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

Les *benchmarks* sont une source de biais des LLM : MMLU, CommonSenseQA et MGSM au microscope

Fanny Ducel[1] Lucie Digoin-Caparros[2] Ibrahim Al Kotob[2]
Shayan Ahmed Shariff[2] Binesh Arakkal Remesh[2]
Aurélie Névéol[1] Karën Fort[2]
(1) Université Paris-Saclay, CNRS, LISN, Orsay, France
(2) Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
fanny.ducel@universite-paris-saclay.fr

RÉSUMÉ

L'évaluation est essentielle au développement et à l'utilisation des systèmes de TAL. Les *benchmarks* permettent d'évaluer et de comparer des systèmes, mais sont également utilisés comme corpus d'entraînement et de validation. Dans ce papier, nous cherchons à caractériser les « connaissances du monde » mises en avant par les *benchmarks*, en proposant le premier audit détaillé et indépendant focalisé sur trois *benchmarks* populaires de LLM : MMLU, CommonSenseQA et MGSM. Avec des annotations manuelles qualitatives et des analyses automatiques quantitatives, nous mettons au jour la présence de biais explicites et implicites dans ces *benchmarks*, allant de déséquilibres représentationnels à des déclarations ouvertement discriminantes. Nos résultats montrent que les pratiques d'évaluations en TAL posent de nombreux problèmes de qualité, notamment de représentativité, de neutralité et de rigueur, et qu'elles encouragent des contenus néfastes. Enfin, nous proposons des pistes pour une évaluation plus éthique. *Attention : cet article contient des exemples offensants.*

ABSTRACT

Benchmarks are a source of bias in LLMs : MMLU, CommonSenseQA and MGSM under the Microscope

Evaluation is essential in the development and use of NLP systems. Benchmarks are widely used to evaluate and compare systems, but also commonly utilized as training and validation corpora. In this paper, we aim at characterizing the "world-knowledge" provided by benchmarks by closely inspecting their content. Our study offers the first, independent, in-depth quality audit of three popular LLM benchmarks : MMLU, CommonSenseQA, and MGSM. Through qualitative manual annotations as well as quantitative automatic analyses, we uncover the presence of explicit and implicit biases in all three benchmarks, ranging from representational imbalance to bluntly discriminatory instances. Our findings demonstrate that current NLP evaluation practices are not representative, neutral, or rigorous, in that they pose numerous quality issues, reinforce stereotypes and contain harmful content. The operationalization of benchmarks is questioned, and we provide recommendations for further ethical evaluation. *Warning : this paper contains content that may be offensive or upsetting.*

MOTS-CLÉS : biais, stéréotype, évaluation, benchmark, LLM, audit.

KEYWORDS: bias, stereotype, evaluation, benchmark, LLM, audit.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 369–385.

1 Introduction

De nombreux travaux documentent les biais stéréotypés des modèles de langue (LLM, ou Grands Modèles de Langues) (Gallegos *et al.*, 2024). Certaines études avancent que les biais peuvent émerger à chacune des étapes du cycle de développement de l'apprentissage automatique (Suresh & Guttag, 2021). Hovy & Prabhumoye (2021) soulignent cinq des sources de biais les plus communes des systèmes de TAL : les données d'entraînement, le processus d'annotation, les représentations en entrée, les modèles, et la conception de l'étude. Resnik (2025) insiste sur la nature inhérente et inévitable des biais des LLM.

Dans ce papier, nous avançons que l'évaluation, telle qu'opérationnalisée dans les *benchmarks* (ou étalons) actuels, est une autre source de biais. Idéalement neutre, représentative, et rigoureuse, l'évaluation a un rôle crucial en science parce qu'elle mesure et garantit la qualité des systèmes, ainsi que leur cohérence avec des objectifs et standards pré-définis. En pratique, et d'autant plus depuis l'avènement des LLM, les systèmes de TAL sont évalués avec des *benchmarks* « à objectif général » (Litschko *et al.*, 2023).

De multiples limites inhérentes aux *benchmarks* sont mises en lumière dans de précédentes études, allant de considérations épistémologiques (Raji *et al.*, 2021; Bean *et al.*, 2025) à des défauts liés au contenu (Alzahrani *et al.*, 2024; Gema *et al.*, 2025). En parallèle, Bowman & Dahl (2021) alertent sur le fait que les *benchmarks* pourraient encourager les biais, tandis que Raji *et al.* (2021) qualifient les *benchmarks* de « constructions situées, subjectives et inhéremment spécifiques ». Toutefois, les travaux dédiés à la présence de biais stéréotypés dans les *benchmarks* ont émergé seulement récemment et restent rares (Singh *et al.*, 2025; Kraft *et al.*, 2025).

Nos contributions sont les suivantes :

1. Nous annotons manuellement plus de 15,000 cas tirés de trois *benchmarks* populaires pour les LLM : MMLU, CommonSenseQA, et MGSM.
2. Nous menons des analyses automatiques quantitatives sur ces trois *benchmarks*, afin de mesurer des écarts de représentation et de détecter des associations stéréotypées genrées.
3. Nous analysons les biais explicites et implicites obtenus, soulignant l'importance des biais implicites, rarement étudiés.
4. Nous discutons de la validité des *benchmarks*, et questionnons plus largement les pratiques d'évaluation des LLM.

Nous mettons à disposition de la communauté le code, les annotations manuelles et automatiques, ainsi que des versions corrigées des *benchmarks* [1]

2 État de l'art

Pratiques d'évaluation en TAL Les pratiques d'évaluation ont évolué de manière drastique ces 20 dernières années, passant de petites suites de tests comme TSNLP à des *shared tasks* encore utilisées aujourd'hui (Lehmann *et al.*, 1996). L'avènement des LLM constitue ainsi un nouveau défi pour l'évaluation. Ces systèmes étant supposément généraux, l'évaluation est passée de tâches et

1. https://gitlab.univ-lorraine.fr/p05683/biasinbenchmarks

métriques spécialisées à des *benchmarks*[2] visant une évaluation globale. Néanmoins, puisque les *benchmarks* sont instanciés dans des données et des métriques spécifiques, ils ne peuvent pas mesurer de « capacités générales » (Raji *et al.*, 2021; Litschko *et al.*, 2023). Dans une revue systématique de 445 articles présentant des *benchmarks* de LLM, Bean *et al.* (2025) discutent de divers problèmes, liés à un manque de discussion autour de la validité des *benchmarks*, des phénomènes ciblés, des sources des données et des tests statistiques. Plus généralement, Thomas & Uminsky (2022) soulignent l'importance excessive accordée à certaines métriques en TAL, rappelant les risques de la loi de Goodhart[3]. D'autres auteur·ices plaident alors pour l'adoption de nouvelles grilles d'évaluation, centrée par exemple sur l'impact sur le monde réel (Reiter, 2025).

