Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Grands modèles de langue pour prédire la santé mentale : une revue exploratoire de la documentation des biais et de l'utilité clinique

Clémentine Bleuze[1] Karën Fort[1] Vincent P. Martin[1] Aurélie Névéol[2]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
clementine.bleuze@univ-lorraine.fr

RÉSUMÉ

L'adaptation de Grands Modèles de Langue (LLM) à des applications en santé mentale se développe dans le domaine du TAL ; et ce alors que les biais encodés (et amplifiés) par ces outils sont de plus en plus documentés, et que les défis liés à leur implémentation clinique restent sous-étudiés. Dans cette revue exploratoire de la littérature, nous décrivons les choix méthodologiques de 201 études exploitant des LLM pour effectuer des prédictions en santé mentale. Nous relevons également les mentions explicites de thèmes en lien avec les biais et l'utilité clinique. Nous montrons une appropriation encore lacunaire de ces notions, souvent confinées à certaines étapes du cycle de développement (*e.g.,* la collecte des données). Nous discutons des implications de nos observations au vu d'hypothétiques déploiements cliniques, et appelons à des réflexions inter-disciplinaires sur le sujet.

ABSTRACT

Large Language Models for Mental Health Prediction : A Scoping Review of Bias and Clinical Utility Documentation.

The use of Large Language Models (LLM) for psychiatric applications is growing in the NLP community, paralleled with increased knowledge of the biases these tools encode (and amplify), as well as challenges raised by clinical implementation. We conduct a scoping review of 201 studies using LLM to produce mental health predictions, and describe their methodological choices. We also identify mentions of themes related to bias and clinical utility within articles. We show that these notions are still incompletely understood, with worries pertaining to precise step of the development pipeline (*e.g.,* data collection). We discuss the implications of our observations regarding future hypothetical deploiements, and call for interdisciplinary reflections on the matter.

MOTS-CLÉS : Grands Modèles de Langue, santé mentale, psychiatrie, biais, applications cliniques.

KEYWORDS: Large Language Models, mental health, psychiatry, bias, clinical applications.

1 Introduction

Le langage naturel constitue un objet d'étude privilégié en santé mentale, où il peut constituer un marqueur de troubles dépressifs (Tølbøll, 2019), de la schizophrénie (Marini *et al.*, 2008; Amblard *et al.*, 2021), ou encore de la maladie d'Alzheimer (Appell *et al.*, 1982). Diverses méthodes issues du Traitement Automatique des Langues (TAL) ont ainsi été développées pour la santé (Demner-Fushman *et al.*, 2009; Abbe *et al.*, 2016; Le Glaz *et al.*, 2021), sous l'hypothèse qu'elles pourraient ultimement

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 287–302.

se traduire par des bénéfices cliniques réels pour les soignant·es et les patient·es, notamment en termes d'aide à la décision clinique et au diagnostic en santé mentale (Fraser *et al.*, 2015; Leroy *et al.*, 2018; Zhang *et al.*, 2022). À l'heure où les Grands Modèles de Langue (LLM) sont présentés comme « incontournables » auprès des scientifiques, des industriels, et du grand public, il convient de rappeler que ces outils peuvent produire des résultats erronés, biaisés, voire dangereux pour les utilisateur·ices (Yang *et al.*, 2024b; Hofmann *et al.*, 2024; Zack *et al.*, 2024; Ducel *et al.*, 2025), et que les études d'impact réel manquent toujours (Reiter, 2025). Par ailleurs, les difficultés concrètes d'implémentation de LLM dans le domaine médical peuvent être sous-estimées (Ong *et al.*, 2024).

Par conséquent, il est actuellement difficile d'établir si les outils de TAL intégrant des LLM sont « prêts » à être déployés pour aider au diagnostic en santé mentale. Si des travaux existants ont inventorié les applications et défis liés à l'intégration d'outils d'IA générative en santé (Moulaei *et al.*, 2024), ou encore des méthodes d'atténuation des biais (Yang *et al.*, 2024a), nous nous proposons d'étudier les applications prédictives des LLM en santé mentale sous le double prisme des *biais* et de l'*utilité clinique*. Nous nous demandons, d'une part, quel est l'état de la recherche actuelle intégrant des LLM pour de la prédiction en santé mentale (QR1) ? D'autre part, dans quelle mesure cette recherche intègre-t-elle des considérations liées aux notions de biais et d'utilité clinique (QR2) ?

2 Méthodologie

2.1 Nature de l'étude et choix terminologiques

Nous effectuons une revue exploratoire de la littérature (*scoping review*), suivant les recommandations PRISMA-ScR (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses* pour les *Scoping reviews*) (Tricco *et al.*, 2018), dont nous avons précédemment publié le protocole dans le dépôt *OSF registry* [1]. Suivant la proposition de Rogers & Luccioni (2024), nous considérons qu'un LLM (i) modélise, et peut générer du texte ; (ii) a été pré-entraîné sur un large corpus textuel [2] ; (iii) peut être utilisé pour de l'apprentissage par transfert ; (iv) utilise une architecture basée sur le transformeur (Vaswani *et al.*, 2017) [3]. Ceci nous permet d'inclure sous l'appellation de *LLM* tout aussi bien des modèles de type encodeur dérivés de BERT (Devlin *et al.*, 2019) que des modèles génératifs basés sur GPT (Radford & Narasimhan, 2018). Nous définissons un *biais* comme la « présence d'erreurs systématiques ou de disparités dans des processus de prise de décision qui affectent certains groupes de la population de manière disproportionnée » (Yang *et al.*, 2024a), et l'*utilité clinique* d'une prédiction comme son « utilité à améliorer l'état des patient·es, à informer la prise de décision médicale, et à optimiser les ressources de santé » (Badrick & Bowling, 2023).

2.2 Identification et sélection des études

Nous considérons cinq sources de littérature représentatives des domaines informatique, biomédical et du TAL : MEDLINE (PubMed), Web of Science, IEEE Xplore, ACM Digital Library, et l'ACL Anthology. Nous recherchons des articles dont le titre et/ou le résumé contiennent au moins un mot-

1. https://osf.io/ygknh/overview?view_only=aac52b70d01f4bcb994ab1533c0bbc74.
2. Ici, c'est le corpus d'entraînement qui est qualifié de *large*, et non pas le nombre de paramètres du modèles. Un seuil indicatif suggéré (bien que nécessairement arbitraire) est d'un milliard de tokens pour un texte en anglais.
3. Les trois premiers critères sont proposés par Rogers & Luccioni (2024), nous ajoutons le dernier.

clé pour chacun des thèmes SANTÉ MENTALE, LLM et PRÉDICTION[4]. Nous effectuons ensuite l'étape de sélection des études pertinentes à l'aide de la plateforme Rayyan (Ouzzani *et al.*, 2016), sur la base des critères d'inclusion suivants :

1. article de recherche original publié entre 2019 et 2024 ;
2. comprenant au moins une tâche de prédiction en lien avec la santé mentale d'individus ;
3. faisant pour cela usage d'au moins un LLM ;
4. exploitant pour cela des données textuelles authentiques (c'est-à-dire, non synthétiques), possiblement en parallèle d'autres modalités (*e.g.,* audio).

La clarté des critères d'inclusion a été mesurée à l'aide de scores d'accord inter-annotateur·ices (n=3) sur un échantillon de 100 études. Nous avons obtenu un accord élevé (kappa de Cohen compris entre 0,65 et 0,81 selon les paires d'annotateur·ices) et discuté ensemble de légères reformulations à apporter (finalisées dans la liste ci-dessus) pour optimiser l'accord et lever les ambiguïtés dans la suite de la phase de sélection.

