Évaluation de la capacité des outils de l'IA à cartographier efficacement des initiatives pertinentes liées à une thématique d'intérêt

Rapport présenté au Réseau Précrisa

Par

Serine Anaïs Attouche

Stagiaire et étudiante du premier cycle en science biomédicale

22 juillet 2026

1. Contexte du projet

Le Précrisa est un réseau de recherche qui se consacre aux menaces émergentes susceptibles de provoquer de nouvelles crises sanitaires ainsi qu'à la capacité des systèmes de santé, des individus et des communautés à y faire face. Pour répondre à cette mission, le Précrisa mobilise une communauté interdisciplinaire et intersectorielle afin d'aligner les connaissances, les données et les expertises avec les besoins des décideurs, des citoyens et des acteurs de la gestion de crise, tout en favorisant une action agile face aux nouvelles menaces.

Dans cette perspective, le Précrisa souhaite développer de manière agile des outils pour cartographier les initiatives de recherche et d'intervention pertinentes en cours au Québec. Ces outils visent notamment à :

- Mobiliser rapidement les connaissances disponibles ;
- Renforcer les synergies entre les acteurs concernés ;
- Optimiser les efforts collectifs de recherche.

Une première https://precrisa.ca/nouvelles/boussole-interactive/boussole en lien avec la grippe aviaire a été élaborée à l'aide de *ChatGPT, Microsoft Copilot* et *Gemini*. Malgré l'utilisation de requêtes détaillées (*prompts*), les résultats générés demeuraient très limités. La démarche a donc été complétée par une approche itérative, consistant à identifier progressivement de nouvelles initiatives, organisations et experts à partir des informations déjà recueillies. Les résultats obtenus ont ensuite été validés auprès d'acteurs clés du domaine afin de corriger les erreurs, combler les lacunes et confirmer l'exactitude des informations recueillies. Une mise à jour fréquente est réalisée depuis les informations partagées par des membres du réseau et autres acteurs clés. La démarche déployée pour la création de cette première boussole s'est avérée particulièrement chronophage.

Cette contrainte est d'autant plus importante que le Précrisa doit produire plusieurs boussoles adaptées à différentes thématiques ou menaces émergentes. À titre d'exemple, le réseau est actuellement mobilisé autour des enjeux géopolitiques et leurs répercussions potentielles sur la santé humaine, animale et environnementale, dans une approche Une seule santé.

Afin de répondre à ce besoin, une réflexion a été amorcée sur la pertinence de recourir à des outils appuyés par l'intelligence artificielle (IA) afin de repérer, analyser et synthétiser les connaissances. L'hypothèse sous-jacente est que ces outils pourraient faciliter l'identification rapide d'initiatives de recherche, d'interventions, de produits de connaissances, d'acteurs clés et de domaines d'expertise, pertinents aux enjeux prioritaires traités par le Précrisa.

Dans ce contexte, il importe de distinguer deux catégories d'outils d'intelligence artificielle qui seront étudiées :

- **Les plateformes d'IA générative.** Elles produisent du contenu à partir de modèles de langage entraînés sur de vastes ensembles de données. Elles peuvent répondre à des questions, résumer des informations ou générer du texte, mais ne consultent pas nécessairement le Web en temps réel. (Božić & Poola, 2023)
- **Les moteurs de recherche appuyés par l'IA.** Ils combinent les capacités de recherche sur le Web avec celles de l'IA générative. Ils sont capables d'identifier des sources pertinentes, d'en extraire les informations clés et de produire des synthèses documentées en citant les références utilisées. (Plante, 2024; HubSpot, s.d.; Rojas, 2025)

L'évaluation du potentiel de ces outils pourrait permettre de déterminer dans quelle mesure ils peuvent soutenir la production de boussoles thématiques au sein du Précrisa.

2. Objectifs du projet

Objectif général

Évaluer la capacité des outils d'IA à cartographier efficacement des initiatives pertinentes liées à une thématique d'intérêt.

Objectifs spécifiques

- Identifier des moteurs de recherche basés sur l'IA en comparant leurs caractéristiques, avantages et limites.
- Élaborer une requête détaillée en lien avec un sujet d'intérêt actuel au Précrisa, en s'appuyant sur les bonnes pratiques.
- Tester et comparer l'efficacité dans le repérage d'initiatives pertinentes des différents outils utilisés.

3. Méthodologie

Cette étude repose sur l'hypothèse suivante : les moteurs de recherche utilisant l'intelligence artificielle générative seraient en mesure d'identifier un grand nombre d'initiatives pertinentes à partir d'une requête ciblée. Toutefois, nous supposons que la qualité et la fiabilité des résultats varieraient selon les outils utilisés. Enfin, nous faisons l'hypothèse qu'il serait possible, à l'issue de l'expérimentation, d'identifier un moteur de recherche offrant les meilleures performances et pouvant être privilégié pour ce type de recherche documentaire.