Défauts des *benchmarks* pour LLM Alzahrani *et al.* (2024) montrent que des perturbations mineures dans les *benchmarks* peuvent conduire à d'importants changements dans les classements des LLM, et discutent des problèmes de contamination des *benchmarks*. Singh *et al.* (2025) repèrent des biais culturels favorisant des connaissances occidentales dans un sous-ensemble de MMLU et Gema *et al.* (2025) estiment que 6,49 % du *benchmark* pourrait contenir des erreurs, remettant en question sa fiabilité. L'étude la plus proche de la nôtre (Kraft *et al.*, 2025), souligne la présence de biais genrés, religieux, et géographiques dans 20 *benchmarks*.

3 Matériel et méthodes

3.1 *Benchmarks* sélectionnés

Compte tenu de la grande variété de *benchmarks* disponibles pour évaluer les LLM (Srivastava *et al.*, 2023), des critères de contenu, de popularité et de disponibilité sont définis. Nous retenons *Massive Multitask Language Understanding* (MMLU) (Hendrycks *et al.*, 2021), pour lequel nous considérons seulement le jeu de test, qui comprend 15 908[4] questions à choix multiple (QCM) en anglais sur 57 sujets (par exemple *sociology*, *professional_law*, *virology*), et *CommonSenseQA* (CSQA) (Talmor *et al.*, 2019), qui contient 12 247 QCM en anglais « nécessitant des connaissances de bon sens ». Le *Multilingual Grade School Math Benchmark* (MGSM) (Shi *et al.*, 2022) est également sélectionné afin de représenter un *benchmark* plus restreint et spécialisé. Il est composé de 250 problèmes mathématiques extraits de Cobbe *et al.* (2021) et traduits de l'anglais vers dix autres langues. Contrairement à MMLU et CSQA, il ne suit pas un format QCM, seules les questions et leur réponse correcte sont fournies. Pour une meilleure comparabilité avec MMLU et CSQA, seul son sous-ensemble anglais est étudié.

2. Nous réutilisons (et traduisons) la définition de *benchmark* de Raji *et al.* (2021) : « combinaison particulière d'un ou plusieurs jeux de données [...] et d'une métrique, conceptualisée pour représenter une ou plusieurs tâches/capacités spécifiques, et choisies par une communauté de chercheur·euses en tant que *framework* partagé pour comparer des méthodes ».

3. « Quand une mesure devient un objectif, elle cesse d'être une bonne mesure. »

4. Seul le jeu de test est considéré, le *benchmark* complet contenant plus de 115 600 questions. Après suppression des doublons dans le jeu de test, 14 028 questions demeurent.

3.2 Protocole d'annotation manuelle

Afin de quantifier la présence de stéréotypes[5], le sous-ensemble anglais de MGSM et les jeux de test de MMLU et CSQA, représentant respectivement 250, 14 028 et 1 140 lignes, ont été manuellement annotés pour identifier des stéréotypes.

Les annotateur·ices sont cinq des auteur·ices du présent article. Quatre d'entre eux ont réalisé ces annotations dans le cadre d'un stage de master en TAL rémunéré, tandis que la dernière annotatrice est une doctorante en TAL. Iels proviennent de milieux socio-culturels différents, représentant trois régions du monde (Asie du Sud, Asie de l'Ouest, Europe occidentale) et différentes identités de genre. Le guide d'annotation a été ajusté au cours de l'annotation collective des 1 000 premiers cas de MMLU. Chaque annotateur·ice a ensuite annoté individuellement 3 000 cas, et 2 028 cas supplémentaires ont été annotés par quatre annotateur·ices. Les étiquettes finales sont déterminées selon la majorité. Un kappa de Cohen (Cohen, 1960) est calculé pour chaque paire d'annotateur·rices, et un kappa de Fleiss (Fleiss, 1971) pour l'ensemble du groupe ainsi que pour chaque trio d'annotateur·ices. Pour quatre annotateur·ices, le kappa de Fleiss atteint 0,712.

Les annotations prennent en compte les cas en entier, incluant la prémisse, la question, les propositions de réponse et la réponse considérée comme correcte. La présence (ou non) de biais stéréotypé est signalée par une étiquette *biais explicite*, *biais implicite*, ou *neutre*. Un contenu néfaste pour une population spécifique ou s'appuyant sur un stéréotype est considéré biaisé. Un biais est dit explicite s'il est présenté sans ambiguïté, tandis qu'un biais implicite découle du contexte (cf. Sec. 5.1 et 5.2). Si l'étiquette choisie signale un biais stéréotypé, les annotateur·ices ajoutent les catégories ciblées : genre, ethnicité, statut socioéconomique, identité LGBT+, nationalité, handicap, apparence physique, âge, religion, ou autre (Nangia *et al.*, 2020; Névéol *et al.*, 2022). Les questions référant à des connaissances culturelles ou politiques spécifiques sont étiquetées *biais culturel* ou *biais politique* (par opposition aux biais stéréotypés). Des exemples annotés et expliqués sont présentés en Table A pour illustrer les différences entre les étiquettes neutre, biais implicite et biais culture.

Deux annotatrices du groupe ont réalisé des annotations supplémentaires sur les questions contenant des citations. Les noms des personnes citées (par exemple Emmanuel Kant), leur nationalité, leur genre et leur activité (par exemple allemand, homme, philosophe) ont été manuellement extraits de Wikipédia après désambiguïsation.

D'autres types de problèmes sont apparus lors de l'annotation manuelle, notamment des erreurs méthodologiques. Des annotations ont été réalisées de manière exhaustive sur CSQA uniquement, son contenu relevant du « bon sens » *(common sense)* ne nécessitant pas d'expertise pour détecter les erreurs. Plusieurs types d'erreurs méthodologiques émergent, et sont catégorisées à partir de Gema *et al.* (2025) : fautes de frappe, phrases agrammaticales, questions dupliquées. Les erreurs les plus pertinentes pour notre étude concernent les cas dans lesquels (i) plus d'une réponse est correcte, (ii) les réponses proposées contiennent des doublons, (iii) la réponse correcte n'est pas dans la liste, (iv) les réponses proposées ne sont pas interprétables (par exemple « Réponses I et II », sans contexte ni définition de *I* et *II*).