2.3 Extraction et analyse

Il n'existe à notre connaissance pas de cadre théorique pensé spécifiquement pour l'analyse de la prise en compte des biais et de l'utilité clinique dans les applications prédictives des LLM en santé mentale. Cependant, nous identifions deux précédents travaux nous semblant pertinents, et selon nous complémentaires. D'une part, Chen *et al.* (2021) déclinent les enjeux éthiques de l'apprentissage automatique en santé tout au long du cycle de développement (*pipeline*) des systèmes considérés, ici modélisé en cinq étapes, depuis leur conception jusqu'à leur déploiement. D'autre part, Hovy & Prabhumoye (2021) identifient cinq sources majeures de biais en TAL, également situées de part et d'autre de ce cycle de développement, ce qui étend la réflexion habituellement centrée sur les données à d'autres sources de biais telle que la phase d'annotation, ou encore les modèles utilisés. Nous proposons d'hybrider ces deux approches pour constituer un cadre de lecture en cinq étapes de développement : (i) Conception de la recherche et sélection du problème, (ii) Collecte des données, (iii) Modélisation du résultat, (iv) Développement du modèle, et (v) Considérations post-déploiement. À chacune de ces étapes, nous attachons une liste d'entités-cibles, caractérisant les informations ciblées lors de l'extraction, et permettant une analyse des biais et de l'utilité clinique en lien avec l'étape considérée. Ceci est récapitulé en Figure 1.

Lors de la phase d'extraction, nous récupérons les méta-données suivantes : titre, liste des auteur·ices, année de publication, journal/conférence. Puis, les entités listées en Figure 1 sont relevées. À ce stade, il importe de préciser que nous relevons aussi bien les affirmations des auteur·ices concernant ce qui a *effectivement* été fait dans leurs travaux (*e.g.,* « F1-score » pour l'entité *évaluation* de l'étape 4), que leurs réflexions sur ce qui *aurait pu* être fait autrement, ou amélioré (*e.g.,* pour cette même entité *évaluation*, nous relèverions une phrase du type « Nous reconnaissons que la métrique de F1-score n'est pas la plus adaptée pour minimiser les faux négatifs »).

Enfin, l'étape d'analyse s'attache à interpréter les résultats issus de la phase d'extraction. Nous nous inspirons de méthodes utilisées en Sciences Humaines et Sociales (Glaser *et al.*, 1998; Smith & Smith, 1977) et créons des catégories porteuses de sens pour les différentes entités ciblées, raffinées dans un processus itératif et agrégatif. Précisons deux choix particuliers de représentation : ce que nous appelons « expertise médicale d'un·e auteur·ice » est une variable binaire valant 1 si l'auteur·ice

4. Les requêtes détaillées sont fournies au lien suivant : https://github.com/ClementineBleuze/bias-clinical-utility-ScR.

compte parmi ses affiliations un diplôme de médecine, un département médical d'une université, ou une entreprise en lien avec le milieu médical, et 0 sinon. Le ratio d'auteur·ices expert·es sur le nombre total d'auteur·ices constitue alors une approximation du niveau d'expertise médical représenté dans un article donné. De plus, nous regroupons les différentes dénominations relevées dans les articles autour des troubles de la santé mentale (*e.g.*, « dépression », « risque de dépression ») sous des étiquettes tirées de la Section II du Manuel diagnostique et statistique des troubles mentaux (DSM)[5] (*e.g.*, « Troubles Dépressifs »). Nous ajoutons également la catégorie « Risque Suicidaire » en raison de sa forte prévalence dans les études identifiées.

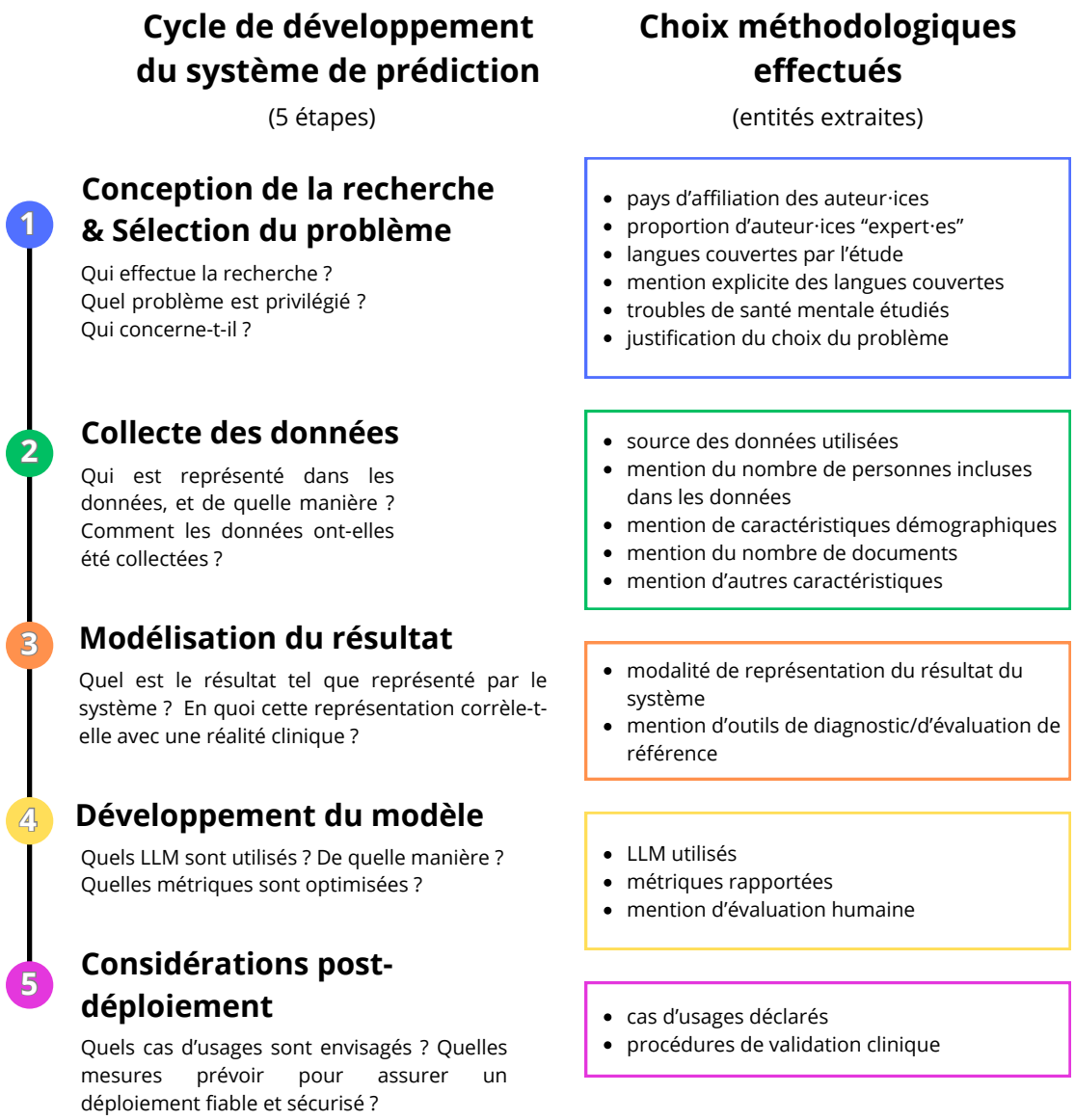

FIGURE 1 – À gauche : notre proposition de cycle de développement (*pipeline*) en cinq étapes pour décrire les systèmes de prédiction de santé mentale à partir de LLM. À droite : entités-cibles relevées lors de la phase d'extraction pour chacune des étapes du cycle.