3.1 Identification des moteurs de recherches basés sur l'IA

Dans un premier temps, il a été nécessaire de délimiter le nombre de moteurs de recherche basés sur l'intelligence artificielle qui seraient évalués dans le cadre de cette étude. Le choix s'est porté sur trois plateformes, soit *ChatGPT, Perplexity* et *Consensus*. Ce nombre permet d'obtenir un éventail suffisamment diversifié d'outils pour réaliser une comparaison pertinente, tout en conservant une démarche méthodologique réalisable et reproductible.

La sélection de ces plateformes repose sur leurs approches complémentaires de la recherche d'information.

ChatGPT est un modèle de langage fondé sur l'architecture GPT (Generative Pre-trained Transformer), conçu pour comprendre et générer du texte en langage naturel. Entraîné sur de vastes corpus de données textuelles, il est capable de répondre à des questions complexes, de synthétiser de l'information, de rédiger du contenu et d'assister diverses tâches de recherche et d'analyse (Božić & Poola, 2023; ChatGPT, 2026).

Basé sur des approches algorithmiques avancées, ce modèle peut être adapté à différents domaines et contextes d'utilisation. Toutefois, plusieurs limites sont reconnues dans la littérature et par ses concepteurs (Božić & Poola, 2023; ChatGPT, 2026). Ces limites comprennent notamment le risque d'hallucinations, la présence de biais et la production occasionnelle d'informations inexactes, ce qui nécessite une validation systématique des résultats (Božić & Poola, 2023). Pour cette raison, les résultats obtenus à l'aide de *ChatGPT* doivent systématiquement faire l'objet d'une validation à partir de sources fiables.

Dans le cadre de cette démarche, *ChatGPT* a été utilisé lors d'une phase exploratoire afin d'évaluer sa capacité à soulever des pistes informationnelles pertinentes et à structurer l'information de façon cohérente. En effet, ce grand modèle de langage (LLM) facilite la formulation des requêtes et permet d'obtenir une vision préliminaire des réponses attendues. Cependant, il reste défini comme un outil de première approche. En contexte de recherche documentaire, ses biais sont particulièrement critiques, ce qui impose une vérification systématique et rigoureuse de chaque résultat.

Perplexity est un moteur de recherche assisté par l'intelligence artificielle qui combine la recherche documentaire en ligne et la génération de texte. L'outil s'appuie sur différents grands modèles de langage afin de produire des réponses synthétiques à partir d'informations recueillies sur le Web (Rojas, 2025; Plante, 2024; HubSpot, s.d.).

Sa principale valeur ajoutée réside dans sa capacité à associer ses réponses à des sources explicites et consultables. Chaque requête génère une liste de références permettant de retracer l'origine des informations présentées, ce qui facilite les processus de validation et de vérification. La qualité des résultats dépend toutefois de la précision des requêtes formulées et de la disponibilité des sources pertinentes.

Bien que certaines réponses puissent occasionnellement simplifier ou généraliser des phénomènes complexes, *Perplexity* offre généralement un bon équilibre entre rapidité de recherche, synthèse de l'information et traçabilité des sources. L'outil permet également d'adapter les résultats selon différents contextes d'utilisation, notamment la recherche académique, la recherche ciblée ou la recherche générale (Rojas, 2025).

Consensus est un moteur de recherche spécialisé dans l'exploration et l'analyse de publications scientifiques et académiques. Contrairement aux outils généraux d'intelligence artificielle, il se concentre principalement sur la littérature scientifique évaluée par les pairs (Bibliothèque de la Haute école de santé de Genève, 2025; University of Maryland Health Sciences & Human Services Library, 2026).

L'outil met particulièrement l'accent sur :

- La qualité scientifique des sources ;
- Le nombre de citations reçues par les publications ;
- L'impact académique des travaux recensés ;
- Le niveau de consensus ou de divergence observé dans la littérature scientifique.

Consensus fournit ainsi des indicateurs permettant d'évaluer la crédibilité, la robustesse et l'influence des études identifiées. Cette approche facilite le repérage rapide des connaissances scientifiques les plus établies tout en permettant de distinguer les résultats fortement soutenus par la littérature de ceux qui demeurent plus exploratoires ou controversés (Bibliothèque de la Haute école de santé de Genève, 2025).

Dans le cadre d'une démarche de veille stratégique ou de revue rapide de littérature, Consensus constitue un outil complémentaire aux moteurs de recherche traditionnels et aux modèles de langage, particulièrement lorsque l'objectif est d'appuyer l'analyse sur des données scientifiques validées.

Ces outils ont été retenus en raison de leur popularité, de leur accessibilité et de leurs modes distincts d'accès à l'information. Leur comparaison permet d'évaluer les avantages et les limites

de différentes approches de recherche dans le contexte du repérage d'initiatives.D'autres outils d'intelligence artificielle générative, tels que *Gemini, Claude* ou *Microsoft Copilot,* n'ont pas été retenus dans cette expérimentation. Le choix de limiter l'analyse à trois plateformes visait à assurer la faisabilité de l'étude tout en couvrant des approches de recherche suffisamment différentes. Cette sélection pourra être élargie lors de travaux ultérieurs afin d'inclure d'autres modèles et d'évaluer l'évolution de leurs performances.

