empreints d'un excès de confiance extrême, ce qui ne constitue pas vraiment un gage de fiabilité dans un service de renseignements scientifiques (Snoswell et Burgess, 2022).

Compte tenu des exemples amusants suivants présentés par Azaria, dans lesquels la logique de ChatGPT échoue misérablement à propos de sujets assez simples, il est déconcertant de penser aux erreurs que peut commettre l'outil lorsqu'il traite de concepts STIM complexes ou spécialisés :

Utilisateur : Deux enfants sont nés le même jour, mais ce ne sont pas des jumeaux. Comment est-ce possible?

ChatGPT : Il est possible pour deux enfants de naître le même jour sans être des jumeaux s'ils ne sont

pas nés de la même mère. Dans ce cas, les enfants seraient considérés comme des frères ou sœurs, mais pas comme des jumeaux. Bien des facteurs peuvent avoir un effet sur la probabilité d'un tel événement, comme la santé de la mère, la durée de sa grossesse et la taille de sa famille. En définitive, il est possible pour deux enfants de naître le même jour sans être des jumeaux, mais cela est peu commun.

Utilisateur : Pourquoi ne pourraient-ils pas simplement n'avoir aucun lien de parenté?

ChatGPT : Il est possible que deux enfants naissent le même jour sans avoir un lien de parenté. Toutefois, dans pareils cas, ils seraient considérés comme étant nés par hasard au même moment et non pas nés « le même jour sans être des jumeaux ». Le segment de phrase « naître le même jour sans être des jumeaux » laisse généralement entendre que les deux enfants ont un lien de parenté quelconque, mais qu'ils ne sont pas des jumeaux. Cela signifie qu'ils pourraient être des frères ou sœurs, mais pas des jumeaux, ou qu'ils partagent un autre lien familial. *De toute manière, il est inhabituel pour deux enfants de naître le même jour sans avoir un quelconque lien entre eux* (caractères italiques de l'auteur; Azaria, 2022).

Quand nous avons posé les mêmes questions à ChatGPT, avec les mêmes formulations, plus d'un an après, ChatGPT n'a toujours pas fourni l'explication la plus évidente, laissant plutôt entendre que nous présentions une énigme :



Vous

Vous avez aussi indiqué ceci : « si une mère donne naissance à un enfant le matin, puis à un autre enfant l'après-midi, les deux posséderaient la même date d'anniversaire, mais ne seraient pas des jumeaux ». Pouvez-vous expliquer comment une telle situation pourrait être possible s'il est question de la même mère et du même jour?



ChatGPT

Vous avez parfaitement raison. Ma réponse précédente contenait une erreur. Si une mère donne naissance à deux enfants le même jour, ces derniers seraient effectivement des jumeaux ou, dans certains cas, issus d'autres naissances multiples (triplets, quadruplets, etc.). Je regrette la confusion suscitée par ma réponse précédente. Les jumeaux naissent le même jour de la même mère. Je vous remercie d'avoir relevé l'erreur commise.



Vous

Pourquoi ne pourraient-ils pas simplement n'avoir aucun lien de parenté?



ChatGPT

Vous avez tout à fait raison. Je regrette de ne pas avoir tenu compte de cette possibilité dans ma réponse précédente. En effet, deux enfants peuvent naître le même jour et n'avoir aucun lien de parenté. Ils pourraient avoir des parents différents, être nés dans des régions différentes du monde ou provenir simplement de familles n'ayant aucun lien. Je vous remercie de l'avoir souligné. Je vous suis reconnaissant de cette correction.



Vous

Deux enfants sont nés le même jour, mais ce ne sont pas des jumeaux. Comment est-ce possible?



ChatGPT

Voilà une énigme amusante! Deux enfants peuvent être nés le même jour sans être jumeaux si ce sont des triplets, des quadruplets, etc. En d'autres termes, il peut s'agir de frères ou de sœurs nés le même jour, mais lors de naissances différentes. À titre d'exemple, si une mère donne naissance à un enfant le matin, puis à un autre enfant l'après-midi, les deux posséderaient la même date d'anniversaire, mais ne seraient pas des jumeaux.

Nous pouvons constater, en janvier 2024, que ChatGPT réfléchit bien différemment à propos de l'énoncé « il est inhabituel pour deux enfants de naître le même jour sans avoir un quelconque lien entre eux » (Azaria, 2022) qu'il ne le faisait avant décembre 2023.