5. Voir codes diagnostics répertoriés dans le DSM-5 : https://www.infodrog.ch/files/content/refbases/DSM-5_Manuel-diagnostique-et-statistique-des-troubles-mentaux.pdf.

3 Résultats

3.1 Études incluses

Nous identifions initialement 2 646 études candidates dans les sources de littérature considérées. Après suppression des doublons, 2 472 études sont triées au vu des critères d'éligibilité, sur la base de leur titre et résumé. Les 263 études restantes sont ensuite filtrées sur la base d'une lecture intégrale. Finalement, 201 études sont incluses dans cette revue, avant de passer par les étapes d'extraction et d'analyse[6]. Les détails du processus sont présentés dans la Figure 2.

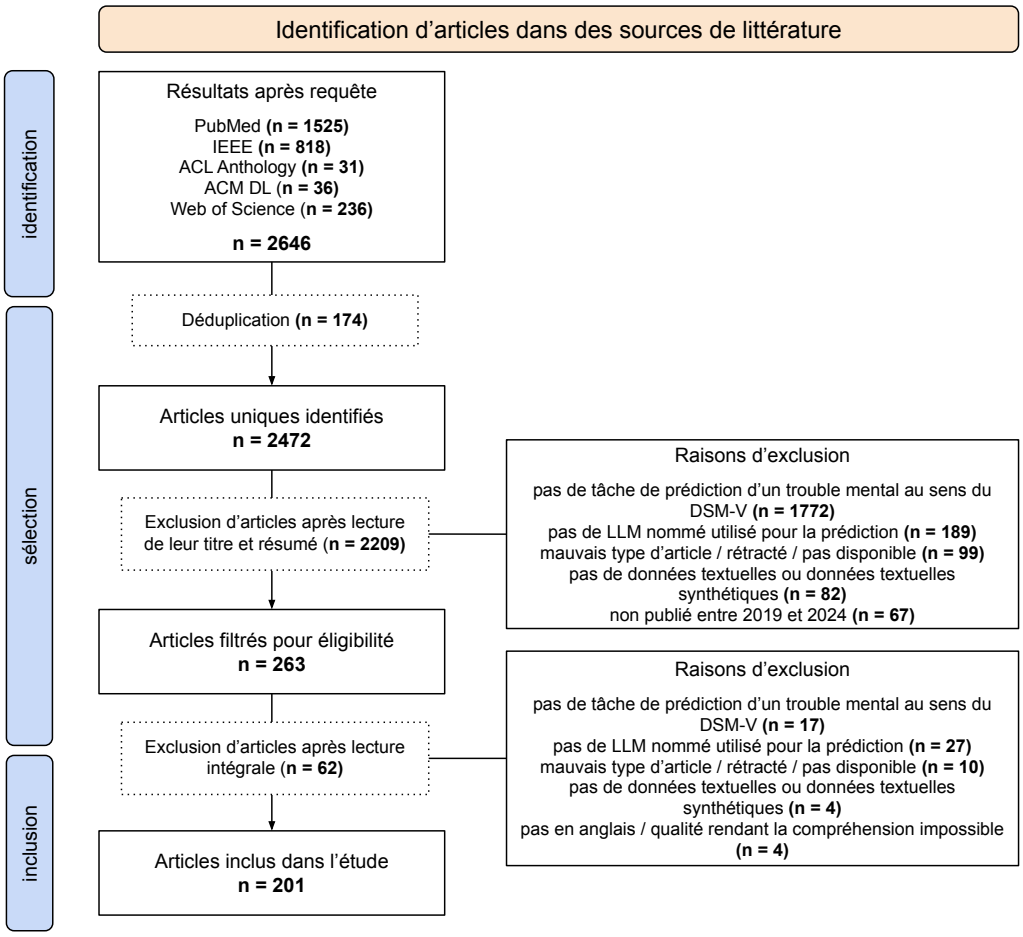

FIGURE 2 – Diagramme de flux PRISMA de l'étude

Bien que l'ensemble de la période 2019-2024 soit considérée, nous notons que 74,1 % (n=149) des études incluses sont publiées entre 2023 et 2024, ce qui semble suivre l'accélération du développement des LLM en TAL. Par ailleurs, IEEE Xplore est la source la plus représentée dans notre corpus (n=124 ; 61,7 %), suivie par PubMed (n=57 ; 28,4 %). L'ACL Anthology ne fournit que 10 articles (5,0 %), à égalité avec le Web of Science et l'ACM Digital Library réunis.

6. La liste des études incluses est fournie au lien suivant : https://github.com/ClementineBleuze/bias-clinical-utility-ScR.

3.2 Tendances générales (QR1)

Dans cette section, nous décrivons les tendances générales observées au sein de 201 études intégrant des LLM pour des prédictions de santé mentale, regroupées par étapes du cycle de développement.

Étape 1 : Conception de la recherche et sélection du problème. Quarante-cinq pays d'affiliation sont identifiés, parmi lesquels la Chine (n=43 ; 21,4 %), les États-Unis (n=42 ; 20,9 %) et l'Inde (n=32 ; 15,9 %) concentrent un grand nombre d'auteur·ices ; nous observons également de fréquentes collaborations internationales (n=54 ; 26,9 %). Le taux moyen d'auteur·ices expert·es du domaine est de 0,22 ($Q_1 = Q_2 = 0,0$; $Q_3 = 0,33$), avec un écart standard de 0,34. La majorité des articles ne comportent aucun·e expert·e (n=125 ; 62,2 %), tandis qu'une minorité est écrite uniquement par des expert·es (n=20 ; 10,0 %). Les Troubles Dépressifs (n=148 ; 73,6 %), le Risque Suicidaire (n=46 ; 22,9 %) et les Troubles Anxieux (n=17 ; 8,5 %) sont les plus représentées des quinze famille de troubles mentaux identifiées (Figure 3) ; tandis que les justifications les plus couramment employées pour appuyer le choix d'un trouble particulier réfèrent aux impacts négatifs sur les individus concernés (n=165 ; 82,1%), à une prévalence importante ou en augmentation (n=170 ; 84,6 %) ou à des difficultés rencontrées par le système de santé « traditionnel » (n=127 ; 63,2 %). Enfin, les études produisent majoritairement des systèmes pensés pour l'anglais (n=153 ; 76,1 %), ce qui est explicité par les auteur·ices dans seulement 51,0 % des cas[7].

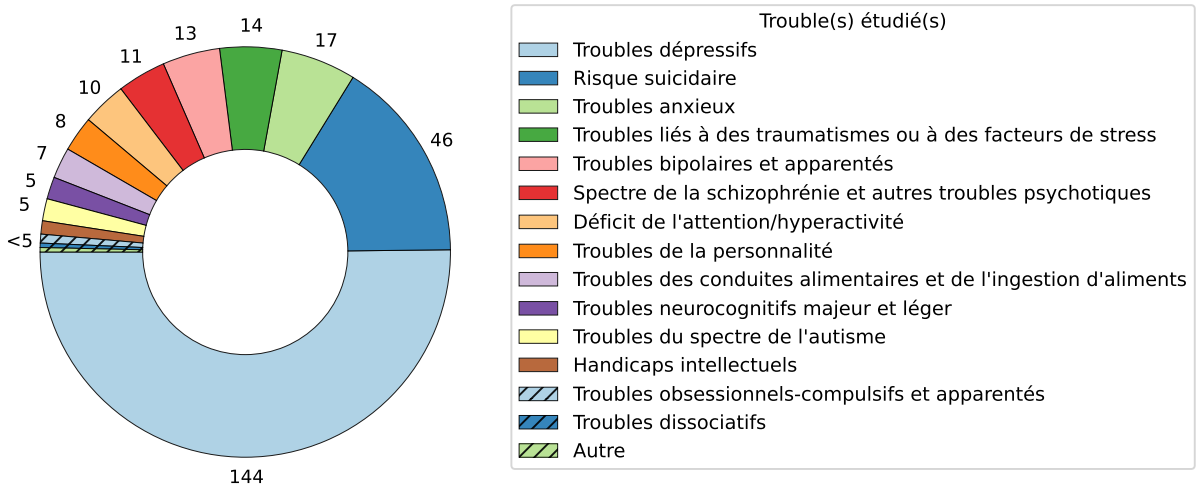

FIGURE 3 – Distribution des troubles mentaux étudiés au sein des articles inclus. Certains articles étudient plusieurs troubles, ce qui explique que la somme des nombres rapportés dépasse 201.

