INESSS
LE SAVOIR PREND FORME

MARS 2024

Rapport méthodologique

Développement et évaluation d'un outil logiciel basé sur GPT-4 pour l'aide au tri de documents dans le cadre de revues de la littérature : une preuve de concept

Une production de l'Institut national d'excellence en santé et en services sociaux (INESSS)

Bureau des données clinico-administratives

Québec

Développement et évaluation d'un outil logiciel basé sur GPT-4 pour l'aide au tri de documents dans le cadre de revues de la littérature : une preuve de concept

Rédaction
Guido Powell
Sergio Cortez Ghio
Hervé Tchala Vignon Zomahoun

Collaboration
Isabelle Boisvert
Carole Campion
Anie Labrecque
Catherine Olivier
Éric Plante
Chantale Provost

Direction
Mike Benigeri

Le contenu de cette publication a été rédigé et édité par l'INESSS.

Membres de l'équipe de projet

Auteurs principaux
Guido Powell, M. Sc.
Sergio Cortez Ghio, Ph. D.
Hervé Tchala Vignon Zomahoun, Ph. D.

Repérage de l'information scientifique
Renaud Lussier, Ph. D., M.S.I.
Mathieu Plamondon, M.S.I.
Lysane St-Amour, M.B.S.I.

Collaborateur et collaboratrices internes
Isabelle Boisvert, Ph. D.
Carole Campion, Ph. D.
Anie Labrecque
Catherine Olivier, Ph. D.
Éric Plante, Ph. D.
Chantale Provost, Ph. D.

Directeur
Mike Benigeri, Ph. D.

Équipe de l'édition

Hélène St-Hilaire
Nathalie Vanier

Sous la coordination de
Catherine Olivier, Ph. D.

Avec la collaboration de
Littera Plus, révision linguistique
Guido Powell, traduction

Dépôt légal

Bibliothèque et Archives nationales du Québec, 2024
ISBN 978-2-550-97186-3 (PDF)

Pour citer ce document : Institut national d'excellence en santé et en services sociaux (INESSS). Développement et évaluation d'un outil logiciel basé sur GPT-4 pour l'aide au tri de documents dans le cadre de revues de la littérature : une preuve de concept. Rapport méthodologique rédigé par Guido Powell, Sergio Cortez Ghio et Hervé Tchala Vignon Zomahoun. Québec, Qc : INESSS; 2024. 32 p.

L'Institut remercie les membres de son personnel qui ont contribué à l'élaboration du présent document.

Conseil scientifique

Membres

M^me^ Roxane Borgès Da Silva, professeure agrégée, Département de gestion, d'évaluation et de politique de santé, École de santé publique, Université de Montréal. Fellow et responsable du pôle Santé, Centre interuniversitaire de recherche en analyse des organisations (CIRANO)

M. François Champagne, professeur titulaire, École de santé publique, Université de Montréal

D^r^ Pierre Dagenais, rhumatologue, Service de rhumatologie ; Centre intégré universitaire de santé et de services sociaux de l'Estrie — Centre hospitalier universitaire de Sherbrooke — Hôpital Hôtel-Dieu. Médecin-conseil, unité d'évaluation des technologies et des modes d'intervention en santé et en services sociaux du CIUSSS de l'Estrie-CHUS. Professeur adjoint, Université de Sherbrooke

M^me^ Edeltraut Kröger, professeure associée, Faculté de pharmacie, Université Laval. Chercheure, Centre d'excellence sur le vieillissement de Québec. Responsable, Regroupement sur les soins de longue durée du Réseau québécois de la recherche sur le vieillissement

D^r^ Marc Rhainds, co-gestionnaire médical et scientifique des activités d'évaluation des technologies et des modes d'intervention, Unité d'évaluation des technologies et des modes d'intervention, CHU de Québec-Université Laval

Présidents et présidentes des comités délibératifs permanents

D^re^ Ewa Sidorowicz, médecin interniste et gestionnaire DSP retraitée, Centre universitaire de santé McGill, Hôpital général de Montréal - CDP - Approches diagnostiques et dépistage

M. Daniel La Roche, gestionnaire en santé et services sociaux, CHU de Québec - Université Laval - CDP - Modes d'intervention en santé

D^re^ Sylviane Forget, gastroentérologue pédiatre – Hôpital de Montréal pour enfants – Centre universitaire de santé McGill - CPD - Remboursement et accès

M. Pierre-Paul Milette, Gestionnaire retraité - (CIUSSS du Centre-Sud de l'île-de-Montréal), CDP - Services sociaux et santé mentale

Lectrice et lecteurs externes

Pour ce rapport, les lectrice et lecteurs externes sont :

M^me^ Tess Berthier, coordinatrice scientifique, Plateforme de REcherche, Données, Intelligence et Santé (PREDIS)

D^r^ Pierre Dagenais, rhumatologue, Service de rhumatologie ; Centre intégré universitaire de santé et de services sociaux de l'Estrie — Centre hospitalier universitaire de Sherbrooke — Hôpital Hôtel-Dieu. Médecin-conseil, unité d'évaluation des technologies et des modes d'intervention en santé et en services sociaux du CIUSSS de l'Estrie-CHUS. Professeur adjoint, Université de Sherbrooke

D^r^ Marc Rhainds, co-gestionnaire médical et scientifique des activités d'évaluation des technologies et des modes d'intervention, Unité d'évaluation des technologies et des modes d'intervention, CHU de Québec-Université Laval

Autres contributions

L'institut tient à remercier aussi les personnes suivantes qui ont contribué à la préparation de ce rapport en fournissant soutien, information et conseils clés :

- Sylvine Carrondo-Cottin, Ph. D.

Déclaration d'intérêts

Toutes les personnes qui ont collaboré à cet état des connaissances ont déclaré leurs intérêts et leurs rôles en tout respect de la Politique de prévention, d'identification, d'évaluation et de gestion des conflits d'intérêts et de rôles des collaborateurs de l'INESSS.

D^r^ Pierre Dagenais et **D^r^ Marc Rhainds** qui ont contribué à la lecture externe sont membres du Conseil scientifique de l'INESSS.

Une des revues de littérature utilisée pour développer et évaluer l'outil d'aide à la décision de tri a été produite et rédigée par un des auteurs principaux du rapport, M. Hervé Tchala Vignon Zomahoun, Ph. D.

Responsabilité

L'Institut assume l'entière responsabilité de la forme et du contenu définitifs de ce document. Ses conclusions et ses recommandations ne reflètent pas forcément les opinions des lecteurs externes ou celles des autres personnes consultées aux fins de son élaboration.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Introduction

Bien qu'essentielles aux productions de l'INESSS, les revues de littérature réalisées à partir d'un repérage documentaire structuré sont coûteuses et chronophages. Le tri de documents scientifiques par la lecture des titres et résumés impose un travail minutieux, laborieux et répétitif. Différents outils semi-automatisés basés sur l'apprentissage automatique sont commercialement disponibles, mais ils demandent une participation importante de l'utilisateur et ne performent pas à un niveau suffisamment élevé pour justifier leur emploi.

Ce rapport décrit le développement et l'évaluation d'un outil basé sur GPT-4, un modèle de langage à grande échelle (LLM) qui vise à automatiser le tri des documents. Grâce à l'immense volume de données textuelles avec lesquelles les LLM sont entraînés, ces modèles peuvent détecter des nuances linguistiques et contextuelles qui leur permettent de classifier des titres et des résumés selon leur pertinence par rapport à un ensemble de critères de sélection spécifiques, et ce, sans avoir recours à l'étiquetage humain.

Méthodologie

Nous avons sélectionné pour l'évaluation de l'outil quatre revues de la littérature qui font partie d'une publication de l'INESSS et correspondent à des critères spécifiques. Pour chaque titre et résumé de document issu des repérages de chaque revue, nous avons demandé au modèle d'évaluer la pertinence du document en fonction des critères d'inclusion et d'exclusion qui avaient été déterminés lors de la réalisation de la revue.

Nous avons employé trois stratégies pour formuler les messages de requête du modèle : (1) une stratégie de base reproduisant une approche employée dans une publication antérieure évaluant le potentiel de GPT-3.5 pour le tri de documents scientifiques, (2) une stratégie sensible modifiant la formulation du message de requête de la stratégie de base pour éviter l'exclusion de documents pertinents, et (3) une stratégie de classement à neuf niveaux (par exemple *almost certain, extremely likely, etc.*) permettant de fixer un seuil décisionnel pour optimiser la performance de l'outil.

Les réponses à ces requêtes ont été comparées aux décisions de sélection des auteurs et auteures des rapports lors de la lecture intégrale des documents. L'objectif était de maximiser la sensibilité tout en ayant une spécificité au minimum « modérée ». Pour la stratégie de classement, nous avons évalué ces deux mesures de performance à différents seuils. Finalement, nous avons aussi estimé le volume de tri potentiellement économisé ainsi que le risque d'exclusions erronées à différents seuils.

Résultats et discussion

Pour la stratégie de base, nous avons obtenu une sensibilité de 92,3 % et une spécificité de 80,4 %. Avec la stratégie sensible, nous avons atteint une sensibilité de 99 % et une spécificité de 55,1 %.

De meilleurs résultats ont été obtenus avec la stratégie de classement en identifiant un seuil d'inclusion (à partir de la classe *unlikely*). En effet, cette stratégie permettait une performance optimale, soit une sensibilité parfaite de 100 % et une spécificité de 57,7 %. Ce seuil conserve donc tous les documents jugés pertinents pour la revue après la lecture intégrale par les auteurs, tout en éliminant plus de la moitié des documents à évaluer. Cela se traduit donc par une économie de volume de tri de 56,3 % en moyenne sans entrainer une perte de document pertinent lors du tri. L'inclusion à partir du seuil supérieur (à partir de la classe *very likely*) pourrait potentiellement permettre d'économiser 63,1 % de volume de tri avec un taux de perte d'environ 1 % des documents pertinents. Cette stratégie permet aussi de classer les documents retenus selon leur niveau de pertinence afin de faciliter le processus de tri.

Évaluations futures

Nous envisageons d'entreprendre des analyses pour évaluer d'autres méthodes proposées dans des articles récents en prépublication qui proposent des LLM en libre accès, l'analyse d'un plus grand nombre de revues ainsi que le développement d'une étude prospective en collaboration avec le personnel professionnel scientifique de l'INESSS afin de valider la robustesse de la performance du modèle.

Enjeux éthiques

Étant donné le partage d'information avec un tiers lors de l'utilisation de GPT-4, il est nécessaire d'exercer un certain jugement concernant les risques entourant la confidentialité des données. De plus, il est aussi important de tenir compte des risques de biais du fait que les LLM sont entraînés sur une sélection de textes provenant d'Internet et de biais pouvant émerger de l'interprétation des critères d'inclusion et d'exclusion par le modèle. Finalement, il existe un enjeu écologique lié à l'importante consommation énergétique nécessaire à l'entraînement de ce type de modèles.