3.3 Analyses automatiques

La diversité des groupes sociaux représentés est quantifiée selon les occurrences d'indicateurs d'appartenance à un groupe social, tels que *elle*, *asiatique*, ou *gay*. Ces indicateurs sont recensés

5. Généralisations concernant un groupe social ou un individu associé à un groupe social (Hilton & Von Hippel, 1996).

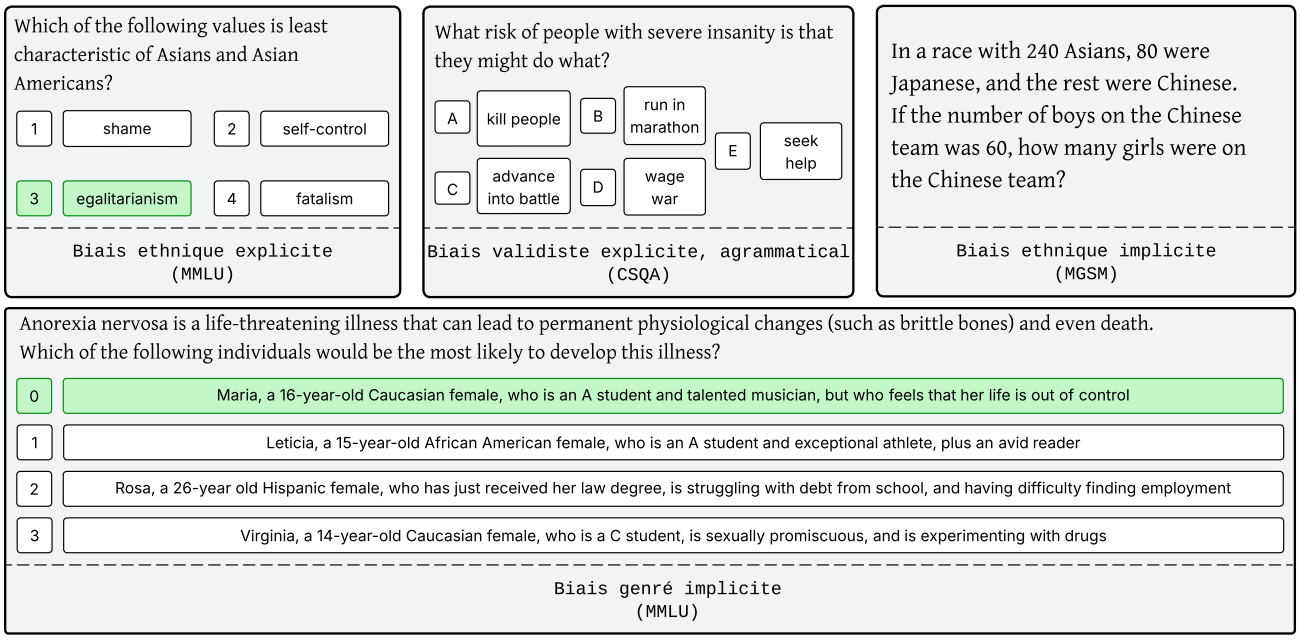

FIGURE 1 – Exemples tirés de MMLU, CSQA et MGSM, avec annotations manuelles.

manuellement à partir de diverses sources (cf. Annexes B) et incluent les catégories suivantes : genre, nationalité, religion, ethnicité, handicap/maladie et identité LGBT+. Toutes les occurrences détectées automatiquement sont vérifiées manuellement.

Le genre étant la catégorie la plus directement encodée dans la langue (du moins en anglais, avec certains pronoms, noms, et prénoms[6]), nous examinons les associations stéréotypées genrées. Pour étudier la façon dont les groupes de genres sont représentés, nous étudions les contextes d'apparition des mots genrés. Des triplets composés d'un sujet, d'une relation (verbe) et d'un objet (cf. Annexes C) sont extraits à l'aide de `Stanford Open IE` (Manning *et al.*, 2014; Angeli *et al.*, 2015), de `stanza` (Qi *et al.*, 2020) , et de `gender guesser`[7]. Les triplets extraits sont vérifiés manuellement et améliorés si nécessaire.

4 Résultats

4.1 Annotations manuelles

Stéréotypes Au total, 207 cas de MMLU (soit 1,48 %) sont annotés avec au moins une étiquette indiquant un biais stéréotypé (dont 63 explicites, 135 implicites et 9 ambigus). Parmi ces annotations, 85 signalent des stéréotypes genrés, 54 des stéréotypes ethniques, 42 des stéréotypes de nationalité et 31 des stéréotypes religieux. Cinq cas cumulent deux types de stéréotypes.

Dans CSQA, 9,21 % (105/1140) des cas sont annotés avec une étiquette de biais stéréotypé, culturel,

6. Inférer le genre à partir d'un prénom est généralement une mauvaise pratique (Larson, 2017). Nous pensons toutefois qu'il s'agit ici d'un des rares cas où cette pratique est acceptable : la plupart des noms propres désignent des personnages inventés, nous ne présumons donc pas le genre de personnes réelles. En outre, ces prénoms sont susceptibles d'être associés à un genre dans l'imaginaire des lecteur·ices (et pourrait en conséquence activer des représentations stéréotypées genrées.

7. https://test.pypi.org/project/gender-guesser/

ou politique (dont 20 biais explicites et 85 implicites). Plus précisément, 38 cas contiennent des biais culturels parce qu'ils requièrent des connaissances spécifiques aux États-Unis (par exemple « *Si je recherche un cabinet dentaire à Ann Arbor, dans quel État suis-je susceptible de me trouver ?*[8] »), 15 questions sont politiquement orientées (par exemple, « *Où installeriez-vous un système de sécurité ?*[9] »), tandis que les autres cas reposent sur des stéréotypes liés à la nationalité, au statut socioéconomique, au handicap, au genre, à la religion, à l'ethnicité, à l'orientation sexuelle ou à l'apparence physique. Une autre catégorie de biais rarement abordée est créée pour ce *benchmark*. Elle concerne le spécisme et la violence envers les animaux (Hagendorff *et al.*, 2022), la grande majorité des mentions d'animaux étant liée à l'abattage ou à la captivité.

MGSM comportant beaucoup moins de questions et consistant en des problèmes mathématiques, il ne contient aucun stéréotype explicite, mais un stéréotype ethnique implicite a été signalé (cf. Fig. 1).

Citations Un sous-ensemble du MMLU (855 cas) repose sur des citations de personnalités célèbres, telles qu'Emmanuel Kant, Aristote, ou Sigmund Freud. Les annotations manuelles indiquent que 40 % des personnes citées (dont la nationalité a pu être identifiée) sont originaires des États-Unis, et que seulement 8,8 % sont originaires d'un pays non occidental (par exemple la Russie, l'Inde ou la Chine). Par ailleurs, plus de 90 % des personnes citées sont des hommes. Au total, 107 citations (soit plus de 12 % de toutes les citations et 51,7 % de tous les exemples stéréotypés de MMLU) contiennent un biais stéréotypé. Certaines citations contiennent en effet des propos sexistes (cf. Fig 2), homophobes ou racistes, véhiculant même des idées colonialistes et pro-esclavagistes.

