Ethics and NLP: 10 years after

Karën Fort, Aurélie Névéol

Actes de la journée d'étude
Journée Éthique et TAL 2024

Avec le soutien de l'ATALA, du GDR TAL et du GDR LIFT 2

Éditeurs

Karën Fort
Sorbonne Université, Loria

Aurélie Névéol
Université Paris-Saclay, LISN

9 Mars 2024
LORIA, 615, rue du Jardin Botanique, 54600 Villers-les-Nancy

Préface

En 2014 eut lieu la première journée scientifique dédiée à l'éthique du TAL, la première journée ATALA Ethique et TAL. 10 ans plus tard, le monde du TAL a subi plusieurs révolutions et l'éthique n'a jamais été aussi présente, dans les médias, nos recherches, les appels des plus grandes conférences, comme dans l'administration du TAL international (comités d'éthique de conférences et d'ACL).

Nous nous proposons de faire le point sur les avancées et les défis à venir le 2 avril 2024, au LORIA (Nancy), avec vous et nos invité·e·s :

— Emily M. Bender (Professeure à l'Université de Washington, Présidente d'ACL),
— Steven Bird (Professeur à Charles Darwin University),
— Alexis Michaud (DR CNRS, LACITO).

Organisatrices

Karën Fort, Aurélie Névéol.

Comité de Programme

Gilles Adda, Maxime Amblard, Jean-Yves Antoine, Philippe Blache, Vincent Claveau, Miguel Couceiro, Fanny Ducel, Cécile Fabre, Benoît Favre, Karën Fort, Anaïs Lefeuvre-Halftermeyer, Gaël Lejeune, Hugues de Mazancourt, Alice Millour, Philippe Muller, Aurélie Névéol.

Édition des actes

Caio Corro.

Remerciements

Nous remercions Anne-Marie Messaoudi, Marie Baron, Caio Corro et Gaël Lejeune pour leur aide dans l'organisation de cette journée. Nous remercions l'ATALA, du GDR TAL et le GDR LIFT 2 pour le soutien financier. Nous remercions les auteurs et autrices des résumés qui rendent cette journée d'étude possible.

Programme de la journée

9h	**Accueil**
9h30	**Introduction** (Karën Fort)
9h50	**Présentations**

— Évaluation éthique de l'impact du numérique : une typologie pour la détermination des risques et de leurs vulnérabilités dans une perspective conséquentialiste (Jean-Yves Antoine, Anaïs Halftermeyer)
— Empreinte carbone des expériences en TAL : les défis de la reproductibilité (Clément Morand, Aurélie Névéol, Anne-Laure Ligozat)
— Harmful NLP : Towards a systemic injustice approach (Marjolein Lanzing, Katrin Schulz)

11h	**Pause café + posters**

— Les biais dans les LLMs, de quoi parle-t-on, quelles pistes pour les détecter et les mitiger ? (Anaïs Bekolo, Emilie Sirvent-Hien, Christèle Tarnec)
— What ChatGPT tells us about ourselves (Aaron Boussidan, Fanny Ducel, Aurélie Neveol, Karën Fort)
— Mesurer les risques de discrimination dans une tâche de détection d'entités nommées (Hugues de Mazancourt, Alice Bruguier, Flavie Nguyen)
— Towards an Ethical Compression of Large Language Models (Irina Proskurina, Guillaume Metzler, Julien Velcin)

12h	**Présentation invitée** **Meaning making with artificial interlocutors and risks of language technology** Emily M. Bender
13h-14h15	**Déjeuner**
14h15	**Présentation invitée** **Must NLP be Extractive ?** Steven Bird
15h15	**Présentations**

— Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le TAL et ses conséquences (Mélanie Jouitteau, Loïc Grobol)
— Cross-Lingual Transfer of Debiasing Techniques (Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, Bart Baesens)
— Normaliser l'IA, une réponse au dilemme éthique des industriels de la langue (Hugues de Mazancourt, Alain Couillault)

16h15	**Pause café**
16h45	**Présentation invitée** **Des insouciances de l'archivisme à une réflexivité constante : considérations éthiques en linguistique de terrain à l'ère du TAL** Alexis Michaud
18h	**Fin**

Table des matières

ÉVALUATION ÉTHIQUE DE L'IMPACT DU NUMÉRIQUE : 10 ANS PLUS TARD, UNE TYPOLOGIE MISE À JOUR POUR LA DÉTERMINATION DES RISQUES ET DE LEURS VULNÉRABILITÉS DANS UNE PERSPECTIVE CONSÉQUENTIALISTE.

Jean-Yves Antoine[12*], **Anaïs Halftermeyer**[21*]
[1]LIFAT, fédération ICVL, Université de Tours
[2]LIFO, fédération ICVL, Université d'Orléans

[*]Adresses mail: Jean-Yves.Antoine@univ-tours.fr, Anais.Halftermeyer@univ-orleans.fr

ABSTRACT

This paper proposes to assess the ethical impacts of digital technologies by means of a risk analysis that relates to a consequentialist vision of ethics. This analysis relies on a typology of risk vulnerabilities that aims to cover all the potential impacts of any digital technology. The typology provides an evaluation grid for an ethical assessment of researches and applications, particularly in AI and NLP. It is structured as a hierarchy of vulnerability where the top classes are corresponding to environmental risks (geophysics and biodiversity), individual risk (physical, cognitive and psychological vulnerabilities) and societal risks (psychosocial and social vulnerabilities).

1 Introduction

La récente apparition dans le débat public des grands modèles de langage et de l'IA générative a mis en exergue de manière particulièrement visible la question des enjeux éthiques posés par l'Intelligence Artificielle. Si l'importance de ces questions se trouve renforcée par l'explosion des applications de l'Intelligence Artificielle, ces enjeux ne sont pourtant pas nouveaux. Il y a ainsi près de 40 ans, Mason (1986) s'interrogeait sur l'ensemble des bonnes propriétés éthiques que se devait de respecter tout concepteur informatique. Les principes PAPA (*Privacy, Accuracy, Property, Accessibility*) qu'il a proposés alors sont toujours d'actualité. Avec l'émergence des techniques d'apprentissage automatique en Intelligence Artificielle, de nouveaux principes éthiques ont été progressivement considérés tels que par exemple l'équité (*fairness*) ou la transparence des modèles élaborés. Au vu de l'accélération des applications relevant de l'Intelligence Artificielle, il semble important que l'évaluation éthique des systèmes numériques (IA comprise) ne se contente pas de suivre l'évolution des techniques et la découverte de problèmes après déploiement. De fait, il semble intéressant que cette évaluation éthique puisse considérer une application numérique d'une manière proche de celle de tout autre artefact technique. Dans cette perspective, nous proposons un cadre d'analyse éthique préventif reposant sur deux idées principales :

- Définition d'une typologie de facteurs de risque guidée par la recherche des objets (environnement, individus, société…) pouvant subir l'impact d'une technologie donnée. Cette analyse de risque est conduite dans une démarche conséquentialiste qui ne doit pas être confondue avec une démarche utilitariste (Mill 1861). Notre propos n'est en effet pas de chercher une maximisation théorique du bien-être permis par la technologie, mais de caractériser des facteurs d'attention qui pourront servir au débat sur le déploiement et les conditions d'usage d'une technologie.

- Analyse éthique agnostique, au sens où la typologie de facteurs de risque est indépendante de l'état de l'art technologie numérique et peut s'appliquer à tout artefact technique[1].

Notre démarche s'éloigne, sans s'opposer, des initiatives institutionnelles promues jusqu'à une date récente en matière d'évaluation de l'impact éthique des technologies numériques. Celles-ci reposaient en effet avant tout sur une vision éthique de type déontologique.

[1]Notons toutefois que notre réflexion a effectivement été guidée par l'évolution du Traitement Automatique des Langues et de l'Intelligence Artificielle.

2 Évaluation éthique de l'Intelligence Artificielle : recommandations et éthique déontologique

L'émergence des techniques d'apprentissage automatique a permis une vraie prise de conscience par la communauté informatique des impacts éthiques de l'Intelligence Artificielle et plus globalement du numérique. Ainsi, l'observation de biais de genre ou d'origine dans les données d'apprentissage et la reproduction des stéréotypes dont ils découlaient a conduit à donner une attention particulière à l'équité des modèles. La capacité de ces modèles à proposer une aide à la décision efficace mais peu explicable a également mis en avant la question de la transparence des modèles. Au fil des problèmes identifiés, les acteurs et actrices du numériques ont ainsi défini un ensemble de principes (*privacy, property, accuracy, fairness, transparency, human-control, accessibility, responsability*...) qui ont été mis en avant, dans une démarche déontologique, comme devant être au centre des préoccupations d'une conception responsable des systèmes. Ces principes ont été regroupés dans une multitude de guides de recommandations, qui ont été aussi bien le fait de scientifiques réunis par des institutions nationales (Dittrich and Kenneally 2012, CNIL 2017) ou transnationales telles que l'UE (Council of the European Union 2024), l'UNESCO ou l'ONU (GEHN IA 2018,...) que par des industriels (The Future of Life Institute 2017)[2]. Jobin et al. (2019) propose un panorama très complet du foisonnement des initiatives de ce type. Plus récemment, ces guides ont été accompagnés d'outils méthodologiques proposant une évaluation du respect de ces recommandations déontologiques sous forme de questionnaires, pouvant conduire à la définition d'un score de qualité éthique. Un exemple parmi d'autres de ce type d'initiatives se retrouve par exemple dans l'Ethical Impact Assessment[3] (EIA) tool proposé par l'UNESCO en 2023, ou bien encore dans l'Outil d'Évaluation de l'Incidence Algorithmique mis en place par l'État canadien[4]. Ces initiatives sont utiles à une conception responsable au sens où elles offrent un outillage méthodologique pratique pour analyser les implications éthiques des technologies que l'on développe. Elles proposent un cadre d'analyse à gros grain qui est également intéressant pour guider le débat et la décision politique et qui se focalise sur les préoccupations identifiées comme majeures par l'ensemble des acteurs et actrices du numérique. Ce type de démarche nous apparaît toutefois insuffisant pour répondre à certaines questions qui ne sauraient être ignorées :

- Les principes déontologiques définis dans ces recommandations sont tous issus de problèmes identifiés sur des technologies déployées. Ils peuvent dès lors passer à côté de points d'attention non encore identifiés ou de techniques émergentes.
- Ces principes mettent en exergue des préoccupations essentielles auxquelles il convient d'être très attentifs. Ils laissent toutefois de côté des préoccupations à grain plus fin qui pourraient de ce fait être ignorées. Les outils d'évaluation déontologique avec calcul de score reposant sur le respect de principes centraux pourrait ainsi laisser une impression erronée de sûreté éthique d'applications potentiellement problématiques.

C'est pour répondre à ces insuffisances que nous proposons un cadre d'évaluation conséquentialiste dont l'objectif est d'éclairer la conception en identifiant des sujets de préoccupation potentiellement ignorés. Ceci pour nourrir un débat éthique et non pas pour proposer une sorte de "certification" éthique. Ne visant pas les mêmes objectifs au-delà d'une conception responsable, approche déontologique et approche conséquentialiste nous apparaissent de ce point de vue complémentaires.

3 Typologie de risques pour une caractérisation conséquentialiste des impacts

L'éthique conséquentialiste suit une approche téléologique qui se focalise non pas sur des principes moraux à respecter, mais sur les conséquences de nos actions (ici, le développement mais également l'usage d'applications numériques). Les principes de précaution et de responsabilité, théorisés par Jonas (1979), relèvent de cette logique et répondent à notre démarche. Notre objectif est de définir un cadre systématique pour l'étude des impacts potentiels d'une technologie. L'idée est de caractériser ces impacts en terme de risques (positifs ou négatifs), à l'instar de l'analyse du risque industriel ou environnemental. Nous proposons ainsi de suivre un cadre normatif standard, comme par exemple les normes européennes (EN 292-1 et EN 1050) "relatives aux risques ayant incidence sur la santé humaine". Dans ce type de démarche, le risque est modélisé comme l'association de trois concepts :

- le **facteur de risque**, qui caractérise l'élément ou le processus susceptible de causer un risque, donc d'être la cause d'une situation indésirable (ici, le développement et/ou l'usage d'une technologie),
- la **vulnérabilité**, qui revient à décrire l'objet du risque, à savoir l'élément qui le subit et d'autre part ses conséquences,

[2] https://futureoflife.org/open-letter/ai-principles/

[3] https://www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence

[4] https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html

- la **criticité**, qui combine l'impact du risque (son effet ou sa gravité, pour reprendre le standard européen détaillé plus haut) avec sa probabilité d'occurrence. La question que l'on se pose ici est l'évaluation de l'impact des technologies que nous développons. Cette évaluation peut être expérimentale (étude statistique sur une population de test) ou subjective et introspective (retour d'expérience d'experts, par exemple). Une analyse sérieuse de la criticité est le plus souvent une entreprise lourde et ne peut être menée que si l'analyse des vulnérabilités a caractérisé préalablement en enjeu suffisamment sensible pour la justifier.

Dans le cadre de nos travaux, notre objectif n'est donc pas de conduire une évaluation complète (par exemple à la manière du calcul d'un "score éthique" utilitariste) mais d'aider les personnes impliquées dans la conception ou le déploiement d'une technologie à caractériser les points de préoccupation qui devront conduire éventuellement à une analyse plus poussée. Notre propos n'est donc pas de proposer une méthode pour estimer la criticité d'un risque éthique, mais de définir de manière exhaustive, les objets de vulnérabilité éthique d'un système. Notre cadre d'analyse se focalise donc sur l'étude des vulnérabilités liés à un système technique et repose sur une typologie hiérarchique de celles-ci, distinguant pour le premier niveau les risques *environnementaux*, *individuels* et *sociétaux*. Les niveaux inférieurs de la typologie concernent les conséquences spécifiques à chaque vulnérabilité. Par exemple, dans le cadre du risque *individuel*, on distingue les impacts (positifs comme négatifs) du système en termes d'impact *physique*, *cognitif* ou *psychique*. Chaque conséquence peut elle-même est subdivisée en sous-types d'impacts, afin de fournir une illustration de chaque branche et feuille pour guider une étude conséquentialiste.

4 Conclusion

Pour élaborer notre typologie de vulnérabilités, nous avons considéré dans un premier temps des nomenclatures médicales (classifications CIF, CIM ou DSM de l'OMS ou de l'APA) ou des études sociales établies en dehors de toute considération liées à l'usage d'artefacts techniques. Ce n'est que dans second temps que nous avons testé la validité de cette catégorisation par l'étude d'études documentant les effets liés à l'usage d'une grande diversité de systèmes numériques. La première version de cette typologie a été proposée il y a désormais 10 ans : Antoine et al. (2014); Lefeuvre et al. (2015); Lefeuvre-Halftermeyer et al. (2016). Nous l'avons réactualisée au fil de la découverte de nouveaux impacts relevés dans la littérature. L'arrivée de l'IA générative pose avec acuité de nouvelles préoccupations non envisagées par le passé. Cette arrivée subite nous a permis de valider la robustesse de notre typologie et nous sommes confiants dans sa capacité à proposer un cadre préventif d'analyse éthique dans le futur. Cette vision conséquentialiste des enjeux éthiques liés à la conception et au déploiement d'une technologie est restée très rare dans le domaine du numérique. Notre typologie se rapproche toutefois de deux propositions récentes de taxonomies d'impacts sociotechniques des systèmes numériques dans le domaine de la vision par ordinateur et/ou du traitement automatique des langues (Katzman et al. 2021; Shelby et al. 2023)[5]. Les taxonomies proposées par Katzman et al. (2021) et Shelby et al. (2023) nous semblent plus liées à la technologie que notre proposition, et sont marquées par un focus particulier sur le renforcement des biais sociaux dans la société. Les revues de la littérature menées dans ces études sont par contre impressionnantes par leur caractère exhaustif et systématisé. Notre intention est donc de reprendre toutes les études relevées dans ces synthèses pour éprouver notre taxonomie, avant d'en faire une publication publique et éventuellement participative sous forme d'une ontologie. Nous aimerions par ailleurs réfléchir à l'implémentation de cette taxonomie dans le cadre d'une analyse de risque menée dans le cadre de l'IA Act européen.

References

Antoine, J.-Y., Labat, M.-E., Lefeuvre, A., and Toinard, C. (2014). Vers une méthode de maîtrise des risques dans l'informatisation de l'aide au handicap. In *Envirorisk'2014, Le forum de la gestion des risques technologiques, naturels et sanitaires*, Gestion des risques naturels, technologiques et sanitaires - Forum Envirorisk'2014, page 9 pages, Bourges, France. Cépaduès.

CNIL (2017). Comment permettre à l'homme de garder la main ? les enjeux éthiques des algorithmes et de l'intelligence artificielle. rapport de synthèse du débat public « ethique et numérique ». Technical report, CNIL.