Conclusion

L'outil proposé, développé en tant que preuve de concept, pourrait permettre de réduire considérablement le volume de tri dans le cadre des revues de la littérature réalisées par l'INESSS, tout en diminuant le risque d'omettre des documents pertinents.
Des évaluations additionnelles seront nécessaires pour confirmer la performance de l'outil.

SUMMARY

Introduction

While essential for INESSS productions, literature reviews conducted through structured literature searches are costly and time-consuming. The screening of scientific articles by reading titles and abstracts requires painstaking, repetitive work. Various semi-automated tools based on machine learning are commercially available but require significant user involvement and do not perform at a sufficiently high level to justify their use.

This report describes the development and evaluation of a document screening tool based on GPT-4, a large language model (LLM). Thanks to the vast amount of textual data with which LLMs are trained, these models can detect linguistic and contextual nuances, enabling them to classify titles and abstracts based on their relevance to a specific set of selection criteria, all without the need for a human labeler.

Methodology

We identified four literature reviews that were part of an INESSS product and met criteria we established for evaluation of the tool. For each title and abstract of articles identified in each review, we asked the model to assess the article's relevance based on the inclusion and exclusion criteria developed for the review.

We employed three strategies to formulate prompts to the model: (1) a basic strategy based on an approach used in a recent publication evaluating the potential of GPT-3.5 for abstract screening, (2) a sensitive strategy modifying the wording of the basic strategy's prompt to avoid excluding relevant documents, and (3) a nine-level ranking strategy (e.g. "almost certain", "extremely likely", etc.) to identify a decision threshold to optimize the tool's performance.

The model's responses to these queries were compared to the authors' decisions at the full-text article selection stage. Our goal was to maximize sensitivity and achieve at least moderate specificity. For the ranking strategy, we evaluated these two performance measures at different thresholds. Finally, we also estimated the potential screening workload savings and the associated risk of erroneous exclusions at different thresholds.

Results and Discussion

For the basic strategy, we achieved a sensitivity of 92.3 % and a specificity of 80.4 %. With the sensitive strategy, we reached a sensitivity of 99 % and a specificity of 55.1 %.

A more significant improvement was obtained with the ranking strategy by identifying an inclusion threshold (starting from the "unlikely" rank) that allowed optimal performance: a perfect sensitivity of 100 % and a specificity of 57.7 %. This threshold retains all articles authors deemed relevant upon reading the full text, while eliminating more than half of the article to be evaluated. These results suggest an average screening workload savings of 56.3 % with no missed relevant articles. Inclusion from the higher threshold (from the "very likely" rank) could potentially save 63.1 % of screening workload with only

a 1 % rate of missed relevant articles. This strategy also allows sorting of the retained documents by relevance rank to facilitate the screening process.

Future Evaluations

We plan to undertake future analyses to evaluate methods of recent preprint articles proposing an open-sourced LLM, to analyze a larger number of reviews, and to develop a prospective study in collaboration with INESSS professionals to ensure the robustness of the model's performance.

Ethical Considerations

Given that use of GPT-4 involves sharing of information with a third party, users must consider the risks surrounding data confidentiality. Furthermore, risks of bias may emerge from the selection of internet corpora on which LLMs are trained as well as from the model’s interpretation of inclusion and exclusion criteria. Finally, there are important ecological concerns related to the energy consumption needed for training such models.

Conclusions

The proposed tool, developed as a proof of concept, shows the possibility of significant reductions in terms of the screening workload for reviews undertaken at INESSS, with minimal erroneous exclusions of relevant articles. Additional evaluations will be needed to further validate the performance of the tool.

SIGLES ET ACRONYMES

API	Interface de programmation d’application
GPT	Transformeur génératif préentraîné
LLM	Modèle de langage à grande échelle
VPP	Valeur prédictive positive

GLOSSAIRE

Apprentissage actif

Type d'apprentissage automatique qui permet une amélioration itérative à la classification de données grâce à l'étiquetage par un humain des éléments priorisés par le modèle (*active learning* en anglais).

Courbe ROC

Graphique représentant la performance d'un modèle de classification, notamment la sensibilité et la spécificité, en fonction de différents seuils de décision. De l'anglais *Receiver Operating Characteristic*.

Étiqueteur humain

Individu ou groupe d'annotateurs humains chargés d'étiqueter ou d'annoter manuellement des données employées dans des tâches d'apprentissage automatique.

Faux négatifs

Documents jugés pertinents par les auteurs d’une revue, mais exclus par l'outil de classification.

Interface de programmation d'application (API)

De l'anglais *Application Programming Interface*. Ensemble de règles et de protocoles qui permettent à différents logiciels et systèmes informatiques de communiquer et d'interagir entre eux. Dans le contexte d'OpenAI, leur API permet de faire fonctionner les modèles GPT de manière coordonnée, facilitant ainsi les évaluations et le développement d'outils par des tiers.

Message de requête *(prompt)*

Instruction donnée à un système d'intelligence artificielle génératif tel que GPT pour obtenir une réponse ou du contenu spécifique.

Modèle de langage à grande échelle (LLM)

De l’anglais *large language model*, aussi appelé grand modèle de langage. C’est un type de modèle d'intelligence artificielle développé avec des techniques d'apprentissage profond, qui a typiquement des milliards de paramètres et un énorme volume de données textuelles d'entraînement. Ces modèles peuvent traiter et générer du texte, permettant ainsi diverses tâches de traitement du langage naturel de manière contextuelle et fluide.

Sélection

L’étape de classification (inclusion vs exclusion) des documents scientifiques accomplie après lecture intégrale des documents. Cette étape est précédée par le tri des documents.

Sensibilité

Mesure de la capacité d'un modèle à identifier correctement les éléments positifs dans un ensemble de données, particulièrement dans le contexte de la classification binaire. Dans le contexte d'un outil de tri de documents scientifiques, la sensibilité indique la capacité du modèle à identifier correctement les documents jugés pertinents par les auteurs des revues.

Seuil de probabilité de pertinence

Classe issue de la stratégie de classement à partir de laquelle un document est considéré comme pertinent pour une revue.

Spécificité

Mesure de la capacité d'un modèle à identifier correctement les éléments négatifs dans un ensemble de données, particulièrement dans le contexte de la classification binaire. Dans le contexte d'un outil de tri de documents scientifiques, la spécificité indique la capacité du modèle à identifier correctement les documents jugés non pertinents par les auteurs des revues.

Transformateurs génératifs préentraînés (GPT)

De l'anglais *Generative Pre-trained Transformer*, ce terme fait référence à une famille de LLM développés par la compagnie OpenAI, dont les versions GPT-3.5 et GPT-4 qui sont à la base de l'agent conversationnel ChatGPT. Ces modèles se distinguent d'autres LLM par leur taille (en nombre de paramètres et en volume de données d'entraînement), par l'approche séquentielle de leur compréhension des données et par leur objectif de prédiction du prochain mot.

Tri

L’étape initiale de classification (pertinent vs impertinent) des documents scientifiques accomplie après lecture des titres et des résumés.

Valeur prédictive positive (VPP)

Mesure de la probabilité qu'un vrai positif (par exemple un document pertinent) soit présent dans les classifications positives d'un modèle de classification.

Volume de tri potentiellement garanti

Proportion des documents exclus par l’outil par rapport au nombre total de documents à évaluer.

INTRODUCTION

Dans le contexte d'une littérature scientifique en croissance exponentielle [Bornmann *et al.*, 2021], la charge de travail associée à la conduite de revues de la littérature s'intensifie continuellement. Au sein de l'Institut national d'excellence en santé et en services sociaux (INESSS), les revues réalisées à partir d'un repérage documentaire structuré sont essentielles à la réalisation de la plupart des produits de connaissance. En effet, ces revues permettent une synthèse exhaustive et reproductible des données probantes. Entre 2019 et 2023, nous estimons que les conseillers en information scientifique du Bureau – Méthodologies et éthique (BME) de l'INESSS ont reçu en moyenne 116 demandes de repérage documentaire structuré chaque année (consultation interne).

Cependant, les revues de la littérature sont coûteuses et chronophages. Parmi les nombreuses étapes inhérentes à la réalisation de ce type de revues, le tri et la sélection des documents pertinents selon des critères d'inclusion et d'exclusion préalablement définis sont des étapes particulièrement longues. Effectivement, elles consistent en un examen approfondi de tous les documents identifiés lors du repérage initial par au moins deux réviseurs travaillant en parallèle : un travail de minutie laborieux et répétitif [Wang *et al.*, 2020].

De ce fait, au cours de la dernière décennie, des outils semi-automatisés d'aide au tri des documents faisant usage d'algorithmes d'apprentissage automatique ont gagné en popularité au sein de la communauté scientifique. Toutefois, la plupart de ces outils fonctionnent par apprentissage actif, ce qui requiert la participation importante et soutenue d'un humain pour attribuer des étiquettes aux documents [Van De Schoot *et al.*, 2021; Ouzzani *et al.*, 2016; Wallace *et al.*, 2012]. De plus, avec ce type d'outil, il est difficile d'identifier le moment optimal pour arrêter l'évaluation des documents durant le tri. Cependant, l'inconvénient le plus notable associé à ces outils est le risque significatif de manquer des documents qui seraient pertinents à la revue. Ce risque varie non seulement d'un outil à l'autre, mais aussi selon le contexte de la recherche [Hamel *et al.*, 2020; Gates *et al.*, 2019].

Puisque l'évaluation et le tri des documents pertinents sont essentiellement des tâches de classification de texte en langage naturel, une avenue prometteuse à cet égard est l'utilisation de modèles de langage à grande échelle (LLM) tel que GPT [Radford *et al.*, 2018]. Contrairement aux algorithmes d'apprentissage automatique traditionnels, les LLM, qui sont entraînés sur d'immenses volumes de textes, peuvent détecter des nuances linguistiques et contextuelles [Liu *et al.*, 2023] et ils sont donc aptes à identifier des textes, tels que des titres et des résumés de documents scientifiques, qui répondent à un ensemble de critères donnés sans avoir recours à un étiqueteur humain.

D'ailleurs, une équipe canadienne a très récemment développé et présenté en prépublication une méthode exploitant le modèle de langage à grande échelle GPT-3.5 d'Open AI pour trier des documents à partir de critères d'inclusion et d'exclusion dans le cadre de revues systématiques de la littérature [Guo *et al.*, 2023]. Cependant, Guo et ses

collaborateurs ont rapporté une sensibilité moyenne de 76 % (variant entre 59 % et 100 %) pour les documents inclus, ce qui signifie qu'en moyenne près d'un quart des documents jugés pertinents par des humains ont été omis par le modèle et qu'il y avait beaucoup de variabilité d'une revue à l'autre.

Inspirés par les travaux de Guo et collaborateurs, nous avons développé un outil logiciel d'aide au tri de documents scientifiques à partir des titres et des résumés dans le cadre de revues de la littérature. Pour ce faire, nous avons non seulement examiné plusieurs stratégies de tri, mais également opté pour un autre LLM, soit GPT-4, le successeur de GPT-3.5. Le présent rapport décrit la méthode de développement de cet outil, sa performance ainsi que les économies de volume de tri qui pourraient potentiellement être réalisées par le personnel professionnel scientifique de l'INESSS en utilisant l'outil.