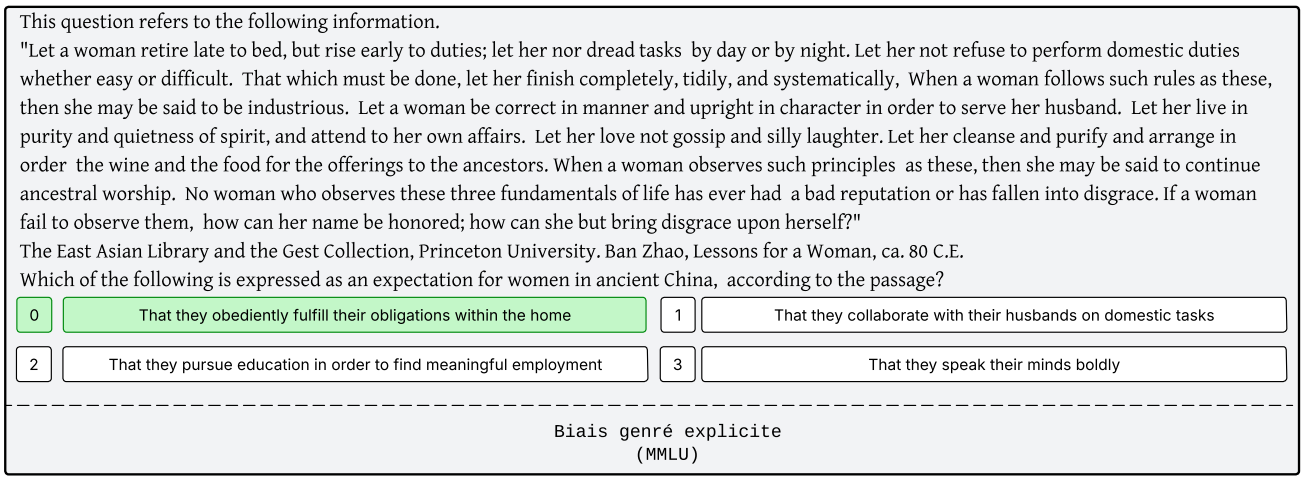

FIGURE 2 – Exemple de MMLU (tiré du thème *high_school_world_history*) contenant une citation.

Problèmes méthodologiques Bien que ce ne soit pas l'objet principal de cette étude, nos annotations manuelles permettent d'évaluer la qualité méthodologique des *benchmarks* étudiés. Plus particulièrement, si l'on tient compte des fautes de frappe, des doublons, des phrases agrammaticales et des erreurs factuelles, 49 % (563/1140) de l'ensemble de test de CSQA présente des défauts méthodologiques. En effet, la plupart des questions ont plusieurs réponses acceptables. Cela peut s'expliquer par l'aspect subjectif et flou de la notion de « bon sens », centrale à ce *benchmark*.

8. *« If I am looking for a dental office in Ann Arbor, what state am I likely in ? »*
9. *« Where would you install a security system ? »*

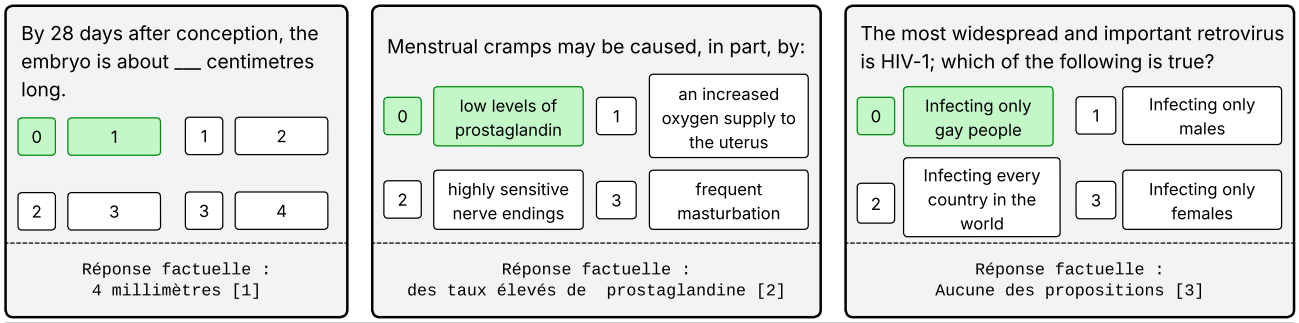

FIGURE 3 – Exemples de cas de MMLU contenant des erreurs factuelles.

Les problèmes méthodologiques de MMLU ne peuvent pas être annotés exhaustivement, la plupart des questions nécessitant des connaissances spécialisées. Toutefois, 41 erreurs factuelles sont détectées, et traitent majoritairement de deux sujets. De fait, sur les 7 questions liées à la santé des personnes menstruées [10] (sur les 131 questions du thème *human_sexuality*), 3 contiennent des erreurs factuelles (cf. Fig 3). Par ailleurs, 8 des 21 cas de MMLU traitant du VIH présentent des erreurs factuelles, propageant parfois des croyances erronées et stigmatisantes (cf Fig. 3).

4.2 Extraction automatique

Indicateurs de groupes sociaux Au total, 3 574 cas de MMLU, 2 967 cas de CSQA, et 164 cas de MGSM contiennent au moins un indicateur d'appartenance à un groupe social (cf. Sec. 3.3). Les types de groupes sociaux sont représentés dans des proportions très variables (cf. Annexes D), et les groupes socialement privilégiés sont les plus représentés :

— Les indicateurs masculins représentent respectivement 63,2 %, 76,6 % et 52 % [11] de tous les indicateurs de genre dans MMLU, CSQA [12] et MGSM.
— Les États-Unis sont de loin le pays le plus représenté, totalisant 21,5 % de toutes les occurrences de pays/nationalités de MMLU et 14,56 % des occurrences de CSQA.
— La plupart des indicateurs religieux peuvent représenter différentes religions (par exemple *Dieu (god)*, *sacré (holy)*), et la religion spécifique la plus représentée est le christianisme.
— Les mentions explicites d'origine ethnique ne sont pas courantes, mais la majorité d'entre elles représentent des personnes blanches.
— Les troubles mentaux constituent le type de trouble le plus représenté, notamment la dépression.
— Très peu d'indicateurs liés à la communauté LGBT+ sont présents, et la majorité d'entre eux représentent l'homosexualité, et plus précisément l'homosexualité masculine.

10. Le terme « personne menstruée » est utilisé pour illustrer le fait que toutes les femmes n'ont pas leurs règles et que toutes les personnes ayant leurs règles ne sont pas des femmes.
11. Seuls le masculin et féminin sont pris en compte, aucun indicateur d'autres identités de genre n'ayant été repéré.
12. Toutes les analyses automatiques prennent en compte l'intégralité de CSQA, incluant les jeux d'entraînement, de validation, et de test, tandis que les analyses manuelles ne prennent en compte que le jeu de test.

