

Multilinguisme et IA au Canada

—

Promesses et limites de l'IA pour la découvrabilité des contenus scientifiques

Marie-Jean Meurs

meurs.marie-jean@uqam.ca

Gaëlle Laperrière

laperriere.gaelle@courrier.uqam.ca

Journée pour la science ouverte au CNRS

25 novembre 2025



Chaire de recherche du Québec
Découvrabilité des contenus
scientifiques en français

Fonds
de recherche
Québec 

CIRST
Centre interuniversitaire
de recherches sur la science
et la technologie

Découvrabilité des contenus scientifiques

Diversité linguistique au Québec et au Canada



Au Québec :

- Le **français**, seule langue officielle
- 84,1% francophones, 14,9% anglophones, 46,4% bilingues Stat. Canada, 2021
- Trois familles de langues autochtones : algonquienne, iroquoienne et eskimo-aléoute (toutes en danger de disparition) BANQ, 2022

Au Canada :

- Une province francophone, le Québec sur 10 provinces et 3 territoires
- Une province officiellement bilingue, le Nouveau-Brunswick
- 22,0% francophones, 76,1% anglophones, 18,0% bilingues Stat. Canada, 2021
- 23,2% autre langue première, plus de 200 langues

Quelles sont les **promesses** et les **limites** de l'IA au service de la découvrabilité des contenus scientifiques ?



Recherche d'information (RI)

1. Rédaction d'une **requête** décrivant l'information cherchée.
2. Sélection par le système d'un ensemble de documents **pertinents**.
3. Classement des documents **par ordre de pertinence** décroissante.

Outils d'IA générative

- ▶ **Génèrent** du texte, des images, en réponse à une **invite de commande**.
- ▶ Utilisent, pour certains, de la RI (génération à enrichissement contextuel(RAG)).

Chercher des informations :

- ▶ maîtriser ses **choix** informationnels
- ▶ maximiser la **diversité** de ses sources
- ▶ conserver son **esprit critique**

⚠ Les outils de génération de contenus ne sont pas conçus pour la recherche d'information

Plateformes de diffusion scientifiques

Internationales, diffusion de contenus principalement **anglophones**



Ex. : Dimensions (GB), Scopus (NL), Ex Libris (Israël), Statista (Allemagne), Web of Science (GB/USA), EBSCO, JSTOR, LexisNexis, Factiva, ProQuest et Google Scholar (USA).

Fonctionnalités utilisant des **outils d'IA** :

- Recherche **multilingue** (moins efficace qu'en anglais)
- Recherche en **langage naturel**
- **Agents conversationnels**
- **Résumé automatique**
- Extraction de **métadonnées**
- **Classification thématique**
- **Statistiques**
- **Autres fonctionnalités spécialisées**
 - ▶ relecture de documents juridiques (LexisNexis)
 - ▶ veille médiatique et économique (Factiva)

Plateformes de diffusion scientifiques

De la francophonie, diffusion de contenus principalement **francophones**



- ▶ **Érudit** (1998, QC) : “Infrastructure numérique [ayant] pour mission de soutenir la publication numérique ouverte et la recherche en sciences humaines et sociales et en arts et lettres.” (multilingue)
- ▶ **HAL** (2001, FR) : "Plateforme en ligne destinée au dépôt et à la diffusion d'articles scientifiques, de thèses ou de rapports techniques. Les publications sont en accès libre, mais pas nécessairement leur utilisation." (multilingue)
- ▶ **Borealis** (2022, CA) : “Dépôt canadien de données de recherche bilingue, multidisciplinaire et sécurisé, soutenu par des bibliothèques universitaires et des établissements de recherche de partout au Canada.”
- ▶ **Cairn.info** (2005, FR) : “Plateforme de référence pour les publications scientifiques francophones, vise à favoriser la découverte d'une recherche de qualité tout en cultivant l'indépendance et la diversité des acteurs de l'écosystème du savoir.”

Plateformes de diffusion scientifiques

Fonctionnalités proposées



Erudit, HAL, Borealis

- ▶ Recherche avancée avec **filtres**
- ▶ Recherche en **texte intégral**
- ▶ Attribution de **DOI** et **permalien**s
- ▶ Formats de **consultation** : PDF, HTML, téléchargement possible
- ▶ Formats de **citation** : BibTeX, RIS, XML, ENW, MLA, APA...
- ▶ Outils d'**alertes** et de **communication** : infolettres, réseaux sociaux...