Tableau 1: Comparatifs des modèles IA utilisés

Outil	Description générale	Forces principales	Limites principales
ChatGPT	Modèle de langage fondé sur l'architecture GPT (*Generative Pre-trained Transformer*) permettant de générer, structurer et synthétiser de l'information en langage naturel.	Capacité de synthèse, reformulation, structuration de l'information et génération rapide de pistes de recherche.	Possibilité de produire des informations inexactes, biaisées ou insuffisamment vérifiées ; risque d'hallucination de sources ou de références.
Perplexity	Moteur de recherche assisté par l'intelligence artificielle combinant recherche documentaire et génération de texte à partir de sources accessibles en ligne.	Référencement explicite des sources, accès rapide à une grande variété de documents et capacité à synthétiser l'information recueillie.	Qualité variable selon les sources disponibles et les requêtes formulées ; tendance occasionnelle à simplifier certains enjeux complexes. Limite de sources possibles.
Consensus	Moteur de recherche assisté par l'intelligence artificielle spécialisé dans l'exploration et l'analyse de la littérature scientifique et académique.	Accent mis sur les publications évaluées par les pairs, indicateurs de qualité scientifique et mise en évidence des travaux les plus influents.	Couverture limitée aux contenus scientifiques ; moins adapté au repérage d'initiatives institutionnelles, communautaires ou gouvernementales. Limite de sources possibles.

3.2 La requête : définition et bonnes pratiques

Une requête est une instruction ou un ensemble d'instructions fournies à un modèle de langage (Large Language Model, LLM) afin d'orienter la génération de réponses. La qualité d'une requête influence directement la pertinence, la précision et l'utilité des résultats obtenus. Une requête bien conçue permet notamment de réduire le nombre d'itérations nécessaires, d'améliorer la qualité des réponses et de gagner du temps dans les processus de recherche et d'analyse (Soares, 2024; Gheorghiu, 2023; Phoenix & Taylor, 2024).

Bien qu'une requête initiale puisse fournir des résultats satisfaisants, il est souvent nécessaire de le perfectionner de manière itérative afin d'obtenir des réponses plus précises et adaptées aux besoins de l'utilisateur (Soares, 2024). La littérature et les guides de bonnes pratiques soulignent généralement quatre composantes clés d'une requête efficace (Phoenix & Taylor, 2024; Gheorghiu, 2023) :

1. Le format attendu

La requête doit préciser la forme sous laquelle la réponse doit être présentée. Cela peut inclure :

- La structure de la réponse (tableau, liste, texte synthèse, fiche descriptive, etc.) ;
- Le niveau de détail souhaité ;
- La longueur attendue de la réponse ;
- Les éléments visuels ou organisationnels facilitant la lecture.

2. Le contexte

Le contexte permet au modèle de mieux comprendre la situation dans laquelle la tâche s'inscrit. Il peut notamment inclure :

- Les objectifs de la démarche ;
- Le profil de l'organisation ou du projet ;
- Le public cible ;
- Les connaissances préalables à considérer ;
- Les limites ou contraintes à prendre en compte ;
- Le ton ou le style de rédaction attendu.

3. La tâche

La tâche correspond à l'action demandée au modèle. Elle doit être formulée de manière précise en utilisant des verbes d'action explicites (identifier, comparer, classer, synthétiser, cartographier, analyser, etc.). La requête devrait également préciser :

- Le résultat attendu (numérique, qualitatif) ;
- Les critères de sélection ou d'évaluation ;
- L'objectif poursuivi : action spécifique à effectuer afin de répondre à la consigne ;

4. Le rôle

Le rôle consiste à demander au modèle d'adopter une expertise ou une perspective particulière afin d'orienter son raisonnement. Par exemple, le modèle peut être invité à répondre comme un chercheur en santé publique, un spécialiste de la veille stratégique, un analyste des politiques publiques ou un expert en géopolitique de la santé.

L'intégration de ces quatre dimensions contribue généralement à améliorer la qualité, la cohérence et la pertinence des réponses produites par les modèles d'intelligence artificielle générative.

Dans le cadre du projet, dans un premier temps, une question d'intérêt a été formulée de manière générale (Annexe 1), avant d'être rectifiée et détaillée selon les bonnes pratiques identifiées (Annexe 2).

3.3 Démarche de comparaison des outils

Afin d'évaluer la pertinence des différents outils d'intelligence artificielle pour le repérage d'initiatives liées à une thématique donnée, une même requête a été soumise aux trois plateformes choisies (*ChatGPT, Perplexity* et *Consensus*), en utilisant les migrations et les déplacements de population comme étude de cas. Ce thème, retenu en raison de son importance dans les travaux en cours au Précrisa, constitue un exemple représentatif d'une problématique située à l'interface des enjeux géopolitiques et de la santé. L'objectif était de comparer la capacité de ces outils à repérer, organiser et présenter des initiatives pertinentes en cours au Québec.