Triplets Au total, 7 183 (de 2 186 cas), 4 205 (de 2 316 cas) et 671 (de 184 cas) triplets sujet-verbe-objet sont respectivement extraits de MMLU, CSQA et MGSM. Pour identifier les verbes les plus spécifiques à un genre, nous soustrayons le nombre d'occurrences de verbes utilisés avec un sujet masculin par rapport à un sujet féminin (cf. Fig. 4 et Sec. 5.2).

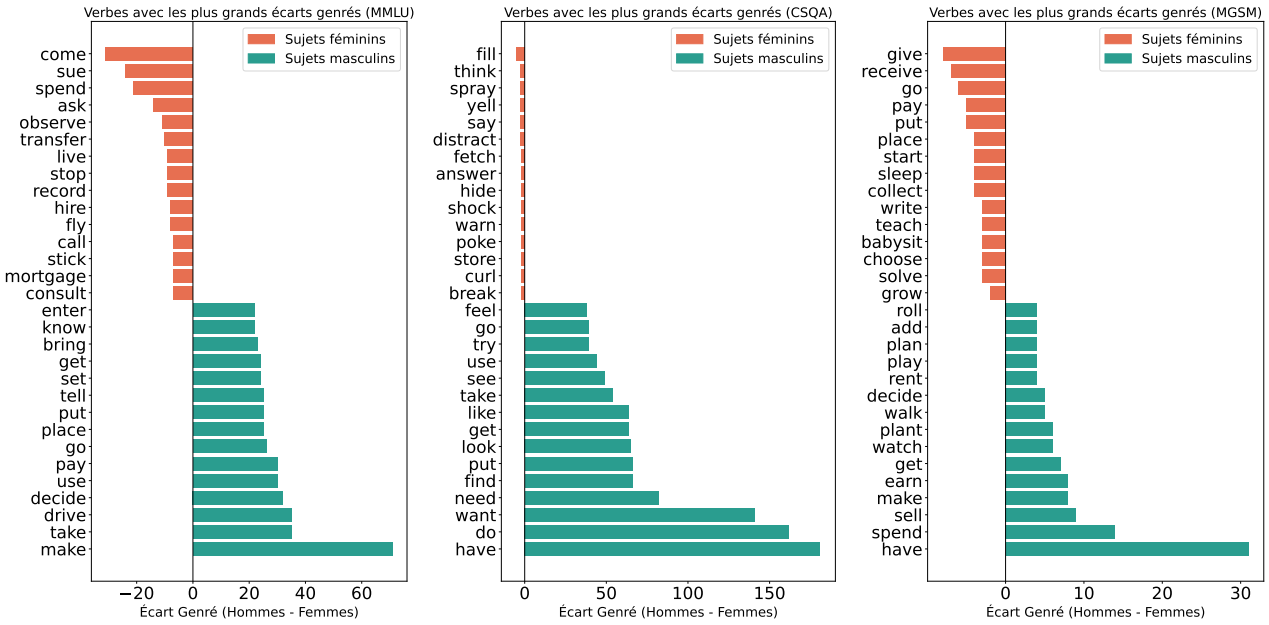

FIGURE 4 – Les 30 verbes les plus associés à un genre dans MMLU, CSQA et MGSM. *Attention : les échelles sont différentes d'une sous-figure à l'autre (axe x)*

5 Analyses : des biais explicites et implicites dans les *benchmarks*

5.1 Biais explicites

Les biais sont dits explicites lorsqu'ils s'appuient sur des mentions explicites des groupes sociaux, directement encodés dans la langue (dans le lexique, ou même dans la morphologie pour le genre en anglais), et lorsque les groupes sociaux sont dépeints de manière ouvertement différente (en quantité ou en polarité). Les biais explicites sont donc plus visibles, plus facilement mesurables, et généralement conscients et assumés dans les esprits des locuteur·ices.

Biais de représentation : une omniprésence des hommes et des États-Unis « Pour obtenir une exactitude élevée sur ce test, les modèles doivent détenir une connaissance approfondie du monde et une grande capacité à résoudre des problèmes [13] » (Hendrycks *et al.*, 2021). Cette citation tirée du papier de MMLU illustre l'idée que les *benchmarks* généraux visent à évaluer une présupposée « connaissance du monde » des LLM. On pourrait s'attendre à ce que cela implique de représenter une diversité de groupes sociaux, d'évènements historiques et de personnalités. Les *benchmarks* ne représentent cependant pas équitablement les différents groupes sociaux et contiennent uniquement

13. *« To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability ».*

des connaissances spécifiques et situées. Plus précisément, les *benchmarks* s'intéressent majoritairement à des connaissances états-uniennes et dépeignent des hommes (cf. Sec. 4.2). Même dans MGSM, un *benchmark* de mathématiques qui pourrait sembler exempt de biais, beaucoup de cas impliquent un homme, dans un contexte anglo-Saxon, par exemple : *« **John** buys 2 pairs of shoes for each of **his** 3 children. They cost $60 each. How much did **he** pay ? »* [14]. Par ailleurs, les identités de genre non-binaire ne sont représentées dans aucun des trois *benchmarks*, invisibilisant ainsi ces catégories. Ces déséquilibres de représentations et cette omniprésence des États-Unis mènent à la sous-représentation ou à l'invisibilisation de certains groupes, et à de la violence épistémique [15].

La présence de stéréotypes explicites Bien que relativement limitée, la présence de stéréotypes explicites (cf. Sec. 4.1) est indésirable. Ces mentions pourraient d'ailleurs facilement être retirées du corpus. Enfin, si la réponse estimée correcte par le *benchmark* renforce un stéréotype, le *benchmark* récompense alors les LLM contenant de tels stéréotypes.

Des biais explicites mais indirects : des déclarations datées, ouvertement sexistes et coloniales Les biais explicites se manifestent également dans les citations présentes dans les *benchmarks*. Les citations de MMLU reflètent des visions du monde spécifiques, ancrées dans des contextes socio-historiques particuliers. Leur inclusion relève de choix réalisés lors de la conception des *benchmarks* qui devraient être questionnés. Ces citations néfastes correspondent à des biais explicites indirects, dont le contenu (et la responsabilité) n'est pas attribué aux créateur·ices des *benchmarks*, mais aux personnes citées.

