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

Sur l'équité politique des grands modèles de langue multilingues

Paul Lerner François Yvon
Sorbonne Université, CNRS, ISIR, 75005, Paris, France
lerner@isir.upmc.fr, yvon@isir.upmc.fr

RÉSUMÉ

Les grands modèles de langue (LLM) sont de plus en plus utilisés, y compris pour des applications à visée politique, mais leur équité dans ces contextes a été peu étudiée. Nous analysons l'équité de la traduction produite par des grands modèles de langues (LLM) pour des discours issus de divers partis politiques. Nous constatons que la qualité de traduction des modèles post-entraînés est corrélée à la perplexité du LLM pré-entraîné correspondant. Ce résultat suggère que les biais politiques des LLM proviennent de leur pré-entraînement, et que ces biais sont peu affectés par le post-entraînement. En nous appuyant sur cette corrélation, nous utilisons les mesures de perplexité, qui ne nécessitent pas de traduction de référence, pour étendre l'étude des biais politiques de plusieurs LLM sur des textes reflétant diverses opinions politiques dans davantage de langues.

ABSTRACT

On the Political Fairness of Multilingual Large Language Models

Large Language Models (LLMs) are increasingly used, including in political applications, but their political fairness has been little studied. We approach this problem by analyzing the fairness of LLM-based translations of speeches from multiple political parties. We find that the translation quality of post-trained LLMs is correlated to the perplexity of the pre-trained LLM counterpart. This result suggest that the political biases of LLMs stem from their pre-training, and are hardly affected by the post-training stage. Building upon this correlation, we extensively assess the political biases of several LLMs on texts bearing various political opinions in even more languages through perplexity measurements, which do not require reference translations.

MOTS-CLÉS : équité, grand modèle de langue, politique, multilinguisme.

KEYWORDS: fairness, large language model, politics, multilingualism.

1 Introduction

Les grands modèles de langue (LLM) sont utilisés quotidiennement par des centaines de millions d'utilisateurs [1] (Milmo *et al.*, 2023) et sont de plus en plus intégrés dans des applications utilisées pour des activités politiques (Small *et al.*, 2023; Tessler *et al.*, 2024; Revel & Penigaud, 2025). Il est donc crucial d'évaluer et de documenter leurs biais afin de les atténuer, au besoin, ou de mieux encadrer leur utilisation (Resnik, 2025). De nombreux travaux portant sur les biais politiques des LLM sont parus récemment, faisant suite à près de dix ans de recherche sur leurs biais liés au sexisme et au racisme (Blodgett *et al.*, 2020; Gallegos *et al.*, 2024; Ducel *et al.*, 2024). Nous passons en revue ces travaux de manière plus approfondie dans la section 2, en notant qu'ils reposent en majorité sur la simulation

1. Top Websites Ranking : https://www.similarweb.com/top-websites/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 576–594.

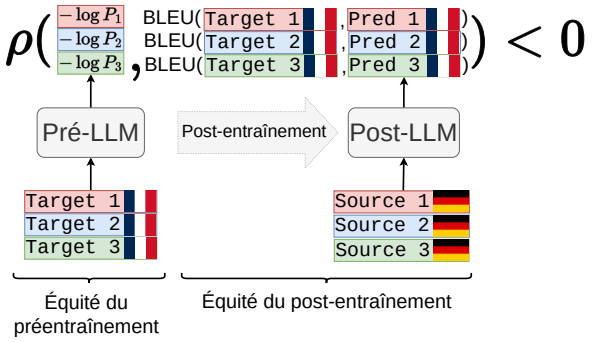

FIGURE 1 – Illustration simplifiée de nos expériences. Étant donné des textes de différents partis politiques (illustrés par différentes couleurs), nous trouvons une corrélation négative entre la log-vraisemblance négative des pré-LLM et les métriques de traduction en aval (e.g., BLEU) des post-LLM correspondants. Cela suggère que les biais politiques des LLM proviennent de leur pré-entraînement, et que ces biais sont peu affectés par le post-entraînement. Nous pouvons donc mesurer directement l'équité via la log-vraisemblance négative.

de réponses à des questionnaires d'opinions habituellement utilisés par des participants humains. Cette méthodologie pose de nombreuses questions (Boelaert *et al.*, 2025) ; en particulier sa robustesse a été mise à l'épreuve dans Ceron *et al.* (2024); Labat *et al.* (2026) qui montrent que les réponses des modèles pré-entraînés (pré-LLM) sont excessivement sensibles à de légères modifications du prompt présentant les questions, conduisant à des réponses quasi aléatoires. Ils recommandent de n'utiliser ces méthodes qu'avec des modèles *post-entraînés* (post-LLM), c.-à-d. ayant subi des étapes ultérieures d'affinage supervisé ou d'alignement. Il est cependant crucial d'évaluer les biais des pré-LLM comme ceux des post-LLM pour mieux comprendre à quel stade d'entraînement les biais politiques se manifestent : lors du pré-entraînement ou du post-entraînement ? Nous estimons que les biais issus du post-entraînement sont facilement explicables et manipulables : il est trivial d'ajuster un LLM pour qu'il donne la réponse y à la question x. Au contraire, les biais issus du pré-entraînement semblent plus profondément enfouis (Resnik, 2025).