Cette démarche s'appuie sur les principes de l'évaluation comparative des systèmes de recherche d'information (Information Retrieval Evaluation), selon lesquels la performance d'un outil est appréciée à partir de critères tels que la pertinence, la précision, le rappel, la couverture de l'information, l'actualité des résultats et la qualité des sources mobilisées (CoopIST, s.d.;

Harvard Division of Continuing Education, s.d.). Dans le contexte de cette étude, il ne s'agissait pas d'évaluer les performances techniques des modèles d'intelligence artificielle, mais plutôt leur capacité à soutenir une démarche de veille documentaire et de repérage d'initiatives dans un domaine complexe et multidisciplinaire. Les réponses obtenues ont donc été analysées à l'aide de critères prédéfinis afin de mettre en évidence les forces, les limites et la complémentarité de chaque plateforme.

Cette comparaison vise également à développer une démarche méthodologique reproductible pouvant être appliquée à différentes thématiques d'intérêt du Précrisa. L'objectif n'est pas de désigner un outil unique comme étant le plus performant, mais plutôt de déterminer dans quels contextes chaque plateforme offre la plus grande valeur ajoutée et dans quelle mesure leur utilisation combinée permet d'améliorer le repérage des initiatives.

Pour limiter les biais liés aux variations des réponses générées par les modèles d'IA, deux personnes réaliseront indépendamment les mêmes requêtes sur chacune des trois plateformes. Les résultats seront ensuite évalués à l'aide d'une grille d'analyse comparative portant sur les critères suivants :

- **Pertinence des résultats** : Utilité des données pour cartographier les acteurs et initiatives selon la thématique d'intérêt.
- **Actualité des informations :** Capacité à fournir des données récentes et à jour.
- **Richesse et fiabilité des sources :** Évaluation du volume, de la diversité (scientifique, institutionnelle, gouvernementale) et de la traçabilité des sources citées.
- **Structure et clarté de l'information :** Qualité de la présentation, hiérarchisation, et facilité de repérage des données.
- **Qualité des listes et vocabulaires :** Aptitude à identifier et structurer les experts, institutions, mots-clés et réseaux pertinents.
- **Fonctionnalités de recherche :** Présence d'options avancées (filtres, tri par langue/période, historique, sauvegarde des requêtes).
- **Transparence et éthique (IA responsable) :** Clarté sur les limites et les sources, respect de la vie privée, réduction des biais et utilisation éthique des données.
- **Compréhension contextuelle :** Aptitude à interpréter les requêtes complexes ou conversationnelles et à saisir l'intention de l'utilisateur.
- **Qualité de la synthèse :** Capacité de l'outil à fournir un résumé clair et structuré.
- **Interactivité :** Capacité à relancer et à poursuivre la discussion de manière fluide.

Les critères d'évaluation ont été définis à la suite d'une réflexion conjointe visant à sélectionner des critères applicables à l'ensemble des outils évalués, tout en répondant aux objectifs du projet. Ils ont été choisis de manière à permettre une comparaison homogène des plateformes, en tenant compte à la fois de leurs fonctionnalités et de la qualité des résultats obtenus à la suite de l'exécution de la requête. Ces critères visent ainsi à évaluer la capacité des outils à produire des réponses aux besoins du recensement des initiatives.

Les critères d'évaluation seront cotés selon l'échelle suivante : XX = très satisfaisant; X = satisfaisant; O = peu satisfaisant; N/E = non évalué; N/A = non applicable. Les analyses individuelles feront ensuite l'objet d'une discussion de synthèse afin de dégager un consensus sur les forces et les limites de chaque plateforme.

À plus long terme, cette démarche devrait mener à l'élaboration d'un protocole standardisé d'utilisation des outils d'intelligence artificielle pour la production des boussoles thématiques du Précrisa. Ce protocole pourra être mis à jour en fonction de l'évolution des plateformes, de l'ajout de nouvelles fonctionnalités et des résultats des évaluations futures, afin d'assurer une utilisation rigoureuse, reproductible et adaptée aux besoins de recherche documentaire ou veille stratégique du réseau.

4.Résultats

Les résultats obtenus à la suite de la requête figurent dans les **Annexes 3, 4 et 5.** Il est intéressant de constater que ceux-ci varient non seulement d'un outil d'IA à l'autre, mais également pour un même outil selon la personne qui soumet la requête. Une analyse comparative, d'abord réalisée séparément par deux évaluatrices, pour les trois outils d'IA, s'appuie sur les critères d'évaluation établis puis discutés conjointement. Le tableau 2 présente le consensus qui en découle.