1 MÉTHODOLOGIE

1.1 Repérage et sélection des revues de la littérature faisant partie d'une publication de l'INESSS

Pour être incluse dans le processus de développement et d'évaluation de la performance de l'outil d'aide au tri de documents scientifiques, une revue de l'INESSS devait correspondre aux critères d'inclusion suivants :

- Au moins un des auteurs de la revue devait être employé par l'INESSS et être disponible pour collaborer au moment du développement de l'outil.
- Il devait s'agir d'une revue réalisée à partir d'un repérage documentaire structuré avec la collaboration d'une conseillère ou d'un conseiller en information scientifique (bibliothécaire).
- La sélection des documents inclus dans la revue devait avoir été complétée en deux étapes ou plus, comprenant au moins une étape de tri par lecture des titres et des résumés ainsi qu'une étape de sélection après lecture intégrale des documents.
- Le processus de sélection des documents devait avoir été suivi par deux personnes de façon indépendante, avec une troisième personne pour trancher s'il n'y avait pas consensus.
- Tous les titres et résumés des documents issus du repérage devaient être disponibles en format « .enl » (bibliothèque EndNote) ou en format tabulaire (.csv, .xlsx, etc.). De plus, chaque document devait être accompagné des décisions relatives au tri et à la sélection des auteurs de la revue (c.-à-d. inclus, exclus ou incertain).
- La revue devait comporter un seul ensemble de critères d'inclusion et d'exclusion. Ces derniers devaient être bien détaillés dans la publication et représenter l'entièreté des critères employés par les auteurs lors du tri et de la sélection des documents.
- 20 documents au minimum devaient avoir été inclus à l'étape du tri de la littérature par lecture des titres et des résumés.

Pour repérer les publications de l'INESSS contenant des revues de la littérature et pouvant servir au développement de l'outil d'aide au tri, nous avons (1) entrepris un échantillonnage de commodité sur le site Web de l'INESSS et (2) demandé aux différentes directions de fournir une liste de leurs publications correspondant à nos critères d'inclusion. Nous avons ensuite contacté au moins une ou un auteur pour chacune des publications retenues afin de vérifier que les documents étaient conformes aux critères d'inclusion définis et d'accéder aux bibliothèques de documents issues des repérages.

1.2 Préparation des données

Le développement de l'outil et toutes les analyses ont été réalisés avec le langage de programmation R [R Core Team, 2023] (v.4.2.2) dans RStudio [Posit Team, 2023] (v.2023.06.1).

1.2.1 Données sur les documents issus des repérages

À partir des bibliothèques de documents issus des repérages, des tables de données comprenant les colonnes suivantes ont été préparées pour chaque revue : (1) titres des documents, (2) résumés des documents, (3) décisions des auteurs à l'étape du tri par la lecture des titres et des résumés ainsi que (4) décisions des auteurs à l'étape de sélection par la lecture intégrale des documents. Les termes employés pour représenter la décision des auteurs ont ensuite été harmonisés. À noter que nous avons traité les documents étiquetés au tri comme « incertains » par les auteurs comme étant inclus. Les documents sans résumé (entre 0,1 % et 8,6 % du total des documents par revue) ont été retirés des tables.

1.2.2 Critères d'inclusion et d'exclusion des revues sélectionnées

Les critères de sélection, habituellement présentés sous forme tabulaire dans les publications de l'INESSS, ont d'abord été transposés en format texte. Les critères faisant référence aux métadonnées des documents, par exemple la langue ou l'année de publication, n'ont pas été intégrés, car ces critères sont typiquement appliqués à l'étape du repérage. Pour assurer une performance optimale du modèle de langage, les critères ont ensuite été traduits du français à l'anglais et, dans certains cas, légèrement clarifiés.

1.3 Développement et exécution des requêtes

1.3.1 Composition des requêtes

Les requêtes soumises à GPT sont divisées en deux types de messages :

(1) Le message utilisateur, qui représente les éléments de conversation avec lesquels le modèle doit interagir. Dans le cadre de ces travaux, c'étaient le titre et le résumé de chaque document contenu dans les tables de données. Ce message était précédé du message système.

(2) Le message système, qui sert à préciser au modèle les types de réponses qui sont attendues ainsi que des instructions supplémentaires. Le message système de chaque requête consistait en une série d'instructions pour indiquer au modèle comment traiter l'information contenue dans le message utilisateur et quel type de réponse produire ainsi que les critères d'inclusion et d'exclusion de chaque revue.

Dans le cadre du développement de l'outil, nous avons testé trois stratégies distinctes afin de déterminer la formulation optimale de la portion des instructions du message système.

1.3.1.1 Stratégie de base

La première stratégie était une reproduction du message système développée par Guo et collaborateurs [Guo *et al.*, 2023]. Le message demandait au modèle de jouer le rôle d'un chercheur et d'exclure les documents qui ne correspondaient pas parfaitement aux critères d'inclusion et d'exclusion spécifiés :

"You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Use the criteria below to inform your decision. If any exclusion criteria are met or not all inclusion criteria are met, exclude the article. If all inclusion criteria are met, include the article.

Only type "included" or "excluded" to indicate your decision. Do not type anything else."

1.3.1.2 Stratégie sensible

La deuxième stratégie était une adaptation de la stratégie de base qui avait pour objectif d'augmenter la sensibilité de la classification. À cet égard, nous avons modifié deux éléments au message système de Guo et collaborateurs. Premièrement, nous avons spécifié que les réponses attendues devaient être *potentially relevan*t ou *definitely irrelevant* au lieu de *included* ou *excluded*. Deuxièmement, nous avons inclus dans le message une directive indiquant d'être aussi sensible que possible en retenant un document donné en cas de doute au sujet de sa pertinence :

"You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Based on the following article title and abstract, and the inclusion criteria provided, determine whether the article may be relevant or should be excluded from the review.

Only type "potentially relevant" or "definitely irrelevant" to indicate your decision.

If you are uncertain, if the article might be relevant for whatever reason, mark it as potentially relevant. Be as sensitive as possible, I do not want to exclude potentially relevant articles."

1.3.1.3 Stratégie de classement

La troisième stratégie étendait la classification binaire (inclusion vs exclusion) à un classement à neuf niveaux de probabilité de pertinence, ou classes. L'objectif de la définition de ces classes était d'éviter de forcer le modèle à trancher en cas d'incertitude. Cette approche nous permettait aussi de fixer un seuil décisionnel plus finement (la classe de probabilité de pertinence à partir de laquelle le document est inclus) afin d'optimiser la performance de l'outil lors de la classification. Les étiquettes pour les classes sont tirées d'un guide sur la communication de l'incertitude pour les évaluations scientifiques [European Food Safety Authority *et al.*, 2019].

“You are a researcher rigorously screening titles and abstracts of scientific papers for inclusion or exclusion in a review paper. Based on the following article title and abstract, and the inclusion criteria provided, determine how likely this article is to be relevant.

Only type one of the following answers to indicate your decision:

“Almost certain”

“Extremely likely”

“Very likely”

“Likely”

“About as likely as not”

“Unlikely”

“Very unlikely”

“Extremely unlikely”

“Almost impossible” ”

1.3.2 Exécution et gestion des réponses

Les requêtes ont été exécutées avec le logiciel RStudio à travers l’interface de programmation d’application (API) d’OpenAI en mode *completion* et elles faisaient appel à la version GPT-4 du modèle de langage. Le paramètre de *temperature*, qui permet l’inclusion d’une réponse moins probable, a été configurée à 0 sans aucune autre modification de paramétrisation du modèle.

Pour chaque revue et chaque stratégie, le message système a été apparié aux messages utilisateur en boucle, de sorte qu’une requête séparée était faite pour chaque document sans aucun historique des requêtes précédentes. Les réponses du modèle de langage ont ensuite été ajoutées aux tables de données contenant les décisions de tri et de sélection des auteurs pour chaque document.

1.4 Évaluation de la performance de classification de l’outil

Pour évaluer la performance de classification de l’outil, nous avons comparé, pour chaque revue et pour chaque stratégie, les décisions du modèle de langage aux décisions du ou des auteurs. La mesure primaire que nous avons retenue est la sensibilité, qui indique la capacité du modèle à identifier correctement les documents jugés pertinents par les auteurs des revues. Nous cherchions à optimiser le modèle pour obtenir une sensibilité supérieure à 95 % afin d’éviter qu’un document pertinent ne soit exclu lorsque l’outil serait utilisé. Nous avons également tenu compte de la spécificité pour évaluer la capacité du modèle à correctement identifier les documents jugés non pertinents par les auteurs et ainsi permettre une réduction du nombre de documents à évaluer. Notre objectif était d’obtenir une spécificité d'au moins 30 %, pourcentage suffisamment large pour assurer l’exclusion de centaines de documents dans la plupart des revues.

Pour la stratégie de classement, nous avons évalué la performance du modèle à différents seuils (niveaux à partir desquels un document était inclus) pour dessiner une courbe ROC *(Receiver Operating Characteristic)*. La courbe ROC illustre à la fois la sensibilité et la spécificité. Cette mesure nous a permis de déterminer le seuil optimal de la stratégie de classement pour permettre une comparaison avec les autres stratégies de classification binaire.

Nous avons également évalué pour toutes les revues la valeur prédictive positive (VPP) à chaque classe, soit la proportion de documents jugés pertinents par les auteurs parmi les documents de la classe.

Afin d'estimer l'économie potentielle moyenne *(E)* de volume de tri à chaque seuil, nous avons ensuite calculé la proportion des documents exclus par l'outil par rapport au nombre total de documents à trier pour chaque revue, à l'aide de la formule suivante :

$$E_s = \frac{VN_s + FN_s}{D_t} \quad (1)$$

où *s* est le seuil de probabilité de pertinence à partir duquel un document est inclus, *VN* est le nombre de vrais négatifs, *FN* est le nombre de faux négatifs et D_t est le nombre total de documents à trier. Ce calcul représente donc la proportion des titres et résumés qui n'ont pas à être triés manuellement par les auteurs du fait qu'ils ont été exclus par l'outil à un seuil donné.

Finalement, nous avons aussi estimé le risque d'exclure un document jugé pertinent à chaque seuil (la proportion de faux négatifs).

Pour l'évaluation des VPP, nous présentons les décisions de l'outil par rapport aux décisions des auteurs pour les étapes de tri et de sélection, tandis que pour les autres évaluations les comparaisons entre l'outil et les auteurs à l'étape du tri se trouvent en annexe seulement. Il convient de noter que l'ensemble de ces évaluations a été réalisé dans le contexte d'une preuve de concept visant à examiner l'utilité et le potentiel de l'outil.

2 RÉSULTATS

2.1 Revues retenues

Nous avons retenu quatre publications de l'INESSS contenant des revues de la littérature réalisées à partir d'un repérage documentaire structuré, qui répondaient à nos critères. Les publications retenues sont détaillées dans le Tableau 1.