Étape 2 : Collecte des données. Les études incluses mobilisent de nombreux jeux de données, soit pré-existants, soit collectés pour l'occasion. Les *posts* d'utilisateur·ices sur les réseaux sociaux constituent une source pour 133 études (66,2 %), principalement *via* Reddit, Twitter, et Sina Weibo ; tandis que 62 (30,8 %) mobilisent des échanges médicaux (*e.g.,* des messages adressés par des patient·es à des soignant·es) ou des entretiens (semi-)directifs (*e.g.,* ceux recueillis dans le DAIC (Gratch *et al.*, 2014)). Les dossiers électroniques de patient·es (*EHR*) sont utilisés dans 11 (5,5 %) études ; ce qui inclut des bases de données publiques telles que ScAN (Rawat *et al.*, 2022) et des jeux de données privés. Dans 122 articles (60,7 %), nous ne trouvons aucune mention du nombre d'individus

7. Dans les cas où la langue n'est pas explicitée, les références aux ressources et/ou aux modèles employés permettent de déduire que les auteur·ices travaillent sur de l'anglais.

représentés dans l'un ou plusieurs des jeux de données utilisés ; ceci s'accentue pour les éléments de profil démographique (âge, sexe/genre, etc.), présents dans seulement 43 articles (21,4 %). Par ailleurs, respectivement 172 (85,6 %) et 127 (63,2 %) articles font mention du nombre de documents ainsi que de détails supplémentaires (année de collecte, distribution des annotations) concernant ces mêmes jeux de données.

Étape 3 : Modélisation du résultat. Les prédictions effectuées par les systèmes de TAL décrits dans notre corpus se déclinent selon plusieurs modalités. Dans 146 (72,6 %) articles, les auteur·ices ont recours à une étiquette binaire (présence/absence) pour qualifier le risque de présenter un trouble donné. D'autres (n=46 ; 22,8 %) utilisent des échelles de sévérité qualitatives (*e.g.,* risque de dépression faible/modéré/sévère) ou quantitatives (n=19 ; 9,5 %), basées notamment sur des questionnaires standardisés. Enfin, neuf articles (4,5 %) fournissent un ensemble de prédictions au niveau des symptômes (*e.g.,* désespoir, auto-dévalorisation). Soixante-quatre (31,8 %) articles déclarent s'appuyer sur des outils d'évaluation, de diagnostic ou de classification de référence en psychiatrie, tels que les questionnaires d'estimation de la dépression PHQ-8/PHQ-9 (*Patient health questionnaire*) (n=30 ; 14,9 %), la Classification Internationale des Maladies (CIM) (n=11 ; 5,5 %), ou le DSM (n=10 ; 5,0 %). Nous notons une diversité des pratiques quant à l'exploitation des questionnaires : certaines études s'en servent pour structurer des entretiens dirigés conduits avec les participant·es, tandis que d'autres s'appuient sur les scores d'auto-administration de participant·es volontaires. Dans les 137 (68,2 %) études restantes, la méthode d'identification des troubles de santé mentale est omise ou bien définie en dehors des outils de références (annotation manuelle des données selon des consignes sur-mesure, heuristiques basées sur des mots-clés, etc.).

Étape 4 : Développement du modèle. Nous identifions 112 LLM distincts mentionnés dans notre corpus, avec une moyenne de 2,5 LLM employés par étude. Nous observons une large prévalence de modèles encodeurs tels que BERT (Devlin *et al.*, 2019) (n=124 ; 61,7 %) et ses dérivés RoBERTa (Liu *et al.*, 2019) (n=49 ; 24,4 %) ou encore DistilBERT (Sanh *et al.*, 2020) (n=25 ; 12,4 %). Cependant, des modèles de type décodeur basés sur GPT (Radford & Narasimhan, 2018) (n=34 ; 16,9 %) ou encore Llama (Touvron *et al.*, 2023) (n=9 ; 4,5 %) sont également employés. Nous notons un emploi minoritaire (n=38 ; 18,9 %) de modèles précédemment entraînés sur des données en lien avec la santé mentale (*e.g.,* MentalBERT et MentalRoBERTa (Ji *et al.*, 2021), BioClinicalBERT (Alsentzer *et al.*, 2019)), au profit de modèles généralistes, non spécialisés. Concernant l'usage qui est fait de ces LLM, il s'agit de produire une représentation des données d'entrée, ensuite fournie à un autre type de modèle, ou, plus souvent, d'effectuer directement la prédiction (*e.g.,* en *promptant* un modèle décodeur, ou en ajoutant une tête de classification à un modèle encodeur). Enfin, nous observons des usages annexes pour effectuer du pré-traitement sur les données d'entrée ou pour générer des explications quant aux prédictions effectuées. Pour l'évaluation, nous relevons des métriques standard en classification/régression (précision, rappel, F-mesure, courbes ROC, coefficients de corrélation, etc.), ainsi que des métriques sur-mesure associées à certaines *shared tasks*. L'évaluation humaine de la qualité des systèmes n'est généralement pas conduite, à l'exception de Wang *et al.* (2023) qui mobilisent des internes de médecine pour évaluer la sécurité, l'utilisabilité et la fluidité de leur outil.

Étape 5 : Considérations post-déploiement. Nous observons une faible propension de la part des auteur·ices à expliciter des cas d'usages concrets et précis pour les systèmes développés. Cependant, l'ambition plus générale de faciliter la détection précoce de troubles mentaux revient fréquemment (n=57 ; 28,4 %). Ceci se décline dans des propositions visant à exploiter les dossiers médicaux de patient·es (*e.g.,* pour anticiper un début de dépression chez des patient·es atteint·es de cancer (de Hond

et al., 2024; van Buchem *et al.*, 2024)), ou encore à surveiller les réseaux sociaux pour suivre des tendances de santé publique (*e.g.,* pour la dépression post-partum (Dhankar & Katz, 2023) ou le risque suicidaire (Boonyarat *et al.*, 2024)). D'autres imaginent également des dispositifs à l'interface entre patient·e et soignant·e (Diniz *et al.*, 2022), ou pour simuler des questionnaires standardisés (Shimamoto *et al.*, 2024). Enfin, nous n'avons pas identifié d'étude présentant une validation clinique du système proposé dans des conditions réalistes, ou proches de conditions cliniques, par exemple à l'aide d'un essai randomisé contrôlé (*RCT*). Ainsi, si Lee *et al.* (2024) comparent la performance de LLM à celle de six cliniciens confirmés pour évaluer le risque suicidaire chez les utilisateur·ices d'une plateforme de télé-santé, les auteur·ices reconnaissent des résultats encore insuffisants pour une utilisation en pratique clinique.