Council of the European Union (2024). IA ACT. https://artificialintelligenceact.eu/fr/ai-act-explorer/.

Dittrich, D. and Kenneally, E. (2012). The menlo report: Ethical principles guiding information and communication technology research. *SSRN Electronic Journal*.

[5] Shelby et al. (2023) propose ainsi une catégorisation des préjudices potentiels des technologies numériques en 5 catégories distinctes : *représentationnels, économiques, variation des performances des systèmes suivant les individus ou les catégories sociales, interpersonnels, sociétaux*. Cette taxonomie est ensuite raffinée sur un second niveau de sous-types de préjudices potentiels. La démarche de Katzman et al. (2021) et celle de Shelby et al. (2023) sont ainsi très proches de la notre, et visent le même objectif : définir une check-list de préoccupations potentielles à étudier systématiquement dans la perspective d'une conception responsable.

GEHN IA (2018). Lignes directrices en matière d'éthique pour une ia de confiance. Rapport du groupe d'experts indépendants de haut niveau sur l'intelligence artificielle constitué par la commission européenne. Technical report, Commission Européenne.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1.

Jonas, H. (trad. fr. 1990; 1979). *Le principe responsabilité*. Champs Flammarion.

Katzman, J., Barocas, S., Blodgett, S. L., Laird, K., Scheuerman, M., and Wallach, H. (2021). Representational harms in image tagging. In *Beyond Fair Computer Vision Workshop at CVPR 2021*.

Lefeuvre, A., Antoine, J.-Y., and Allegre, W. (2015). Ethique conséquentialiste et traitement automatique des langues : une typologie de facteurs de risques adaptée aux technologies langagières. In *Atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), conférence TALN'2015*, Actes de la 1e Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), Caen (France), pages 53–66, Caen, France.

Lefeuvre-Halftermeyer, A., Govaere, V., Antoine, J.-Y., Allegre, W., Pouplin, S., Departe, J.-P., Slimani, S., and Spagnulo, A. (2016). Typologie des risques pour une analyse éthique de l'impact des technologies du TAL [typology of risks for an ethical analysis of the impact of NLP technologies]. *Traitement Automatique des Langues*, 57(2):47–71.

Mason, R. (1986). *Four ethical issues of the information age*. Mis Quarterly.

Mill, J. (rééd. 1998; 1861). *Utilitarianism*. Oxford University Press.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23. ACM.

The Future of Life Institute (2017). Asilomar AI Principles. https://futureoflife.org/open-letter/ai-principles/.

EMPREINTE CARBONE DES EXPÉRIENCES EN TAL : LES DÉFIS DE LA REPRODUCTIBILITÉ

Clément Morand[1] **Aurélie Névéol**[1] **Anne-Laure Ligozat**[1,2]
(1) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique
(2) ENSIIE
prenom.nom@lisn.upsclay.fr

1 Introduction

Les dégâts environnementaux du Traitement Automatique des Langues (TAL) peuvent être très importants (Strubell *et al.*, 2019; Luccioni *et al.*, 2023). Afin de mieux les comprendre et mieux les maîtriser, (Henderson *et al.*, 2020; Bender *et al.*, 2021) notamment ont proposé de systématiquement calculer et indiquer des évaluations de dégâts environnementaux pour chaque modèle. Subséquemment, des évaluations d'empreinte carbone sont apparues dans des articles expérimentaux plus classiques (Bannour *et al.*, 2021; Cattan *et al.*, 2021, 2022; Dinarelli *et al.*, 2022).

Les méthodologies de calcul de l'empreinte carbone sont généralement fondées sur la mesure ou l'estimation de la consommation des équipements sur lesquels les modèles sont entraînés ou déployés. Des outils de mesure comme CodeCarbon (Schmidt *et al.*, 2022) ou d'estimation comme Green Algorithms (Lannelongue *et al.*, 2021) peuvent être utilisés.

Cependant, le manque d'informations, en particulier sur le matériel précis utilisé, ne permet pas toujours de vérifier les calculs d'empreinte carbone. Dans ce travail, nous avons évalué la reproductibilité des calculs d'empreinte carbone de plusieurs expériences de TAL, en nous appuyant sur l'outil MLCA, que nous présenterons en 2.1.

2 Expériences de reproductibilité

La question de la reproductibilité des résultats n'est pas nouvelle et a déjà été explorée. Cohen *et al.* (2018) par exemple définissent trois niveaux différents de reproductibilité. Le premier est la capacité à reproduire exactement la même *valeur* ; ce niveau est rarement réalisable car les processus ne sont pas toujours déterministes. Le second est la capacité à obtenir des résultats proches de ceux présentés, il s'agit du niveau *résultat*. Nos expériences se situent à ce niveau. Le troisième niveau est celui de la *conclusion*, c'est-à-dire la capacité à arriver aux mêmes conclusions avec la ré-itération d'un processus expérimental, ce qui inclut une interprétation de *résultats* obtenus expérimentalement.

Nous avons analysé les résultats de 5 études de TAL : Strubell *et al.* (2019) et Luccioni *et al.* (2023) évaluent l'empreinte carbone de l'entraînement de modèles de langue, Bannour *et al.* (2021) la reconnaissance d'entité nommées en français, Dinarelli *et al.* (2022) la compréhension de la parole en anglais et en français et Cattan *et al.* (2022) la reconnaissance du français parlé. Ces études ont été choisies car elles comportent une évaluation de l'empreinte carbone des expériences, et en outre soit donnent suffisamment d'informations pour tenter de reproduire les calculs, soit ont été faites par des collègues que nous connaissions et à qui nous pouvions demander des informations supplémentaires.

2.1 Outil

Les calculs d'empreintes réalisés dans les cinq études choisies ont été effectués à l'aide d'outils différents par chacun des groupes d'auteur·ice·s et ont porté sur plusieurs aspects de l'impact liés aux équipements utilisés dans les expérimentations : les impacts liés à l'utilisation (toutes les études) et les impacts liés à la production du matériel (uniquement (Luccioni *et al.*, 2023)).

Pour nos expériences, nous avons cherché à utiliser un cadre de mesure commun. Nous nous sommes appuyés sur l'outil MLCA (Morand, 2023). Cet outil prend en compte les phases de production (extraction des matières premières et fabrication) et d'usage des équipements utilisés pour faire tourner un programme, ainsi que trois indicateurs environnementaux, empreinte carbone, épuisement des ressources énergétiques et épuisement des ressources abiotiques (métaux) (Bruijn *et al.*, 2002). Dans le travail présenté ici, nous utilisons les résultats d'empreinte carbone issus de la fabrication et de l'usage des équipements, qui permettent de couvrir l'ensemble des mesures réalisées dans les travaux que nous cherchons à reproduire. L'estimation des impacts de la fabrication s'appuie sur la méthodologie et les données de Boavizta (2021), et l'ajout des impacts des GPU sur celles de Gröger *et al.* (2021) et les données de Loubet

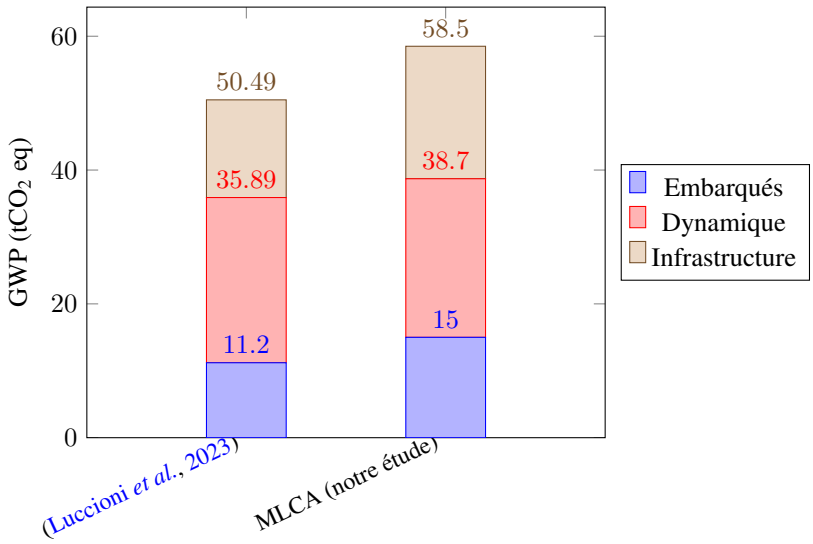

FIGURE 1 – Comparaison des estimations obtenues dans notre étude avec les résultats présentés dans la table trois de Luccioni *et al.* (2023) sur les différentes sources d'émissions (Les impacts embarqués sont ceux de la production du matériel attribuables à la tâche. La consommation dynamique est celle qui est due au matériel qui fait tourner le calcul. Le reste correspond à la consommation de l'infrastructure.)

et al. (2023). Les impacts de l'usage utilisent la méthodologie et les données de l'outil en ligne Green Algorithms (Lannelongue *et al.*, 2021) [1].

2.2 Résultats

Dans l'ensemble, la reproduction des résultats des différentes études s'est avérée beaucoup plus difficile que prévu. Nous avons rencontré plusieurs difficultés dans cette démarche. À moins qu'un réel effort ne soit fait pour permettre la réplication des résultats présentés dans un article, il est la plupart du temps très difficile de trouver toutes les informations nécessaires pour effectuer des estimations et reproduire ces résultats. Si le matériel sur lequel ces résultats ont été produits n'est pas détaillé, il est impossible de reproduire les expériences et de vérifier la qualité des résultats présentés.

La plupart du temps, nous avons pu réaliser des expériences grâce aux auteur·ice·s, qui nous ont donné des indications sur la configuration matérielle de leurs expériences que nous ne pouvions pas trouver dans les manuscrits (3 études sur 5). Nous avons également exploité les informations matérielles disponibles sur les centres de données utilisés. Lorsque les informations matérielles sont données, le modèle précis peut manquer dans notre base de données. Lorsque nécessaire, nous avons choisi un modèle proche du matériel réel en terme de configuration matérielle. par exemple l'étude du modèle BLOOM ((Luccioni *et al.*, 2023)) utilisait des informations sur les impacts de la production d'un serveur légèrement différent de celui utilisé alors que nous nous sommes basés sur les informations de configurations de la partition de Jean Zay utilisée pour entraîner le modèle afin de réaliser notre estimation.

Nous avons obtenu pour plusieurs expériences des estimations d'impact un peu plus élevées que celles des articles. En effet, les outils d'estimation comme celui que nous avons utilisé tendent à produire des résultats plus élevés que les outils de mesure (Jay *et al.*, 2023) (les résultats de Cattan *et al.* (2022), Dinarelli *et al.* (2022) et Strubell *et al.* (2019) ont été obtenus par des mesures) et en outre notre outil d'estimation a considéré une utilisation des équipements à 100% par défaut puisque le taux d'utilisation n'était généralement pas précisé. Nous avons réussi à reproduire les résultats de Jay *et al.* (2023), Dinarelli *et al.* (2022), Strubell *et al.* (2019), et Luccioni *et al.* (2023) au niveau *résultat* et ceux de Cattan *et al.* (2022) et Bannour *et al.* (2021) uniquement au niveau *conclusion*. Par exemple, pour l'étude de BLOOM nous obtenons les estimations d'impact présentées dans la figure 1.

Même lorsque nous disposions d'informations pour réaliser nos estimations avec suffisamment de précision pour espérer obtenir les résultats escomptés, nous avons été confrontés à plusieurs reprises à des incohérences dans les données présentées dans les résultats. Ce fut par exemple le cas pour certains résultats présentés dans Bannour *et al.* (2021) où nous avons observé des incohérences (facteur d'émission différent de celui de la France, du à la détermination auto-

1. http://calculator.green-algorithms.org/

matique erronée du facteur d'émission par l'outil Carbon Tracker) et inconsistances (facteur d'émission non constant) dans le facteur d'émission utilisé pour convertir la consommation d'électricité en empreinte carbone. Ce fut aussi le cas dans Cattan *et al.* (2022) où les résultats présentés dans le manuscrit sont plusieurs ordres de grandeur plus élevés que l'ordre grandeur attendu. Après avoir souligné les problèmes posés par les données présentées dans Cattan *et al.* (2022), les auteur·ice·s ont mené de nouvelles expériences pour résoudre les problèmes liés à leurs données et nous avons pu reproduire ces nouveaux résultats. Les expériences menées dans Jay *et al.* (2023), non commentées ici car hors TAL, ont été plus faciles à reproduire grâce au matériel supplémentaire fourni. Néanmoins, à cause d'une erreur de de recopie dans le supplementary material, les entrées exactes pour certaines expériences n'étaient pas indiquées et des hypothèses ont dû être formulées afin de reproduire des résultats exacts, ce qui rappelle la difficulté de rendre ses résultats reproductibles, même quand des efforts importants sont déployés par les auteurs et autrices.

3 Discussion

Par rapport aux niveaux de conclusions évoqués en introduction, nos expériences ont tenté de reproduire des expériences au niveau *résultat* : nous avons généralement essayé de retrouver des résultats obtenus avec une méthode différente. Parfois, face à des valeurs anormales dans un manuscrit, nous ne pouvions reproduire les résultats qu'au niveau de la *conclusion*, c'est-à-dire en constatant qu'une expérience plus longue a des effets plus importants qu'une expérience plus courte sur le même matériel.

Digan *et al.* (2020) ont proposé une liste de recommandations pour assurer un haut niveau de reproductibilité des résultats présentés dans un article. Parmi les règles parfois non respectées dans les manuscrits sur lesquels nous avons travaillé, on peut citer (R03) 'System metadata (e.g. RAM, CPU, OS, etc.)' et (R04) 'Record parameters of tools' qui compliquent la reproductibilité ou (R28) 'Absence of manual steps' qui pourraient expliquer des incohérences entre les tableaux ou des valeurs anormales dans un tableau.

Ce que nous avons observé sur les expériences de TAL rejoint les conclusions d'autres études sur la reproductibilité des évaluations environnementales, en particulier du numérique. Pasek *et al.* (2023) soulèvent que dans l'évaluation des dégâts environnementaux du secteur du numérique, le périmètre choisi induit des variations importantes des résultats. Pour les expériences de TAL, de telles variations peuvent par exemple être l'inclusion ou non de la production (extraction des matières premières et fabrication) du matériel, qui peut représenter une part significative des impacts (Luccioni *et al.*, 2023). Comme discuté dans Lippert (2016) ou encore Pasek *et al.* (2023), on observe une grande variabilité géographique et temporelle des facteurs d'émissions, ce qui est confirmé pour le numérique par Clément *et al.* (2020). Pour le TAL, les facteurs d'intensité carbone de l'électricité varient grandement en fonction de la localisation géographique (Lannelongue *et al.*, 2021) mais aussi temporellement (par exemple en fonction de la demande, des vents et de l'ensoleillement qui influent sur la part de production électrique renouvelable). Enfin, dans l'étude de la création de bilans carbone des entreprises, Lippert (2016) montre que de nombreuses décisions pas toujours traçables sont impliquées dans l'obtention d'un chiffre. Dans le cas des évaluations de modèles de TAL, cela peut se traduire dans le choix d'un matériel proche de celui utilisé car inexistant dans base de données ou encore dans le choix quant à la façon d'obtenir le taux d'utilisation des unités de calculs.

Au final, on constate que les problèmes de reproductibilité des résultats s'appliquent également aux évaluations des impacts environnementaux des expériences de TAL. L'introduction d'une méthodologie standardisée pour détailler les différents paramètres des évaluations environnementales des expériences de TAL (par exemple s'assurer de donner toutes les informations demandées par le calculateur Green Algorithms ainsi que la version de l'outil) pourrait grandement en améliorer la reproductibilité.

Références

BANNOUR N., GHANNAY S., NÉVÉOL A. & LIGOZAT A.-L. (2021). Evaluating the carbon footprint of NLP methods : a survey and analysis of existing tools. In N. S. MOOSAVI, I. GUREVYCH, A. FAN, T. WOLF, Y. HOU, A. MARASOVIĆ & S. RAVI, Éds., *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, p. 11–21, Virtual : Association for Computational Linguistics. doi:10.18653/v1/2021.sustainlp-1.2.

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.

BOAVIZTA (2021). Numérique et environnement : Comment évaluer l'empreinte de la fabrication d'un serveur, au-delà des émissions de gaz à effet de serre˜ ?

BRUIJN H., DUIN R., HUIJBREGTS M. A. J., GUINEE J. B., GORREE M., HEIJUNGS R., HUPPES G., KLEIJN R., KONING A., VAN OERS L., SLEESWIJK A. W., SUH S. & DE HAES H. A. U. (2002). *Handbook on Life Cycle*

Assessment - Operational Guide to the ISO Standards. Springer Dordrecht. doi:10.1007/0-306-48055-7.

CATTAN O., GHANNAY S., SERVAN C. & ROSSET S. (2022). Benchmarking transformers-based models on french spoken language understanding tasks. In *INTERSPEECH 2022*, Incheon, South Korea. HAL : hal-03715340.