Tableau 1 Détail des publications retenues pour le développement et l'évaluation de l'outil d'aide au tri de la littérature

Première ou premier auteur	Année	Titre	Nombre de documents		
			Repérés*	Inclus à l'étape du tri	Inclus à l'étape de la sélection
Plante, É.	2020	*Le régime cétogène dans le traitement de l'épilepsie réfractaire*	1 542	119	31
Provost, C.	2023	*Principes et critères de sélection des gènes pour le diagnostic moléculaire des maladies en génétique constitutionnelle par séquençage de nouvelle génération (SNG)*	594	34	7
Shun, P.	2023	*Obstacles et facilitateurs rencontrés par les femmes vivant avec une déficience physique (DP), une déficience intellectuelle (DI) ou un trouble du spectre de l'autisme (TSA) lorsqu'elles ont recours à des services périnataux*	1 392	109	43
Zomahoun, H.	2022	*Enjeux liés à l'implantation d'un système de soutien à la décision clinique visant la prescription d'un examen diagnostique*	737	118	23

* Seuls les articles accompagnés d'un résumé ont été comptabilisés.

Toutes les publications retenues étaient des rapports scientifiques d'un état des connaissances publiés entre 2020 et 2023. La publication de Plante et collaborateurs présentait les résultats d'une revue systématique concernant l'efficacité, l'innocuité et l'adhésion au traitement avec le régime cétogène pour traiter l'épilepsie réfractaire en comparaison avec le traitement habituel sans un régime particulier [Institut national d'excellence en santé et en services sociaux (INESSS), 2020]. Les critères d'inclusion de la revue étaient élaborés selon le cadre PICOT *(Population*, *Intervention*, *Comparator*, *Outcome* et *Time)* et se distinguaient par leur niveau de détail et de nuance par rapport à ceux des autres revues retenues.

Provost et collaborateurs présentaient les résultats d'une revue rapide de la littérature grise sur les principes et les critères d'encadrement du choix des gènes à analyser en fonction de leur pertinence clinique pour établir le diagnostic des maladies en génétique constitutionnelle [Institut national d'excellence en santé et en services sociaux (INESSS), 2023a]. À noter que, contrairement aux autres revues retenues, les auteurs avaient adopté le format de critères d'inclusion PIPOH *(Population*, *Intervention*, *Professions*, *Outcome* et *Healthcare system)*.

La revue rapide retenue suivante était celle de Shun et collaborateurs [Institut national d'excellence en santé et en services sociaux (INESSS), 2023b]. Leur travail consistait à examiner les obstacles et les facilitateurs auxquels font face les femmes vivant avec une déficience physique ou intellectuelle ou un trouble du spectre de l'autisme lorsqu'elles ont recours à des services périnataux. Les critères d'inclusion de cette revue présentaient un certain chevauchement avec les critères d'exclusion, qui étaient souvent formulés comme l'inverse ou le complément des critères d'inclusion de la même catégorie.

En appliquant une méthodologie de revue rapide, Zomahoun et collaborateurs ont précisé les obstacles et les facilitateurs liés à l'implantation d'un système de soutien à la décision clinique visant la prescription d'examens diagnostiques [Institut national d'excellence en santé et en services sociaux (INESSS), 2022]. Dans leur rapport, les critères d'inclusion et d'exclusion ont été décrits selon l'approche PEOSS *(Population*, *Exposure*, *Outcome*, *Study design* et *Setting*).

Les critères de sélection complets tirés directement des revues retenues et leur forme traduite et adaptée pour l'intégration aux messages utilisateur des requêtes sont présentés à l'annexe A.

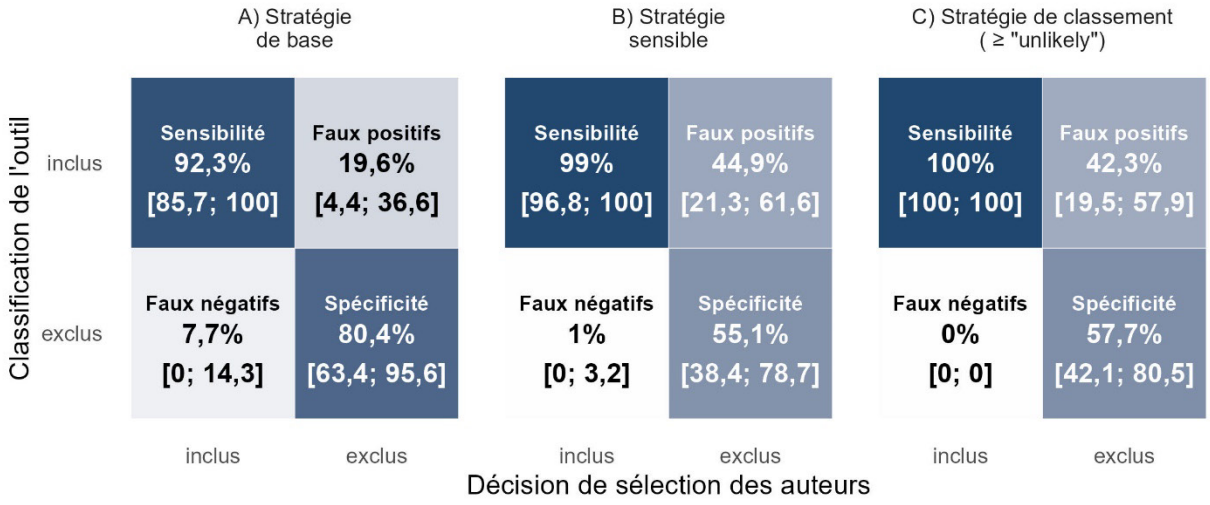
2.2 Performance de classification de chaque stratégie

2.2.1 Stratégie de base

En évaluant les inclusions de l'outil pour l'ensemble des publications, la stratégie de base [Guo *et al.*, 2023] était la moins performante (Figure 1) avec une sensibilité moyenne de 92,3 %, variant entre 85,7 % (6/7) et 100 %. Pour ce qui est d'exclure les documents qui avaient été jugés non pertinents par les auteurs, cette stratégie était la plus performante avec une spécificité moyenne de 80,4 %, qui variait entre 63,4 % (453/714) et 95,6 %

(1290/1349). Pour les mesures de la performance à l'étape du tri, voir l'annexe B. Voir l'annexe C pour les mesures par publication.

Figure 1 Comparaison de la classification des documents par l'outil aux décisions de sélection des auteurs pour les trois stratégies



2.2.2 Stratégie sensible

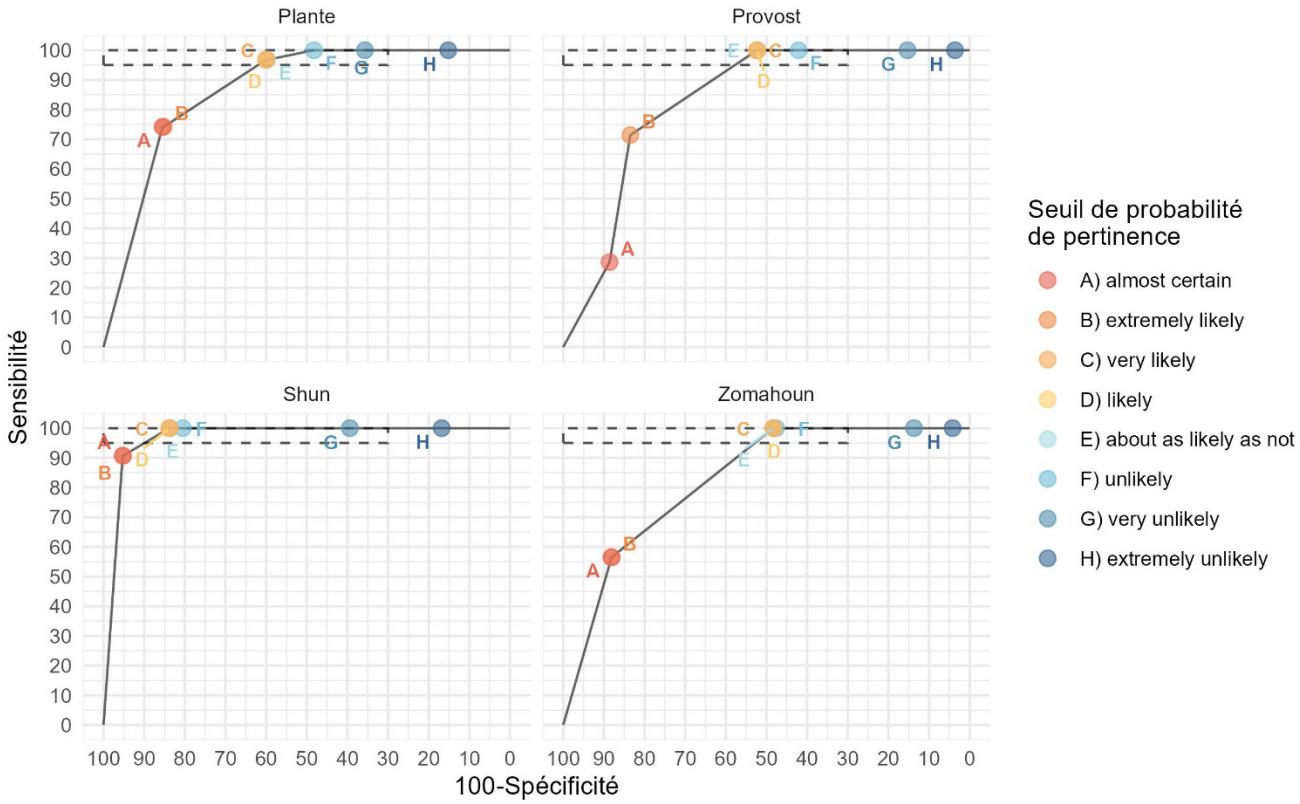
La stratégie sensible affichait une meilleure performance que la stratégie de base pour identifier les documents jugés pertinents par les auteurs (Figure 1). Cette plus grande sensibilité, de 99 % en moyenne, variait entre 96,8 % (30/31) et 100 %, et elle avait comme coût une baisse de la spécificité moyenne qui était de 55,1 %, variant entre 38,4 % (274/714) et 78,7 % (1 061/1 349).

2.2.3 Stratégie de classement

Pour la stratégie de classement, nous rapportons dans la figure 1 la performance de l'outil lorsque les documents classés par l'outil comme *unlikely* ou plus probables d'être pertinents étaient considérés comme inclus. Cette approche visait à obtenir une plus grande sensibilité que la stratégie précédente, tout en optimisant la spécificité. En fixant ce seuil, la sensibilité de la stratégie de classement était de 100 % pour toutes les revues. La spécificité moyenne était de 57,7 %, variant entre 42,1 % (247/587) et 80,5 % (1086/1349).

La figure 2 présente une courbe ROC qui illustre la performance de l'outil à différents seuils de probabilité de pertinence. Les résultats de la classe *almost impossible* ne sont pas présentés sur la figure, puisque la spécificité était de 0 % pour chaque revue. Une performance optimale se traduit par une position à l'intérieur des boîtes en pointillés, donc un classement très sensible et modérément spécifique. L'évaluation de la courbe ROC pour l'étape de tri se trouve à l'annexe D.