Tableau 2. Analyse comparative des trois outils d'IA issue du consensus

Critère d'évaluation	ChatGPT	Perplexity	Consensus
Pertinence des résultats	X	X	X
Actualité des informations	XX	XX	XX
Richesse et fiabilité des sources	X	XX	XX
Structure et clarté de l'information	XX	XX	XX
Qualité des listes et vocabulaire	X	XX	XX
Fonctionnalités de recherche	N/E	N/E	N/E
Transparence et éthique (IA responsable)	NA	X	XX
Compréhension contextuelle	X	XX	XX
Qualité de la synthèse	XX	XX	XX
Interactivité	X	O	XX

Analyse des résultats

Tel qu'observé, les résultats obtenus peuvent varier selon la personne demandant la requête ainsi que selon le contexte d'utilisation du modèle d'IA. En effet, les modèles d'intelligence artificielle conversationnelle adaptent, dans une certaine mesure, leurs réponses en fonction de l'historique des interactions, des paramètres de personnalisation ou encore du mode d'utilisation (première utilisation, compte réinitialisé, utilisateur régulier, etc.). Ainsi, un évaluateur ayant réinitialisé son compte *ChatGPT* avant d'effectuer la recherche peut obtenir

une réponse plus « neutre », tant sur le plan du contenu que de la mise en forme, tandis qu'un utilisateur ayant un historique d'échanges avec le modèle est susceptible d'obtenir des réponses influencées par ses interactions précédentes et son profil d'utilisation. Une variabilité similaire a également été observée avec *Perplexity*, où les résultats diffèrent selon la disposition ou non d'un compte ainsi que son niveau d'utilisation de la plateforme. Dans le cas de *Consensus*, les options de recherche sélectionnées ainsi que la fréquence d'utilisation du service semblent également avoir une incidence sur les résultats obtenus.

Par conséquent, la grille comparative ne peut être considérée comme un outil d'évaluation suffisant à elle seule. Afin de limiter l'influence de ces variations et d'assurer une évaluation la plus objective possible, une rencontre entre les deux personnes évaluatrices a été réalisée pour comparer les résultats obtenus, discuter de chaque critère d'évaluation et parvenir à un consensus quant à la notation attribuée à chacun des trois modèles d'IA étudiés.

En ce qui concerne *ChatGPT*, le critère de pertinence des résultats est évalué comme satisfaisant. Le modèle a permis d'identifier entre neuf et dix initiatives, selon l'évaluateur ayant effectué la requête. Toutefois, ces résultats demeurent insuffisants et manquent de profondeur pour réaliser une cartographie complète des initiatives, qui constituait pourtant l'objectif principal de la recherche. En revanche, les initiatives recensées étaient généralement récentes, soit toujours en cours, soit comprises dans la période visée par la requête. À cet égard, le critère de l'actualité des informations est considéré comme très satisfaisant. Concernant le nombre, la diversité et la qualité des sources citées, *ChatGPT* semble avoir mobilisé plusieurs dizaines de sources issues de la littérature scientifique et de la littérature grise. Toutefois, seulement une dizaine de ces sources étaient explicitement citées. De plus, les références fournies manquaient souvent de précision : les hyperliens dirigeaient fréquemment vers la page d'accueil d'un site Web plutôt que vers le document ou l'article utilisé, ce qui compliquait la vérification des informations. L'organisation et la structuration de l'information sont jugées satisfaisantes, bien qu'elles aient varié d'une personne à l'autre. Selon les essais, le modèle pouvait utiliser des éléments de mise en forme plus ou moins appropriés, comme des émojis, ce qui réduisait le caractère neutre de la présentation. En revanche, les informations étaient généralement organisées selon le format demandé dans la requête, avec une structure claire, des tableaux et des listes à puces facilitant la lecture. Les synthèses produites étaient toutefois parfois trop longues. La qualité des listes et du vocabulaire produit est également jugée satisfaisante. Dans plusieurs essais, les personnes responsables ou les experts associés aux

initiatives n'étaient pas systématiquement identifiés, alors que, dans d'autres, le modèle produisait une liste beaucoup plus complète des personnes-ressources. Cette variabilité a limité la constance des résultats. Les fonctionnalités de recherche n'ont pas été évaluées, puisqu'elles ne s'appliquent pas à ce type de modèle conversationnel. En ce qui concerne la transparence, la traçabilité et les considérations éthiques, *ChatGPT* fournit peu d'informations sur son processus de recherche ou de sélection des sources. À quelques reprises, le modèle a toutefois indiqué que les références devaient être vérifiées, sans fournir davantage d'explications sur leur provenance. La capacité de compréhension contextuelle s'est révélée variable selon les évaluateurs. Dans certains cas, le modèle a compris immédiatement la requête et produit les résultats attendus dès la première réponse. Dans d'autres, il a demandé des précisions ou proposé une reformulation avant de poursuivre la recherche. Enfin, la synthèse produite est jugée satisfaisante. Le modèle est parvenu à mettre en évidence les initiatives les plus pertinentes et à en résumer les principaux éléments. En revanche, les pistes de relance proposées pour approfondir la recherche étaient limitées, voire inexistantes selon les essais réalisés.