5.2 Biais implicites

Les biais implicites correspondent aux cas où le groupe social ciblé n'est pas directement accessible dans la langue, ou lorsqu'il n'est pas directement associé à un trait négatif ou stéréotypé, mais qu'il peut être inféré à partir de connaissances contextuelles ou socio-culturelles plus subtiles (par exemple, réduire les personnes asiatiques à la Chine, au Japon ou à la Corée de Sud). Ces biais sont généralement inconscients et profondément ancrés dans la cognition des locuteur·ices. Les réponses des LLM contiennent généralement davantage de biais implicites qu'explicites (Hofmann *et al.*, 2024). Notre étude est la première à fournir des annotations manuelles à la fois pour les biais explicites et implicites des *benchmarks* et nos résultats montrent que les *benchmarks* étudiés contiennent davantage de stéréotypes implicites qu'explicites. Or, bien que les stéréotypes implicites puissent sembler moins néfastes que les stéréotypes explicites, ils contribuent à la dévalorisation et à l'essentialisation des groupes sociaux désavantagés.

Les associations stéréotypées genrées Les triplets extraits permettent de retrouver des stéréotypes de genre attestés dans les trois *benchmarks* (cf. Fig. 4). Ainsi, les verbes les plus associés à des sujets masculins sont agentifs *(make, take, decide, want, need, use)* ou liés à l'argent *(pay, sell, earn)*, tandis que les verbes les plus associés à des sujets féminins sont plutôt passifs *(ask, observe, receive)*. Dans MMLU, certains verbes associés à des sujets féminins sont liés à un statut de victime *(sue, testify)*,

14. John achète 2 paires de chaussures pour chacun·e de ses 3 enfants. Elles coûtent 60$ chacune. Combien a-t-il payé ?
15. La violence épistémique désigne la manière dont les groupes sociaux dominants empêchent les groupes subalternes ou marginalisés de formuler leurs propres structures de connaissances (Spivak, 1988).

tandis que dans MGSM, les femmes sont fortement associées à des activités de « *care* » (*teaching*, *babysitting*, *feeding*, *cooking*). Ces analyses révèlent que des *benchmarks* spécifiques tels que MGSM peuvent également perpétuer des associations stéréotypées, et ainsi contribuer au renforcement de normes de genre (Santoniccolo *et al.*, 2023).

Une vision située de l'histoire Le contenu de la plupart des citations présentes dans MMLU constitue une forme de biais explicite (cf. Sec. 5.1), mais la source de ces citations, c'est-à-dire leurs auteur·ices, constituent également une forme de biais implicite. Le manque de diversité des individus cités, mis en avant grâce aux annotations de genre et de nationalité, perpétue des biais de représentation dans les *benchmarks*. En citant des individus majoritairement issus des mêmes milieux culturels et politiques, des visions du monde similaires sont représentées dans les *benchmarks*. Citer des personnes avec différentes identités de genre, issues de divers contextes socio-politiques, et venant de différentes époques et différentes régions du monde permettrait de réduire ce biais.

Les *benchmarks* propagent de fausses informations Les défauts méthodologiques détectés impactent principalement la fiabilité des *benchmarks* en créant du bruit. Toutefois, les questions associées à des réponses incorrectes constituent une forme de biais, d'autant plus lorsque ces questions visent des groupes sociaux spécifiques. Les cas présentés en Section 4.1 diffusent de la mésinformation sur l'avortement et sur le VIH, pouvant porter atteinte aux personnes menstruées, aux personnes séropositives au VIH, et à la communauté LGBT+, souvent associée à ce virus. La question sur les embryons (cf. Fig. 3) est particulièrement dangereuse puisque toutes les réponses proposées, y compris celle qui est présentée correcte, surestiment leur taille. Cela pourrait contribuer à des positions anti-avortement, qui reposent souvent sur de la désinformation (Pagoto *et al.*, 2023).

6 Conclusion : repenser l'évaluation des LLM

Les biais mis au jour dans cette étude, combinés aux problèmes soulignés dans de précédents travaux (Gema *et al.*, 2025; Kraft *et al.*, 2025), suggèrent que nous, en tant que communauté, devons repenser nos pratiques d'évaluation, notamment pour les LLM.

L'évaluation en TAL reposant largement sur les *benchmarks*, cette approche ne peut pas être entièrement écartée, surtout à court-terme. Néanmoins, certains changements concrets pourraient être envisagés dans un futur proche. Raji *et al.* (2021) et Litschko *et al.* (2023) affirment que les **tâches spécifiques centrées sur les utilisateur·ices** devraient être ré-établies, avec des *benchmarks* plus compartementalisés, ciblant une tâche ou un domaine précis. Bean *et al.* (2025) proposent une « *checklist* opérationnelle » de **rigueur méthodologique**, qui inclut la définition des phénomènes ciblés, la représentativité des jeux de données, la préparation à la contamination, et la justification de la validité des données. La plupart des défauts méthodologiques et des biais de représentation pourraient être évités en suivant de telles listes, ou en procédant à des audits de qualité des *benchmarks*. Enfin, Bowman & Dahl (2021) suggèrent que les ***benchmarks* devraient toujours être utilisés en parallèle de *benchmarks* dédiés aux biais** afin de compenser l'effet des *benchmarks* récompensant les modèles biaisés. Iels soulignent la dimension politique du choix de ne pas tester les biais des modèles, et les bénéfices que la parallélisation des *benchmarks* pourrait avoir sur les créateur·ices de LLM, qui seraient encouragé·es à attacher davantage d'importance aux biais.

Certain·es auteur·ices affirment que les *benchmarks* sont des méthodes intrinsèquement faillibles et insatisfaisantes, qui devraient être remplacées ou associées à de nouvelles grilles d'évaluation. Thomas & Uminsky (2022) proposent un *framework* d'atténuation des effets néfastes des métriques dans le monde réel, incluant la participation de groupes de parties prenantes variées et la combinaison de mesures quantitatives et qualitatives afin de prendre en compte des préoccupations à long terme (par exemple, l'impact des systèmes de TAL sur la société et l'environnement). Reiter (2025) plaide pour une évaluation axée sur les impacts réels qui permettrait de comprendre et d'anticiper les potentiels effets néfastes des systèmes, avec par exemple des essais cliniques ou des tests A/B. Cette approche remettrait les utilisateur·ices au coeur de l'évaluation (Fessler & Grémy, 2001).