Figure 2 Performance de classification des documents en appliquant la stratégie de classement à différents seuils de probabilité de pertinence

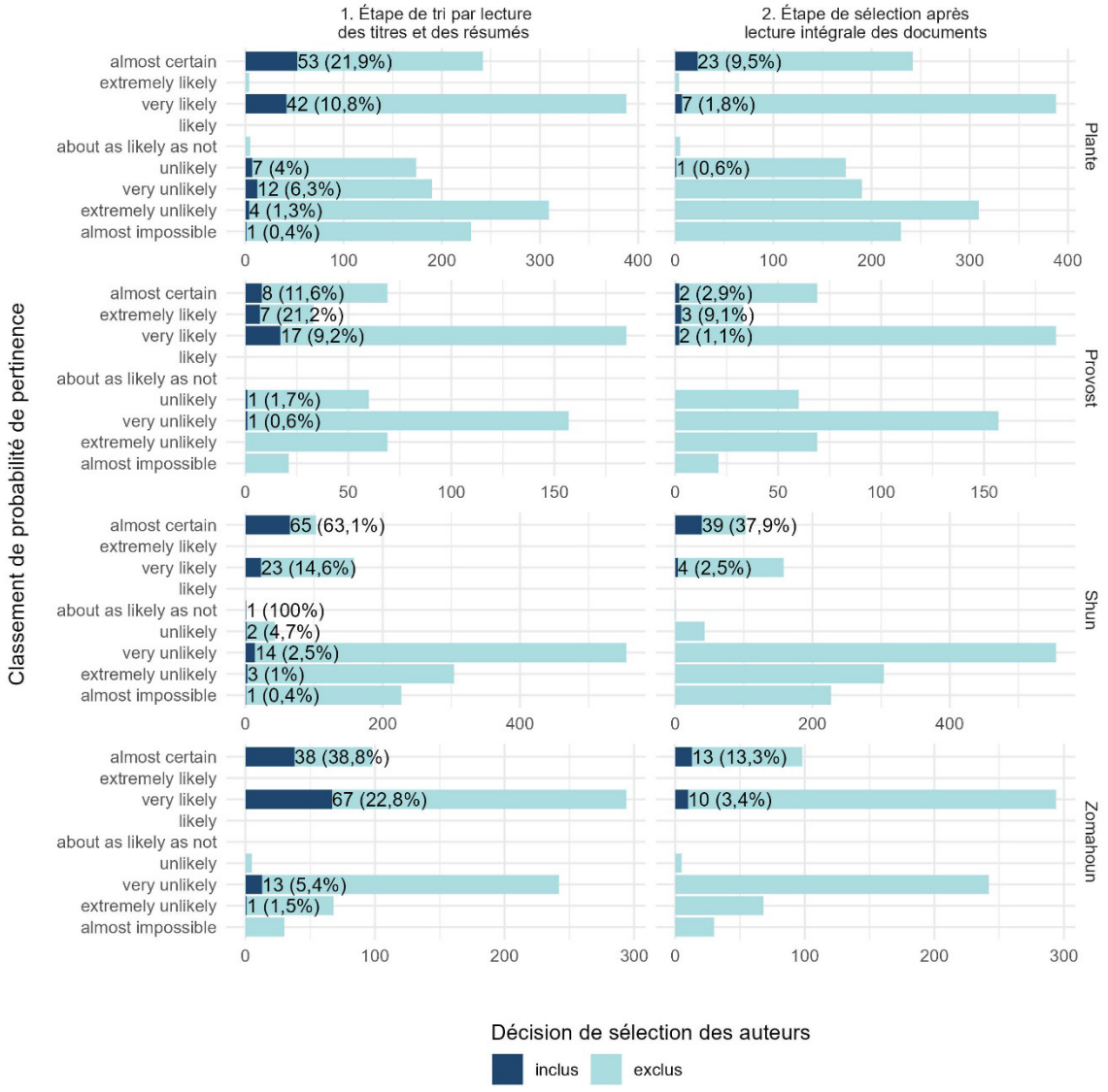


L'outil atteignait une sensibilité de 100 % pour les 4 revues ainsi qu'une spécificité de plus de 40 % pour les mêmes revues en se servant du seuil *unlikely* et plus probable. Le prochain seuil plus élevé *(about as likely as not)* diminuait la sensibilité pour une des revues (Plante) à 96,8 % (une exclusion erronée), mais il augmentait la spécificité à plus de 50 % pour toutes sauf une revue (Zomahoun) qui avait une spécificité de 48,3 %. Employer le seuil inférieur *(very unlikely)* faisait diminuer la spécificité pour deux des revues (Provost et Zomahoun) à moins de 20 %. L'outil performait exceptionnellement bien pour la revue de Shun, dont la spécificité était de plus de 80 % à quatre différents seuils tout en préservant une sensibilité de 100 %.

À la figure 3, nous présentons les VPP associées à chaque classe de probabilité de pertinence pour les étapes du tri et de la sélection. Pour toutes les revues, la probabilité de trouver un document jugé pertinent par les auteurs parmi les documents d'une classe donnée est généralement plus élevée que celle des classes inférieures. Les VPP sont plus élevées pour l'étape du tri, car il y a une plus grande proportion de documents jugés pertinents par les auteurs à cette étape. On constate aussi que la proportion des documents jugés pertinents par les auteurs au moment du tri, qui sont aussi jugés pertinents à l'étape de la sélection, est généralement plus élevée pour les classes supérieures.

Cette figure met également en évidence les différences de distribution des classes. À travers les revues, on constate que les classes *extremely likely* et *about as likely as not* sont rarement attribuées et que la classe *likely* n'apparaît jamais. On remarque aussi que, comparativement aux autres revues, les documents contenus dans Shun et collaborateurs sont classés en grande majorité plus bas que le seuil de *unlikely*, ce qui permet une VPP plus élevée au seuil plus probable de *almost certain*.

Figure 3 Distribution et valeurs prédictives positives des classes de probabilité de pertinence issues de la stratégie de classement



Finalement, le tableau 2 dépeint un portrait des économies potentielles en termes de volume de tri, en plus des risques d'omission de documents pertinents lorsque l'outil est utilisé. La classe *likely* n'est pas présentée, car il n'y a eu aucun document attribué à cette classe, tandis que la classe *almost impossible* n'est pas présentée, car elle ne contient pas d'exclusions. On voit qu'avec le seuil de probabilité de pertinence *unlikely* et plus probable, l'économie potentielle de volume de tri se situe à 56,3 %, et ce, en ne manquant aucun document jugé pertinent par les auteurs des revues. Aux seuils de *about as likely as not* et *very likely*, de plus grandes économies de volume de tri (62,9 % et 63,1 %, respectivement) sont accompagnées d'un risque très bas d'omission de documents pertinents (1 %). À partir du seuil *extremely likely*, le risque d'omission est nettement plus élevé (23,1 %). Il doit aussi être noté que le coût de l'utilisation de GPT-4 varie entre 6 et 16 CAD par revue, selon le nombre de documents et nombre de mots dans les titres et les résumés.

Tableau 2 Économies potentielles de volume de tri et risques d'omission de documents pertinents pour différents seuils de probabilité de pertinence

Seuil de probabilité de pertinence (classe minimale d'inclusion)	Pourcentage moyen de volume de tri potentiellement économisé (%)* [min; max]	Pourcentage moyen de documents pertinents potentiellement manqués (%)† [min; max]
almost certain	88,0 [84,3; 92,6]	26,0 [9,3; 71,4]
extremely likely	87,1 [82,8; 92,5]	23,1 [9,3; 43,5]
very likely	63,1 [46,8; 81,2]	1,0 [0; 3,2]
about as likely as not	62,9 [46,8; 81,1]	1,0 [0; 3,2]
unlikely	56,3 [41,6; 78]	0,0 [0,0; 0,0]
very unlikely	29,5 [13,3; 38,1]	0,0 [0,0; 0,0]
extremely unlikely	11,9 [3,5; 16,3]	0,0 [0,0; 0,0]

* Le volume de tri potentiellement économisé est représenté par la proportion de tous les documents qui sont exclus par l'outil.
† Les documents pertinents sont ceux qui ont été inclus par les auteurs à l'étape de la sélection.

DISCUSSION

Dans ce rapport, nous avons décrit le développement et évalué la performance d'un outil logiciel basé sur le modèle de langage à grande échelle GPT-4. Cet outil vise à rendre plus efficace la conduite de revues réalisées à partir de repérages documentaires structurés en assistant le personnel professionnel scientifique pour le tri de documents à partir des titres et résumés. Nous atteignons un niveau de sensibilité très élevé en appliquant une stratégie de classification binaire, et une sensibilité parfaite avec une approche de seuils de probabilité de pertinence.

Avec cette dernière, nous sommes en mesure d'identifier un seuil de classification qui conserve tous les documents jugés pertinents par les auteurs tout en éliminant en moyenne plus de la moitié des documents à évaluer. Les classes de probabilité de pertinence associées à chacun des documents restants peuvent ensuite faciliter le processus de tri par le personnel professionnel, puisque la probabilité d'identifier un document pertinent (VPP) est généralement plus élevée pour les classes supérieures. Les professionnels scientifiques auraient donc plusieurs options quant à l'utilisation de l'outil. Au seuil le plus sensible, l'outil pourrait servir à éliminer une grande portion des documents à évaluer et ordonner ceux restants par niveau de pertinence. Autrement, si une personne utilisatrice cherchait à compléter une revue plus rapidement tout en acceptant un certain niveau de risque d'omission de documents potentiellement pertinents, un seuil moins sensible permettrait d'éliminer une plus grande proportion de documents potentiellement non pertinents lors du tri, ce qui pourrait se traduire par une plus grande économie de temps. À noter que la VPP et l'économie de volume de tri sont aussi dépendantes de la précision du repérage.

Une autre manière d'utiliser l'outil pourrait être la division de la tâche de tri, par exemple en lui donnant le rôle d'un deuxième lecteur, ou en le laissant trancher en cas de désaccord entre deux professionnels. Une étude prospective serait intéressante pour estimer les économies de temps associées aux économies de volume de tri, pour des validations additionnelles ainsi que pour déterminer comment utiliser l'outil de manière optimale à l'INESSS.

Avantages des LLM et innovations récentes

Dans le cadre de nos travaux, nous avons reproduit et amélioré l'approche de Guo et collaborateurs [Guo *et al.*, 2023]. Avec notre stratégie sensible, nous avons obtenu une performance plus élevée en modifiant quelques mots dans la requête et en structurant clairement les critères d'inclusion et d'exclusion de chaque revue. Cette adaptation met en évidence l'avantage des LLM par rapport aux outils de tri commercialement disponibles, qui sont basés sur l'apprentissage automatique. Non seulement pouvons-nous adapter une méthodologie existante et obtenir une sensibilité plus élevée, mais des améliorations de la performance sont possibles sans avoir recours à un étiquetage actif par le personnel professionnel pour entraîner *de novo* ou optimiser l'entraînement du modèle.