CATTAN O., SERVAN C. & ROSSET S. (2021). On the usability of transformers-based models for a french question-answering task. In G. ANGELOVA, M. KUNILOVSKAYA, R. MITKOV & I. NIKOLOVA-KOLEVA, Éds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, p. 244–255 : INCOMA Ltd.

CLÉMENT L.-P. P.-V., JACQUEMOTTE Q. E. & HILTY L. M. (2020). Sources of variation in life cycle assessments of smartphones and tablet computers. *Environmental Impact Assessment Review*, **84**, 106416.

COHEN K. B., XIA J., ZWEIGENBAUM P., CALLAHAN T., HARGRAVES O., GOSS F., IDE N., NÉVÉOL A., GROUIN C. & HUNTER L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. In N. C. C. CHAIR), K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

DIGAN W., NÉVÉOL A., NEURAZ A., WACK M., BAUDOIN D., BURGUN A. & RANCE B. (2020). Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, **28**(3), 504–515. doi:10.1093/jamia/ocaa261.

DINARELLI M., NAGUIB M. & PORTET F. (2022). Toward low-cost end-to-end spoken language understanding. In H. KO & J. H. L. HANSEN, Éds., *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, p. 2728–2732 : ISCA. doi:10.21437/INTERSPEECH.2022-10702.

GRÖGER J., LIU R., STOBBE L., DRUSCHKE J. & RICHTER N. (2021). *Green Cloud Computing : lebenszyklusbasierte Datenerhebung zu Umweltwirkungen des Cloud Computing : Abschlussbericht*. Umweltbundesamt.

HENDERSON P., HU J., ROMOFF J., BRUNSKILL E., JURAFSKY D. & PINEAU J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, **21**(1). doi:10.5555/3455716.3455964.

JAY M., OSTAPENCO V., LEFEVRE L., TRYSTRAM D., ORGERIE A.-C. & FICHEL B. (2023). An experimental comparison of software-based power meters : focus on cpu and gpu. In *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, p. 106–118. doi:10.1109/CCGrid57682.2023.00020.

LANNELONGUE L., GREALEY J. & INOUYE M. (2021). Green algorithms : Quantifying the carbon footprint of computation. *Advanced Science*, **8**(12), 2100707. doi:https ://doi.org/10.1002/advs.202100707.

LIPPERT I. (2016). Failing the market, failing deliberative democracy : How scaling up corporate carbon reporting proliferates information asymmetries. *Big Data & Society*, **3**(2), 2053951716673390. doi:10.1177/2053951716673390.

LOUBET P., VINCENT A., COLLIN A., DEJOUS C., GHIOTTO A. & JEGO C. (2023). Life cycle assessment of ict in higher education : a comparison between desktop and single-board computers. *The International Journal of Life Cycle Assessment*, p. 1–19. doi:https ://doi.org/10.1007/s11367-022-02131-z.

LUCCIONI A. S., VIGUIER S. & LIGOZAT A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176b parameter language model. *Journal of Machine Learning Research*, **24**(253), 1–15.

MORAND C. (2023). Evaluation of the environmental impacts of Natural Language Processing methods.

PASEK A., VAUGHAN H. & STAROSIELSKI N. (2023). The world wide web of carbon : Toward a relational footprinting of information and communications technology's climate impacts. *Big Data & Society*, **10**(1), 20539517231158994. doi:10.1177/20539517231158994.

SCHMIDT V., GOYAL-KAMAL, COURTY B., FELD B., AMINE S., KNGOYAL, ZHAO F., JOSHI A., LUCCIONI S., LÉVAL M., BOGROFF A., DE LAVOREILLE H., LASKARIS N., CONNELL L., WANG Z., SABONI A., CATOVIC A., BLANK D., STECHLY M., ALENCON, JPW, BOOKS M., SWADIK S., M. H., COUTAREL M., POLLARD M., MCCARTHY C., HUSOM E. J., VICENTE F. & TAE J. (2022). mlco2/codecarbon : v2.1.4. doi:10.5281/zenodo.7049269.

STRUBELL E., GANESH A. & MCCALLUM A. (2019). Energy and policy considerations for deep learning in NLP. In A. KORHONEN, D. R. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1 : Long Papers*, p. 3645–3650 : Association for Computational Linguistics. doi:10.18653/v1/p19-1355.

HARMFUL NLP: TOWARDS A SYSTEMIC INJUSTICE APPROACH

Marjolein Lanzing[1], **Katrin Schulz**[2]
[1]philosophy, UvA
[2]ILLC, UvA

In recent years we have seen many cases of machine learning applications that show unfair biased behaviour towards particular groups or individuals. This led to growing societal concerns about harmful discrimination and reproduction of structural inequalities once these technologies become institutionalised in society. Identifying and resolving algorithmic bias and achieving algorithmic fairness has therefore become an important research field.

However, it has been argued that in AI research and policy, the remedies against harmful algorithms are often narrowly framed as addressing technological design problems, rather than complex, structural, social-political problems. This leads to a highly technocentric (and technocratic) approach towards algorithmic fairness A. Balayn (2021), and to the development of technological solutions that leave the actual causes of the problem unchallenged. As a result, the solutions provide only superficial symptom-fixes with no sense of real control of the problem.

In this paper we do two things. In the first part we present the results of a systematic literature review that we carried out to test the claim that in AI research harmful behaviour of AI technology is narrowly framed as a technological problem in need of a technological solution. For this we focus on the field of Natural Language Processing (NLP). We filtered the the ACL Anthology library, one of the biggest platforms for research papers on the study of computational linguistics and natural language processing, for papers that study the negative societal impact of NLP technology. From the 207 papers on this topic we found for the year 2022 we selected randomly 50 papers and coded them with respect to three questions: (i) how does the field frame the problem of negative societal impact of NLP technology, (ii) what do the authors describe as the causes for the negative impact, and (iii) what solutions are proposed for the problem. To summarise the results, our study confirms that the field dominantly adopts a techno-centric and technocratic approach to negative societal impact of NLP technology. In line with results of Blodgett et al. (2020), who specifically looked at the discussion bias in studies of language technology, we find that problem of negative societal impact is poorly conceptually framed and misses normative grounding. When it comes to the discussion of the causes of harmful behaviour of NLP technology, we do see that authors look beyond technology and point to underlying societal issues (for instance how bias in society can lead to biased language technology). However, the causal reasoning stops as soon as a technology-external source for the problem is identified. In that sense the causal framing of the problem is still narrow. Finally, we notice that even though authors do discuss causes of negative societal impact beyond simple malfunctions of technology, by far most papers are devoted to technological fixes and refrain from addressing the social-political or political-economical questions around NLP applications.

One might say that these findings are not surprising given that we looked at a sample of papers written by AI engineers and researchers. One might even argue that a techno-centric approach is fine as long as it solves the problem. However, the narrow causal framing of the problem blindsights researchers to the unavoidable limitations of a technological fix. The negative societal impact of AI technology inherently rests in the fact that this technology is situated in and interacts with society. From that perspective framing AI research as only concerned with technological issues is shortsighted. It should always be an interdisciplinary effort that looks at the interaction with society as well.

In the second part, we make a start on developing an alternative conceptual framing of the problem of harm and injustice in NLP. To challenge the techno-centric approach we propose to expand the conceptual tools to understand harmful NLP by creating a framework that also addresses the institutional, political, social and structural dimensions of the problem. We propose that this alternative framing should center around the notion of systemic injustice Haslanger (2022, 2023). Focusing on systemic instead of individual harms creates space to consider more complex mechanisms underlying patterns of algorithmic injustice. It also changes the type of mechanisms we are looking for: a systematic injustice approach would reveal a social-political context that is harmful as opposed to individual actions of AI systems or particular design choices made in the development of this technology.

Reframing the problem leads naturally also to a reframing of the solution. In particular, we will argue that this new framework will widen the NLP repertoire to address these harms beyond techno-centric solutions (such as debiasing) to include also legal, policy, institutional and organisational measures.

A final important advantage of a systematic injustice approach to societal harms is that it would open the debate on AI and fairness to be broadened to include societal harms and risks beyond bias such as climate justice, (human) exploitation and dependency on corporate technocrats. In fact, we claim that in many cases the problem of bias is

connected and intertwined with other material, social and political injustices, which the current technological fixes completely ignore.

The new framing proposed here will help to create awareness among NLP researchers concerned with harmful NLP and hopefully also raise so far unexplored conceptualisation and solutions of harmful NLP. Finally, it will challenge existing power relations in which these conceptualisation and solutions 'as techno-fixes' are left to the expertise and therefore discretion of powerful and wealthy technology companies.

References

A. Balayn, S. G. (2021). Beyond debiasing. tegulating ai and its inequalities. Technical report, European Digital Rights (EDRi), Brussel, Belgium.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Haslanger, S. (2022). Systemic and structural injustice: Is there a difference? *Journal of Social Philosophy*, 53:512 – 534.

Haslanger, S. (2023). Systemic and structural injustice: Is there a difference? *Philosophy*, 98:1 – 27.

LES BIAIS DANS LES LLMS, DE QUOI PARLE-T-ON, QUELLES PISTES POUR LES DÉTECTER ET LES MITIGER ?

Anaïs Bekolo[1], Emilie Sirvent-Hien[2], and Christèle Tarnec[3]

[1]Université Paris Cité, Orange (stage)
[2,3]Orange

ABSTRACT

Depuis la sortie de chatGPT en novembre 2022, les usages et les déploiements en production de système à base de grands modèles de langages (LLM) se multiplient. Pour autant les problèmes de fiabilité sont loin d'être résolus et de nombreux exemples de biais (des générations automatiques de lettres de recommandations particulièrement genrées par exemple (Wan, Pu, et al., 2023)) alertent sur le besoin de développer une approche responsable. L'objectif de cet article est de faire un état des lieux sur la notion de biais pour les LLM, comment les détecter, les évaluer et les pistes pour les mitiger. En particulier pour les LLMs, l'architecture des Transformers permet d'envisager les biais selon une classification de biais intrinsèques ou extrinsèques. Les méthodes d'évaluation des biais les plus connues reposent sur des benchmarks, dont la validité est parfois discutée. Dans le cas d'un modèle déjà pré-entraîné, les biais peuvent être en partie mitigés soit lors de re-entraînements du modèle (*memory-efficient fine-tuning, reinforcement learning*) soit directement dans le prompt (*role play*) . Mais les biais observés au moment de la prédiction notamment peuvent avoir des origines diffuses, difficiles à situer et donc à mitiger. Par conséquent, plutôt que de les envisager de façon générique, à l'échelle du modèle, l'évaluation et la mitigation des biais doivent être traitées selon un cas d'usage précis et son contexte pour pouvoir être déployés opérationnellement.

Depuis la sortie de chatGPT en novembre 2022, les usages et les déploiements en production de système à base de grands modèles de langages (LLM) se multiplient. Pour autant les problèmes de fiabilité sont loin d'être résolus et de nombreux exemples de biais (des générations automatiques de lettres de recommandations particulièrement genrées par exemple [40]) alertent sur le besoin de développer une approche responsable. L'objectif de cet article est de faire un état des lieux sur la notion de biais pour les LLM, comment les détecter, les évaluer et les pistes pour les mitiger.

Davat (2023) [10] définit le biais comme une divergence par rapport à une norme. Cette norme pouvant être de différentes natures, l'auteur identifie trois grands types de biais : le biais méthodologique, le biais cognitif et le biais socio-historique, chacun de ces types de biais étant respectivement une divergence selon un standard scientifique, selon une décision dite « rationnelle » et selon une société idéale. Ainsi, un même préjudice peut être le fait de plusieurs types de biais différents. Dans le cas classique d'une performance inégale d'un traducteur automatique pour deux langues données par exemple, la sous-représentation d'une des deux langues dans les données d'entraînement représente à la fois un biais méthodologique (biais de représentation) et à la fois un biais socio-historique (hiérarchisation des langues).

De fait, dans la littérature éthique en *machine learning*, les biais sont souvent liés à la *fairness* et traités selon les préjudices sociaux (*social harms*) qu'ils induisent. On en distingue deux types : les préjudices d'allocation (*allocative harms*), qui se manifestent lorsqu'un système octroie des ressources et des opportunités de façon inégale à différentes populations, et les préjudices de représentation (*representational harms*) que sont les stéréotypes perpétués à l'encontre de certaines populations et les variations de performance des systèmes selon ces mêmes populations [3, 37].

Au regard de cette taxonomie, les LLMs sont un cas d'étude intéressant lorsqu'il s'agit d'observer des préjudices de représentation. En effet, beaucoup de travaux ont été menés pour étudier et mesurer les préjudices d'allocations engendrés par des systèmes de classification utilisés à grande échelle [21] et la plupart des métriques traditionnelles de *fairness* sont des métriques d'évaluation de classification [3, 30, 39] car l'évaluation de ce genre de systèmes repose sur des valeurs discrètes et facilement quantifiables.

Les modèles de langue eux, par définition, sont des modèles statistiques dont le but est de représenter le langage naturel et ils sont utilisés pour résoudre différentes tâches (*downstream tasks*). On distingue trois types de modèle

basé sur l'architecture du Transformers : les encoders-decoders (BART-type), les autoencoders (BERT-type) et les autorégressifs (GPT-type), les modèles autorégressifs étant ceux dont les performances sont les meilleurs dans l'état de l'art ; cette prépondérance s'explique par leur taille et leur capacité à substituer toute tâche à une tâche unique de complétion/génération de texte.

Les modèles autorégressifs occupent une place particulière dans le paysage de la littérature de la *fairness*, d'abord parce que l'architecture des Transformers permet d'envisager les biais selon une classification nouvelle : on considère alors les biais selon qu'ils sont intrinsèques ou extrinsèques [18, 7]. Ainsi, un biais intrinsèque correspond au biais contenu dans la façon dont le LLM modèle encode ses représentations quand un biais extrinsèque reflète un biais dans les décisions du modèle au moment de la prédiction.

Mais cette place particulière s'explique aussi par le fait que les LLMs sont utilisés de façon générique («*General-purpose* » [28]), et donc on peut faire l'hypothèse d'un transfert de leurs biais inhérents dans leurs différentes applications spécifique. En effet, on observe entre le pré-entraînement d'un LLM et son déploiement une longue chaîne de *fine-tuning* successifs, d'adaptation du modèle, le long de laquelle les biais peuvent se multiplier et se renforcer, faisant de leur évaluation et leur mitigation une tâche plus complexe encore [19].

Ainsi, concernant leur **évaluation**, la classification intrinsèque/extrinsèque permet d'envisager l'évaluation des biais dans les LLMs de différentes manières. Comme évoqué plus tôt, ils peuvent être observés dans les représentations du LLMs (intrinsèque) et cela de deux façons différentes : en observant les proximités des représentations dans l'espace scalaire [22, 8] ou les différences dans la distribution de probabilité des tokens sur des tâches de complétion de texte (Li et al., 2023). En ce qui concerne les biais extrinsèques, on évalue les modèles selon différentes populations sur des tâches définies, avec les métriques associées. Il peut s'agir d'une tâche de complétion de texte, tâche « essentielle » des LLMs, qui s'évalue généralement avec des mesures de toxicité, de sentiment et de regard [11] ou encore avec des benchmarks qui s'appuient sur des jeux de données [23] ; mais il peut également s'agir d'une tâche « substituée » plus classique, de classification [31] ou de résolution de coréférence par exemple [44].

Dans un premier temps, le fait de pouvoir utiliser les LLMs pour des tâches prédictives pour lesquelles ceux-là n'ont pas été pré-entrainés (la classification) soulève les mêmes questions d'évaluation de *fairness* que celles qui traversent la littérature depuis plusieurs années déjà : sur quel critère statistique considérons-nous que nous sommes équitables ? C'est l'enjeu de l'incompatibilité de certaines métriques de *fairness* selon la définition choisie, qui ne s'appuient pas sur la même condition fondamentale d'indépendance statistique [30, 3].

Deuxièmement, cet éventail de possibilités d'applications nécessite également un éventail de méthodes d'évaluation. Gallienne et Poibeau (2023) [12] évoquent la notion de « classes d'application » pour mettre en lumière la nécessité d'avoir une méthode d'évaluation de *fairness* spécifique au cas d'usage précis pour lequel un LLM donné sera utilisé. On parle d'un biais de mesure lorsque la mesure utilisée pour évaluer une construction abstraite ne reflète pas complètement la réalité [27, 36] et à ce titre, on peut évoquer les études qui démontrent que les métriques d'évaluation de biais intrinsèques ne permettent pas nécessairement d'avoir une compréhension au moment de la prédiction du modèle [13].

