# Empreinte carbone des expériences en TAL : les défis de la reproductibilité

Clément Morand, Anne-Laure Ligozat, Aurélie Névéol

# 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 *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

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

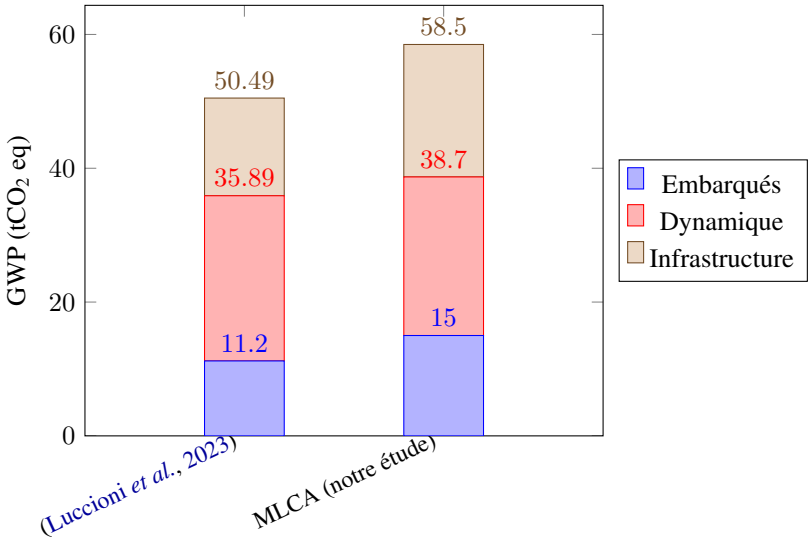


FIGURE 1 – Comparaison des estimations obtenues dans notre étude avec les résultats présentés dans la table trois de Luccioni *et al.* (2023) sur les différentes sources d'émissions (Les impacts embarqués sont ceux de la production du matériel attribuables à la tâche. La consommation dynamique est celle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

*et al.* (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 automatique 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.