3.3 Documentation des biais et de l'utilité clinique (QR2)

Après avoir décrit la manière dont les auteur·ices de notre corpus procèdent pour effectuer des prédictions de santé mentale à l'aide de LLM, nous rendons compte dans cette section des réflexions qu'ils et elles soulèvent quant aux possibles biais ou à l'utilité clinique de leur démarche. Il est important de soulever que ceci peut inclure des limitations effectivement reconnues par les auteur·ices quant à leur travail, ou au contraire des affirmations que les biais ont été pris en compte et mitigés, ou que l'utilité clinique est assurée. En bref, nous mesurons systématiquement le nombre de *mentions* de thèmes liés aux biais et à l'utilité clinique (à quel point ces sujets sont-ils abordés ?), sans nous avancer avec précision sur leur contenu. Chacun des thèmes identifiés dans au moins un des articles a été associé à une étape du cycle de développement, puis recensé dans l'ensemble du corpus (Figure 4).

Nous identifions des mentions de thèmes liés aux biais ou à l'utilité clinique dans 164 (81,6 %) articles. Les étapes du cycle de développement les plus discutées sont la *Collecte des données* (n=108 ; 53,7 %), le *Développement du modèle* (n=71 ; 35,3 %) et la *Conception de la recherche et sélection du problème* (n=60 ; 29,8 %). Plus rarement, des thèmes liés aux étapes de *Définition des critères d'évaluation* et de *Considérations post-déploiement* sont évoquées dans respectivement 45 (22,4 %) et 31 (15,4 %) études. Notons que la discussion est confinée à une unique étape du cycle de développement dans 75 (37,3 %) articles, tandis que respectivement 47 (23,4 %), 27 (13,4 %), et 10 (5,0 %) articles en couvrent deux, trois et quatre. Cinq articles (2,5 %) évoquent des thèmes issus de l'ensemble des cinq étapes du cycle de développement.

Étape 1 : Conception de la recherche et sélection du problème. Des mentions relevées à cette étape identifient la langue couverte par une étude comme source de biais, soit que les auteur·ices reconnaissent travailler exclusivement sur de l'anglais, soit qu'ils et elles soulignent la rareté des travaux sur leur langue de travail. D'autres appuient l'utilité clinique de leur travaux pour des populations particulièrement exposées aux troubles de la santé mentale (*e.g.,* des femmes enceintes/en post-partum (Bartal *et al.*, 2024), des vétérans (Zuromski *et al.*, 2024)). Enfin, certains affirment travailler sur des troubles sous-étudiés, comme Juhng *et al.* (2023) qui affirment que « bien moins de travaux en TAL se sont concentrés sur la détection des troubles anxieux que des troubles dépressifs ».

Étape 2 : Collecte des données. Plusieurs caractéristiques des jeux de données sont mentionnées comme source de biais : des classes déséquilibrées (*e.g.,* bien plus de patient·es classifié·es « non schizophrènes » que « schizophrènes ») - parfois ré-équilibrées avec des techniques d'échantillonnage ou d'augmentation, un manque de diversité démographique parmi les personnes représentées (*e.g.,* en termes de genre, d'ethnicité, de statut social), ou encore le manque de données pour constituer un jeu

de taille satisfaisante. Certaines études soulèvent que le diagnostic clinique nécessite des données continues dans le temps (ce qui s'oppose aux jeux de données « ponctuels »), tandis que d'autres évoquent les biais présents dans les données elles-mêmes (*e.g.*, introduits par l'auteur·ice d'une note clinique).

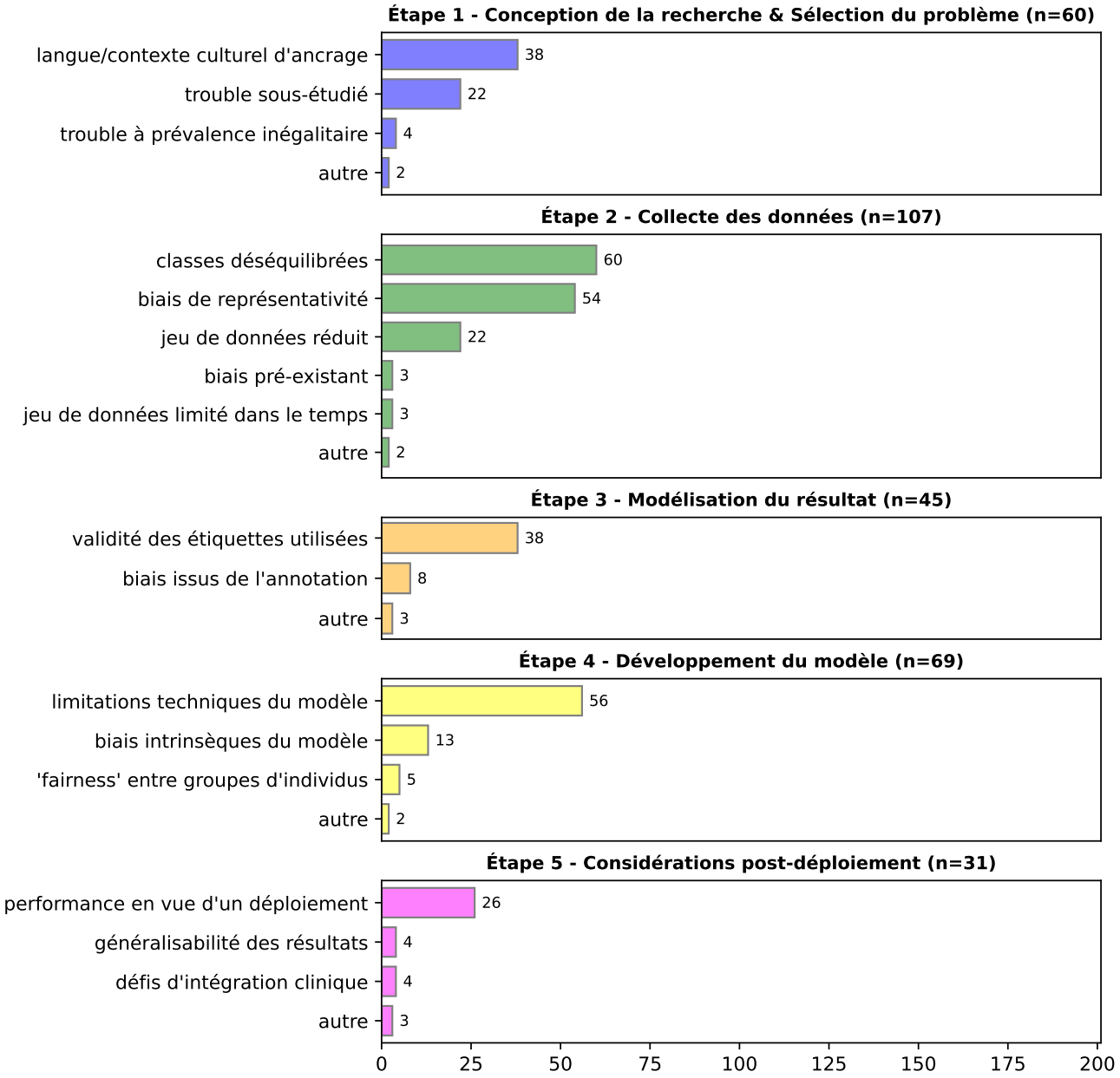

FIGURE 4 – Thèmes en lien avec les biais/l'utilité clinique des systèmes développés, et nombre d'études incluses les abordant, par étape du cycle de développement. Les thèmes associés à une étape et abordés dans moins de 3 études sont fusionnés dans un thème "autre".

Étape 3 : Modélisation du résultat. La validité (en termes d'interprétation clinique) des étiquettes prédites par le système est le principal défi mentionné à cette étape, notamment à l'encontre des prédictions binaires, et des étiquettes issues d'auto-diagnostics. Les erreurs et approximations potentiellement induites lors de la phase d'annotation sont également marginalement évoquées.