Du fait de leur simplicité de configuration, les LLM sont de plus en plus évalués pour le tri de documents dans le cadre de revues de la littérature. Depuis la fin de nos analyses, deux autres articles parus en prépublication ont documenté la performance de LLM pour le tri de documents à partir des titres et résumés. Syriani et collaborateurs ont employé dans leurs requêtes un message système semblable à celui de notre stratégie sensible pour classifier les documents contenus dans cinq revues [Syriani *et al.*, 2023]. En effet, les auteurs demandent, eux aussi, au modèle d'être indulgent *(lenient)* et ils indiquent qu'ils et elles préfèrent inclure trop de documents que de risquer d'en manquer. Leur évaluation démontre la supériorité de cette stratégie avec la version 3.5 de GPT comparativement à d'autres modèles d'apprentissage automatique, bien que la sensibilité moyenne obtenue ne soit pas optimale (74 %). Il est intéressant de constater que des stratégies de construction de requêtes similaires émergent indépendamment dans différentes équipes.

Le deuxième article, par Robinson et collaborateurs, réalise une plus grande évaluation du tri avec le modèle GPT-3.5 en exploitant des milliers de revues de la bibliothèque Cochrane [Robinson *et al.*, 2023]. Les auteurs comparent aussi ce modèle à des modèles d'apprentissage automatique et offrent un LLM libre d’accès développé sur mesure pour le tri de documents scientifiques, qui est assez léger pour être exécuté directement sur un ordinateur local. En s'inspirant de la structure des requêtes de Syriani et collaborateurs, les auteurs obtiennent une sensibilité variable (de 82 % à 96 %), autant avec GPT-3.5 qu’avec leur LLM sur mesure. Par contre, Syriani et collaborateurs mentionnent que leur modèle est entraîné et testé sur une banque d’articles préfiltrés, c’est-à-dire qu’un tri à partir des titres et des résumés a déjà été fait. Ce type d’outil n’est donc potentiellement pas approprié pour répondre aux besoins de l’INESSS.

Limites

D’abord, il est essentiel de préciser que ce travail, dans son intégralité, sert de preuve de concept pour l'utilisation des LLM à des fins d’automatisation du tri de documents scientifiques dans le cadre de revues de la littérature. Cette affirmation prend tout son sens lorsqu’on tient compte des limites de cette étude.

La première limite réside dans le fait que quatre revues sont insuffisantes pour évaluer adéquatement la robustesse de l’outil d’aide au tri des documents. Les revues incluses dans le présent rapport n’ont pas été sélectionnées aléatoirement, mais en suivant un ensemble de critères spécifiques pour permettre le développement et l’évaluation de l’outil. Il est donc difficile de dire si ces résultats seront transposables à tous les types de repérages et de revues. Étant donné nos résultats, il apparaît évident que la manière de cadrer les critères de sélection des documents par rapport aux objectifs des auteurs risque de compromettre la performance de l'outil. Dans le cas de la revue de Shun, par exemple, les critères d'exclusion très détaillés ont possiblement permis d’obtenir une spécificité plus élevée. Ces variations sont liées à une limite plus générale de l'utilisation des LLM, celle de la vulnérabilité des résultats associée à des modifications légères dans la formulation des requêtes.

Une autre limite est la validité des décisions de tri du personnel professionnel scientifique, que nous traitons dans nos évaluations comme l'étalon-or, bien qu'elles puissent possiblement être erronées.

Ensuite, une limite difficile à contourner est la dépendance de l'outil par rapport au modèle GPT, ce qui rend la performance future et les coûts d'opération de l'outil susceptible à des changements qui seraient apportés sans préavis par OpenAI. En effet, GPT est propriétaire, de source fermée et comporte des processus de mise à jour opaques [Robinson *et al.*, 2023]. En théorie, il serait nécessaire de procéder à l'évaluation de l'outil toutes les fois que des changements mineurs ou majeurs du GPT seraient effectués. Néanmoins, il est à noter que ces enjeux liés à la performance future pourraient éventuellement être résolus par le développement d'un LLM libre d'accès qui pourrait être utilisé localement.

Enjeux éthiques et environnementaux

Finalement, divers enjeux éthiques sont à retenir avec un outil basé sur l'utilisation de GPT. D'abord, GPT comme outil de tri de documents exige un partage d'information avec un tiers, dans ce cas-ci OpenAI. Même si cette information ne comprend pas nécessairement des renseignements personnels, les travaux menés par l'INESSS revêtent un caractère confidentiel avant leur diffusion publique sur le site Web de l'organisme [Gouvernement du Québec, 2023]. Il est donc nécessaire d'exercer un certain jugement pour en assurer la confidentialité.

Il y a aussi un risque que l'utilisation de GPT renforce certains biais sociaux dus au fait que le modèle est entraîné à partir de textes trouvés sur Internet, qui sont eux-mêmes potentiellement biaisés. Les réponses du modèle risquent notamment de refléter des biais inhérents aux publications scientifiques qui favorisent des résultats positifs ou d'autres biais sociodémographiques [Hosseini et Horbach, 2023; Zack *et al.*, 2023].

De plus, la consommation énergétique des LLM est considérable. Une analyse de la version GPT-3 du modèle estime que l'entraînement seul aurait engendré 552 tonnes d'émissions de dioxyde de carbone [Patterson *et al.*, 2021]. La version GPT-4 est entraînée sur 10 fois plus de paramètres et 3 fois plus de données. Même si les besoins de consommation d'une requête, estimés seulement à quelques grammes de CO_2, sont beaucoup moins importants [Tomlinson *et al.*, 2023], l'enjeu environnemental demeure significatif et justifie l'évaluation de modèles alternatifs.

CONCLUSION ET RECOMMANDATIONS

En somme, nous avons démontré dans ce rapport, qui sert de preuve de concept, le potentiel d'un outil basé sur un modèle de langage à grande échelle pour assister le tri de documents scientifiques dans le cadre de revues de la littérature réalisées à partir d’un repérage documentaire structuré. Malgré certains enjeux éthiques et de performance future liés à l'utilisation d’un modèle tel que GPT, l'outil réduit considérablement le travail nécessaire pour évaluer et trier des documents tout en minimisant le risque d'omettre des documents pertinents. Des projets futurs et des évaluations prospectives pourront permettre d’étudier davantage la performance et l'utilisation sécuritaire de l'outil avec un plus grand nombre de revues.

RÉFÉRENCES

Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications 2021;8(1):1-15.

European Food Safety Authority, Hart A, Maxim L, Siegrist M, Von Goetz N, da Cruz C, et al. Guidance on communication of uncertainty in scientific assessments. EFSA Journal 2019;17(1):e05520.

Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, Hartling L. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Systematic reviews 2019;8(1):1-11.

Gouvernement du Québec. Loi sur l'Institut national d'excellence en santé et en services sociaux. RLRQ 2023:Chap I-13.03.

Guo E, Gupta M, Deng J, Park Y-J, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models. arXiv preprint arXiv:230500844 2023;

Hamel C, Kelly S, Thavorn K, Rice D, Wells G, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening–impact on reviewer-relevant outcomes. BMC medical research methodology 2020;20:1-14.

Hosseini M et Horbach SP. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review. Research Integrity and Peer Review 2023;8(1):4.

Institut national d'excellence en santé et en services sociaux (INESSS). Principes et critères de sélection des gènes pour le diagnostic moléculaire des maladies en génétique constitutionnelle par séquençage de nouvelle génération - Rapport en soutien au déploiement du service de séquençage. Québec, QC : État des connaissances, Rédigé par Chantale Provost et Catherine Gravel; 2023a.

Institut national d'excellence en santé et en services sociaux (INESSS). Obstacles et facilitateurs rencontrés par les femmes vivant avec une déficience physique (DP), une déficience intellectuelle (DI) ou un trouble du spectre de l'autisme (TSA) lorsqu'elles ont recours à des services périnataux. Québec, QC : État des connaissances, Rédigé par Priscilla Lam Wai Shun et Sabrina Servot; 2023b.

Institut national d'excellence en santé et en services sociaux (INESSS). Enjeux liés à l'implantation d'un système de soutien à la décision clinique visant la prescription d'un examen diagnostique. Québec, QC : État des connaissances, Rédigé par Hervé Tchala Vignon Zomahoun; 2022.

Institut national d'excellence en santé et en services sociaux (INESSS). Le régime cétogène dans le traitement de l'épilepsie réfractaire. Québec, QC : État des connaissances, Rédigé par Éric Plante et Valérie Garceau; 2020.

Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. Meta-Radiology 2023:100017.

Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews 2016;5:1-10.

Patterson D, Gonzalez J, Le Q, Liang C, Munguia L-M, Rothchild D, et al. Carbon emissions and large neural network training. arXiv preprint arXiv:210410350 2021;

Posit Team. RStudio: Integrated Development Environment for R. Boston, MA : Posit Software, PBC; 2023.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria : R Foundation for Statistical Computing; 2023.

Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018;

Robinson A, Thorne W, Wu BP, Pandor A, Essat M, Stevenson M, Song X. Bio-SIEVE: Exploring Instruction Tuning Large Language Models for Systematic Review Automation. arXiv preprint arXiv:230806610 2023;

Syriani E, David I, Kumar G. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. arXiv preprint arXiv:230706464 2023;

Tomlinson B, Black RW, Patterson DJ, Torrance AW. The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans. arXiv preprint arXiv:230306219 2023;

Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence 2021;3(2):125-33.

Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. Proceedings of the 2nd ACM SIGHIT international health informatics symposium 2012:819-24.

Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PloS one 2020;15(1):e0227742.

Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. medRxiv 2023:2023.07. 13.23292577.

ANNEXE A

Détails des critères de sélection de chaque revue retenue

Critères originaux de la revue de Plante

PICO	CRITÈRES D'INCLUSION	CRITÈRES D'EXCLUSION
Intervention	Alimentation par un régime cétogène en complément à un traitement pharmacologique ou non	Traitement à faible indice glycémique
Comparateur	**Efficacité :** ▪ Traitement pharmacologique seul (avec alimentation standard) ou en combinaison avec un autre régime **Innocuité et observance/persistance :** ▪ Traitement pharmacologique seul (avec alimentation standard) ou en combinaison avec un autre régime ▪ Alimentation standard sans traitement pharmacologique ▪ Autre régime sans traitement pharmacologique	Sans groupe comparateur (p. ex. : avant-après)
Paramètres évalués (*Outcomes*)	**Efficacité :** ▪ Taux de réduction des crises d'épilepsie de 100 % (élimination complète des crises) ▪ Taux de réduction des crises d'épilepsie > 90 % ▪ Taux de réduction des crises d'épilepsie > 50 % ▪ Effet sur la fréquence des crises ▪ Effet sur la sévérité des crises ▪ Amélioration de la qualité de vie ▪ Amélioration des capacités cognitives et comportementales **Innocuité :** ▪ Effets indésirables, y compris les effets sur le profil lipidique **Observance et persistance :** ▪ Taux d'adhésion	Efficacité d'un régime cétogène en soutien au traitement pharmacologique de l'épilepsie non réfractaire à la médication
Types de publications	RS, essais comparatifs à répartition aléatoire (ECRA), études de cohorte et études cas-témoin	Éditorial, thèse, mémoire, lettre à l'éditeur, résumé de congrès, études cliniques non publiées
Qualité méthodologique	Jugée adéquate en priorisant la rigueur d'élaboration et le risque de biais	Jugée inadéquate

Critères traduits et adaptés de la revue de Plante

* À la suite de discussions avec l'auteur, une modification a été apportée aux critères originaux. Notamment nous avons précisé les questions de recherche (efficacité, innocuité) auxquelles s'appliquaient différents critères où ce n'était pas spécifié originalement.