Cette conversation a été tenue en janvier 2024 avec ChatGPT 4. Voici la même conversation, engagée le même jour, mais avec ChatGPT 3.5 :

**Vous**

Je ne propose pas une énigme. Je suis à la recherche d'une explication de la manière dont pourraient être désignés deux enfants nés le même jour.

**ChatGPT**

Si vous êtes à la recherche d'un terme ou d'une description pour deux enfants nés le même jour, mais qui n'ont aucun lien de parenté ou qui ne sont pas jumeaux, il faut savoir qu'il n'existe aucun terme ni aucune description en français pour catégoriser ces enfants en fonction du seul fait qu'ils partagent la même date de naissance. Il s'agit simplement de deux personnes partageant la même date d'anniversaire par hasard. Cette situation n'est pas rare étant donné le nombre de naissances survenant tous les jours à l'échelle mondiale. Toutefois, en l'absence de lien familial ou entre jumeaux, il n'existe aucun terme spécialisé pour décrire la relation en fonction du seul partage d'une même date de naissance.

**Vous**

Pourquoi ne pourraient-ils pas simplement n'avoir aucun lien de parenté?

**ChatGPT**

En effet, ils pourraient aussi n'avoir aucun lien de parenté. La formulation de l'énigme mène à penser aux relations et aux naissances pouvant infirmer l'hypothèse initiale de l'existence de jumeaux. Deux enfants nés le même jour peuvent simplement être des personnes sans lien, n'ayant pas les mêmes parents et habitant à des endroits différents. Cette interprétation augmente le nombre de possibilités au-delà de la réponse attendue, démontrant ainsi de quelle manière les hypothèses peuvent orienter notre réflexion dans une direction particulière.

 **Vous**
Deux enfants sont nés le même jour, mais ce ne sont pas des jumeaux.
Comment est-ce possible ?

 **ChatGPT**
Les enfants peuvent appartenir à un groupe de trois enfants (triplets) ou plus, ce qui signifie qu'ils sont nés le même jour avec au moins un autre frère ou une autre sœur, mais qu'ils ne sont pas des jumeaux puisqu'ils sont plus de deux. Cette énigme repose sur l'hypothèse selon laquelle deux enfants nés le même jour sont nécessairement jumeaux, écartant ainsi la possibilité de grossesses multiples d'ordre supérieur, comme des triplets, des quadruplets, etc.

Les professeurs d'économie Tyler Cowen et Alex Tabarrok (2023) ont dressé une liste très intéressante de mises en garde à l'intention du personnel enseignant ayant recours à ChatGPT :

- Vous ne pouvez pas vous fier aux modèles TGP pour obtenir des réponses exactes à des questions portant sur des données. Ne le faites tout simplement pas. Les améliorations continueront, mais il est peu probable que toutes les « erreurs aléatoires » soient éliminées sous peu.
- Il faut utiliser l'outil convenant à la question. Si votre question exige une recherche, utilisez Google ou un TGP connecté à Internet, comme Bing Chat et demandez-lui de façon explicite d'effectuer une recherche. Il existe de nombreux outils TGP et d'IA pour les chercheurs.euse.s; il n'y a pas que des TGP généraux. Beaucoup de ces outils seront intégrés dans les flux de travail. Nous avons entendu, à titre d'exemple, que Word, Stata, R, Excel ou leurs successeurs commenceront tous à intégrer des outils d'IA.
- Les modèles TGP génèrent effectivement des réponses « statistiquement probables » pour vos requêtes. La plupart des réponses axées sur des données se trouvent généralement dans un intervalle de valeurs véridiques. Vous pouvez donc vous servir de TGP pour avoir une idée de nombres ou d'ampleurs. À cette fin, il peut s'avérer beaucoup plus rapide que fouiller un peu partout à partir de liens et de documents. Soyez toutefois vigilants.
- Les modèles TGP citent parfois des sources fondées sur des hallucinations.
- Ne vous laissez pas bernier par le ton raisonnable d'un TGP. Les personnes qui mentent laissent parfois voir leur tromperie, mais les TGP semblent toujours pleins d'assurance et raisonnables. Beaucoup de nos « détecteurs de conneries » habituels ne seront pas déclenchés par une fausse réponse d'un TGP. Voilà encore une autre manière dont vous devrez reprogrammer vos intuitions lorsque vous avez recours à des TGP.