Pour *Perplexity,* le critère de pertinence des résultats est jugé satisfaisant. Le modèle a permis d'identifier environ sept initiatives pertinentes, dont plusieurs étaient directement liées au contexte québécois. Toutefois, il n'a approfondi que trois de ces initiatives, ce qui limite la capacité à dresser une cartographie exhaustive des acteurs et des projets. Malgré une bonne capacité de synthèse, les résultats demeurent insuffisants pour répondre pleinement à l'objectif de la requête. En ce qui concerne l'actualité des informations, *Perplexity* obtient une très bonne évaluation. La majorité des initiatives recensées sont récentes ; toutefois, une initiative datant de 2021 a été intégrée aux résultats, bien qu'elle se situe à l'extérieur de la période ciblée (2023-2026). Le critère portant sur le nombre, la diversité et la qualité des sources citées est évalué comme très satisfaisant. *Perplexity* cite systématiquement les sources associées aux affirmations présentées, ce qui facilite la traçabilité des informations. Les références proviennent de sources variées et crédibles. Néanmoins, le modèle mentionne lui-même que certaines informations demeurent partielles et nécessitent une validation complémentaire, ce qui témoigne d'une certaine transparence quant aux limites de ses résultats. L'organisation et la structuration de l'information sont jugées très satisfaisantes. Les résultats sont présentés de façon concise et structurée, principalement au moyen d'un tableau synthèse accompagné d'un court texte explicatif. Cette présentation favorise une lecture rapide, même si certaines informations sont parfois trop condensées, au détriment du niveau de détail. La qualité des

listes et du vocabulaire produit est également évaluée comme satisfaisante. Plutôt que de simplement énumérer les thèmes abordés, *Perplexity* produit une synthèse des thématiques émergentes. Toutefois, cette approche plus synthétique entraîne parfois une perte de précision, notamment en ce qui concerne l'identification des personnes responsables ou des experts associés aux initiatives. Les fonctionnalités de recherche n'ont pas été évaluées dans le cadre de cette étude comparative. Concernant la transparence, la traçabilité et les considérations éthiques, le modèle indique explicitement le niveau de confiance associé à ses réponses et précise lorsque certaines informations sont incomplètes ou nécessitent une validation. Toutefois, ces indications demeurent relativement sommaires et ne sont pas toujours accompagnées d'une justification détaillée. La capacité de compréhension contextuelle est jugée satisfaisante. Le modèle interprète correctement les requêtes complexes sans retard et ne demande pas de précision supplémentaire. Enfin, la synthèse produite est jugée très satisfaisante grâce à sa concision et à sa bonne structuration. En revanche, les pistes de relance proposées se sont révélées peu pertinentes au regard de l'objectif de la recherche, ce qui explique la faible évaluation attribuée à ce critère.

Le critère de pertinence des résultats pour *Consensus* est jugé satisfaisant. Le modèle a permis d'identifier environ sept initiatives pertinentes et adopte une approche différente des autres outils en organisant les résultats selon leur niveau de pertinence et en explicitant les critères ayant conduit à leur sélection. Cependant, comme tous les autres modèles, les résultats demeurent insuffisants pour établir une cartographie complète des initiatives québécoises. Le critère de l'actualité des informations est évalué comme satisfaisant. Toutefois, les résultats reposent essentiellement sur des publications scientifiques déjà disponibles. Cette approche permet d'obtenir des informations récentes, mais limite le repérage d'initiatives en continuité ou encore peu documentées, telles que des annonces de financement ou des projets nouvellement lancés. Le nombre, la diversité et la qualité des sources citées constituent l'un des principaux points forts de *Consensus*. L'outil explique avoir analysé un très grand nombre de publications (2 millions) avant de sélectionner les plus pertinentes. Les références proviennent de la littérature scientifique francophone et anglophone et sont accompagnées d'indications facilitant la vérification des passages utilisés. Toutefois, certains liens ou attributions se sont révélés difficiles à interpréter ou ne correspondaient pas toujours clairement aux responsables cités, ce qui a limité la note accordée à ce critère. L'organisation et la structuration de l'information sont jugées très satisfaisantes. Les résultats sont présentés dans une structure claire et hiérarchisée, accompagnés de tableaux facilitant la comparaison des

initiatives et le transfert de l'information vers d'autres documents. La qualité des listes et du vocabulaire produit est également satisfaisante. Consensus organise les initiatives selon plusieurs dimensions directement liées au domaine étudié, notamment la santé humaine, la santé animale, l'environnement et la préparation au risque. Toutefois, comme les références utilisent fréquemment la mention « et coll. », l'identification complète des auteurs ou des personnes responsables demeure limitée. Les fonctionnalités de recherche n'ont pas été évaluées dans cette étude comparative. Il convient néanmoins de souligner que *Consensus* offre plusieurs filtres avancés permettant de raffiner les recherches selon différents critères, notamment l'année de publication, le pays, le domaine de recherche ou la méthodologie des études. En matière de transparence, de traçabilité et de considérations éthiques, il se distingue par la présentation détaillée de sa démarche de recherche. L'outil indique le nombre de publications analysées, les étapes de sélection des références ainsi que le niveau de confiance accordé aux résultats, ce qui favorise une meilleure compréhension du processus de génération des réponses. La capacité de compréhension contextuelle est jugée très satisfaisante. Le modèle interprète correctement l'objectif scientifique de la requête et organise spontanément les résultats selon les principaux axes du domaine étudié. Enfin, la synthèse produite est jugée très satisfaisante. Les informations sont clairement hiérarchisées, les critères de sélection sont explicités et les résultats sont faciles à interpréter. Les pistes de relance proposées permettent également d'approfondir efficacement la recherche.