Étape 4 : Développement du modèle. De nombreuses limitations techniques des LLM sont mentionnées, qui peuvent impacter négativement l'utilité clinique des systèmes (*e.g.,* le défi du traitement de longs documents, les coûts computationnels). Quelques études mentionnent les biais intrinsèquement portés par les LLM. Seules cinq évoquent la notion d'équité entre groupes (*fairness*), dont certaines rapportent des résultats désagrégés en fonction de l'âge, du sexe/genre ou de l'ethnicité des personnes incluses (Huang *et al.*, 2022; de Hond *et al.*, 2024; van Buchem *et al.*, 2024).

Étape 5 : Considérations post-déploiement. La performance des systèmes développés dans un contexte de déploiement hypothétique occupe la majorité des mentions relevées ; notons qu'à cette occasion des jugements négatifs comme positifs sont émis[8]. Les autres mentions concernent la généralisabilité de résultats ponctuels à des situations et/ou populations différentes, ou encore des difficultés liées à l'intégration en routine clinique (*e.g.,* au vu des contraintes réglementaires sur la sécurité des données de santé).

4 Discussion

Un champ de recherche relativement peu diversifié, présentant de multiples biais. Notre revue permet de mettre en évidence un profil prototypique d'étude exploitant des LLM pour prédire la santé mentale : focalisée sur les troubles dépressifs, en anglais, *via* des *posts* Reddit, effectuant des prédictions binaires à l'aide d'un modèle dérivé de BERT, en l'absence de validation clinique ; là où une plus grande diversité d'approches semblerait davantage productive. Nous retrouvons des biais documentés par ailleurs en TAL, tels que l'omniprésence de la langue anglaise (au détriment de langues moins dotées) (Hovy & Prabhumoye, 2021; Bender, 2019; Ducel *et al.*, 2022), déclarée dans seulement 51,0 % des cas. Par ailleurs, le recours massif aux réseaux sociaux (perçus comme des sources faciles d'accès, à grande échelle) questionne quant à la qualité et à la représentativité des données extraites (Chen *et al.*, 2021), et risque d'effacer les individus (au profil démographique rarement connu) ainsi que leur consentement à l'utilisation de telles données (Benton *et al.*, 2017). Nous remarquons également qu'il est rare pour les auteur·ices de motiver explicitement leur choix de certains LLM au détriment d'autres ; en particulier, la spécialisation d'un modèle (ici en santé mentale) ou sa propension à exacerber les biais ne semblent pas déterminants, ni son coût computationnel, pourtant en lien direct avec son implémentabilité réelle et son impact environnemental (Ahmed *et al.*, 2023; Morand *et al.*, 2024). Enfin, si les auteur·ices font mention d'un certain nombre de thèmes en lien avec les biais ou l'utilité clinique de leur système, ceci reste peu systématisé à l'ensemble du cycle de développement. Les thèmes privilégiés semblent témoigner d'une vision des biais issus des données, mitigés par des approches techniques ; ce qui peut se révéler insuffisant pour traiter de ces problèmes en profondeur (Hofmann *et al.*, 2024; Resnik, 2025).

8. Par exemple, Matero *et al.* (2022) recommandent de ne pas utiliser (à ce stade) leur système en pratique clinique, tandis que Bartal *et al.* (2024) sont confiant·es du potentiel de leur modèle à « s'intégrer sans accroc en routine de soins obstétriques ».

La nécessité d'une approche inter-disciplinaire pour accroître l'utilité clinique. Nos observations mettent en évidence une faible intégration d'auteur·ices expertes du domaine médical (37,8 % des cas), là où l'interdisciplinarité peut pourtant accroître la validité des résultats (Littmann *et al.*, 2020). Ainsi, une expertise clinique permet de définir adéquatement des termes comme « dépression »[9], parfois assimilés de manière ambigüe et réductrice dans certains articles à un simple « sentiment négatif » ou à un comportement suicidaire. Il importe également pour d'hypothétiques usages cliniques de préciser si les systèmes développés sont entraînés sur la base d'auto-diagnostics exprimés sur les réseaux sociaux, ou bien de diagnostics établis par des psychiatres dans un parcours de soin. En parallèle des recommandations émises par Chen *et al.* (2021) (auditer les systèmes développés, mesurer systématiquement l'impact sur différentes populations, etc.), des cadres théoriques interdisciplinaires tels que le V3 (Goldsack *et al.*, 2020), notamment utilisé par l'Agence européenne des médicaments, peuvent constituer une source d'inspiration. Celui-ci appuie la nécessité d'une validation clinique, logicielle, et analytique des systèmes numériques, et s'est enrichi récemment d'un critère de validation de l'utilisabilité (Bakker *et al.*, 2025), qui ré-oriente la réflexion vers l'usage clinique et, ultimement, les patient·es supposé·es en bénéficier.

Limitations. Bien qu'intégrant 201 études, notre revue demeure nécessairement non-exhaustive, et limitée à des travaux rédigés en anglais, publiés entre 2019 et 2024. Il est possible que l'inclusion de littérature plus récente puisse faire émerger de nouveaux phénomènes (par exemple, l'augmentation des LLM génératifs par rapports aux modèles encodeurs basés sur BERT) ; nous supposons cependant que les tendances globales esquissées demeurent valides, en tant qu'elles sont représentatives d'une période couvrant 6 années de recherche. Notre but étant de décrire un large panorama de la littérature sur les prédictions de santé mentale assistées par LLM, nous n'avons pas procédé à une évaluation de la qualité scientifique des articles inclus. Par ailleurs, la notion de « niveau de preuve » (*level of evidence*) s'applique difficilement au domaine du TAL, et les outils existants pour d'autres disciplines (par exemple le QUADAS-2) ne sont pas adaptés à notre cas d'étude. En raison de la taille conséquente du corpus inclus, la phase d'extraction n'a pas pu être assurée de manière indépendante par plusieurs auteur·ices, ce que nous avons compensé au mieux en nous réunissant régulièrement pour échanger sur des éléments méthodologiques.

5 Conclusion

Les Grands Modèles de Langue sont de plus en plus utilisés en TAL pour simuler la prédiction de troubles psychiatriques. À l'aide d'un cadre de lecture suivant le cycle de développement des systèmes proposés, nous avons décrit un domaine émergent aux tendances méthodologiques marquées, notamment vers la prédiction des troubles dépressifs sur des données issues des réseaux sociaux. Si les auteur·ices font généralement mention de thèmes en lien avec les biais et l'utilité clinique, ceci demeure non systématique. En particulier, les considérations post-déploiement et études d'impact manquent, ce qui nous conduit à questionner l'idée que les LLM révolutionnent actuellement la santé mentale. En ce qui concerne l'utilisation de méthodes de TAL pour des applications médicales, nous défendons une vue combinée des notions de biais et d'utilité clinique, requérant des efforts inter-disciplinaires, pour remettre les soignant·es et patient·es au cœur de la réflexion.

9. Ce concept est particulièrement ambigu ; il peut s'agir dans le thésaurus MeSH (*Medical Subject Headings* ; voir https://meshb.nlm.nih.gov/) d'un symptôme ou bien d'un trouble, avec plusieurs niveaux de sévérité.

Remerciements

Ce travail a bénéficié du soutien du projet ANR InExtenso (ANR-23-IAS1-0004).

Références

ABBE A., GROUIN C., ZWEIGENBAUM P. & FALISSARD B. (2016). Text mining applications in psychiatry : a systematic literature review. *International Journal of Methods in Psychiatric Research*, **25**(2), 86–100. DOI : https ://doi.org/10.1002/mpr.1481.

AHMED M. I., SPOONER B., ISHERWOOD J., LANE M., ORROCK E. & DENNISON A. (2023). A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*, **15**(10), e46454. DOI : 10.7759/cureus.46454.