Enfin, les questions que posent l'évaluation des productions des modèles autorégressifs restent ouvertes [29] d'un point de vue général étant donné que l'évaluation standard d'un LLM repose sur ses performances sur des benchmarks données, qui peuvent elles-mêmes être biaisées ou incomplètes [24, 5] et l'évaluation de la *fairness* d'un modèle avec des benchmarks n'échappe pas à cette règle [6]. C'est en partie ce que souligne Davat lorsqu'elle préconise de parler de « re-biasage » plutôt que de « débiaisage ». En effet, les LLMs encodent un point de vue dominant qui n'est pas exempt de biais [4, 14, 35], point de vue qui peut être modifié à l'aide de *fine-tuning* entre autres [17, 32].

En outre, on compte parmi ces questions, l'enjeu de l'intégration d'une évaluation de *fairness* aux évaluations de performance, pour avoir un benchmark d'évaluation plus complète [18, 11]. C'est la proposition notamment dans le benchmark Stereoset de la métrique de *l'idealized CAT score* qui allie un score de performance sur la tâche de *language modeling* et un score de *fairness* de réponse stéréotypée [25].

Au même titre que les méthodes d'évaluation, on peut observer une classification des **méthodes de mitigation** selon le lieu de manifestation du biais : de façon intrinsèque ou extrinsèque. Dans l'utilisation majoritaire des LLMs aujourd'hui, ces derniers sont toutefois considérés comme des « *black boxs*», récupérées après un pré-entraînement général, avec ou sans couches d'entraînement ajoutées en amont (*fine-tuning*), rendant ainsi certaines techniques de mitigation intrinsèques caduques (*pre-processing/in-processing*), au moins en ce qui concerne le pré-entraînement. Pour autant, il existe différentes techniques de mitigation applicables sur des modèles pré-entraînés. On peut les considérer selon la nécessité ou non de passer par des étapes d'entraînement supplémentaires.

Ainsi, il est possible de fine-tuner un LLM modèle avec des nouvelles données qu'on considérera « non-biaisées ». Ce *fine-tuning* peut être intégrale ou modulaire (*memory-efficient/parameter fine-tuning*), cette dernière option étant préférablement choisie pour des raisons de coût et de conservation des performances du modèle -pour éviter le catastrophe forgetting en d'autres termes. On compte parmi les méthodes les plus usitées, les méthodes de prompt-tuning, de prefix-tuning, d'adapteurs et de reparamétrisassions [38, 43]. Dans certains cas même, on peut apparenter les méthodes de *parameter-efficient fine-tuning* à du *machine unlearning* [9].

A cette étape d'entraînement supervisée peut s'ajouter une étape d'apprentissage par renforcement [26] pour introduire des préférences de réponses alignées avec les valeurs humaines [42]. Au cours de cette étape, il est possible de fine-tuner le modèle avec des données annotées par des humains (*Reinforcement Learning from Human Feedback*) ou par le LLM lui-même qui a recours à un « auto-diagnostique » de ses propres contenus générés grâce à l'aide de principes constitutifs (*Reinforcement Learning from AI Feedback*) [2]. A noter qu'au-delà de l'apprentissage par renforcement, cette capacité d'auto-diagnostique peut également être utilisée pour agir sur la distribution de probabilités après génération de texte en elle-même [20, 33] mais que cette technique de mitigation ne requiert pas de nouvel entraînement et opère plutôt une méthode de censure du modèle, plutôt que de re-biaisage.

Finalement et dans la même veine que cette idée de censure par l'auto-diagnostique, le «*role-play* » [34] peut être envisagé comme une façon de contraindre la vision du monde que le modèle adopte et d'orienter ses biais. Le prompting en lui-même « révèle les forces et les biais des LLMs » [31]. En effet, on observe dans cette étude que le modèle performe mieux une tâche lorsqu'on demande au système de prendre le point de vue d'un expert mais également qu'il existe des variations de performance lorsqu'il s'agit de prendre le point de vue de différents groupes sociaux. Dans le même registre, d'autres études montrent également que l'intensité du préjudice de représentation varie selon le rôle endossé par le modèle [41, 16]. Cependant, on note avant tout que ces expériences de *role-play* révèlent d'abord les biais inhérents des modèles et que selon ces travaux et ceux de Hubinger et al., (2024) [15], les biais observés au moment de la prédiction notamment peuvent avoir des origines diffuses, difficiles à situer et donc à mitiger.

En synthèse, plutôt que de les envisager de façon générique, à l'échelle du modèle, l'évaluation et la mitigation des biais doivent être traitées selon un cas d'usage précis et son contexte. C'est ce que les entreprises devront déployer [1].

Références

[1] IA non biaisée : les entreprises sont-elles prêtes ? - Hello Future Orange, October 2021.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI : Harmlessness from AI Feedback, December 2022. arXiv :2212.08073 [cs].

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery.

[5] Guillaume Le Berre. *Vers la mitigation des biais en traitement neuronal des langues*. phdthesis, Université de Lorraine ; Université de Montréal, June 2023.

[6] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian Salmon : An Inventory of Pitfalls in Fairness Benchmark Datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. arXiv :2108.07258 [cs].

[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334) :183–186, April 2017.

[9] Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast Model Debias with Machine Unlearning, November 2023. arXiv :2310.12560 [cs].

[10] Ambre Davat. What is the meaning of "bias" in AI ?

[11] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models : A Survey, September 2023. arXiv :2309.00770 [cs].

[12] Romane Gallienne and Thierry Poibeau. Quelques observations sur la notion de biais dans les modèles de langue. In *18e Conférence en Recherche d'Information et Applications\backslash\backslash16e Rencontres Jeunes Chercheurs en RI\backslash\backslash30e Conférence sur le Traitement Automatique des Langues Naturelles\backslash\backslash25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 1–13. ATALA, 2023.

[13] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias, June 2021. arXiv :2012.15859 [cs].

[14] Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose Language Counts as High Quality ? Measuring Language Ideologies in Text Data Selection, January 2022. arXiv :2201.10474 [cs].

[15] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper Agents : Training Deceptive LLMs that Persist Through Safety Training, January 2024. arXiv :2401.05566 [cs].

[16] Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large Language Models as Superpositions of Cultural Perspectives, November 2023. arXiv :2307.07870 [cs].

[17] Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing Biases and the Impact of Multilingual Training across Multiple Languages, May 2023. arXiv :2305.11242 [cs].

[18] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A Survey on Fairness in Large Language Models, August 2023. arXiv :2308.10149 [cs].

[19] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs : a Survey and Guideline for Evaluating Large Language Models' Alignment, August 2023. arXiv :2308.05374 [cs].

[20] Ambri Ma, Arnav Kumar, and Brett Zeligson. Diagnosing and Debiasing Corpus-Based Political Bias and Insults in GPT2, November 2023. arXiv :2311.10266 [cs].

[21] Lauren Kirchner Surya Jeff Larson Mattu, Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm.

[22] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders, March 2019. arXiv :1903.10561 [cs].

[23] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models, April 2022. arXiv :2110.08527 [cs].

[24] Ramaravind Kommiya Mothilal, Shion Guha, and Syed Ishtiaque Ahmed. Towards a Non-Ideal Methodological Framework for Responsible ML, January 2024. arXiv :2401.11131 [cs].

[25] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet : Measuring stereotypical bias in pretrained language models, April 2020. arXiv :2004.09456 [cs].

[26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. arXiv :2203.02155 [cs].

[27] Samir Passi and Solon Barocas. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 39–48, New York, NY, USA, January 2019. Association for Computing Machinery.

[28] Oliver Radley-Gardner, Hugh Beale, and Reinhard Zimmermann, editors. *Fundamental Texts On European Private Law*. Hart Publishing, 2016.

[29] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy : Behavioral Testing of NLP Models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[30] Boris Ruf and Marcin Detyniecki. Towards the Right Kind of Fairness in AI, September 2021. arXiv :2102.08453 [cs].

[31] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-Context Impersonation Reveals Large Language Models' Strengths and Biases, November 2023. arXiv :2305.14930 [cs].

[32] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect ?, March 2023. arXiv :2303.17548 [cs].

[33] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-Diagnosis and Self-Debiasing : A Proposal for Reducing Corpus-Based Bias in NLP, September 2021. arXiv :2103.00453 [cs].

[34] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-Play with Large Language Models, May 2023. arXiv :2305.16367 [cs].

[35] Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. Aligning with Whom ? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks, November 2023. arXiv :2311.09730 [cs].

[36] Harini Suresh and John V. Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, October 2021. arXiv :1901.10002 [cs, stat].

[37] The Artificial Intelligence Channel. The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017, December 2017.

[38] Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. Efficient Methods for Natural Language Processing : A Survey, March 2023. arXiv :2209.00099 [cs].

[39] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden, May 2018. ACM.

[40] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "Kelly is a Warm Person, Joseph is a Role Model" : Gender Biases in LLM-Generated Reference Letters, December 2023. arXiv :2310.09219 [cs].

[41] Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. Are Personalized Stochastic Parrots More Dangerous ? Evaluating Persona Biases in Dialogue Systems, November 2023. arXiv :2310.05280 [cs].

[42] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. From Instructions to Intrinsic Human Values – A Survey of Alignment Goals for Big Models, September 2023. arXiv :2308.12014 [cs].

[43] Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing Parameter-Efficient Modules with Arithmetic Operations, December 2023. arXiv :2306.14870 [cs].

[44] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution : Evaluation and Debiasing Methods. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

WHAT CHATGPT TELLS US ABOUT OURSELVES

Aaron Boussidan[1], **Fanny Ducel**[2], **Aurélie Névéol**[2], **Karën Fort**[3]
(1) Université Gustave Eiffel, LIGM (France)
(2) Université Paris-Saclay, CNRS, LISN (France)
(3) Sorbonne Université / LORIA (France)
aaron.boussidan@univ-eiffel.fr, fanny.ducel@lisn.fr

March 29, 2024

1 Introduction

Large Language Models are known to generate various biases, and the scientific community has started to tackle them. There are two main approaches to detect and evaluate stereotypical biases: intrisically or extrinsically. Intrisic methods aim at uncovering internal biases, located within the model's representations (Bolukbasi *et al.*, 2016; Guo & Caliskan, 2021), whereas extrinsic methods focus on biases that are produced during inference, i.e. the output of a use-case (De-Arteaga *et al.*, 2019; Parrish *et al.*, 2022). Since our goal is to work on ChatGPT, we are compelled to opt for the second approach as OpenAI did not release enough information on the training corpora or the internal model's representations. More precisely, this experiment is an attempt at measuring ChatGPT's biases about various countries in the world. With the help of prompting, generative models can easily be used for other classical NLP tasks. In this experiment in particular, we prompted ChatGPT 3.5 to perform sentiment analysis, by getting it to assign a "positivity score" to every country of the world. We then generated a choropleth (a map where countries are each assigned a color associated to their score) with ChatGPT's rating of each country. The experiment has been conducted in 4 different languages: French, English, Russian and Arabic, to assess potential variations between languages. We briefly present our methodology, and then detail our results, their interpretation, and we conclude by discussing the design of the experiment and its inherent biases.

2 Methodology

We used the publicly available version of ChatGPT, using the web interface, between February 21, 2024 and March 6, 2024. For each language, we used the same prompts. Translations into languages other than English were generated using DeepL, and checked for fluency and correctness by native or bilingual speakers. We queried ChatGPT to assign a score between -10 (very negative) and +10 (very positive) to different words:

> Hi ! I will give you words, and I would like you to assign scores between -10 and 10 to those words, based on how positive they are

We then provide it with a list of countries[1], asking it to assign a score to each of them:

> Please score the following words, and give your results in a .csv file: ['Afghanistan', 'Albania', ...'Zambia', 'Zimbabwe','eSwatini']

We then generated maps where each country was assigned a color on a red (most negative) to green (most positive) scale. All of our code, as well as our prompts will be made available upon acceptation.

3 Results and first interpretations

3.1 A Western perspective?

Looking at the results and data for the four languages (See Figures 1 to 4), a clear global trend appears: North American and European countries are given higher scores, while African countries are given the worst ones. South American and Asian countries tend to be in the medium range. In Oceania, Australia and New-Zealand are evaluated more positively than their neighbours. Some countries are notable outliers: North Korea, Venezuela, Belarus, and Myanmar, for example.

Comparing our generated maps with GDP [2] or Internet access maps for the world shows some notable similarities. This suggests that the scores obtained seem to reflect the opinions of the users and writers of the Web, who positively perceive themselves and their fellow users. Indeed, we know that Western countries have better access to the Internet[3]

[1] We used the list of countries provided by the Python library *Geopandas*.
[2] https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD
[3] https://ourworldindata.org/grapher/share-of-individuals-using-the-internet?time=2021&tab=map

and that, at least for Wikipedia, the majority of users who express themselves online are Western men who are either in their mid-20s or retired[4]. We can easily see similar trends between these aforementioned statistics and the results of our experiment: ChatGPT reproduces, and maybe amplifies, what people write on the Internet.

The model outputs often mention these scores as reflecting a "general opinion". However, no particular way of scoring countries was integrated to our initial prompt: we only asked how "positive" some "words" are. When asked to score countries individually, the model usually explains how these scores are linked to the "economic level" of the country, its "reputation" or some of its policies: these criteria of scoring have been purely inferred by the model. For example, when prompted to elaborate about the score for Uruguay, the model produces the following explanation:

> Uruguay (Score: 5): Uruguay is perceived positively for its stable democracy, progressive social policies (including legalizing cannabis), strong rule of law, environmental sustainability efforts, and relatively high quality of life in Latin America.

Among other things, framing cannabis legalisation as a positive factor to evaluate a country is a political choice. When prompted about the United Arab Emirates, the model produces a list of qualities of the countries, which does not include social policies, that had been factored in previously.

> United Arab Emirates (Score: 7): The UAE is viewed positively due to its rapid development, modern infrastructure, economic diversification beyond oil, tourism initiatives, and ambitious projects such as the Burj Khalifa and Palm Islands.

These outputs also seem to be outdated, as across all languages, Russia is given a positive score, whereas the context of the Russia-Ukraine war is known to have negatively impacted people's view of the country[5]. The training data of ChatGPT 3.5 allegedly stops in January 2022, a month before Russia's invasion of Ukraine and the beginning of the war. This reminds us that the model reproduces a snapshot of the world's perception of other countries, in a given time and space.

3.2 Variation between languages

We can also observe some differences across languages. Prompts in French lead to a more negative score for African countries, especially the ones that France colonized[6]. On the contrary, when prompted in English, the model gives positive scores to the Commonwealth nations (e.g. India, Australia). For Arabic and Russian, the results are quite similar and more homogeneous, with a more positive overall average. The rare countries that have negative scores are also in Africa and, especially for the Russian version, in the Middle East.

We can propose different hypotheses to explain the changes in results based on the language used in the prompt. On one hand, it could be related to the cultural differences that are encapsulated in the training corpora. On the second hand, it could be the reflection of the generations' linguistic quality and of the use of toxicity filters and Reinforcement Learning with Human Feedback (RLHF), that differ based on the language.

This experiment is, of course, not a systematic large-scale approach at those biases, but rather a first approach, supported by a visualization method. In particular, the model is non-deterministic in its core: country scores vary from conversation to conversation, and even the average score among all countries varies from conversation to conversation. We have tried to counter-act this by normalising the data, to make it comparable, and by trying to have a larger sample size for English, but a more large scale approach should be taken to further pursue this experiment.

Finally, it might be misguided to take these results as showing that ChatGPT " has an opinion" on countries, or "favors" some over others. When asked to explain its scores, the model gives explanation for why it has chosen its score. A positivity score is fundamentally 1-dimensional, and by forcing the model to choose, we are somewhat forcing discrepancies between countries.

3.3 Variation between iterations

Since ChatGPT is non-deterministic, we replicated the experiment on English 15 times, on different user accounts, using different conversation tabs and shuffling the order of the countries in our prompt. Our goal was to observe whether or not scores change drastically across iterations. We then computed the averages, medians, and standard deviations for each country. Our results show that the majority of countries obtain very similar scores across iterations. Nonetheless, it is not the case for 30 countries (See Table A), which present significant standard deviation (>2, when the mean standard deviation is 1.35), e.g. notable score differences across trials. The concerned countries are mostly

[4] See the example of Wikipedia editors: https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia%3F

[5] https://www.pewresearch.org/global/2023/07/10/overall-opinion-of-russia/

[6] https://fr.wikipedia.org/wiki/Afrique_fran%C3%A7aise

situated in Africa (e.g. Sierra Leone, Zimbabwe), or, to a lesser extent, in South/Southeast Asia (e.g. Bangladesh, Myanmar). The rest are situated in Eastern-European (Ukraine, Belarus and Russia), the Near East (Lebanon, Saudi Arabia) or South America (Venezuela, Cuba). We can imagine that the notable differences of these 30 countries are related to a less important quantity of data about them, or to very polarized opinions in corpora.

4 A flawed and biased experiment

The presented experiment and results are actually fundamentally flawed, as they encapsulate our (the authors') own biases. As presented by Hovy & Prabhumoye (2021), there are five sources of bias in NLP, including the research design.