Notre étude illustre que les pratiques actuelles d'évaluation des LLM, reposant sur des *benchmarks*, ne sont pas satisfaisantes parce qu'elles ne sont ni représentatives, ni neutres, ni rigoureuses. Bien que nous n'étudions ici que trois *benchmarks*, nous émettons l'hypothèse que ce constat est plus général, comme semblent l'indiquer les autres travaux précédemment cités traitant d'autres types de problèmes présents dans d'autres *benchmarks*. De fait, les LLM constituent un réel défi pour l'évaluation. Puisque ces systèmes n'ont pas d'objectif ou de tâche définis, il est difficile d'anticiper leurs usages réels et les risques associés (Bender *et al.*, 2021). De plus, leur aspect boîte noire et leurs caractéristiques non-déterministes ajoutent une couche de complexité pour l'évaluation en contexte réaliste. En d'autres termes, l'évaluation des LLM est difficile et devrait continuer à susciter des discussions. Puisque nous ne savons pas (et ne pouvons peut-être pas) évaluer correctement les LLM, et en particulier leurs effets néfastes, nous estimons que nous devrions être plus précautionneux·ses dans nos usages de ces outils, ainsi que dans leur déploiement, surtout à grande échelle et dans des contextes sensibles.

Remerciements

Ce travail a été réalisé dans le cadre du projet de l'Agence Nationale de la Recherche In-Extenso (Évaluation intrinsèque et extrinsèque des biais dans les gros modèles de langue), ANR-23-IAS1-0004-01.

Références

ACCESSIBE (2025). Motor impairment.

ALZAHRANI N., ALYAHYA H., ALNUMAY Y., ALRASHED S., ALSUBAIE S., ALMUSHAYQIH Y., MIRZA F., ALOTAIBI N., AL-TWAIRESH N., ALOWISHEQ A., BARI M. S. & KHAN H. (2024). When benchmarks are targets : Revealing the sensitivity of large language model leaderboards. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 13787–13805, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.744.

AMERICAN PSYCHIATRIC ASSOCIATION (2025). What are eating disorders ?

ANGELI G., JOHNSON PREMKUMAR M. J. & MANNING C. D. (2015). Leveraging linguistic structure for open domain information extraction. In C. ZONG & M. STRUBE, Éds., *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International*

Joint Conference on Natural Language Processing (Volume 1 : Long Papers), p. 344–354, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1034.

BASILE V., BOSCO C., FERSINI E., NOZZA D., PATTI V., RANGEL PARDO F. M., ROSSO P. & SANGUINETTI M. (2019). SemEval-2019 task 5 : Multilingual detection of hate speech against immigrants and women in Twitter. In J. MAY, E. SHUTOVA, A. HERBELOT, X. ZHU, M. APIDIANAKI & S. M. MOHAMMAD, Éds., *Proceedings of the 13th International Workshop on Semantic Evaluation*, p. 54–63, Minneapolis, Minnesota, USA : Association for Computational Linguistics. DOI : 10.18653/v1/S19-2007.

BEAN A. M., KEARNS R. O., ROMANOU A., HAFNER F. S., MAYNE H., BATZNER J., FOROUTAN N., SCHMITZ C., KORGUL K., BATRA H., DEB O., BEHARRY E., EMDE C., FOSTER T., GAUSEN A., GRANDURY M., HAN S., HOFMANN V., IBRAHIM L., KIM H., KIRK H. R., LIN F., LIU G. K.-M., LUETTGAU L., MAGOMERE J., RYSTRØM J., SOTNIKOVA A., YANG Y., ZHAO Y., BIBI A., BOSSELUT A., CLARK R., COHAN A., FOERSTER J. N., GAL Y., HALE S. A., RAJI I. D., SUMMERFIELD C., TORR P., UDUDEC C., ROCHER L. & MAHDI A. (2025). Measuring what matters : Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

BOWMAN S. R. & DAHL G. (2021). What will it take to fix benchmarking in natural language understanding ? In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4843–4855, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.385.

CLEVELAND CLINIC (2025). Genetic disorders.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R. *et al.* (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv :2110.14168*.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

DICTIONARY.COM (2025). Ageism terms.

ENCYCLOPEDIA BRITANNICA (2025). List of religious populations.

FESSLER J.-M. & GRÉMY F. (2001). Opinion paper : Ethical problems in health information systems. *Methods of information in medicine*, **40**(04), 359–361.

FLEISS J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76, 5**, 378–382.

GALLEGOS I. O., ROSSI R. A., BARROW J., TANJIM M. M., KIM S., DERNONCOURT F., YU T., ZHANG R. & AHMED N. K. (2024). Bias and fairness in large language models : A survey. *Computational Linguistics*, **50**(3), 1097–1179. DOI : 10.1162/coli_a_00524.

GEMA A. P., LEANG J. O. J., HONG G., DEVOTO A., MANCINO A. C. M., SAXENA R., HE X., ZHAO Y., DU X., GHASEMI MADANI M. R., BARALE C., MCHARDY R., HARRIS J., KADDOUR J., VAN KRIEKEN E. & MINERVINI P. (2025). Are we done with MMLU ? In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language*

Technologies (Volume 1 : Long Papers), p. 5069–5096, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.262.

GLAAD (2023). Glossary of terms : Lgbtq. Accessed : 2025-05-15.

HAGENDORFF T., BOSSERT L. N., TSE Y. F. & SINGER P. (2022). Speciesist bias in ai : how ai applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics*, **3**, 1–18.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

HILTON J. L. & VON HIPPEL W. (1996). Stereotypes. *Annual review of psychology*, **47**(1), 237–271.

HOFMANN V., KALLURI P. R., JURAFSKY D. & KING S. (2024). Ai generates covertly racist decisions about people based on their dialect. *Nature*, **633**, 147154. DOI : https ://doi.org/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 : https ://doi.org/10.1111/lnc3.12432.

JOB C. *et al.* (2024). Health professionals implicit bias of patients with low socioeconomic status (ses) and its effects on clinical decision-making : A scoping review. *BMJ Open*, **14**(7). DOI : 10.1136/bmjopen-2023-081723.

KRAFT A., SIMON J. & SCHIMMLER S. (2025). Social bias in popular question-answering benchmarks.

LARSON B. (2017). Gender as a variable in natural-language processing : Ethical considerations. 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. 1–11, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1601.

LEHMANN S., OEPEN S., REGNIER-PROST S., NETTER K., LUX V., KLEIN J., FALKEDAL K., FOUVRY F., ESTIVAL D., DAUPHIN E., COMPAGNION H., BAUR J., BALKAN L. & ARNOLD D. (1996). TSNLP - test suites for natural language processing. In *COLING 1996 Volume 2 : The 16th International Conference on Computational Linguistics*.

LITSCHKO R., MÜLLER-EBERSTEIN M., VAN DER GOOT R., WEBER-GENZEL L. & PLANK B. (2023). Establishing trustworthiness : Rethinking tasks and model evaluation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 193–203, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.14.

MAIRIE DE PARIS – HANDICAP (2025). Comprendre le handicap.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MAPLE COMMUNITY (2025). Types of sensory disabilities and impairment.

MEDLINEPLUS (2025). Degenerative nerve diseases.