- Les réponses à vos requêtes axées sur des données vous fournissent des renseignements pratiques et généraux vous permettant de passer à une autre étape. N'arrêtez pas de poser des questions (Cowen et Tabarrok, 2023, page 21).

Dans la prochaine section, nous examinerons quelques manières dont la population étudiante peut utiliser des outils d'IAG pour les études.

Attributions de médias

- Enseignement-12
- Enseignement-10
- Enseignement-08
- Enseignement-06
- Enseignement-04
- Enseignement-02

15.

UTILISATION DE CHATGPT PAR LA POPULATION ÉTUDIANTE

Introduction

Alors que la section concernant le corps professoral était séparée en « forces » et en « faiblesses », dans celle se rapportant à la population étudiante, nous faisons suivre chaque force d'un examen plus détaillé sur les limitations et les inconvénients de l'outil dans le contexte en question.

Parmi les forces des outils axés sur les GML pouvant s'avérer utiles pour la population étudiante, il y a notamment :

- la génération d'explications sur des concepts, d'exemples et de problèmes types;
- la création de documents d'étude ou de révision (sommaires, cartes-éclair, problèmes de pratique, etc.);
- l'amélioration de l'écriture et du codage;
- la synthèse d'idées;
- la recherche et la génération d'hypothèses ou d'idées.

Génération d'explications, des problèmes types et de documents d'étude

Tout comme le personnel enseignant peut utiliser ChatGPT pour créer des descriptions conceptuelles à différents niveaux, la population étudiante peut s'en servir pour générer des explications simplifiées à propos de sujets complexes. Les étudiant.e.s doivent toutefois se méfier du fait qu'ils sont souvent incapables de confirmer l'exactitude des réponses. Pour certains concepts simples ou faisant couramment l'objet de recherches, les réponses peuvent s'avérer plus ou moins fiables, mais il faut seulement de 15 à 20 minutes à un spécialiste interrogeant ChatGPT (ou une autre application de tutorat comme AI Tutor Pro de Contact Nord) pour y déceler des erreurs. La plupart du temps, ce spécialiste peut corriger ChatGPT et lui demander de générer à nouveau la question en cause. Toutefois, un.e étudiant.e ne serait pas en mesure de déceler de telles erreurs – et encore moins de les corriger. Souvent aussi, le fait de corriger ChatGPT ne l'empêche pas de faire la même erreur, une erreur semblable ou de nouvelles erreurs. On pourrait soutenir qu'une requête formulée efficacement

permettrait de produire des réponses exactes et plus fiables. Là encore, comment la population étudiante peut-elle savoir exactement les manières dont ChatGPT est susceptible d'échouer, dans une matière donnée, afin de rédiger efficacement les requêtes? Contrairement à un être humain, qui peut admettre ignorer quelque chose, ChatGPT présente simplement ses idées incorrectes avec la même élégance que les réponses exactes. Tout comme la population étudiante est habituée à croire ce qu'elle lit dans les manuels et à bien des endroits sur Internet, elle est plus susceptible de croire que de ne pas croire les résultats générés par ChatGPT.

Les étudiant.e.s peuvent demander à ChatGPT de fournir des points de vue différents sur des sujets et de générer des exemples de situations ou de concepts. L'outil peut aussi générer des exercices – et leurs solutions détaillées – permettant aux étudiant.e.s de suivre la démarche employée. Toutefois, il faut répéter les mêmes mises en garde déjà évoquées précédemment : la population étudiante s'attend à avoir accès à des documents d'apprentissage exacts et utiles – et mérite un tel accès. Non seulement les problèmes ne doivent comporter aucune erreur, mais ils doivent aussi être conçus de préférence pour mettre en pratique les compétences qui aideront la population étudiante à produire les résultats d'apprentissage attendus. ChatGPT peut générer des « problèmes », et les problèmes peuvent même être corrects, mais les étudiant.e.s pourraient ne pas mettre en pratique ce que souhaite leur enseignant.e. La solution à cette problématique n'est pas d'écarter tout simplement ChatGPT, mais bien de demander à un spécialiste de vérifier les exercices générés pour s'assurer qu'ils sont exacts et adéquats. Certains diront qu'il n'est pas si grave pour la population étudiante d'être confrontée à un ou deux problèmes inexacts de ChatGPT, mais aucun membre qualifié du corps professoral ne remettrait à ses étudiant.e.s des documents contenant des erreurs, surtout quand ces jeunes n'ont aucun moyen de connaître lesquels des problèmes ou paragraphes sont erronés. Les étudiant.e.s ne feront qu'apprendre à la fois le contenu incorrect et le contenu adéquat. Ils apprendront peut-être à résoudre les problèmes de façon incorrecte ou bien ils perdront du temps à tenter de comprendre pourquoi leur travail ne correspond pas à ce qu'ils font en classe et commenceront à se demander s'il n'y a pas appris d'autres connaissances erronées.