In our case, some biases and questionable choices can be found at different levels. First of all, the goal and choice of the experiment itself can raise questions such as: *What are we truly asking the model? What are we trying to measure? What does the provided score mean?* This last question is especially important, as the used prompts only ask for a "score" on a "word", and we decided to interpret it as the perceived positivity of a country.

We can also raise questions about the provided (over)interpretation of the outputs. ChatGPT is not transparent about its training, so we can not know where the outputs come from and we can not give definite answers. All the explanations are purely hypothetical, relying on the authors' internal biases and opinions, and could easily be twisted: we could make the results, hence ChatGPT, say anything. This can be used as a reminder that language models do not understand nor produce meaning, but the readers do (Bender & Koller, 2020).

Finally, the choices of implementation that were made (or were not made, when letting default parameters) are also important. The chosen scores scale can be questioned (-10 to 10), as well as the colors of the maps. These colors, red and green, have connotations and invite readers to interpret green as positive and red as negative. Besides, they reduce the accessibility for color-blind people. Moreover, the maps' design, especially their orientation (Europe in the center, North countries on top, South countries on the bottom), and their projection (Mercator) also represent some political and cultural values (Graham & Dittus, 2022).

In other words, every step of this experiment implied its load of decisions. Some decisions may seem shallow at first sight, whereas they have important consequences and implications. We can hypothesize that different design choices would have led to different results (even more as ChatGPT is non-deterministic and outputs vary between iterations and between small prompt changes), hence different interpretations and an overall different paper.

References

Bender E. M. & Koller A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5185–5198, Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.463.

Bolukbasi T., Chang K.-W., Zou J. Y., Saligrama V. & Kalai A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, **29**.

De-Arteaga M., Romanov A., Wallach H., Chayes J., Borgs C., Chouldechova A., Geyik S., Kenthapadi K. & Kalai A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 120–128. arXiv:1901.09451 [cs, stat], doi:10.1145/3287560.3287572.

Graham M. & Dittus M. (2022). *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press.

Guo W. & Caliskan A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, p. 122–133, New York, États-Unis: Association for Computing Machinery. doi:10.1145/3461702.3462536.

Hovy D. & Prabhumoye S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, **15**(8), e12432. _eprint: https://online.library.wiley.com/doi/pdf/10.1111/lnc3.12432, doi:10.1111/lnc3.12432.

Parrish A., Chen A., Nangia N., Padmakumar V., Phang J., Thompson J., Htut P. M. & Bowman S. (2022). BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, p. 2086–2105, Dublin, Ireland: Association for Computational Linguistics. doi:10.18653/v1/2022.findings-acl.165.

A Maps for each language and English scores

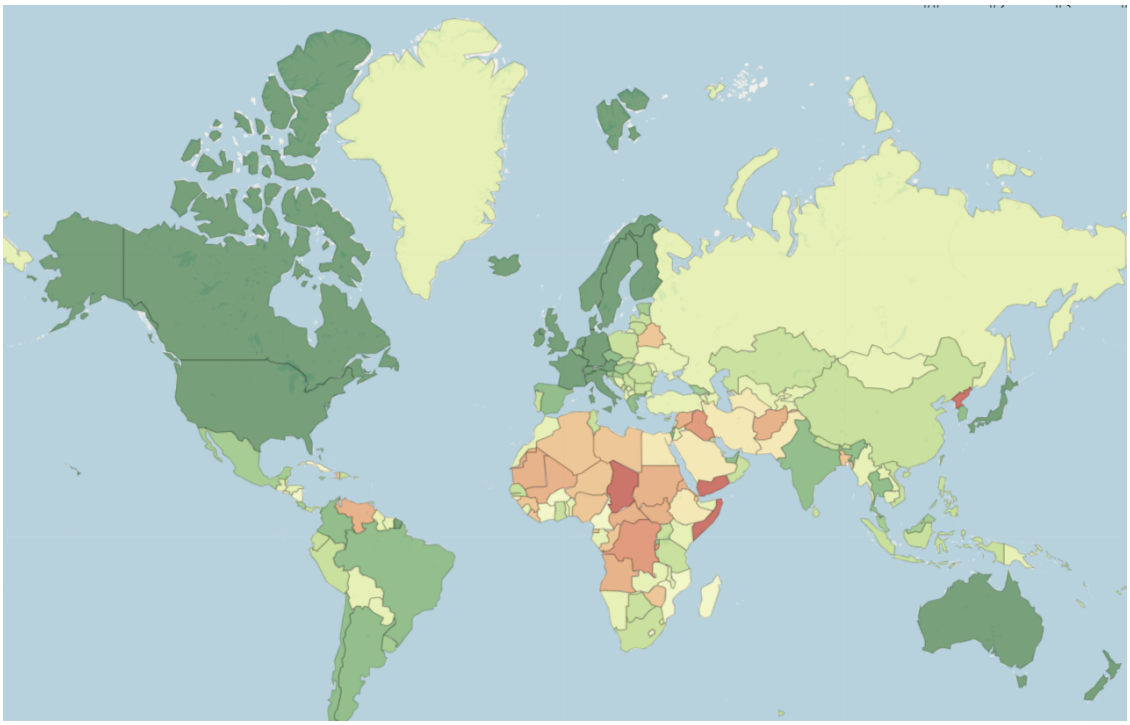

Figure 1: Map of countries score, English version, initial version

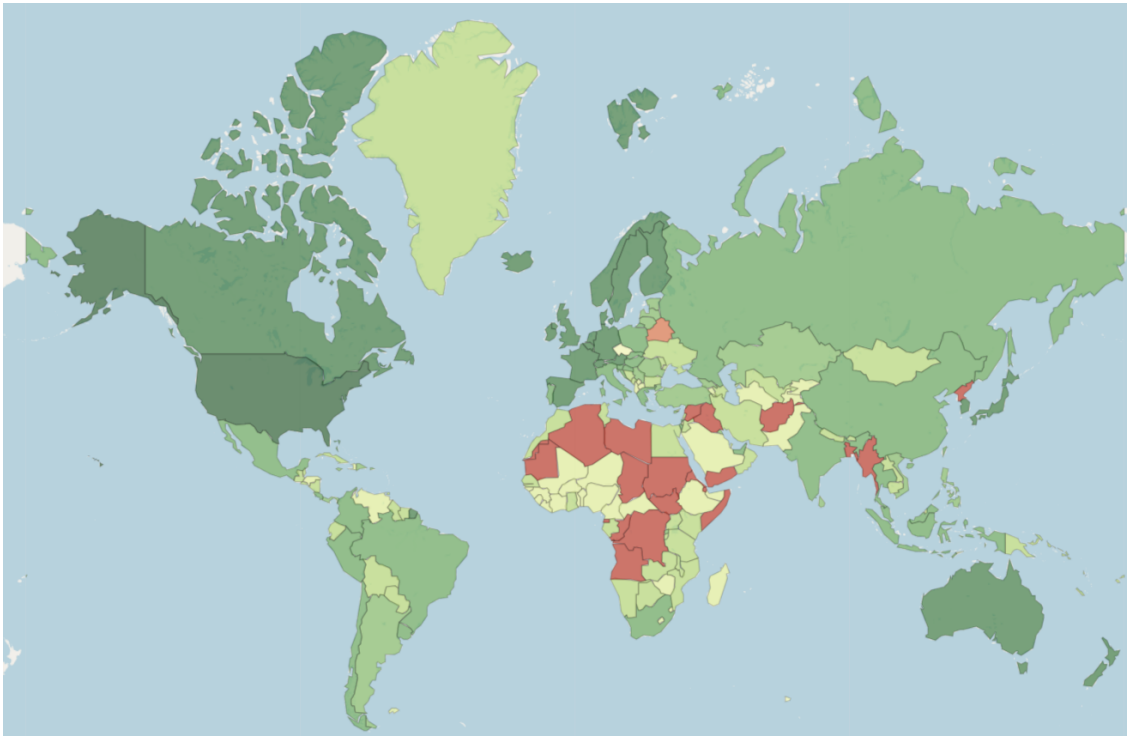

Figure 2: Map of countries score, French version

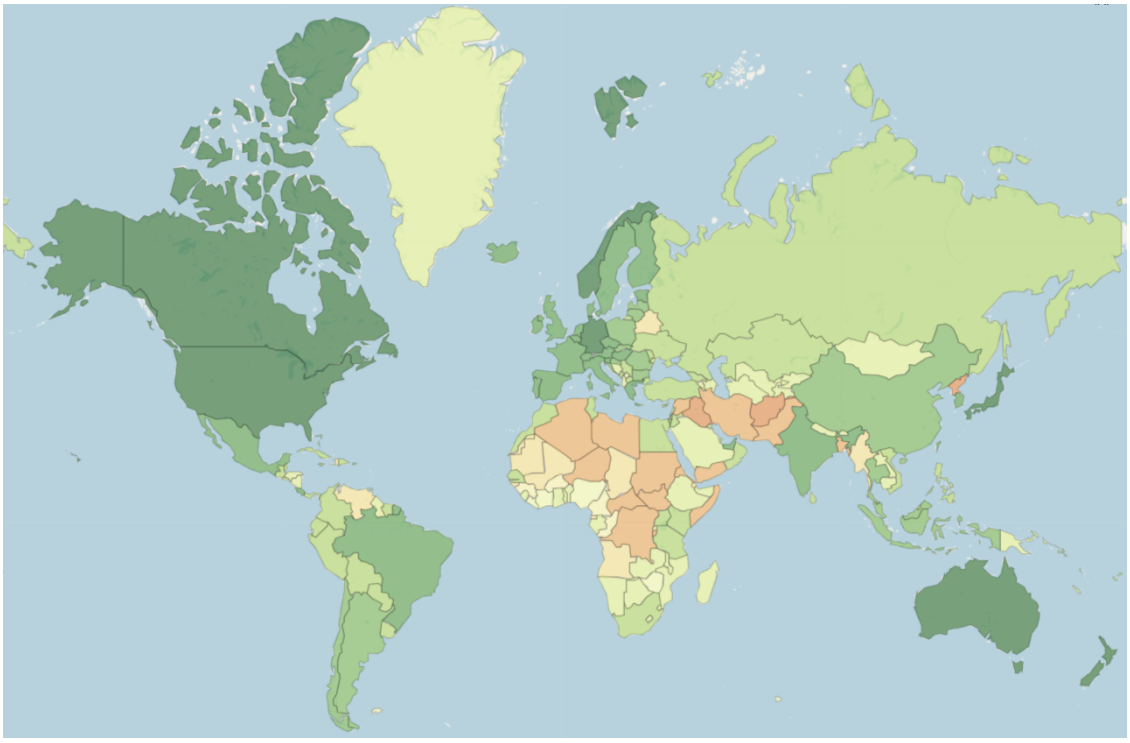

Figure 3: Map of countries score, Russian version

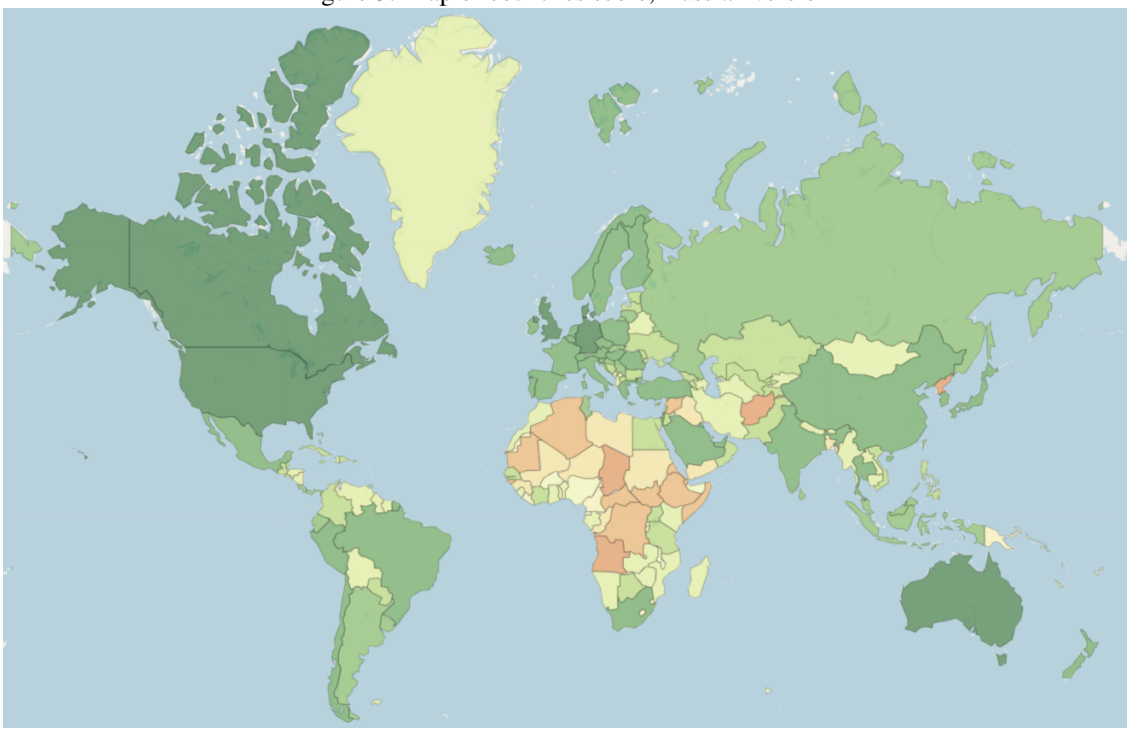

Figure 4: Map of countries score, Arabic version

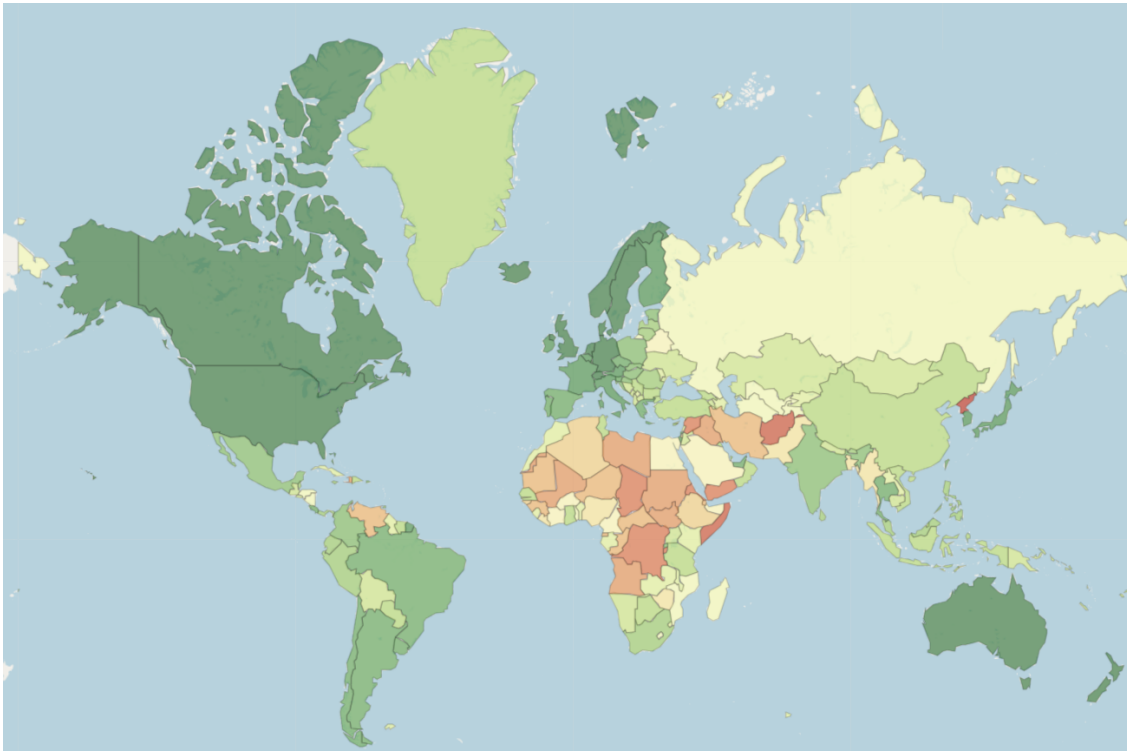

Figure 5: Map of countries score, English version, average score of 15 tries

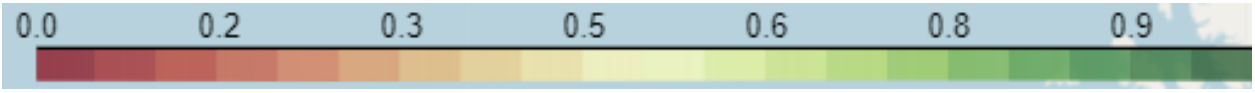

Figure 6: Scale reflecting the scores, normalized between 0 and 1. It is the same for all maps.