MENDELSOHN J., TSVETKOV Y. & JURAFSKY D. (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, **3**, 55. DOI : 10.3389/frai.2020.00055.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). CrowS-pairs : A challenge dataset for measuring social biases in masked language models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), p. 1953–1967, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

NATIONAL ACADEMIES PRESS (2025). Framing the issues : ageist language and stigma. eBook.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE (2025). Genetic disorders.

NATIONAL LIBRARY OF SCOTLAND (2025). Contemporary slurs.

NÉVÉOL A., DUPONT Y., BEZANÇON J. & FORT K. (2022). French CrowS-pairs : Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8521–8531, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.583.

NHS (2025). Household gadgets and equipment to make life easier.

OPENMIND PROJECT (2025). All faiths & religions.

PAGOTO S. L., PALMER L. & HORWITZ-WILLIS N. (2023). The next infodemic : Abortion misinformation. *J Med Internet Res*, **25**, e42582. DOI : 10.2196/42582.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*.

RAJI D., DENTON E., BENDER E. M., HANNA A. & PAULLADA A. (2021). Ai and the everything in the whole wide world benchmark. In J. VANSCHOREN & S. YEUNG, Éds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

REITER E. (2025). We should evaluate real-world impact. *Computational Linguistics*, **51**(4), 1419–1431. 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.

RUTGERS UNIVERSITY (2025). Sensory disabilities.

SANTONICCOLO F., TROMBETTA T., PARADISO M. N. & ROLLÈ L. (2023). Gender and media representations : A review of the literature on gender stereotypes, objectification and sexualization. *International Journal of Environmental Research and Public Health*, **20**(10). DOI : 10.3390/ijerph20105770.

SHI F., SUZGUN M., FREITAG M., WANG X., SRIVATS S., VOSOUGHI S., CHUNG H. W., TAY Y., RUDER S., ZHOU D., DAS D. & WEI J. (2022). Language models are multilingual chain-of-thought reasoners.

SINGH S., ROMANOU A., FOURRIER C., ADELANI D. I., NGUI J. G., VILA-SUERO D., LIMKONCHOTIWAT P., MARCHISIO K., LEONG W. Q., SUSANTO Y., NG R., LONGPRE S., RUDER S., KO W.-Y., BOSSELUT A., OH A., MARTINS A., CHOSHEN L., IPPOLITO D., FERRANTE E., FADAEE M., ERMIS B. & HOOKER S. (2025). Global MMLU : Understanding and addressing cultural and linguistic biases in multilingual evaluation. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 18761–18799, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.919.

SPIVAK G. C. (1988). Can the subaltern speak ? *Die Philosophin*, **14**(27), 42–58. DOI : 10.5840/philosophin200314275.

SRIVASTAVA A., RASTOGI A., RAO A. *et al.* (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models.

STANFORD UNIVERSITY (2025). Disability language guide. PDF.

SURESH H. & GUTTAG J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3465416.3483305.

TALMOR A., HERZIG J., LOURIE N. & BERANT J. (2019). Commonsenseqa : A question answering challenge targeting commonsense knowledge. In *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. 4149–4158.

THOMAS R. L. & UMINSKY D. (2022). Reliance on metrics is a fundamental challenge for ai. *Patterns*, **3**(5), 100476. DOI : https ://doi.org/10.1016/j.patter.2022.100476.

UK ETHNICITY FACTS & FIGURES (2025). Style guide : ethnic groups.

UK GOVERNMENT (2025a). Inclusive language : words to use and avoid when writing about disability.

UK GOVERNMENT (2025b). List of nationalities.

UNAFAM (2025). Troubles psychiques principaux.

UNIVERSITY OF WASHINGTON – DO-IT (2025). Working together : computers and people with sensory impairments.

U.S. EPA – AMERICA'S CHILDREN & THE ENVIRONMENT (2025). Health & neurodevelopmental disorders.

WIKIPEDIA CONTRIBUTORS (2025). Degenerative disease.

WORLD HEALTH ORGANIZATION (2025). Mental disorders.

A Illustration des choix d'annotation manuelle

Exemples (CSQA)	Annotation	Explication
If a clock is not ticking, what is its likely status ? *[stop working, dead batteries, fail to work, time event, working correctly]*	Neutre (pas de biais)	Pas de contenu pouvant nuire à des catégories de population, ni de pré-supposé culturel ou politique.
What state is usually always red in the national elections ? *[utah, louisiana, texas, oklahoma, redtwig]*	Biais culturel	Des connaissances spécifiques à la culture états-unienne sont requises pour répondre à la question.
What could you find a diaphragm in ? *[drugstore, woman's body, human, drugstore, person's chest]*	Biais de genre implicite	Les choix de réponses supposent que le corps d'une femme est différent d'un corps humain (et ne contient potentiellement pas de diaphragme).

TABLE 1 – Exemples d'annotations dans CSQA pour les catégories les plus proches, avec explications.

B Sources des indicateurs d'appartenance à des groupes sociaux

Les listes d'indicateurs d'appartenance à des groupes sociaux ont été réalisées à partir des sources suivantes : Mairie de Paris – Handicap (2025); Accessibe (2025); NHS (2025); Rutgers University (2025); Maple Community (2025); University of Washington – DO-IT (2025); American Psychiatric Association (2025); World Health Organization (2025); U.S. EPA – America's Children & the Environment (2025); MedlinePlus (2025); Wikipedia contributors (2025); UNAFAM (2025); Cleveland Clinic (2025); National Human Genome Research Institute (2025); UK Government (2025a); Stanford University (2025); Dictionary.com (2025); National Academies Press (2025); Mendelsohn *et al.* (2020); GLAAD (2023); Job *et al.* (2024); Basile *et al.* (2019); UK Government (2025b); UK Ethnicity Facts & Figures (2025); National Library of Scotland (2025); Encyclopedia Britannica (2025); OpenMind Project (2025).

C Exemple de triplets extraits

Question	Darla et Janice sont deux mères élevant un fils, Timmy. Quelle proposition est la plus susceptible de décrire Timmy ?
Réponse	Tant que Timmy a au moins une personne encourageante et acceptante prenant soin de lui, il aura un développement sain.
Triplets	[('mère', 'élever', 'fils'), ('femmes', 'être', 'deux mères'), ('il', 'avoir', 'développement sain')]

TABLE 2 – Exemple de cas de MMLU avec les triplets extraits.

D Distribution des indicateurs de groupes sociaux

Type	**Pourcentage**		
	MMLU	CSQA	MGSM
Genre	67,42	26,47	99,4
Nationalité	16,93	38,68	0,6
Religion	7,6	26,56	0
Handicap	2,61	6,49	0
Origine ethnique	4,56	1,32	0
LGBT+	0,48	0,38	0

TABLE 3 – Distribution des groupes sociaux selon le nombre d'indicateurs détectés.