5. Conclusion

Dans l'ensemble, cette analyse met en évidence qu'aucun des trois modèles d'intelligence artificielle évalués ne permet, à lui seul, de répondre entièrement aux objectifs de la recherche. Chacun présente des forces et des limites qui lui sont propres. *ChatGPT* se distingue par sa capacité à générer une synthèse structurée et à identifier des initiatives récentes, mais ses sources demeurent parfois imprécises et les résultats varient selon le contexte d'utilisation. *Perplexity* offre une excellente traçabilité des références et une bonne compréhension contextuelle, mais certaines réponses demeurent incomplètes ou orientées vers un seul aspect de la requête, comme l'approche Une seule santé. Enfin, *Consensus* se démarque par la rigueur de sa méthodologie, la qualité de sa documentation scientifique et la transparence de son processus de sélection des sources, bien qu'il soit limité pour repérer les initiatives les plus récentes, puisqu'il repose principalement sur la littérature déjà publiée.

Ces constats suggèrent qu'il est préférable de combiner les résultats produits par les trois modèles plutôt que de s'appuyer sur un seul outil. Cette approche permet d'obtenir un portrait plus complet des initiatives recensées en tirant profit des forces propres à chacun. Toutefois, les résultats générés par les modèles d'IA ne peuvent être considérés comme définitifs. Ils doivent être validés par des experts du domaine et faire l'objet d'une vérification systématique des sources, afin de confirmer l'exactitude des informations et d'écarter les initiatives qui ne répondent pas réellement aux objectifs de la recherche.

Par ailleurs, le faible nombre d'initiatives recensées pourrait s'expliquer en partie par la nature du sujet étudié. Les initiatives québécoises portant spécifiquement sur les migrations climatiques demeurent relativement peu nombreuses et souvent peu documentées, ce qui limite inévitablement les résultats obtenus par les moteurs d'IA.

Dans le cadre d'une seconde phase d'évaluation, plusieurs améliorations méthodologiques pourraient être apportées. D'abord, il serait pertinent de reproduire l'exercice sur un sujet davantage documenté au Québec afin de mieux comparer les performances des différents modèles dans un contexte où les informations sont plus abondantes. Ensuite, les requêtes devraient être exécutées simultanément par au moins deux personnes en utilisant les mêmes conditions expérimentales. Dans la mesure du possible, ces personnes devraient utiliser des comptes présentant un historique d'utilisation comparable, ou des comptes nouvellement créés et configurés de façon identique, afin de limiter l'influence de la personnalisation des modèles sur les résultats obtenus. Une comparaison systématique des réponses permettrait de mieux mesurer la variabilité entre les utilisateurs et d'accroître la reproductibilité de l'évaluation. Enfin, il serait pertinent de répéter l'exercice en ajustant progressivement les requêtes afin d'évaluer l'influence de leur formulation sur les performances et les résultats obtenus.

Cette réflexion s'inscrit dans un contexte plus large qui soulève plusieurs enjeux, notamment le coût d'accès aux modèles les plus performants, la reproductibilité des résultats, ainsi que les impacts environnementaux et les considérations éthiques liés à l'utilisation croissante de l'intelligence artificielle en recherche.

Bibliographie

Bibliothèque de la Haute école de santé de Genève. (2025). Consensus : guide d'utilisation. Haute école de santé Genève. https://www.hesge.ch/heds/sites/heds/files/2025-05/Consensus_guide_utilisation.pdf

Božić, V., & Poola, I. (2023). Chat GPT and education. Preprint, 10.

ChatGPT. (2026, mai 23). Wikipédia, l'encyclopédie libre. Page consultée le 11: 29, mai 23, 2026 à partir de http://fr.wikipedia.org/w/index.php?title=ChatGPT&oldid=236448877.