Word	Avg Score	Std dev
Afghanistan	-6.40	1.24
Albania	2.40	0.83
Algeria	-2.13	2.10
Angola	-4.60	1.35
Antarctica	1.60	1.80
Argentina	6.20	1.47
Armenia	2.40	0.99
Australia	8.40	0.91
Austria	7.67	0.49
Azerbaijan	1.47	0.99
Bahamas	6.27	1.10
Bangladesh	-1.87	2.56
Belarus	-0.80	2.68
Belgium	7.07	0.80
Belize	4.40	0.91
Benin	1.67	1.35
Bhutan	4.20	1.21
Bolivia	2.53	1.13
Bosnia and Herz.	2.67	1.35
Botswana	3.60	1.30
Brazil	6.60	1.18
Brunei	4.00	1.13
Bulgaria	3.73	1.33
Burkina Faso	-0.07	2.15
Burundi	-5.20	1.08
Cambodia	2.20	1.15
Cameroon	-0.20	1.47
Canada	8.60	0.51
Central African Rep.	-3.67	2.09
Chad	-5.40	1.35
Chile	6.27	0.80
China	3.20	1.78
Colombia	5.33	1.23
Congo	-3.13	1.13
Costa Rica	6.20	0.68
Côte d'Ivoire	-0.27	1.71
Croatia	5.40	0.74
Cuba	1.27	2.02
Cyprus	4.47	0.83
Czechia	6.40	0.74
Dem. Rep. Congo	-5.53	1.06
Denmark	8.13	0.35
Djibouti	-3.33	0.90
Dominican Rep.	3.53	0.99
Ecuador	4.33	1.11

Word	Avg Score	Std dev
Egypt	-0.27	2.66
El Salvador	2.13	0.92
Eq. Guinea	-3.73	1.44
Eritrea	-5.07	1.33
Estonia	5.13	0.74
eSwatini	1.87	1.85
Ethiopia	-2.40	1.92
Falkland Is.	1.47	1.19
Fiji	3.80	0.77
Finland	7.93	0.26
France	7.80	0.41
Gabon	1.27	1.67
Gambia	1.47	1.06
Georgia	3.47	1.19
Germany	8.60	0.51
Ghana	3.27	1.16
Greece	6.20	0.77
Greenland	3.13	1.41
Guatemala	2.87	0.99
Guinea	-3.20	1.66
Guinea-Bissau	-2.80	1.61
Guyana	1.93	1.03
Haiti	-4.47	0.83
Honduras	0.73	1.22
Hungary	5.67	0.82
Iceland	8.20	0.41
India	5.60	1.06
Indonesia	4.00	1.00
Iran	-3.33	1.50
Iraq	-4.80	0.77
Ireland	7.53	0.52
Israel	6.07	0.59
Italy	7.33	0.49
Jamaica	4.93	0.88
Japan	7.60	0.63
Jordan	2.27	1.71
Kazakhstan	2.67	1.40
Kenya	1.93	1.16
Kosovo	1.13	1.13
Kuwait	3.73	1.53
Kyrgyzstan	1.60	0.91
Laos	1.53	1.19
Latvia	4.40	0.99
Lebanon	1.13	2.07
Lesotho	0.27	0.80

Word	Avg Score	Std dev
Liberia	-1.80	2.11
Libya	-4.40	1.12
Lithuania	4.67	0.90
Luxembourg	7.00	0.53
Madagascar	0.87	1.51
Malawi	1.00	0.85
Malaysia	4.87	0.92
Mali	-4.13	1.30
Mauritania	-3.73	1.03
Mexico	5.00	0.93
Moldova	1.80	1.08
Mongolia	2.20	1.01
Montenegro	2.53	1.06
Morocco	1.93	1.58
Mozambique	0.33	1.63
Myanmar	-1.20	2.54
N. Cyprus	1.40	1.68
Namibia	2.13	0.74
Nepal	2.53	0.92
Netherlands	8.13	0.35
New Caledonia	2.13	1.13
New Zealand	8.53	0.52
Nicaragua	0.07	1.16
Niger	-3.80	1.61
Nigeria	-1.93	1.62
North Korea	-7.53	1.25
North Macedonia	2.47	0.92
Norway	8.20	0.41
Oman	3.40	1.18
Pakistan	-1.53	2.45
Palestine	-1.67	1.99
Panama	4.93	1.44
Papua New Guinea	2.13	1.60
Paraguay	3.27	1.28
Peru	4.60	1.12
Philippines	4.53	1.30
Poland	5.87	1.06
Portugal	6.47	0.92
Puerto Rico	5.20	0.86
Qatar	4.67	1.91
Romania	4.87	1.19
Russia	0.40	2.47
Rwanda	5.13	1.92

Word	Avg Score	Std dev
S. Sudan	-4.93	1.91
Saudi Arabia	-0.67	2.47
Senegal	3.73	2.12
Serbia	3.67	1.29
Sierra Leone	1.13	3.70
Slovakia	5.60	0.74
Slovenia	6.47	0.99
Solomon Is.	2.73	1.58
Somalia	-6.67	1.29
Somaliland	-0.80	3.08
South Africa	4.33	2.38
South Korea	7.47	0.64
S. Antarc. Lands (FR)	1.60	1.80
Spain	6.93	0.46
Sri Lanka	4.53	1.36
Sudan	-4.07	1.28
Suriname	3.13	1.30
Sweden	8.13	0.52
Switzerland	8.53	0.52
Syria	-5.33	1.11
Taiwan	6.13	0.92
Tajikistan	0.33	2.02
Tanzania	3.00	1.93
Thailand	6.00	1.00
Timor-Leste	1.47	2.50
Togo	0.80	2.98
Trinidad and Tobago	4.60	1.59
Tunisia	3.33	2.19
Turkey	3.40	1.92
Turkmenistan	0.00	2.04
Uganda	2.80	2.34
Ukraine	2.13	3.20
United Arab Emirates	6.67	0.90
United Kingdom	8.13	0.35
U.S.A.	8.33	0.82
Uruguay	6.00	1.51
Uzbekistan	0.87	1.96
Vanuatu	2.60	2.87
Venezuela	-3.20	2.96
Vietnam	4.00	2.30
W. Sahara	-4.73	1.22
Yemen	-5.47	2.13
Zambia	1.93	2.79
Zimbabwe	-1.60	3.18

MESURER LES RISQUES DE DISCRIMINATION DANS UNE TÂCHE DE DÉTECTION D'ENTITÉS NOMMÉES

Hugues de Mazancourt
CTO, Datapolitics
hugues@datapolitics.fr

Flavie Loshouarn
Lead developper, Datapolitics
flavie@datapolitics.fr

Alice Bruguier
NLP engineer, Datapolitics
alice@datapolitics.fr

29 mars 2024

ABSTRACT

La reconnaissance d'entités nommées (NER) est une tâche classique en NLP, parfois sujette à des biais liés au corpus d'apprentissage [4], [3] complexes à mesurer [1], [2]. Nous présentons ici l'adaptation d'une méthode de détection de biais et son évaluation sur deux tâches de reconnaissance de noms de personnes dans les chaînes d'analyse de textes politiques.

1 Introduction

Créée en 2021, Datapolitics est une start-up ayant pour but de produire un système d'analyse de données intelligent lié à la vie démocratique française. Elle fournit à ses clients un outil de veille et d'analyse. L'un des niveaux d'analyse fourni par le logiciel consiste à annoter les prises de position par le nom du locuteur. Les prises de position considérées sont soit des prises de parole directes (dans des procès-verbaux de délibérations des mairies) ou des paroles rapportées dans des articles de presse. D'une manière générale, dans l'univers Datapolitics, le locuteur est celui qui exprime une opinion dans l'espace public. Une identification correcte est importante pour suivre l'évolution de la position d'un homme ou d'une femme politique et d'être au plus près de ses opinions. Elle permet aussi de pondérer les opinions détectées en fonction du pouvoir de décision de celui qui émet l'opinion.

Datapolitics collecte ces différents types de textes en continu pour produire une veille quotidienne.

2 Motivation

Datapolitics a été fondé dans un objectif d'aide à la démocratie. Son but est de rendre la vie politique française accessible à tous, permettant une meilleure représentativité et une meilleure capacité d'expression des causes et des personnes. L'invisibilisation de groupe d'appartenance ethnique est un problème social contre lequel Datapolitics veut lutter. Il nous a donc semblé important d'évaluer nos propres modèles de langue sous cet angle et de nous assurer de l'absence de biais discriminatoire dans notre traitement des données. Il s'agit avant tous de pouvoir certifier que Datapolitics n'introduit ni n'aggrave des inégalités sociales.

Dans cette étude, nous évaluerons l'importance de l'origine du nom dans la qualité de sa détection. Est-ce que nous détectons mieux des noms de personnes françaises que de ressortissants d'autres nationalités ? Est-ce que le corpus d'apprentissage, quasiment-exclusivement issus de personnes de nationalité française, pourrait avoir un impact sur une éventuelle application à des acteurs étrangers ?

3 Contexte de l'étude

Notre détection de locuteur s'effectue sur deux contextes différents, conduisant à deux tâches de difficultés distinctes en termes de NLP. La première consiste à détecter les locuteurs dans un article de presse (*qui exprime l'opinion ?*) et la seconde à détecter les auteurs de prise de parole dans des procès verbaux de communautés territoriales (mairies, conseils d'arrondissement, etc.). Pour chacune de ces deux tâches (nommées **presse** et **mairie**, resp.), nous avons constitué des corpus manuellement annotés (resp. 600 et 800 paragraphes) pour identifier le locuteur. L'apprentissage s'est ensuite effectué avec l'outil SpaCy, en affinant le modèle de reconnaissance de noms de personne.

Il faut noter que ces corpus, exclusivement en français, sont par nature relativement réguliers : ils mentionnent les noms de personne essentiellement sous la forme *Prénom Nom*, *Nom Prénom* ou *M/Mme Nom*, ce qui facilite la tâche d'évaluation, comme on le verra.

4 Méthode d'évaluation

Nous avons constitué des corpus synthétiques à partir de données non-issues de l'apprentissage. Nous avons ensuite remplacé les noms identifiés manuellement comme *LOCUTEUR* dans notre corpus de test par d'autres noms de

différentes origines afin de valider l'impact sur la reconnaissance automatique. Il s'agit d'une extension de la méthode présentée en [5] qui se limitait aux prénoms.

4.1 Sélection des noms

Nous avons choisi les pays d'origine en fonction de l'importance de leur immigration vers la France. Nos choix ont porté sur :

- En Afrique du nord, le Maroc, l'Algérie, la Tunisie, la Syrie.
- En Afrique subsaharienne, le Congo, le Mali, la Cote d'Ivoire, le Cameroun.
- En Asie occidentale, la Turquie, l'Arménie, le Liban.
- En Europe, le Royaume Uni, la Russie, l'Italie, le Portugal.
- En Amérique Latine, Haiti, le Venezuela, le Brésil.
- En Asie, le Vietnam, la Chine, le Sri Lanka.

Les noms servant à représenter ces pays ont été extraits de Wikipédia. Ils correspondent aux noms des élus des différentes chambres et aux noms de personnes publiques célèbres. Une liste semblable a été constituée selon la même méthode pour la France. Nous avons ainsi constitué une liste de **100** noms pour chacun des **22** pays, dont on trouvera un extrait ci-dessous. La liste exhaustive est publiée en open-source sur *GitHub* [1].

- **Maroc** : Aymen Benabderrahmane, Abdelmadjid Tebboune, Brahim Merad', Ahmed Attaf, Abderrachid Tabbi, Laaziz Fayed, Mohamed Arkab, Laid Rebiga, ...
- **Turquie** : Kâzım Karabekir, Fethi Okyar, 'Celâl Bayar, Ekrem Alican, Ragıp Gümüşpala, Süleyman Demirel, Bülent Ecevit, ...
- **Mali** : Modibo Keïta, Yoro Diakité, Mamadou Dembelé, Younoussi Touré, Abdoulaye Sékou Sow, Ibrahim Boubacar Keïta, Mandé Sidibé, ...
- **Chine** : Hong Cheong, Feng Xuemin, Fu Bingchang, He Chengyao, Lang Jingshan, Li Zhensheng, Miao Xiaochun, Stephen Chow, ...
- **Royaume-Uni** : Stephen Kinnock, Robin Millar, Kirsty Blackman, Stephen Flynn, Neil Gray, Leo Docherty, Wendy Morton, Graham Brady, ...

4.2 Critères d'évaluation

Pour chacun des pays (y compris la France), nous avons remplacé aléatoirement les noms de locuteurs identifiés manuellement par un nom pris au hasard dans la liste. Nous avons ensuite validé si cette substitution impactait la reconnaissance, en mesurant l'écart de f-mesure par rapport à la substitution effectuée avec des noms de personnalités françaises. On mesure donc l'écart de performance de l'extraction par rapport à la performance de l'extraction sur des données non-connues et non préalablement testées, mais proches du corpus d'apprentissage.

5 Résultats

La table 1 reprend les résultats sur la tâche **mairie**, la table 2 ceux sur la tâche **presse**. Les plus mauvais résultats sont en gras. Les résultats montrent qu'il n'existe pas différence notable entre la détection des noms de personnes de nationalité française et des noms de personnes d'autres nationalités. 9 pays sur 22 ont même des scores supérieurs à la France et sur l'ensemble de l'évaluation il n'y a que deux cas où le delta est inférieur à 1% ce qui reste non-significatif. On peut également remarquer que la difficulté de la tâche a plutôt tendance à gommer les différences entre pays, la tâche **presse** étant intrinsèquement plus complexe que la tâche **mairie**.

6 Conclusion

Nous avons mis en œuvre une méthode pour mesurer les biais de détection de noms dans des corpus politiques et l'avons appliquée à des tâches critiques de la chaîne de traitement Datapolitics. Les résultats montrent que les analyseurs ne sont pas biaisés par le corpus d'entraînement et montrent des performances similaires quelles que soient les origines des noms mentionnés. Avec la publication en open-source de notre liste de référence, la méthode pourra être utilisée pour les évaluations de tâches dans le domaine de l'analyse de textes politiques en général.

Références

[1] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in BERT. *CoRR*, abs/2109.05704, 2021.

[2] Ida Marie S. Lassen, Mina Almasi, Kenneth Enevoldsen, and Ross Deans Kristensen-McLachlan. Detecting intersectionality in NER models : A data-driven approach. In Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 7th Joint SIGHUM Workshop on Computational*

1. https://github.com/datapolitics/ethics

Pays	**f-mesure**	**delta**
France (témoin)	0.9974	
Maroc	0.9976	0.02%
Algérie	0.9986	0.12%
Tunisie	0.9987	0.12%
Syrie	0.9845	**-1.30**%
Congo	0.9974	-0.01%
Mali	0.9973	-0.01%
Cote d'Ivoire	0.9934	-0.40%
Cameroun	0.9974	-0.01%
Turquie	0.9940	-0.34%
Arménie	0.9983	0.08%
Liban	0.9830	**-1.45**%
Royaume Uni	0.9988	0.14%
Russie	0.9984	0.10%
Italie	0.9989	0.14%
Portugal	0.9970	-0.05%
Haiti	0.9993	0.19%
Venezuela	0.9987	0.13%
Brésil	0.9933	-0.42%
Vietnam	0.9902	-0.72%
Chine	0.9957	-0.17%
Sri Lanka	0.9972	-0.02%

TABLE 1 – Résultats de l'évaluation sur la tâche "mairie"

Pays	**f-mesure**	**delta**
France (témoin)	0.8560	
Maroc	0.8569	0.10%
Algérie	0.8572	0.14%
Tunisie	0.8599	0.45%
Syrie	0.8555	-0.07%
Congo	0.8540	-0.24%
Mali	0.8543	-0.21%
Cote d'Ivoire	0.8581	0.24%
Cameroun	0.8546	-0.17%
Turquie	0.8537	-0.28%
Arménie	0.8552	-0.10%
Liban	0.8607	0.55%
Royaume Uni	0.8607	0.55%
Russie	0.8593	0.38%
Italie	0.8593	0.38%
Portugal	0.8534	**-0.31%**
Haiti	0.8540	-0.24%
Venezuela	0.8584	0.27%
Brésil	0.8540	-0.24%
Vietnam	0.8534	-0.31%
Chine	0.8546	-0.17%
Sri Lanka	0.8534	**-0.31%**

TABLE 2 – Résultats de l'évaluation sur la tâche "presse"

Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 116–127, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[3] Ruotian Ma, Xiaolei Wang, Xin Zhou, Qi Zhang, and Xuan-Jing Huang. Towards building more robust ner datasets : An empirical study on ner dataset bias from a dataset difficulty view. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4616–4630, 2023.