Pour évaluer les biais politiques des pré-LLM et des post-LLM, nous analysons la corrélation entre la perplexité de modèles pré-entraînés sur des textes produits par divers partis politiques avec les métriques de traduction de leur équivalent post-LLM sur 420 paires de langues (section 4.2), en nous appuyant sur les travaux de Lerner & Yvon (2026). Nous observons une corrélation négative : si un LLM assigne une faible probabilité au texte d'un parti politique donné, il sera peu susceptible de le regénérer, y compris dans une autre langue, de sorte que la traduction sera de mauvaise qualité, même après post-entraînement. Ce résultat est cohérent sur cinq LLM. Cette observation nous permet d'étudier l'équité des pré-LLM et des post-LLM sur des corpus comparables en calculant leur perplexité, car les jeux de données parallèles annotés avec des métadonnées politiques sont rares. Nous constatons que la perplexité varie significativement selon les partis pour la grande majorité des 36 langues étudiées sur deux jeux de données (section 5.3). Nous menons ensuite une étude ciblée sur le portugais sur les deux jeux de données (section 5.4) et agrégeons les perplexités des familles de partis à travers les langues (section 5.5) pour constater que (i) les partis sociaux-démocrates sont légèrement mieux traités que les partis d'extrême-droite et nationalistes mais (ii) le biais n'est pas graduel car la gauche radicale et le centre sont également mal représentés, ce qui fournit une analyse plus fine que la grossière dichotomie gauche-droite des méthodes basées sur des questionnaires (section 2) et est cohérent avec les résultats de Lerner & Yvon (2026).

2 Travaux connexes

Méthodes basées sur des questionnaires Feng *et al.* (2023); Rozado (2023); Santurkar *et al.* (2023); Motoki *et al.* (2024); Potter *et al.* (2024) simulent avec des post-LLM le passage de questionnaires conçus pour des participants humains (et anglophones) comme par exemple le *Political Compass Test*[2], un test populaire sur Internet qui donne des scores selon deux axes en fonction de l'accord avec des affirmations telles que « *Sex outside marriage is usually immoral* » : économique (gauche-droite) et social (libertaire-autoritaire). Ces travaux s'accordent globalement sur le fait que les post-LLM « penchent à gauche ». Röttger *et al.* (2024); Ceron *et al.* (2024); Boelaert *et al.* (2025) ont montré indépendamment le manque de robustesse de cette méthode par questionnaire, car la réponse extraite du LLM varie selon la formulation exacte du prompt, par exemple utiliser « do you *agree* or *disagree* » vs. « do you *disagree* or *agree* » peut conduire à des réponses différentes. Durmus *et al.* (2024); Helwe *et al.* (2025) ont recours à la traduction automatique pour traduire des questionnaires en anglais dans plusieurs langues. Toutefois, la TA peut (i) contenir des erreurs, y compris en déformant l'opinion exprimée (Shafiabadi & Yvon, 2026) ; (ii) être biaisée en faveur d'un parti politique (Lerner & Yvon, 2026). Néanmoins, Durmus *et al.* (2024) constatent que l'interrogation des post-LLM en anglais, russe, chinois ou turc conduit toujours à des réponses occidentales, c'est-à-dire que les réponses sont cohérentes entre les langues. Au contraire, Helwe *et al.* (2025) constatent que « la langue a une forte influence », bien qu'ils ne contrôlent pas les effets parasites du prompt décrits ci-dessus. Labat *et al.* (2026) ont évalué conjointement la variation causée par la langue et la formulation du prompt. Ils ont constaté que (i) il n'était pas possible d'évaluer le biais des pré-LLM par des questionnaires, car la réponse variait trop selon la formulation du prompt ; (ii) l'effet de la langue dépendait de la question.