INCLUSIONS:

Population:

- *For articles on Efficacy: People suffering from refractory epilepsy*
- *For articles on Safety and Tolerability: All indications*

Intervention:

- *Dieting with a ketogenic diet in addition to pharmacological treatment or not*

Comparator:

- *For articles on Efficacy: Pharmacological treatment alone (with standard diet) or in combination with another diet*
- *For articles on Safety and adherence/persistence: Pharmacological treatment alone (with standard diet) or in combination with another diet, Standard diet without pharmacological treatment, other diet without pharmacological treatment*

Outcomes:

- *For articles on Efficacy: 100 % reduction rate of epilepsy seizures (complete elimination of seizures), Reduction rate of epilepsy seizures > 90 %, Reduction rate of epilepsy seizures > 50 %, Effect on the frequency of seizures, Effect on the severity of seizures, Improvement in quality of life, Improvement in cognitive and behavioral abilities*
- *For articles on Safety: Adverse effects, including effects on the lipid profile*
- *For articles on Adherence and persistence: Adherence rate*

Types of publications:

- *Systematic Reviews, Randomized Controlled Trials (RCTs), cohort studies, and case-control studies*

EXCLUSIONS:

Intervention:

- *Low glycemic index treatment*

Comparator:

- *Without comparator group (e.g.: before-after)*

Outcomes:

- *For articles on Efficacy: Efficacy of a ketogenic diet in support of pharmacological treatment of epilepsy not refractory to medication*

Types of publications:

- *Editorial, thesis, dissertation, letter to the editor, conference abstract, unpublished clinical studies*

Critères originaux de la revue de Provost

PARAMÈTRE	SPÉCIFICATION
Population à qui s'adresse l'intervention	Patient / individu / fœtus suspecté d'une condition génétique (héritée ou *de novo*)
Intervention	Analyse par SNG de l'ADN germinal d'un gène, d'un panel de gènes ciblé ou d'un panel virtuel à partir de l'exome ou du génome complet
Professionnels à qui s'adressent les travaux	Personnel des laboratoires de diagnostic moléculaire, pathologistes moléculaires, généticiens et autres professionnels qui collaborent au SNG à des fins diagnostiques
Retombées et résultats d'intérêt (de l'anglais *Outcome*)	Principes ou critères qui guident le choix des gènes à analyser lors du diagnostic moléculaire de maladies génétiques en fonction des niveaux de preuve scientifique de leur utilité clinique
Milieu de soins (de l'anglais *Health care setting*)	Milieux cliniques comparables à celui du Québec, notamment avec un système public de soins et services de santé (sans s'y restreindre)

Critères traduits et adaptés de la revue de Provost

INCLUSIONS

Population:

- *Patient/individual/fetus with suspected genetic condition (inherited or de novo)*

Intervention:

- *Next-generation sequencing analysis of germinal DNA from a gene, a targeted gene panel, or a virtual panel from the exome or the entire genome.*
- *Professionals targeted: Molecular diagnostic laboratory staff, molecular pathologists, geneticists, and other professionals who collaborate on Next-Generation Sequencing for diagnostic purposes.*

Outcome:

- *Principles or criteria guiding the choice of genes to be analyzed during molecular diagnosis of genetic diseases based on the levels of scientific evidence of their clinical utility.*
- *Healthcare setting: Clinical environments comparable to that of Quebec, particularly with a public health care and services system (but not limited to it).*

Critères originaux de la revue de Shun

	CRITÈRES D'INCLUSION	CRITÈRES D'EXCLUSION
Temporalité	Grossesse Accouchement Période postnatale jusqu'à ce que l'enfant atteigne l'âge de 1 an	Études portant uniquement sur la période de planification de la conception Études portant uniquement sur l'expérience de parents d'enfants de plus de 1 an
Milieux d'intervention (*Settings*)	Tous milieux offrant des services périnataux	Études ne faisant pas référence au contexte des services périnataux
Type de document	Études primaires, à devis : ▪ qualitatif ▪ quantitatif ▪ mixte	Revues de littérature Thèses Lettres à l'éditeur Commentaires Rapports d'organisations savantes ou d'organismes
Date	Études publiées à partir de 2012	Études publiées avant 2012
Langue	Anglais ou français	Autres qu'anglais et français
Pays	Pays membres de l'OCDE	Pays non membres de l'OCDE

Critères traduits et adaptés de la revue de Shun

INCLUSIONS

Population:

- *Women, 18 years of age and older, living with a physical disability (PD) or intellectual disability (ID) or an autism spectrum disorder (ASD)*

Intervention:

- *Perinatal care or services*

Comparator:

- *NA*

Outcomes:

- *Barriers or facilitators reported by women, their relatives or providers of perinatal care and services.*

Time:

- *Pregnancy, labour, postnatal period until the child reaches the age of 1 year.*

Settings:

- *All settings offering perinatal care and services.*

Types of publications:

- *Primary studies, either qualitative, quantitative or mixed design.*

Country:

- *OECD member countries.*

EXCLUSIONS

Population:

- *Adolescents (less than 18 years old). Studies in which participants represent a specific subpopulation of women living with PD, ID, or ASD (e.g., a study focusing only on women with spina bifida).*

Intervention:

- *Studies with no references to perinatal care or services*

Comparator:

- *NA*

Outcomes:

- *Studies making no reference to barriers or facilitators reported by women, their relatives or providers of perinatal care and services.*

Time:

- *Studies focusing solely on the preconception planning period. Studies focusing solely on the experience of parents with children over 1 year old.*

Settings:

- *Studies that do not refer to the context of perinatal services.*

Types of publications:

- *Literature reviews, theses, letters to the editor, commentaries, reports from scholarly organizations or institutions*

Country:

- *Non OECD members*

Critères originaux de la revue de Zomahoun

	Critères d'inclusion	Critères d'exclusion
Population à l'étude	Les individus pouvant contribuer aux soins ou à la gestion des soins ou encore bénéficier des soins : gestionnaires de système de santé, professionnels de la santé, patients et proches aidants	Aucun
Intervention/ facteur d'exposition	Tout facteur pouvant être un obstacle ou un facilitateur de l'implantation d'un SSDC visant la prescription d'un examen diagnostique. Le SSDC pouvait être techniquement couplé ou non à un prescripteur électronique. Cela inclut à la fois les facteurs évalués en pré ou postimplantation. Le SSDC est défini comme un système informatisé/application développé pour soutenir une prise de décision clinique basée à la fois sur les données probantes, les données cliniques et les caractéristiques du patient.	• SSDC basé sur l'intelligence artificielle • SSDC ayant une fonction qui ne vise pas explicitement la prescription d'un examen diagnostique • SSDC destiné uniquement à l'usage ou à l'autogestion du patient
Comparateur	Sans objet	Sans objet
Résultat	La fidélité à l'emploi d'un SSDC, son acceptabilité, son adoption, sa pertinence, sa faisabilité, son adaptabilité, sa portée *(penetration)* et sa pérennisation	Les résultats de l'évaluation économique ne sont pas considérés.
Devis d'étude	Toute étude observationnelle, y compris les études transversales, de cohortes et de cas-témoins. Les études mixtes et qualitatives sont également considérées pour inclusion.	Revues de littérature* de toute nature, lettres à l'éditeur, et commentaires.
Milieu d'étude	Tout milieu de soins	Ø

Critères traduits et adaptés de la revue de Zomahoun

INCLUSIONS

Population:

- *Individuals who can contribute to care or care management, or benefit from care: health system managers, healthcare professionals, patients, and caregivers.*

Intervention:

- *Any factor that can be a barrier or facilitator to the implementation of a clinical decision support system (CDSS) aimed at prescribing a diagnostic test. The CDSS could be technically coupled or not with an electronic prescriber. This includes both pre or post-implementation evaluated factors.*
- *The CDSS is defined as a computerized system/application developed to support clinical decision making based on both evidence-based data, clinical data, and patient characteristics.*

Outcome:

- *The fidelity to the use of a CDSS, its acceptability, its adoption, its relevance, its feasibility, its adaptability, its scope (penetration), and its sustainability*

Study design:

- *Any observational study, including cross-sectional studies, cohort studies, and case-control studies. Mixed-methods and qualitative studies are also considered for inclusion.*

EXCLUSIONS

Intervention:

- *CDSS based on artificial intelligence; CDSS having a function that does not explicitly aim at prescribing a diagnostic test; CDSS intended only for the use or self-management by the patient.*

Outcome:

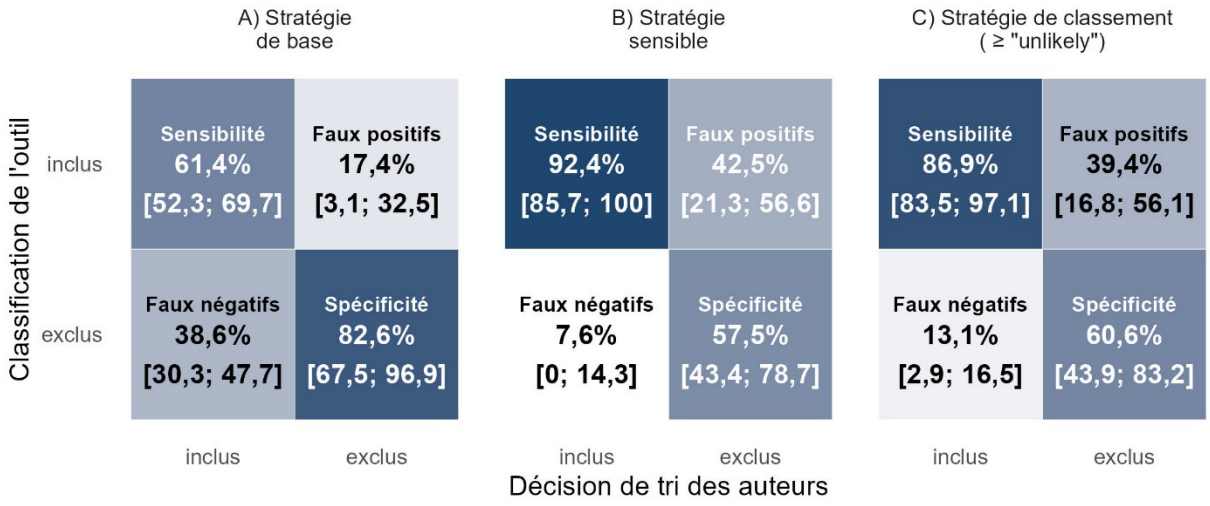
- *The results of economic evaluations are not considered*

Study design:

- *Literature reviews of any kind, letters to the editor, and comments*.