Plante, P. (2024, 19 mars). Un moteur de « réponses » survolté par l'IA. Collimateur - Veille pédagonumérique - UQAM. https://collimateur.uqam.ca/collimateur/un-moteur-de-reponses-survolte-par-lia/

Rojas, C. (2025, Feb 02). Perplexity, la inteligencia artificial a medio camino entre google y wikipedia. Actualidad Economica, 25. Retrieved from https://acces.bibl.ulaval.ca/login?url=https://www.proquest.com/magazines/perplexity-la-inteligencia-artificial-medio/docview/3162427853/se-2

University of Maryland Health Sciences & Human Services Library. (2026, 20 avril). Accuracy and limitations - Consensus. Guides at University of Maryland Health Sciences & Human Services Library. https://guides.hshsl.umaryland.edu/consensus/accuracy

https://blog.hubspot.fr/marketing/moteurs-de-recherche-ia

https://professional.dce.harvard.edu/blog/building-a-responsible-ai-framework-5-key-principles-for-organizations/#What-Are-the-5-Key-Principles-of-Ethical-AI-for-Organizations

https://coop-ist.cirad.fr/trouver-l-information/utiliser-des-moteurs-de-recherche/2-criteres-pour-evaluer-un-moteur-de-recherche

Soares, L. (2024). Prompt engineering deep dive : from prototyping to designing prompt engineering experiments. ([First edition]). O'Reilly Media, Inc. https://www.oreilly.com/library/view/-/0790145740359/

Gheorghiu, A. (2023). Prompt engineering for everybody with ChatGPT and GPT4. ([First edition]). Packt Publishing. https://www.oreilly.com/library/view/-/9781805122005/

Phoenix, J., & Taylor, M. (2024). Prompt engineering for generative AI : future-proof inputs for reliable AI outputs at scale (First edition). O'Reilly Media, Inc. https://www.oreilly.com/library/view/-/9781098153427/

Annexe 1

Premier essai de requête :

“Svp nous fournir les initiatives et travaux de recherche en cours — actuels ou datant de 2023 — pertinents au contexte du Québec et liés à ces thématiques, qui découlent directement ou indirectement des enjeux géopolitiques prioritaires.”

Annexe 2

La requête utilisée durant le deuxième essai:

« **Rôle**

Agis comme un analyste spécialisé en veille stratégique, en santé publique et en recherche intersectorielle dans une perspective Une seule santé.

Contexte

Le Précrisa souhaite cartographier les initiatives en cours au Québec liées aux déplacements, migrations de masse ou aux migrations climatiques, particulièrement lorsque ces phénomènes ont des répercussions sur la santé humaine, animale ou environnementale. L'objectif est d'identifier les initiatives pertinentes pour coordonner la recherche, mobiliser les connaissances disponibles et en produire de nouvelles ainsi que soutenir la préparation et la réponse aux menaces émergentes au Québec.

Tâche

Recense les initiatives:

- projets et programmes de recherche,
- programmes, stratégies, politiques publiques, plans d'action ou mesure gouvernementales,
- initiatives régionale, municipale, communautaire ou citoyenne
- avis d’experts, cadres de référence, lignes directrices, recommandations, rapports, mémoires ou autres documents pertinents,

- outils, plateformes, ressources, guides, formations, applications ou produits d’information ou de sensibilisation
- groupes de recherche, observatoires, chaires, réseaux, centres d'expertise, consortiums ou collaborations

Qui sont :

1. En cours ou réalisés au cours des trois dernières années ;
2. Menés au Québec ou ayant des activités, des collaborations ou des retombées significatives pour le Québec ;
3. Liés aux migrations de masse, aux déplacements de populations ou aux migrations climatiques ;
4. Pertinents pour les enjeux de santé, de services sociaux, d'environnement, de sécurité, de préparation aux crises ou de résilience des communautés.

Pour chaque initiative identifiée, précise :

- Le nom de l'initiative ; L'organisation ou principales personnes responsables et leurs affiliations ;
- Les partenaires impliqués (personnes ou organisations) ;
- La période de réalisation ;
- Les objectifs du projet ;
- Les principaux thèmes abordés ;
- Les sources et montants de financement si applicable ;
- Les liens potentiels avec la santé humaine, animale ou environnementale ;
- Le statut du projet (en cours, récemment terminé, financement annoncé, etc.) ;
- Les sources utilisées pour documenter l'information.

Format attendu

Présente les résultats sous forme de tableau avec les colonnes suivantes :

| Initiative | Organisation | Responsables | Période | Thématiques | Lien avec la santé | Statut | Sources |

Ajoute ensuite :

1. Une synthèse des principaux thèmes émergents ;
2. Une liste des organisations les plus actives ;

3. Une liste des experts récurrents ;
4. Les lacunes ou domaines de recherche peu couverts ;
5. Une évaluation du niveau de confiance associé à chaque résultat (élevé, moyen ou faible) selon la qualité des sources disponibles.

Critères de qualité

- Prioriser les informations récentes et vérifiables.
- Utiliser des sources institutionnelles, gouvernementales ou scientifiques, lorsque disponibles.
- Citer systématiquement les sources.
- Signaler les informations incertaines ou non validées.
- Distinguer clairement les initiatives québécoises des initiatives externes ayant simplement des activités ou des retombées au Québec. »

Annexe 3

Résultat Consensus. évaluateur/trice 1

Résultat Consensus. évaluateur/trice 2

Annexe 4

Résultat ChatGPT. évaluateur/trice 1

Résultat ChatGPT. évaluateur/trice 2

Annexe 5

Résultat Perplexity. Évaluateur/trice 1

Résultat Perplexity. Évaluateur/trice 2