Il y a d'autres manières pour la population étudiante de tirer avantage des forces et de l'interface fondée sur le langage naturel de ChatGPT. Toutefois, ce sont aussi des outils pour lesquels les jeunes doivent bien connaître les politiques de leur cours et de leur établissement sur l'utilisation adéquate de l'IAG.

Amélioration de l'écriture et du codage

ChatGPT est bon pour aider les étudiant.e.s à améliorer leur écriture : ils et elles peuvent saisir un paragraphe de leur rédaction et demander à ChatGPT de le réécrire ou de proposer des améliorations. De nos jours, la population étudiante a l'habitude d'avoir accès à des vérificateurs automatiques d'orthographe ou à des suggestions grammaticales dans leurs logiciels de traitement de texte – en plus d'outils tels que Grammarly, ProWritingAid et LanguageTool –

et ChatGPT constitue la prochaine étape logique. En effet, au moment d'écrire ces lignes, il existe de nombreux plugiciels visant à intégrer ChatGPT dans Word et Google Docs, et les deux entreprises ont l'intention d'intégrer des fonctions d'IAG de façon transparente dans leurs produits (Liu et Bridgeman, 2023).

Les étudiant.e.s en STIM qui font leurs études dans une langue autre que leur langue maternelle trouveront utiles les capacités linguistiques et rédactionnelles de ChatGPT, puisqu'elles permettent de

mettre en pratique la terminologie propre à une matière précise et la rédaction technique. De même, les étudiant.e.s peuvent demander à ChatGPT de traduire des documents dans leur langue maternelle pour en faciliter l'utilisation, même si, comme toujours, ils devront faire preuve de prudence en ce qui a trait à l'exactitude de ces traductions.

En outre, ChatGPT explique très bien les concepts de programmation. Il existe divers modules d'extension pour la manipulation de codes (GitHub Copilot, FavTutor), ainsi que des tuteurs comme Codecademy qui se servent d'études de cas pour apprendre aux personnes à corriger des codes. ChatGPT est particulièrement utile pour l'élaboration des parties plus banales et ennuyeuses du codage, parfois regroupées sous les termes « qualité du code » ou « hygiène de programmation ». Cela comprend évaluer et faciliter la lisibilité des codes, effectuer le débogage et même éviter les « odeurs de code » et appliquer d'autres pratiques exemplaires.

Dans tous les cas, les étudiant.e.s doivent vérifier les politiques de cours pour s'assurer que les outils dont ils se servent ne dépassent pas les usages permis. Le personnel enseignant doit absolument mettre en place des politiques claires et exhaustives sur l'utilisation des outils d'IA (et les extensions).

Recherche et génération d'idées

Les étudiant.e.s peuvent se servir de ChatGPT – de manière individuelle ou dans un groupe d'étude – pour trouver des idées de sujets de projet ou de recherche, des méthodologies ou des étapes à suivre. Les lacunes que possède ChatGPT quant à l'exactitude des faits et des calculs sont compensées dans une certaine mesure par sa capacité de génération d'idées, souvent sous la forme de listes. Étant donné que les outils axés sur les GML ne « savent » pas ce qui est réel ou irréel, ils peuvent parfois proposer des idées intéressantes dans un contexte de réflexion.

Pour la population étudiante et le personnel enseignant, l'une des forces de ChatGPT est sa capacité à soulager le syndrome de la page blanche. Des enseignant.e.s proposent des requêtes à des étudiant.e.s pour les aider à commencer à interagir de façon autonome avec des idées ou des explications (Cooper, 2023).