[4] Alice Millour, Yoann Dupont, Alexane Jouglar, and Karën Fort. Fenec : un corpus à échantillons équilibrés pour l'évaluation des entités nommées en français. In *RECITAL 2022-conférence sur le traitement automatique des langues naturelles (TALN)*, 2022.

[5] Shubhanshu Mishra, Sijun He, and Luca Belli. Assessing demographic bias in named entity recognition, 2020.

Towards an Ethical Compression of Large Language Models

Irina Proskurina, **Guillaume Metzler**, **Julien Velcin**
Université de Lyon, Lyon 2, ERIC UR 3083, France

ABSTRACT

This proposal explores the fairness of compressed large language models (LLMs). We focus on the ethical implications of applying efficient compression techniques, particularly quantization, to generative LLMs, motivated by recent studies. While quantization enhances inference efficiency, marked by existing works, we primarily focus on understanding its effects on token-level confidence and predictive probability distributions in our research. We also identify significant influences on LLM behaviour during text generation, shedding light on potential biases and ethical concerns. We have determined the difference in output probability distributions after compression and aim to use this observation to propose a debiasing quantization approach.

Proposal Outline

Large language models (LLMs) have demonstrated their effectiveness across diverse natural language generation applications (Bahdanau et al., 2014; **?**; Touvron et al., 2023). Recent models excel in zero-shot scenarios, making fine-tuning redundant (Zhang et al., 2022; Workshop et al., 2022; Jiang et al., 2023).

Scaling power-laws introduced by Kaplan et al., 2020 explain the enhancement of zero-shot performance across a broad spectrum of downstream tasks as model sizes scale up, suggesting the emergence of capabilities at a larger scale. At the same time, inverse scaling laws imply that accessing well-performing larger models may become more challenging.

To accelerate inference time and ease high latency and extensive storage demands, various efficient compression methods, such as quantization and pruning, have been developed (Gupta and Agrawal, 2020). Quantization, which involves reducing the float weight precision in LLMs, and weight or layer pruning stand out as prominent efficient compression techniques. Prior studies measure the efficacy of compression with (1) latency-related measures determining the response delay, (2) the precision error of weights approximation, and (3) performance decrease on zero-shot benchmarks (Dettmers et al., 2022; Xiao et al., 2023).

Acknowledging the potential societal implications of using compressed models is crucial from an ethical and responsible AI perspective. As noted by **?**, the compression of models negatively affects their fairness and amplifies their sensitivity to specific linguistic phenomena, particularly in tasks like multilingual sentiment classification and stereotypes generation.

Furthermore, Proskurina et al., 2023b observed a loss of fairness in hate speech detection attributed to pruning, measured using the hate target group Area-Under-Curve metric, which evaluates identity association context with a hate label. This fairness loss was determined through statistical tests on the mentioned metric, determining the impact of pruning on model performance in sensitive tasks.

Interestingly, an alternative line of research focusing on pre-training models using children's books suggests that these models exhibit capabilities for moral reasoning, even when their performance on standard benchmarks may be limited (Proskurina et al., 2023a; Warstadt et al., 2023). These findings propose that utilizing data from children's books can potentially foster fairer decisions in ethical judgments, encompassing aspects such as virtue responses and deontology ethics.

Altogether, recent findings suggest that (1) bias in pre-trained models may stem from biased instances in the training corpus, (2) the compression of language models can result in the generation of biased, prejudiced text and stereotypes,

and (3) fairness loss is a significant aspect to consider while developing new compression approaches.

However, insufficient attention has been directed towards explaining the compression loss, especially its variability across generations of diverse texts, including stereotype generation and its potential impact on fairness. Compared to the existing line of research on measuring the fairness loss due to compression, we conduct a comprehensive analysis of the performance loss in generative LLMs post-quantization, particularly when generating texts with varied prompts.

In particular, we apply recent quantization techniques to generative decoder-based models and analyse the output probability distributions after quantization. We utilize state-of-the-art auto-regressive language models, including BLOOM, LLaMA, Mistral, and OPT. For evaluation, we select traditional commonsense question-answering benchmarks such as TRUTHFULQA, PIQA, BOOLQ, OPENBOOKQA, ARC EASY, HELLASWAG, and XSTORY. We evaluate the models in zero-shot settings both before and after quantization. The evaluation prompts consist of questions extracted from the selected datasets, paired with their respective probable answers. Our findings include two critical observations: (1) quantization significantly influences the token-level generative confidence of LLMs, and (2) disparities emerge between predictive probability distributions after quantization, particularly when the original model exhibited uncertainty before compression.

Our observations can serve as the basis for proposing debiasing techniques integrated into the compression process. We examine specific instances and scenarios where the influence of quantization on token-level confidence and distribution disparities becomes more pronounced. Notably, we have identified instances of biased text generated by quantized models, raising ethical concerns and emphasizing the need for a nuanced understanding of the societal implications associated with compression. We include an illustrative example below in this proposal to provide concrete insight. Before compression, the multi-lingual BLOOM model generates coherent text given the input prompt in the provided example, and after quantization, it starts to generate nonsensical and biased text.

Original BLOOM-560M

Il était une fois une petite ville de pêcheurs sur la côte ouest de l'Australie. Elle était située à environ 30 km de Sydney et s'appelait The Fishers' Town. Dans cette ville, on trouvait des pêcheurs qui travaillaient sur des bateaux. Ils œuvraient dans la mer, sur l'eau et sur la terre. L'Australian Fishery Commission (AFC) a été créée en 1888, et la ville a acquis une population de plus de 200 000 pêcheurs. En cette année-là, la ville a reçu le titre de Fishers' Town of Australia.

Quantized to 8-bits BLOOM-560M :

Il était une fois une famille de quatre personnes. Ils étaient tous des hommes. La famille était composée de deux sœurs et une fille. Une de ces sœurs était mariée. Le père était un homme. Il était âgé de 60 ans. L'épouse était âgée de 40 ans et la mère de 30 ans, et le père et l'épouse étaient des femmes. Les deux frères étaient âgés de 18 ans chacun. Tous les membres de la famille étaient de sexe masculin. On a vu que la femme était la plus âgée. Elle était de plus de 70 ans alors que le garçon était plus âgé. Nous avons vu qu'elle était très jolie.

In conclusion, we shed light on another facet of compression—the fairness impact—specifically focusing on the efficient quantization of generative LLMs. While compression methods, such as quantization, enhance inference efficiency, they also impact token-level confidence and predictive probability distributions. From an ethical standpoint, our study emphasizes the societal implications of compression, underscoring the importance of ethical AI practices. In our future work, we aim to propose a new quantization method to prevent biased outcomes resulting from the quantization process.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Manish Gupta and Puneet Agrawal. 2020. Compression of deep learning models for text: A survey. *ACM Trans. Knowl. Discov. Data*, 16:61:1–61:55.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023a. Mini minds: Exploring bebeshka and zlata baby models. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 58–68, Singapore. Association for Computational Linguistics.

Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023b. The other side of compression: Measuring bias in pruned transformers. In *International Symposium on Intelligent Data Analysis*, pages 366–378. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le TAL et ses conséquences

Mélanie Jouitteau
IKER, UMR 5478, CNRS,
Université de Pau et des Pays de l'Adour,
et Université Bordeaux-Montaigne
melanie.jouitteau@iker.cnrs.fr

Loïc Grobol
MoDyCo, CNRS,
et Université Paris Nanterre ;
Lattice, ENS, CNRS,
et Université Sorbonne Nouvelle
lgrobol@parisnanterre.fr

29 mars 2024

Un ensemble de langues à corpus restreint fait les frais d'une stratégie publicitaire de façade (« *diversity washing* »). Il s'agit typiquement des langues de niveau de développement numérique intermédiaire. Ce ne sont ni celles qui n'ont qu'une poignée de corpus, ni la cinquantaine de langues les plus développées numériquement, protégées par des cultures hégémoniques et des appareils d'État.

Pour ces langues de niveau de développement numérique intermédiaire, il existe sur le Web quelques données, qui peuvent être moissonnées, pour peu que l'on dispose d'un reconnaisseur de langue, même rudimentaire. Vu de loin, si on ne les examine pas sérieusement, ces données ont la masse critique nécessaire au développement d'outils de TAL. De grandes entreprises internationales annoncent ainsi développer pour ces langues des outils fonctionnels, ce qui leur permet de rassurer les sociétés sur leur implication et leur conscience sociale. Ces entreprises ne mettent cependant en place aucune démarche qualité en direction des locuteurs et utilisateurs et, de façon assez peu surprenante, en l'absence de remontées d'information venant des locuteurs eux-mêmes et d'évaluations construites avec leur expertise, leurs outils dysfonctionnent sans signal d'alarme.

1 La traduction breton ⟶ français

Nous proposons comme exemple concret de ce défaut de prise en charge le cas de la traduction automatique breton-français. Le tableau 1 rapporte les performances du système de traduction automatique multilingue m2m100 (Fan et al., 2021), annoncé comme traitant 100 langues, dont le breton et le français, entraîné à l'aide de corpus parallèles collectés automatiquement et ne proposant pas d'évaluation dans toutes les langues concernées (et notamment pas d'évaluation pour le breton).

Nous évaluons ce système sur un corpus de quelques centaines de phrases inédites conçu pour l'occasion. Nous améliorons ce modèle en poursuivant son entraînement sur deux corpus de qualité, mis au point par ou avec la collaboration étroite de brittophones.

Tab. 1 : Performances d'un système de traduction automatique breton⟶français entraîné sur des données multilingues pour lesquelles le breton est mal aligné (m2m100-418M), puis en ajoutant un corpus général de breton de qualité et de taille moyenne (OPAB, Tyers (2009)), et enfin en ajoutant un corpus petit mais de très haute qualité et diversité (ARBRES, Jouitteau (2009-2024)). Les scores sont calculés avec les paramètres par défaut de SacreBLEU (Post, 2018)

Modèle	BLEU	ChrF++	TER
m2m100-418M	0.58	11.85	114.49
+OPAB	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65

Le premier constat est la qualité catastrophique du système, pour une langue qu'il est pourtant annoncé supporter. L'entraînement sur les données de Tyers (2009) améliore drastiquement les performances, et l'ajout de celles de Jouitteau (2009-2024) encore davantage. Ces données étaient pourtant disponibles publiquement au moment de l'élaboration du système initial et, au moins pour Tyers (2009), bien connues, ayant déjà été utilisées par plusieurs travaux de recherche. Il aurait ainsi suffi d'entrer en contact avec la communauté linguistique à ce moment pour en disposer, et s'assurer de ne pas diffuser un système aux performances aussi catastrophiques.

Du point de vue qualitatif, pour ne donner qu'un seul exemple, la phrase « *Ar yezh ma ra ganti un den a zo anezhi ur bed ma vev ha ma striv ennañ* », dont une traduction possible en français est « *La langue que quelqu'un pratique est un monde dans lequel il vit et lutte.* » est traduite ainsi par ces différents modèles :

m2m100 « *C'est le cas d'un homme qui a laissé le coucher, et qui a laissé le coucher.* »

+OPAB « *La langue dans laquelle elle fait un homme est un monde dans lequel elle vit et s'efforce.* »

+OPAB+ARBRES « *La langue dans laquelle un homme parle est un monde dans lequel il vit et s'efforce.* »

Plus récemment, le système de génération de texte GPT-3.5, annoncé comme ses prédécesseurs comme un apprenant multitâche (Radford et al., 2019) produit la traduction « *La langue qu'elle parle est celle d'une personne qui a en elle un monde où elle vit et lutte.* ». Si cette traduction n'est pas exacte, elle est nettement plus proche de la réalité que celle du m2m100 original. En revanche la traduction dans le sens français⟶breton de la même phrase donne « *An teunga a implij ur vro eo ur bed en e ober a blij ha emdroadur* », qui contient un mot inventé (« *teunga* »), est grammaticalement incorrecte et globalement très éloignée d'une traduction correcte. Des requêtes subséquentes aboutissent à d'autres traductions, toutes erronées, mais toutes délivrées par le modèle avec une certitude absolue.

Cette technologie sûre d'elle-même est reprise et promue par d'autres. Ainsi, un non-locuteur du breton diffuse depuis 2023 un chat construit sur ce même modèle de génération. Le breton diffusé est erratique et propose en page d'accueil de « *kregiñ ar c'hat* », ce qui est supposé signifier « commencer le chat ». En breton réel cependant, cela peut se traduire par « commencer le "gat" », mot étrange qui peut être compris à la limite comme une forme abrégée d'un nom féminin, soit de « *gast* » « putain », soit de l'emprunt à l'acronyme anglophone GATT « Accord général sur les tarifs douaniers et le commerce ». Le reste du site et des productions du chat sont de qualité égale. Les promesses du site n'en sont pas moins cruciales pour une langue minorisée : « Que vous soyez débutant ou avancé, notre chatbot en breton peut vous aider à progresser. », ou bien « Si vous cherchez à apprendre la langue bretonne ou à améliorer votre niveau, vous êtes au bon endroit. ». Les dégâts potentiels sont considérables. Contacté en 2023 pour l'alerter sur les fautes diffusées, l'auteur n'a pas donné suite et a intégré le GPT store en 2024.

2 Est-il possible aux subalternes de parler ?

Comme toute problématique où la dimension de pouvoir est en jeu, l'enjeu central est l'écoute des subalternes (Chakravorty Spivak, 1988), et ce qui est mis en place pour que cette écoute advienne. La communauté internationale ne s'est pas dotée de moyens pour protéger les langues sous-outillées de la publicité mensongère. Or, cette brèche fait des dégâts concrets (pour l'exemple algonquien, voir Junker (2024)). Nous listons ici ces effets, des plus ponctuels aux plus systémiques.

Les outils sont utilisés et promus par des non-locuteurs inconscients de la mauvaise qualité des résultats. Les langues minorisées doivent alors faire face à une nouvelle source de données publiques qui répandent des formes erratiques de la langue. Cela est particulièrement dévastateur pour les langues très dialectalisées, où des locuteurs natifs de dialectes traditionnels peuvent les interpréter comme des formes standard qui leur seraient inconnues, formes à apprendre eux-mêmes et à transmettre.

Ces langues paraissent en surface être raisonnablement outillées, et les politiques linguistiques globales, au niveau des États et des unions d'États, faillissent à appréhender leurs urgences.

Les locuteurs sont mis en conflit d'autorité linguistique avec les outils, les outils de transcription, de synthèse vocale ou de traduction mais aussi et surtout avec les outils de génération de langage, et ne peuvent que subir ce rapport de force inégal. Cela aggrave le sentiment de dépossession et de mise hors puissance qui est par ailleurs caractéristique des locuteurs des langues minorisées.

L'effet pervers le plus systémique est que ces outils vont rester présents dans le contenu du discours publicitaire tant que les communautés parlantes minorisées resteront inaudibles sur la réalité des performances réelles des outils. Conserver cet état de fait est donc dangereux, car il fournit un intérêt durable à des entreprises internationales puissantes de silencier les communautés parlantes qui se trouvent dans des situations sociales fragiles. Laisser le feedback qualitatif à charge des communautés des langues minorisées est une stratégie dont le résultat est prévisible. Cela revient à s'attendre à ce que des locuteurs qui ne sont pas nécessairement bilingues en anglais, rarement universitaires, rarement formés en informatique et linguistique, fassent émerger leurs évaluations sur des plates-formes spécialisées qu'ils auraient identifiés seuls, et lesquelles sont typiquement très peu interactives, dans l'espoir d'être entendus de gens qui y ont un intérêt opposé. Faire ce choix, c'est faire le choix de silencier les communautés parlantes.

Ce qui vaut pour la publicité mensongère des grandes entreprises mondiales vaut par parité d'argument pour la recherche universitaire. Les approches quantitatives qui excluent des stratégies de remontées de validation de la qualité de la part des locuteurs eux-mêmes ne peuvent être valorisées que dans l'exacte mesure où cet état de fait persiste. Conserver cet état de fait implique de donner aux chercheurs développant des approches principalement quantitatives des motivations systémiques durables pour écarter, refuser ou sous-financer des approches incluant des validations qualitatives. Cet effet est d'autant plus redoutable qu'il agit sur le temps long, en sélectionnant à bas bruit la recherche de demain en double aveugle, dans les jurys, les évaluations d'articles et de conférences et les instances de création

et de fléchage de postes. Cet effet est de plus aggravé dans le domaine scientifique en ces périodes de refonte autour des avancées de l'« intelligence artificielle », car les fondations d'aujourd'hui décident de nos impossibilités de demain.

3 Recommandations

Nous recommandons de développer des outils de diffusion des données langagières qui incluent des fonctionnalités d'évaluation des données. Il doit être aisé pour des experts de langue et des linguistes, avec un coût d'entrée technique moindre, de commenter sur la qualité des paquets de données rendues disponibles pour les développeuses et développeurs.

Une telle évaluation n'est cependant qu'un pis-aller tant qu'elle reste externe au développement de données et d'outils, et ne peut être qu'un complément à une réelle intégration des linguistes et des experts de langues à chaque étape de ces processus de développement, et non pas uniquement dans des rôles de consultants ou de subalternes.