ALSENTZER E., MURPHY J., BOAG W., WENG W.-H., JINDI D., NAUMANN T. & MCDERMOTT M. (2019). Publicly available clinical bert embeddings. In A. RUMSHISKY, K. ROBERTS, S. BETHARD & T. NAUMANN, Éds., *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 72–78, Minneapolis, Minnesota, États-Unis : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1909.

AMBLARD M., MUSIOL M. & REBUSCHI M. (2021). Discourse coherence - From psychology to linguistics and back again. In M. AMBLARD, M. MUSIOL & M. REBUSCHI, Éds., *(In)coherence of discourse - Formal and Conceptual issues of Language*, p. 1–17. Springer. Language, Cognition and Mind, DOI : 10.1007/978-3-030-71434-5_1, HAL : hal-02269640.

APPELL J., KERTESZ A. & FISMAN M. (1982). A study of language functioning in alzheimer patients. *Brain and Language*, **17**(1), 73–91. DOI : 10.1016/0093-934X(82)90006-2.

BADRICK T. & BOWLING F. (2023). Clinical utility – information about the usefulness of tests. *Clinical Biochemistry*, **121–122**, 110656. DOI : 10.1016/j.clinbiochem.2023.110656.

BAKKER J. P., BARGE R., CENTRA J., COBB B., COTA C., GUO C. C., HARTOG B., HOROWICZ-MEHLER N., IZMAILOVA E. S., MANYAKOV N. V., MCCLENAHAN S., MOTOLA S., PATEL S., PAUN O., SCHOONE M., SEZGIN E., SWITZER T., TANDON A., VAN DEN BRINK W., VAIRAVAN S., VANDENDRIESSCHE B., VRIJENS B. & GOLDSACK J. C. (2025). V3+ extends the V3 framework to ensure user-centricity and scalability of sensor-based digital health technologies. *NPJ Digit. Med.*, **8**(1), 51. DOI : 10.1038/s41746-024-01322-2.

BARTAL A., JAGODNIK K. M., CHAN S. J. & DEKEL S. (2024). Ai and narrative embeddings detect ptsd following childbirth via birth stories. *Scientific Reports*, **14**(1), 8336. DOI : 10.1038/s41598-024-54242-2.

BENDER E. (2019). The #BenderRule : On naming the languages we study and why it matters. *The Gradient*, **14**(1).

BENTON A., COPPERSMITH G. & DREDZE M. (2017). Ethical research protocols for social media health research. In D. HOVY, S. SPRUIT, M. MITCHELL, E. M. BENDER, M. STRUBE & H. WALLACH, Éds., *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, p. 94–102, Valence, Espagne : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1612.

BOONYARAT P., LIEW D. J. & CHANG Y.-C. (2024). Leveraging enhanced bert models for detecting suicidal ideation in thai social media content amidst covid-19. *Information Processing & Management*, **61**(4), 103706. DOI : 10.1016/j.ipm.2024.103706.

CHEN I. Y., PIERSON E., ROSE S., JOSHI S., FERRYMAN K. & GHASSEMI M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, **4**(1), 123–144. DOI : 10.1146/annurev-biodatasci-092820-114757.

DE HOND A., VAN BUCHEM M., FANCONI C., ROY M., BLAYNEY D., KANT I., STEYERBERG E. & HERNANDEZ-BOUSSARD T. (2024). Predicting depression risk in patients with cancer using multimodal data : Algorithm development study. *JMIR medical informatics*, **12**, e51925. DOI : 10.2196/51925.

DEMNER-FUSHMAN D., CHAPMAN W. W. & MCDONALD C. J. (2009). What can natural language processing do for clinical decision support ? *Journal of Biomedical Informatics*, **42**(5), 760–772. DOI : 10.1016/j.jbi.2009.08.007.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota, États-Unis : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DHANKAR A. & KATZ A. (2023). Tracking pregnant women's mental health through social media : an analysis of reddit posts. *JAMIA Open*, **6**(4), ooad094. DOI : 10.1093/jamiaopen/ooad094.

DINIZ E. J. S., FONTENELE J. E., DE OLIVEIRA A. C., BASTOS V. H., TEIXEIRA S., RABÊLO R. L., CALÇADA D. B., DOS SANTOS R. M., DE OLIVEIRA A. K. & TELES A. S. (2022). Boamente : A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. *Healthcare*, **10**(4), 698. DOI : 10.3390/healthcare10040698.

DUCEL F., FORT K., LEJEUNE G. & LEPAGE Y. (2022). Do we name the languages we study ? the #BenderRule in LREC and ACL articles. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 564–573, Marseille, France : European Language Resources Association.

DUCEL F., HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2025). "women do not have heart attacks !" gender biases in automatically generated clinical cases in french. In *Findings of the Association for Computational Linguistics : NAACL 2025*, p. 7145–7159, Albuquerque, Nouveau Mexique, États-Unis. DOI : 10.18653/v1/2025.findings-naacl.398.

FRASER K. C., MELTZER J. A. & RUDZICZ F. (2015). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, **49**(2), 407–422. DOI : 10.3233/JAD-150520.

GLASER B. G., STRAUSS A. L. *et al.* (1998). Grounded theory. *Strategien qualitativer Forschung. Bern : Huber*, **4**.

GOLDSACK J. C., CORAVOS A., BAKKER J. P., BENT B., DOWLING A. V., FITZER-ATTAS C., GODFREY A., GODINO J. G., GUJAR N., IZMAILOVA E., MANTA C., PETERSON B., VANDENDRIESSCHE B., WOOD W. A., WANG K. W. & DUNN J. (2020). Verification, analytical validation, and clinical validation (v3) : the foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). *NPJ Digit. Med.*, **3**(1), 55. DOI : 10.1038/s41746-020-0260-4.

GRATCH J., ARTSTEIN R., LUCAS G., STRATOU G., SCHERER S., NAZARIAN A., WOOD R., BOBERG J., DEVAULT D., MARSELLA S., TRAUM D., RIZZO S. & MORENCY L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources*

and Evaluation (LREC'14), p. 3123–3128, Reykjavik, Islande : European Language Resources Association (ELRA).

HOFMANN V., KALLURI P. R., JURAFSKY D. & KING S. (2024). Ai generates covertly racist decisions about people based on their dialect. *Nature*, **633**(8028), 147–154. DOI : 10.1038/s41586-024-07856-5.

HOVY D. & PRABHUMOYE S. (2021). Five sources of bias in natural language processing. *Language and linguistics compass*, **15**(8), e12432. DOI : 10.1111/lnc3.12432.

HUANG G., SHEN W., LU H., HU F., LI J. & LIU H. (2022). Multimodal depression detection based on factorized representation. In *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, p. 190–196, En ligne. DOI : 10.1109/HDIS56859.2022.9991717.

JI S., ZHANG T., ANSARI L., FU J., TIWARI P. & CAMBRIA E. (2021). Mentalbert : Publicly available pretrained language models for mental healthcare. arXiv :2110.15621 [cs], DOI : 10.48550/arXiv.2110.15621.

JUHNG S., MATERO M., VARADARAJAN V., EICHSTAEDT J., V GANESAN A. & SCHWARTZ H. A. (2023). Discourse-level representations can improve prediction of degree of anxiety. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 1500–1511, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-short.128.

LE GLAZ A., HARALAMBOUS Y., KIM-DUFOR D.-H., LENCA P., BILLOT R., RYAN T. C., MARSH J., DEVYLDER J., WALTER M., BERROUIGUET S. *et al.* (2021). Machine learning and natural language processing in mental health : systematic review. *Journal of medical Internet research*, **23**(5), e15708.