Cairn.info Rapport d'activités, juin 2025

- ▶ **Architecture** matérielle/logicielle **dédiée à l'IA** et **hébergée localement**
→ déployer des **modèles ouverts** en local
- ▶ **SophIA** (RAG, questions en langage naturel, réponses basées "uniquement" sur les contenus Cairn.info, accès direct aux extraits des publications)
- ▶ **Génération de textes alternatifs** des images hébergées



- ▶ Agents conversationnels
 - recherche plus **naturelle** et **intuitive**
- ▶ Résumé automatique
- ▶ Extraction d'informations
- ▶ Traduction automatique
 - suppression des **exigences linguistiques**
 - diffusion multilingue **simultanée** des contenus publiés
 - facilitation des **échanges scientifiques** inter-lingues
- ▶ Génération automatique de bibliographies
 - formulation de requêtes en langage **naturel**
 - obtention **rapide** des résultats
- ▶ **⚠** Nombreux **risques** et **limites** à considérer pour un déploiement et une utilisation responsables et sécuritaires

Risques et limites de l'IA générative

Fiabilité des résultats



Données d'entraînement, nuances, inférences

- ▶ langues dominantes dans les données d'entraînement “multilingues”
→ faibles performances pour des langues à faible ressources
- ▶ données d'entraînement trop généralistes
→ lacunes de vocabulaire technique
- ▶ données d'entraînement moissonnées générées par IA
→ perte de diversité lexicale
→ dégradation progressive des performances
- ▶ difficulté à restituer les nuances linguistiques et à gérer les négations
→ mauvaise interprétation des résultats
- ▶ difficulté à traduire et traiter les néologismes et concepts émergents
- ▶ invention de contenus

Risques et limites de l'IA générative

Reproduction et amplification des biais



Biais linguistiques et d'indexation

- ▶ référencement **incomplet** des contenus scientifiques
- ▶ prédominance des publications **anglophones**
- ▶ sur-représentation des revues **nord-américaines** et **britanniques**
- ▶ sous-représentation systématique de certaines **disciplines**

Biais sociaux

- ▶ ambiguïtés lexicales anglophones
→ émergence et perpétuation de **biais d'accords**
- ▶ absence de normes établies pour l'écriture inclusive
→ **traduction imprécise** d'écrits scientifiques
- ▶ variations dialectales inégalement représentées
→ **disparition** progressive des **dialectes**

Risques et limites de l'IA générative

Impacts sur les personnes utilisatrices



Perte de compétences

- ▶ délégation excessive menant à la **perte d'indépendance intellectuelle**
- ▶ diminution de l'**esprit critique** et de la **capacité d'analyse** des résultats
- ▶ perte de familiarité avec les **supports primaires de recherche d'information**

Propriété intellectuelle

- ▶ reformulations sans attribution, plagiat, production de citations erronées...
- ▶ manque de **transparence** concernant les données d'entraînement
→ empêche l'évaluation du respect de la propriété intellectuelle

Souveraineté des données

- ▶ dépend du cadre juridique du pays où l'outil est **développé** et **déployé**
→ ne correspond pas toujours aux **standards de protection optimaux**

Risques et limites de l'IA générative

Impacts environnementaux



- Coût en matières premières : cartes graphiques, espace de stockage...
- Coût énergétique : entraînement et utilisation des modèles

 Luccioni *et al.* (2025) **Entraînement = défi environnemental majeur.**

 Patterson *et al.* (2021) Jusqu'à 626 000 kg de CO2 générés lors de l'**entraînement de certains modèles** ($\approx$ 250 vols New York - Pékin).

 Luccioni *et al.* (2024) **Coût énergétique** d'une requête sur ChatGPT jusqu'à 10 fois supérieur à celui d'une recherche sur un moteur de recherche.

Avant de miser massivement sur ces technologies, il est **crucial** d'évaluer rigoureusement leur **efficacité** et leur réelle **valeur ajoutée** pour la science.