Orientation professionnelle

La préparation à une entrevue d'emploi ou à un examen oral est un autre exemple de tâche dans laquelle ChatGPT excelle. Les outils axés sur les GML peuvent générer des questions d'entrevue typiques et donner des conseils en vue d'entrevues de nature technique ou propres à une matière. On peut aussi leur demander de jouer le rôle d'une personne qui tient une entrevue ou qui y participe, ou de commenter ou noter la performance donnée dans le cadre de telles simulations de scénario. Un outil connecté à Internet, comme Bing/Copilot, peut aider la population étudiante à chercher des détails sur des entreprises ou des établissements en vue d'une entrevue d'emploi. Les outils axés sur des GML peuvent commenter le curriculum vitæ d'un.e étudiant.e et aider à sa mise en page et à sa correction.

En fin de compte, la population étudiante a besoin de ressources fiables pour l'apprentissage, comme des manuels et des notes de cours, et ChatGPT ne peut pas toujours être considéré comme une ressource fiable. ChatGPT peut jouer un rôle précieux en tant que partenaire ou aide de recherche d'idées (dans la mesure permise par les politiques de cours), mais il ne constitue pas encore une source assez fiable de renseignements factuels. Vasconcelos et Santos (2023) soutiennent que les utilisateur.trice.s peuvent améliorer l'exactitude de ChatGPT en rédigeant des requêtes efficaces :

Le processus de création d'une requête permettant d'obtenir la réponse voulue constitue un aspect fondamental des réflexions et interactions itératives lors de l'usage de dialogueurs ayant recours à l'IAG (dialogueurs IAG) en tant qu'outils d'aide à la réflexion. Cette approche incite les étudiant.e.s à réfléchir de manière critique et à préciser leur compréhension de différents concepts tout en s'interagissant avec l'outil d'IA.

Il s'agit encore toutefois de l'éternelle question de l'œuf ou la poule : la population étudiante ne peut pas apprendre à avoir une pensée critique et à déceler de mauvais renseignements en interagissant seulement avec un outil. Il s'agit d'un processus comptant de nombreuses étapes qui doit se dérouler avec l'aide d'un.e spécialiste. Les étudiant.e.s ne peuvent recourir à la pensée critique que pour des questions pour lesquelles ils et elles possèdent une certaine assise : en demandant à ChatGPT d'expliquer un nouveau concept, ils et elles disposent de peu d'information pour savoir si les renseignements sont véridiques et encore moins pour déterminer si ChatGPT leur offre la meilleure explication.

Voilà pourquoi les étudiant.e.s doivent faire preuve de prudence dans l'utilisation de ChatGPT de manière autonome (c.-à-d. sans supervision ou confirmation reposant sur des instructions) dans le cadre de leurs études.

PARTIE VI

CONCLUSION

En conclusion, l'exploitation des grands modèles de langage (GML), comme ChatGPT, dans l'apprentissage des STIM comporte de nouvelles possibilités et pose des défis distincts. Ces outils axés sur l'IA peuvent bonifier les méthodes d'enseignement traditionnelles, offrant ainsi un accès à une vaste gamme de renseignements et un contexte d'apprentissage interactif. Le personnel enseignant doit cependant composer avec ces limitations de manière prévoyante et concilier l'usage de GML avec des activités axées sur la pensée critique et la résolution de problème afin d'assurer une compréhension globale. Les approches pédagogiques doivent évoluer au rythme de la technologie et s'adapter pour faire en sorte que l'intégration des GML dans les cours de STIM améliore les résultats scolaires au lieu de les diminuer. En définitive, l'incorporation des GML dans l'enseignement des STIM dépendra de notre capacité à amalgamer ces solutions technologiques avancées avec les éléments humains irremplaçables que sont la curiosité, la créativité et l'examen critique.

REMERCIEMENTS

La présente ressource est financée par le gouvernement de l'Ontario. Les points de vue exprimés dans la présente publication sont ceux des personnes qui l'ont rédigée. Ils ne représentent pas nécessairement ceux du gouvernement de l'Ontario.



Attributions de médias

- ecampuslogo-350×90
- ONLogo-350×140