LEE C., MOHEBBI M., O'CALLAGHAN E. & WINSBERG M. (2024). Large language models versus expert clinicians in crisis prediction among telemental health patients : Comparative study. *JMIR mental health*, **11**, e58129. DOI : 10.2196/58129.

LEROY G., GU Y., PETTYGROVE S., GALINDO M. K., ARORA A. & KURZIUS-SPENCER M. (2018). Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders : Development, evaluation, and application. *Journal of Medical Internet Research*, **20**(11), e10497. DOI : 10.2196/10497.

LITTMANN M., SELIG K., COHEN-LAVI L., FRANK Y., HÖNIGSCHMID P., KATAKA E., MÖSCH A., QIAN K., RON A., SCHMID S., SORBIE A., SZLAK L., DAGAN-WIENER A., BEN-TAL N., NIV M. Y., RAZANSKY D., SCHULLER B. W., ANKERST D., HERTZ T. & ROST B. (2020). Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence*, **2**(1), 18–24. DOI : 10.1038/s42256-019-0139-8.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. DOI : 10.48550/arXiv.1907.11692.

MARINI A., SPOLETINI I., RUBINO I. A., CIUFFA M., BRIA P., MARTINOTTI G., BANFI G., BOCCASCINO R., STROM P., SIRACUSANO A., CALTAGIRONE C. & SPALLETTA G. (2008). The language of schizophrenia : An analysis of micro and macrolinguistic abilities and their neuropsychological correlates. *Schizophrenia Research*, **105**(1), 144–155. DOI : https ://doi.org/10.1016/j.schres.2008.07.011.

MATERO M., HUNG A. & SCHWARTZ H. A. (2022). Evaluating contextual embeddings and their extraction layers for depression assessment. In J. BARNES, O. DE CLERCQ, V. BARRIERE, S. TAFRESHI, S. ALQAHTANI, J. SEDOC, R. KLINGER & A. BALAHUR, Éds., *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, p.

89–94, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.wassa-1.9.

MORAND C., LIGOZAT A.-L. & NÉVÉOL A. (2024). Bracing for impact : on-going digitalization of healthcare requires urgent characterization of impact on environment and beyond. In *Undone Computer Science*, Nantes, France : Guillaume Munch-Maccagnoni and Chantal Enguehard and Maël Pégny and Marc Anderson. HAL : hal-04579545.

MOULAEI K., YADEGARI A., BAHARESTANI M., FARZANBAKHSH S., SABET B. & AFRASH M. R. (2024). Generative artificial intelligence in healthcare : A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics*, **188**, 105474. DOI : 10.1016/j.ijmedinf.2024.105474.

ONG J. C. L., CHANG S. Y.-H., WILLIAM W., BUTTE A. J., SHAH N. H., CHEW L. S. T., LIU N., DOSHI-VELEZ F., LU W., SAVULESCU J. *et al.* (2024). Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, **6**(6), e428–e432.

OUZZANI M., HAMMADY H., FEDOROWICZ Z. & ELMAGARMID A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, **5**(1), 210. DOI : 10.1186/s13643-016-0384-4.

RADFORD A. & NARASIMHAN K. (2018). Improving language understanding by generative pre-training. Pre-print.

RAWAT B. P. S., KOVALY S., YU H. & PIGEON W. (2022). Scan : Suicide attempt and ideation events dataset. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1029–1040, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.75.

REITER E. (2025). We should evaluate real-world impact. *Computational Linguistics*, p. 1–13. DOI : 10.1162/coli.a.18.

RESNIK P. (2025). Large language models are biased because they are large language models. *Computational Linguistics*, **51**(3), 885–906. DOI : 10.1162/coli_a_00558.

ROGERS A. & LUCCIONI A. S. (2024). Position : Key claims in llm research have a long tail of footnotes. In *Proceedings of the Forty-first International Conference on Machine Learning*, Vienne, Autriche : JMLR.org.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2020). Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. arXiv :1910.01108 [cs], DOI : 10.48550/arXiv.1910.01108.

SHIMAMOTO M., ISHIZUKA K., OHTANI K., INADA T., YAMAMOTO M., TACHIBANA M., KIMURA H., SAKAI Y., KOBAYASHI K., OZAKI N. & IKEDA M. (2024). Machine learning algorithm-based estimation model for the severity of depression assessed using montgomery-asberg depression rating scale. *Neuropsychopharmacology Reports*, **44**(1), 115–120. DOI : 10.1002/npr2.12404.

SMITH J. M. & SMITH D. C. (1977). Database abstractions : Aggregation and generalization. *ACM Transactions on Database Systems (TODS)*, **2**(2), 105–133. DOI : 10.1145/320544.320546.

TØLBØLL K. B. (2019). Linguistic features in depression : a meta-analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, **4**(2), 39.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models. arXiv :2302.13971 [cs], DOI : 10.48550/arXiv.2302.13971.

TRICCO A. C., LILLIE E., ZARIN W., O'BRIEN K. K., COLQUHOUN H., LEVAC D., MOHER D., PETERS M. D., HORSLEY T., WEEKS L., HEMPEL S., AKL E. A., CHANG C., MCGOWAN

J., STEWART L., HARTLING L., ALDCROFT A., WILSON M. G., GARRITTY C., LEWIN S., GODFREY C. M., MACDONALD M. T., LANGLOIS E. V., SOARES-WEISER K., MORIARTY J., CLIFFORD T., TUNÇALP & STRAUS S. E. (2018). Prisma extension for scoping reviews (prisma-scr) : Checklist and explanation. *Annals of Internal Medicine*, **169**(7), 467–473. DOI : 10.7326/M18-0850.

VAN BUCHEM M. M., DE HOND A. A. H., FANCONI C., SHAH V., SCHUESSLER M., KANT I. M. J., STEYERBERG E. W. & HERNANDEZ-BOUSSARD T. (2024). Applying natural language processing to patient messages to identify depression concerns in cancer patients. *Journal of the American Medical Informatics Association : JAMIA*, **31**(10), 2255–2262. DOI : 10.1093/jamia/ocae188.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, New York, États-Unis : Curran Associates Inc.

WANG X., LIU K. & WANG C. (2023). Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, p. 532–536, Dali, Chine. DOI : 10.1109/CCIS59572.2023.10263217.

YANG Y., LIN M., ZHAO H., PENG Y., HUANG F. & LU Z. (2024a). A survey of recent methods for addressing ai fairness and bias in biomedicine. *Journal of Biomedical Informatics*, **154**, 104646. DOI : 10.1016/j.jbi.2024.104646.

YANG Y., LIU X., JIN Q., HUANG F. & LU Z. (2024b). Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, **4**(1), 176. DOI : 10.1038/s43856-024-00601-z.

ZACK T., LEHMAN E., SUZGUN M., RODRIGUEZ J. A., CELI L. A., GICHOYA J., JURAFSKY D., SZOLOVITS P., BATES D. W., ABDULNOUR R.-E. E., BUTTE A. J. & ALSENTZER E. (2024). Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care : a model evaluation study. *The Lancet. Digital Health*, **6**(1), e12–e22. DOI : 10.1016/S2589-7500(23)00225-X.

ZHANG T., SCHOENE A. M., JI S. & ANANIADOU S. (2022). Natural language processing applied to mental illness detection : a narrative review. *NPJ digital medicine*, **5**(1), 46.

ZUROMSKI K. L., LOW D. M., JONES N. C., KUZMA R., KESSLER D., ZHOU L., KASTMAN E. K., EPSTEIN J., MADDEN C., GHOSH S. S., GOWEL D. & NOCK M. K. (2024). Detecting suicide risk among u.s. servicemembers and veterans : a deep learning approach using social media data. *Psychological Medicine*, p. 1–10. DOI : 10.1017/S0033291724001557.