Évolution des pratiques de recherche documentaire

Communautés de recherche



 Batista *et al.* (2024) *Communautés étudiantes et enseignantes*

- utilisation variable selon les disciplines
- inquiétudes de plagiat

 Saúde *et al.* (2025) *Communautés étudiantes*

- appréciation des avantages immédiats
- sous-estimation des effets négatifs

 Gracey *et al.* (2025) *Étudiant·e-s aux cycles supérieurs, UNB*

- utilisation pour des tâches spécifiques (révision, remue-méninges)
- méfiances sur la fiabilité des résultats et leurs biais
- inquiétudes de plagiat et de dépendance excessive
- préoccupations éthiques au sujet des données et de la transparence

► **frustration face aux divergences dans le corps professoral**

- absence de politiques institutionnelles claires et cohérentes
- besoin de formations efficaces à ces outils : avantages, limites, usage

Évolution des pratiques de recherche documentaire

Communautés de recherche au Québec



 Conseil sup. de l'éducation et Commission de l'éthique en science et en technologie. (2024)

*“Intelligence artificielle générative en enseignement supérieur :
Enjeux pédagogiques et éthiques”*

- ! délégation du travail des auxiliaires de recherche
→ empêche leur intégration progressive au milieu de recherche

 Données internes de l'université Laval (QC)

- adoption en STM supérieure aux SHS
- utilisation comme “aide à la recherche” pour 65 % des 4 500 étudiant·e·s aux cycles supérieurs
- formations par les bibliothèques très populaires
- engouement du corps professoral à développer des compétences de recherche hybrides chez les communautés étudiantes

Évolution des pratiques de recherche documentaire

Communautés académiques au Québec



- La majorité des universités hébergent des **guides** sur l'IA générative.
→ termes techniques, notions de base, liste des outils et leurs limites...
- La majorité des bibliothèques universitaires offrent des **formations** et **ateliers** pour les communautés étudiantes et le personnel encadrant :
 - ▶ utilisation responsable des outils
 - ▶ bonnes pratiques
 - ▶ exploration critique de leurs promesses et limites
 - ▶ **!** personnel de formation non-expert
- Certaines universités offrent des **abonnements** à des outils.
 - ✗ finance les entreprises plutôt que les établissements publics
 - ✗ fait obstacle à la souveraineté des données universitaires
 - ✗ très peu de transparence car outils commerciaux

i Ministère de l'Enseignement supérieur : recensement des formations disponibles dans les établissements et leurs communautés d'ici **fin 2026**.

Recommandations

Au niveau des universités



- 👍 **Former** les personnes étudiantes, chercheures et enseignantes **aux enjeux de l'usage de l'IA générative** pour la découvrabilité des contenus scientifiques multilingues.
- 👍 Assurer que les formations sont données par des personnes **expertes** dans le domaine de la RI et de l'IA.
- 👍 Renforcer l'apprentissage des pratiques de recherche documentaire auprès de **chercheur·euse·s expérimenté·e·s**.
- 👍 En tant qu'institution de recherche, soutenir ses communautés dans la mise à disposition d'**outils libres installés localement**, afin de préserver la **propriété intellectuelle** et la **sécurité des données**.

Recommandations

Au niveau des plateformes de diffusion



- 👍 Développer l'**interopérabilité des plateformes de diffusion scientifique** (ex. : améliorer l'indexation des revues savantes, avec surveillance de la qualité des transferts de données réalisés vers les agrégateurs de contenus scientifiques).
- 👍 Développer des **outils d'IA** pour une meilleure découvrabilité des contenus scientifiques, autour de trois axes principaux :
 - Outils **sémantiques**, agnostiques à la langue
 - **Traduction des plateformes de diffusion**
 - ciblée en fonction des besoins réels
 - automatique, professionnelle, ou hybride
 - Génération automatique de **métadonnées multilingues**

Adoption **lente** et **mesurée** des fonctionnalités d'IA
⇒ contrôle et efficacité

Recommandations

Aux niveaux national et international



- 👍 Aligner les **politiques de science ouverte** (FRQ, France, autres pays francophones) en les **adaptant** aux défis amenés par les nouvelles technologies.
- 👍 Imposer aux projets financés par des fonds publics des quotas de publication de travaux scientifiques et/ou leurs communications dans la **langue officielle** de nos pays **non-anglophones**.
- 👍 Soutenir la création de **manuels terminologiques** pour l'accroissement de la littératie multilingue afin d'améliorer la compréhension de la **communication savante** habituellement anglophone.
- 👍 Développer des services partagés de **traduction scientifique** ouverte, avec mise en commun d'une **terminologie à jour et validée**, de moteurs de traduction entraînés dans tous les domaines de la science, avec l'accompagnement de **personnes expertes**.



Questions ?