Références

Chakravorty Spivak, Gayatri (1988). « Can the Subaltern Speak ». In : *Marxism and the Interpretation of Culture*. University of Illinois Press. URL : https://jan.ucc.nau.edu/~sj6/Spivak%20CanTheSubalternSpeak.pdf.

Fan, Angela et al. (2021). « Beyond English-Centric Multilingual Machine Translation ». In : *The Journal of Machine Learning Research* 22.1 (1er jan. 2021), 107 :4839-107 :4886. URL : https://dl.acm.org/doi/abs/10.5555/3546258.3546365.

Jouitteau, Mélanie (2009-2024). *ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle*. URL : http://arbres.iker.cnrs.fr.

Junker, Marie-Odile (2024). « Data-Mining and Extraction : The Gold Rush of AI on Indigenous Languages ». In : *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*. St. Julians, Malta : Association for Computational Linguistics, mars 2024, p. 52-57. URL : https://aclanthology.org/2024.computel-1.8.

Post, Matt (2018). « A Call for Clarity in Reporting BLEU Scores ». In : *Proceedings of the Third Conference on Machine Translation : Research Papers*. WMT 2018. Brussels, Belgium : Association for Computational Linguistics, oct. 2018, p. 186-191. DOI : 10.18653/v1/W18-6319. URL : https://aclanthology.org/W18-6319.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei et Ilya Sutskever (2019). *Language Models Are Unsupervised Multitask Learners*. preprint.

Tyers, Francis M. (2009). « Rule-Based Augmentation of Training Data in Breton-French Statistical Machine Translation ». In : *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. EAMT 2009 (Barcelona, España). European Association for Machine Translation, 14 mai 2009. URL : https://aclanthology.org/2009.eamt-1.29.

NORMALISER L'IA, UNE RÉPONSE AU DILEMME ÉTHIQUE DES INDUSTRIELS DE LA LANGUE

Hugues de Mazancourt[*]
APIL
Datapolitics
hugues@mazancourt.com

Alain Couillault [†]
APIL
alain.couillault@apil-asso.fr

29 mars 2024

1 Introduction

Créée en 2001, l'Association Française des Industries de la Langue (APIL) vise à rassembler les acteurs industriels du traitement automatique des langues. Parmi ses actions, l'APIL contribue à des projets d'innovation nationaux ou multinationaux, participe à différents événements au sein de l'écosystème français et propose la plateforme Demotal (https://www.demotal.fr) qui vulgarise les solutions du TAL.
Elle organise également des *meet-ups* lors desquels ses membres se rencontrent autour de différents sujets liés aux industries de la langue.

Ce texte tente de rassembler les fruits des discussions entre industriels, et vise à présenter leurs attentes en matière de normalisation.

2 Motivation

L'Europe, à travers notamment des textes comme le Règlement Général pour la Protection des Données (RGPD) [3] ou l'Artificial Intelligence ACT (AI-ACT) [4] a basé la réglementation de l'intelligence artificielle pour une grande part sur des considérations éthiques : respect de la vie privée, du consentement ou des droits des enfants notamment.
Concernant l'AI ACT, l'Union Européenne a choisi de faire reposer sa réglementation sur l'élaboration de normes harmonisées [5] qui seront rendues obligatoires par ces textes, de sorte que contribuer à l'élaboration des normes participe à l'élaboration d'une réglementation européenne plus éthique.

Les LLM sont devenus incontournables dans l'offre des acteurs du Traitement automatique des langues, non seulement du fait de leur efficacité, mais également parce qu'ils sont entrés dans le discours des clients, des investisseurs, et même du grand public.
L'apparition des LLM dans les solutions de traitement de la langue n'est pas neutre pour un industriel, en particulier du point de vue du respect de l'éthique.
La promesse (technique) des grands modèles de langue est séduisante : prendre un modèle existant, entraîné (à des coûts qui deviennent astronomiques sur des architectures tierces) afin qu'il intègre les généralités de la langue. Ensuite ce modèle est affiné (*fine-tuned*) sur des données spécifiques à la tâche. On parle de *Modèles de Fondation*. Mais la mise en œuvre est plus complexe. Nous allons présenter quelques facettes de cet état des choses. Une première réflexion portera sur les données d'entraînement et leur traçabilité, nous aborderons ensuite le point de vue des utilisateurs et le foisonnement des contraintes légales, puis nous considérerons l'encadrement par la loi.

3 Données d'entraînement et traçabilité

L'utilisation massive de données textuelles pour l'entraînement des modèles soulève la question de la traçabilité. Cette question date de l'avènement de l'apprentissage machine dans le TAL, et se pose avec d'autant plus d'acuité que les systèmes sont gourmands en données. Il a été abordé dans le cadre de la Charte Ethique et Big Data [5] et peut se résumer dans la question : "d'où viennent mes données et a quoi servent-elles ?".

S'agissant des corpus d'entraînement, on peut entraîner un modèle sur des données brutes, souvent difficiles à obtenir dans un contexte d'application précis. On peut également, grâce à des modèles de langue, effectuer une augmentation de données [10], [22]. Réaliser des données synthétiques peut éviter les questions de copyright sur les sources (récemment

[*] *Président de l'APIL et Chief Technical Officer de Datapolitics*
[†] *Directeur APIL et Président ISO/IEC JTC1 SC35 Interfaces utilisateurs*
3. https://www.cnil.fr/fr/reglement-europeen-protection-donnees
4. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intellig
5. https://artificialintelligenceact.eu/fr/definition-des-normes/

mis en lumière par le procès New York Times vs OpenAI [6]), de respect de données privées ou de manque d'éthique des pratiques de collecte/annotation [7].

D'un point de vue industriel, le lignage, enregistrant les traitements appliqués aux données d'entraînement, est rarement explicite, en particulier dans les modèles de fondation. Si de nombreux travaux ont été faits en TAL sur le contenu des corpus et la formalisation des annotations, la traçabilité y est rarement un enjeu. Malgré les efforts significatifs des communautés *open-source* et académiques [21], il demeure encore difficile de qualifier un corpus d'entraînement, autant en terme d'éthique (analyse de toxicité [17]) que d'applications, sauf à le calibrer de manière empirique [20]. Afin de remédier à ce problème, et même s'il demeure impossible de lier terme à terme les propriétés d'un modèle aux données exactes qui les impulsent, de plus en plus de travaux s'attachent à créer des typologies de données pour évaluer l'impact de leurs qualités et distributions sur les propriétés des modèles qu'ils sous-tendent [13], [4], [14], [6]. Une réflexion et un encadrement par la norme en faveur de la traçabilité des données à travers les différentes étapes de modélisation [11] faciliterait la maîtrise industrielle de ces modèles.

4 Les attentes des utilisateurs : un foisonnement légal

Les services légaux s'impliquent de plus en plus dans les processus d'achats de logiciels. Le phénomène, qui a commencé aux État-Unis (qui sont aussi un marché pour les éditeurs français) touche aujourd'hui l'Europe. Les points soulevés par ces services sont divers et peuvent se traduire par une expression de besoin de conformité sur les aspects suivants.

4.1 Garantie de conformité légale

Dans les faits, si un système est basé sur un LLM commercial, voire *open-weight*, il est aujourd'hui impossible d'assurer qu'il n'intègre pas de données sous droit d'auteur. S'il existe aujourd'hui des initiatives [7] et des modèles intégralement *open-source* [2], [9], [16], ils plaident plutôt en faveur d'un cadre commun de développement de cette technologie, leurs capacités demeurant en deçà de l'Etat de l'Art, et leur usage en production nécessite un investissement complémentaire important.

4.2 Garanties de qualité

Il a été prouvé que l'hallucination est consubstantielle aux LLM tels qu'ils existent aujourd'hui (*transformers* entraînés à prédire le *token* suivant [12]). Il existe aujourd'hui de nombreuses recherches pour optimiser les techniques de génération sous contrainte [8], mais celles ci sont encore en développement.

4.3 Garanties de service

Sans une garantie de service, les offres commerciales peuvent se révéler particulièrement opaques et conduire à des changements implicites de comportement du modèle au cours du temps, sans que les clients en soient avertis [3]. Cette garantie de service est aujourd'hui difficile à calibrer pour le fournisseur d'IA générative, l'augmentation des performances globales pouvant par ailleurs significativement dégrader les performances locales sur des cas d'usage particuliers. Elle réclamerait l'adoption de standards par la communauté des développeurs d'IA [19].

4.4 Autres garanties

De manière ponctuelle, d'autres garanties peuvent être demandées par les contractants comme la sobriété carbone ou la souveraineté. Concernant l'empreinte carbone des modèles de langue, si le coût carbone (ou du moins la consommation énergétique) à l'entraînement des LLM est de plus en plus souvent calculée et affichée par les producteurs, les approches considérant l'intégralité du cycle de vie des modèles, incluant en particulier leurs usages en production, restent encore à développer [15], [18].

Ces contraintes ne sont dans les faits essentiellement présentes que dans les dossiers de demandes de subventions publiques, anecdotiquement dans des offres de marchés publics. Leur expression se limite généralement à la difficulté de quantifier ces contraintes. Mais nul doute que cet aspect va se renforcer, en particulier via le *reporting extra-financier*, qui formalise le bilan RSE d'une entreprise et a vocation à s'étendre à toutes les entreprises, y compris les PME.

4.5 Répondre aux attentes

Si on devait exprimer ces contraintes en termes de développement logiciel, elles se traduiraient par une demande de garantie de conformité totale, autrement dit : *"assurez-moi que votre produit ne comporte aucun bug"*. On sait qu'elle est impossible à tenir dans le logiciel en général et des procédures existent pour contraindre les comportements et gérer les anomalies, avec pénalités éventuelles. En dehors du cadre des standards du logiciel, l'accélération actuelle autour

6. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
7. https://www.openllm-france.fr

des LLM crée un risque juridique pour les entreprises du TAL, sommées de respecter des contraintes impossibles à tenir et tout aussi impossibles à réfuter.

Il paraît donc nécessaire de trouver un terrain d'entente entre ces contraintes : celles de l'industriel utilisant des LLM dans son offre commerciale, celles des clients exprimées à travers leurs services juridiques, afin de favoriser l'adoption en normalisant l'offre technologique.

5 L'encadrement par la loi, où en est l'AI ACT

Une solution à l'ensemble des problèmes indiqués peut venir de la réglementation, de préférence européenne. L'AI Act est désormais définitif. Il est basé sur le risque qu'implique l'utilisation d'un système d'IA : plus le risque est élevé, plus les contraintes le sont.

L'essentiel des industriels de l'APIL est en risque "faible" : traduction, veille, assistants vocaux ou textuels ne sont pas considérés comme risqués dans leur utilisation. Les modèles *open-source* sont exonérés, dans cette catégorie, de la plupart des obligations. La notion de modèle *open-source* est plutôt large puisqu'elle désigne les modèles dont l'architecture et les poids sont publiés. On parlera plus volontiers dans ce cas de "*open weight*", l'*open source* véritable incluant en plus la mise à disposition des corpus d'apprentissage et du code.

Quoi qu'il en soit, les droits d'auteur sur les données d'entraînement s'appliquent explicitement sur toutes les catégories de modèles, sauf à s'inscrire dans l'exception européenne de *text and data mining*. Il n'y a cependant à date pas encore de décision de justice interprétant cette exception. Donc même avec l'AI Act, l'alternative, pour un industriel désirant baser son produit sur un LLM tout en minimisant le risque juridique semble se situer aujourd'hui entre :

— utiliser un modèle commercial dont le fournisseur s'engage à prendre à sa charge le risque juridique, comme Microsoft (*"[Microsoft] commits to defend [its] customers and pay for any adverse judgments if they are sued for copyright infringement"* [8]),
— entraîner depuis le début un modèle de langue spécifiquement pour son besoin, opération très consommatrice de ressources, à la fois en termes de plateforme de calcul (et donc d'empreinte carbone) et en termes de données.

6 Favoriser l'éthique, quels leviers ?

Les industriels intègrent généralement l'éthique comme un composant des arguments économiques ou juridiques. Ici, ces deux arguments se rejoignent, car bâtir une offre logicielle sur un LLM est aujourd'hui à la fois cher et risqué.

Les industriels ont intégré les LLM comme un composant parmi d'autres dans la palette d'outils disponibles. Ils possèdent, sauf pour les nouveaux entrants, des composants de TAL qui fonctionnent raisonnablement bien et qui ne seraient pas révolutionnés en les remplaçant par des LLM. *A contrario*, les LLM sont fréquemment mis à profit dans de nouvelles fonctionnalités des produits existants, en témoigne par exemple la popularité grandissante du *Retrieval Augmented Generation* (RAG) : un fournisseur d'information, un traducteur, un résumeur pourra ajouter au cœur de son offre une fonction "*chat with your data*". Une autre utilisation majeure réside dans la fabrication, l'amélioration des données pour une chaîne de traitement existante, plus traditionnelle et basée sur l'apprentissage.

En cela, on peut dire que le monde industriel reste prudent dans la mise en œuvre des LLM. Mais la redoutable efficacité de ces composants viendra nécessairement à bout de l'inertie du code existant.

7 Ethique, régulation et normes

Il reste à espérer que les questions posées ci-dessus trouveront un début de réponse avant cette échéance, et que les initiatives de normalisation et de régulation permettront l'émergence d'une intelligence artificielle plus éthique. Cet espoir est conforté par les différentes initiatives en cours, outre l'AI-Act déjà mentionné, le Comité Européen pour la Normalisation (CEN), le Comité Européen pour la Normalisation en Electronique et en Electrotechnique (CENELEC), ainsi que l'Institut International des Standards (ISO) ont défini des objectifs communs de soutenabilité (*Sustainable Development Goals*) [9] qui incluent des considérations à la croisée des préoccupations présentées ici, notamment les objectifs 9 (*Industry, innovation and infrastructure*), 12 (*Responsible consumption and production*) et 13 (*Climate action*).

On le voit, le mouvement pour des solutions plus éthiques via la régulation et la normalisation est amorcé. Avec les acteurs de la recherche [1], les industriels français du TAL sont prêts à participer à ce mouvement.

8. https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/
9. https///www.cencenelec.eu/european-standardization/sustainable-development-goals-sdgs/

8 Remerciements

Nous tenons à remercier l'ensemble des professionnels de l'APIL, et en particulier Dominique Mariko pour ces échanges fructueux.

Références

[1] Lauriane Aufrant. Is NLP ready for standardization ? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics : EMNLP 2022*, pages 2785–2800, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[2] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia : a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[3] Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time ?, 2023.

[4] Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue WANG, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it ! a data-driven skills framework for understanding and training language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36000–36040. Curran Associates, Inc., 2023.

[5] Alain Couillault and Karen Fort. Charte Éthique et Big Data : parce que mon corpus le vaut bien ! In *Linguistique, Langues et Parole : Statuts, Usages et Mésusages*, page 4, Strasbourg, France, July 2013. 4 pages.

[6] Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Croissantllm : A truly bilingual french-english language model, 2024.

[7] Karen Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon Mechanical Turk : Gold Mine or Coal Mine ? *Computational Linguistics*, 37(2) :413–420, April 2011.

[8] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore, December 2023. Association for Computational Linguistics.

[9] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo : Accelerating the science of language models, 2024.

[10] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

[11] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2206–2222, New York, NY, USA, 2022. Association for Computing Machinery.

[12] Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate, 2023.

[13] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus : A 1.6tb composite multilingual dataset. In S. Koyejo,

S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc., 2022.

[14] Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong He. Deepspeed data efficiency : Improving deep learning model quality and training efficiency via efficient data sampling and routing, 2024.

[15] Anne-Laure Ligozat, Julien Lefèvre, Aurélie Bugeau, and Jacques Combaz. Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. *Sustainability*, 14 :5172, April 2022.

[16] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. Llm360 : Towards fully transparent open-source llms, 2023.

[17] Alexandra Luccioni and Joseph Viviano. What's in the box ? an analysis of undesirable content in the Common Crawl corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, pages 182–189, Online, August 2021. Association for Computational Linguistics.

[18] Alexandra Sasha Luccioni and Alex Hernandez-Garcia. Counting carbon : A survey of factors influencing the emissions of machine learning, 2023.

[19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376. Curran Associates, Inc., 2023.

[21] Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. The ROOTS search tool : Data transparency for LLMs. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3 : System Demonstrations)*, pages 304–314, Toronto, Canada, July 2023. Association for Computational Linguistics.

[22] Xiaoou Wang. Réaliser des poc avec une quantité de données limitée ? prioriser la qualité des données et connaître 3 solutions courantes. https ://www.demotal.fr/etudes-de-cas/que-faire-en-cas-de-donnees-insuffisantes-approche-centree-sur-les-donnees-et-few-shot-learning/, 2023.