ANNEXE B

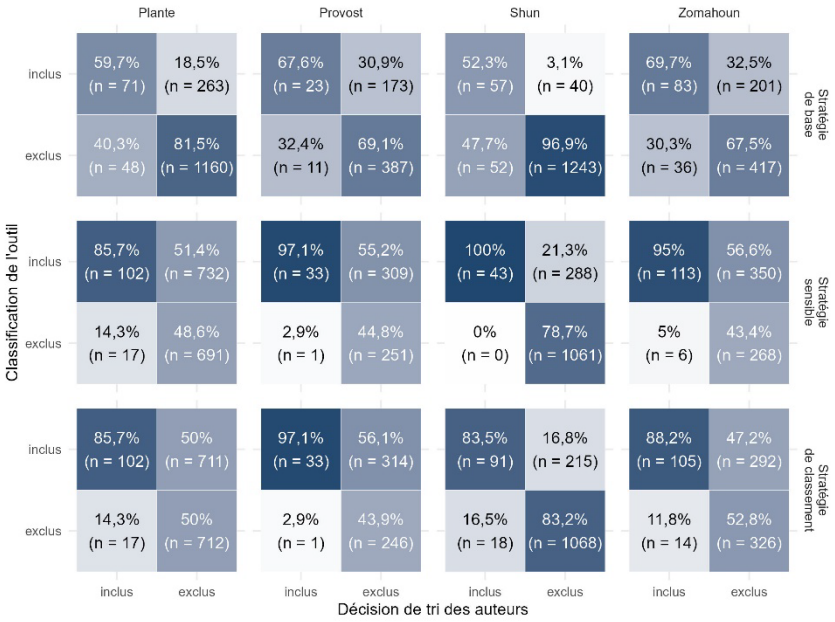
Comparaison de la classification des documents par l'outil aux décisions de tri des professionnel(le)s pour les trois stratégies



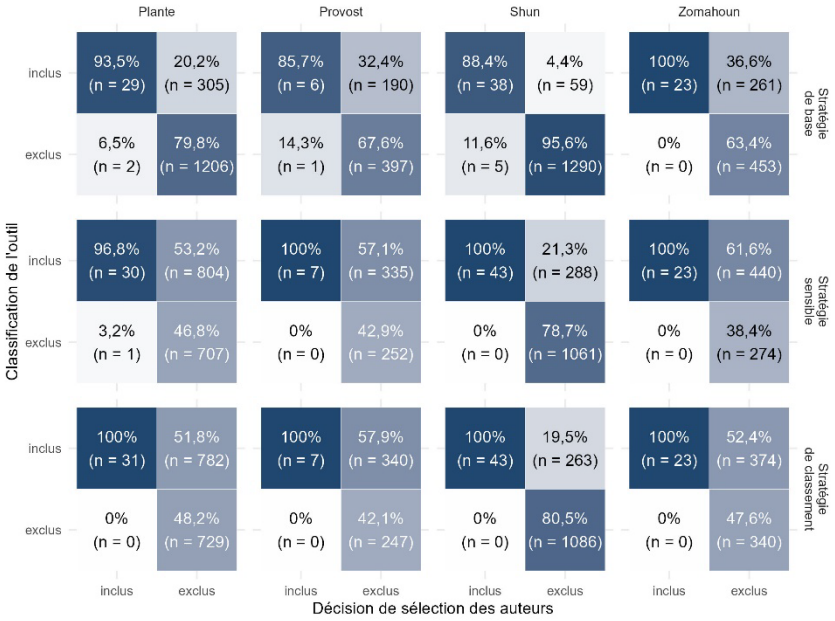
ANNEXE C

Comparaison de la classification des documents par l'outil aux décisions de tri et de sélection des professionnel(le)s pour les trois stratégies et par revue

1. Étape de tri par lecture des titres et des résumés

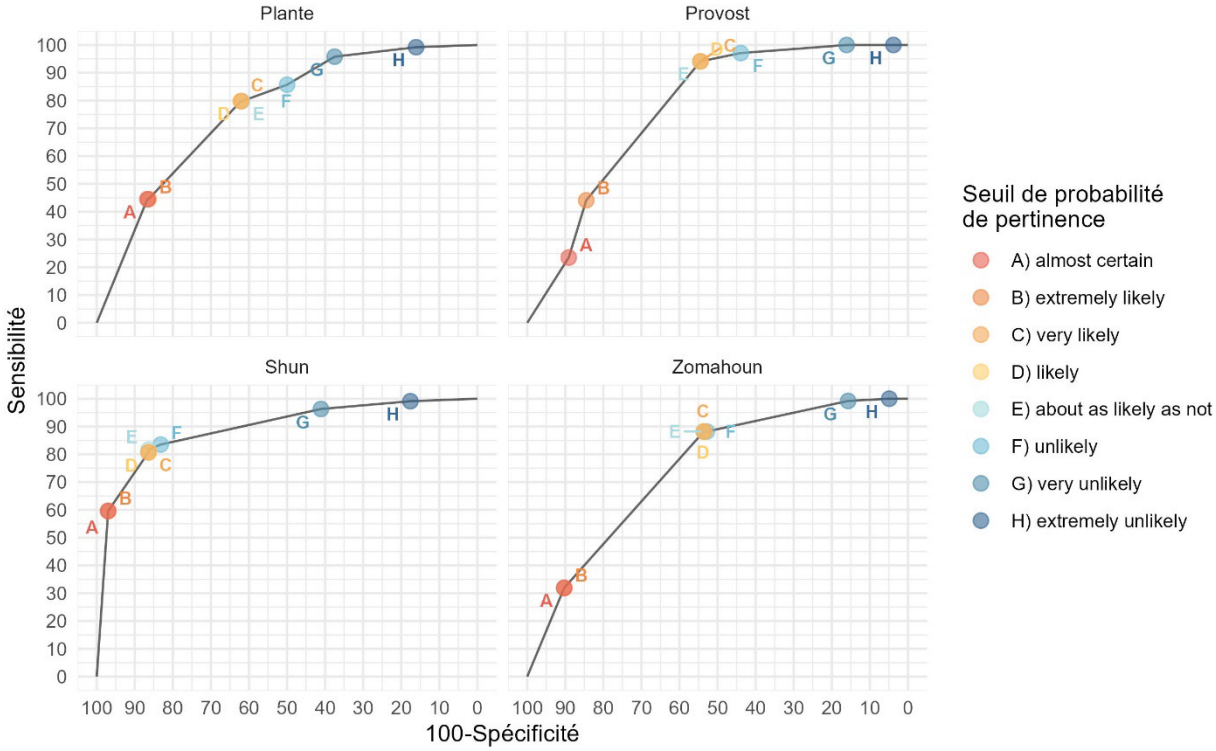


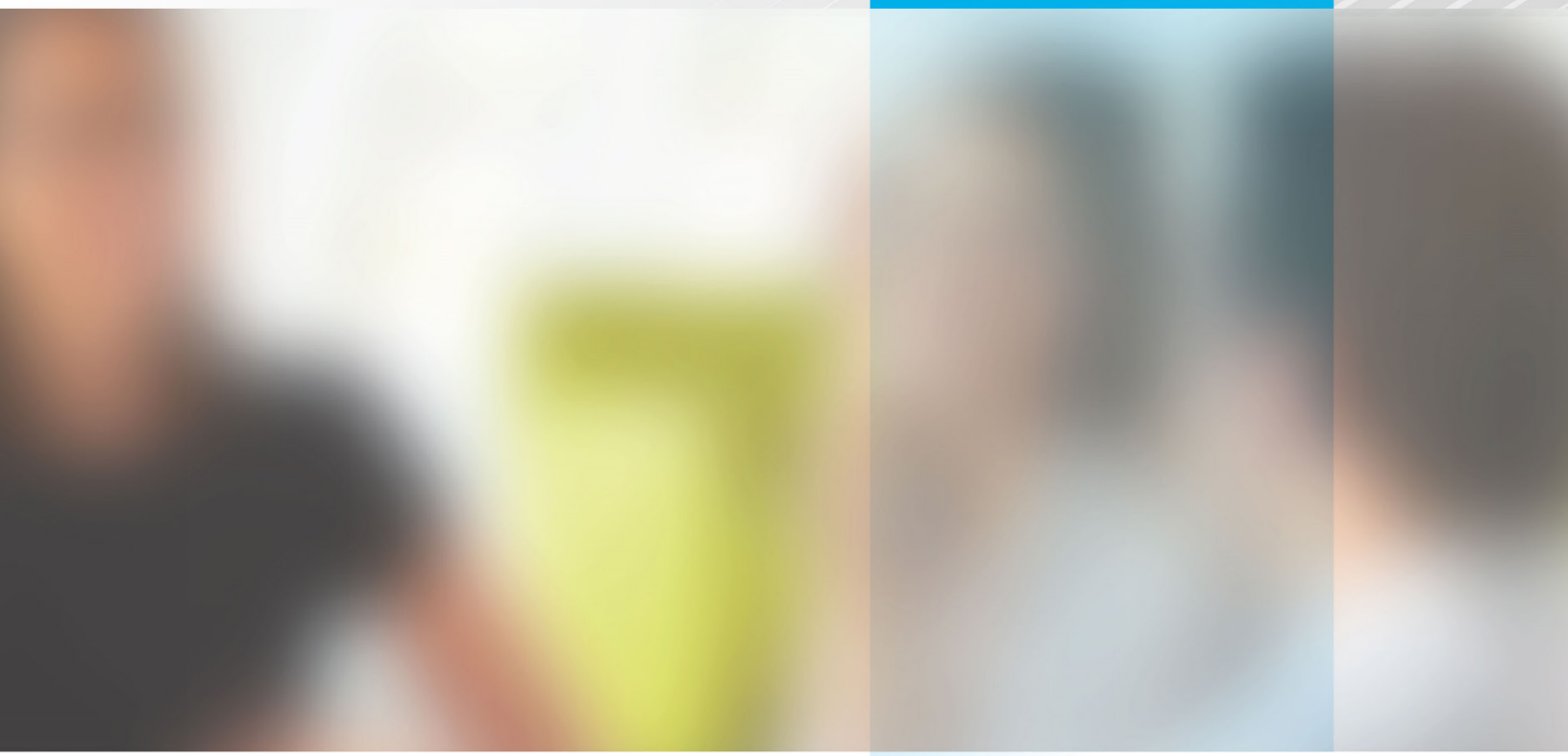
2. Étape de sélection après lecture intégrale des documents



ANNEXE D

Performance de classification des documents en utilisant la stratégie de classement, comparé à l'étape de tri par lecture des titres et résumés





Siège social

2535, boulevard Laurier, 5e étage
Québec (Québec) G1V 4M3
418 643-1339

Bureau de Montréal

2021, avenue Union, 12e étage, bureau 1200
Montréal (Québec) H3A 2S9
514 873-2563
inesss.qc.ca

Institut national d'excellence en santé et en services sociaux
Québec