Autres méthodes Potter *et al.* (2024) mènent d'autre part une expérience d'interaction humain-ordinateur. Ils constatent qu'après avoir débattu avec un post-LLM en anglais, les partisans de Trump penchent davantage pour Biden. Röttger *et al.* (2026) évaluent la position des post-LLM lorsqu'on leur demande de générer des textes en anglais sur des sujets débattus. Ils constatent que ces modèles s'alignent avec les Démocrates américains, par opposition aux Républicains. Ceron *et al.* (2025) étudient le contenu politique des données de préentraînement et de post-entraînement des LLM en anglais. À l'aide d'un classifieur de contenu gauche-droite, ils trouvent une majorité de données penchant à gauche, ce qui est cohérent avec les résultats discutés ci-dessus. Lerner & Yvon (2026) étudient l'équité de la TA des post-LLM par rapport à l'affiliation politique du locuteur dans les débats du Parlement européen. Ils constatent que les partis traditionnels de gauche et de droite sont favorisés par rapport à la gauche radicale et à l'extrême droite. Cependant, l'effet n'est pas graduel car le parti vert est également mal traduit. Nous calculons la corrélation entre la perplexité des pré-LLM et les métriques de traduction en aval de leur équivalent post-LLM (tirées de Lerner & Yvon, 2026).

3 Méthodes

3.1 Équité politique

Suivant Lerner & Yvon (2026), nous nous intéressons aux biais politiques incorporés par les LLM en formulant cette question comme une question d'équité, c.-à-d. en posant comme idéal un modèle

2. https://www.politicalcompass.org/

capable de manipuler (reformuler, résumer, traduire) des textes avec un niveau de performance indépendant de l'orientation partisane qu'ils expriment. Lerner & Yvon (2026) proposent ainsi une analyse comparative des performances en traduction automatique[3] de plusieurs LLM à partir des métriques automatiques sBLEU (Papineni *et al.*, 2002; Chen & Cherry, 2014) et COMET (Rei *et al.*, 2022) de TA pour 420 directions de traduction et 8 partis politiques. Comme ces métriques ne sont pas directement comparables entre langues, ils proposent d'agréger les scores de chaque parti sur toutes les paires de langues en utilisant la méthode Borda (McLean, 2019). Chaque paire de langues fournit un classement des K partis basé sur leur score sBLEU (ou COMET). Le dernier parti obtient 0 point, l'avant-dernier 1 point, etc. jusqu'au premier qui obtient $K-1$ points. Ces scores sont ensuite moyennés sur toutes les paires de langues pour obtenir un score final. Un modèle idéalement équitable classe au même rang chaque parti dans chaque paire de langues et atteint donc un score de $\frac{K-1}{2}$.

Pour remonter à la source de ces biais, nous comparons les performances de traduction avec celles mesurées plus directement par des métriques de vraisemblance (section 3.2), dans leur version conditionnelle et inconditionnelles, et pour les modèles pré- et post-entraînés. Ces mesures sont réalisés sur les textes originaux, sans nécessiter de traduction supplémentaire. Nous réalisons les mesures avec des paires de LLM (pré-LLM, post-LLM) de trois familles de modèles multilingues : Gemma-3-4B (Team, 2025), Qwen3 (Yang *et al.*, 2025), en deux tailles (4B et 8B), et Llama-3.1 (Grattafiori *et al.*, 2024), également en deux tailles (8B et 70B). Les différentes tailles permettent d'évaluer l'effet de la taille du modèle sur l'équité, tandis que la diversité de familles de modèles permet de s'assurer que les résultats sont robustes. Comme dans Lerner & Yvon (2026), EuroLLM (Martins *et al.*, 2024) et Salamandra (Gonzalez-Agirre *et al.*, 2025) n'ont pas été considérés car ils sont entraînés sur EuroParl (Koehn, 2005), qui se chevauche avec 21-EuroParl.

3.2 Log-vraisemblance et perplexité

Étant donné un texte t constitué d'une séquence de L tokens $(t_1, t_2, \ldots, t_L)$, où t_i est un symbole appartenant à un vocabulaire fini (par exemple, de mots, sous-mots ou caractères), un (grand) modèle de langue définit une distribution de probabilité $P(t)$, décomposée récursivement comme un produit de termes $P(t_i|t_{<i})$, correspondant à la probabilité de t_i étant donné le contexte $t_{<i}$ formé par les tokens précédents. La log-vraisemblance négative de la séquence entière t est définie par $\mathrm{NLL}(t) = -\log_2 P(t) = \sum_{i=1}^{L} -\log_2 P(t_i|t_{<i})$, avec t_0 un token spécial marquant le début de la séquence[4]. La NLL (aussi appelée entropie croisée) correspond à la fonction objectif que les LLM minimisent lors du pré-entraînement. La NLL est mesurée en bits et s'interprète comme la quantité d'information nécessaire pour encoder t sous P. Si t est la traduction d'un texte source s, il est simple de calculer la NLL conditionnellement à la source comme $-\log_2 P(t|s)$ en faisant précéder chaque contexte $t_{<i}$[5] par s. La NLL, somme de termes positifs, augmente avec L : les textes plus longs sont moins probables que les textes plus courts. La méthode la plus courante pour en tenir compte est de calculer la moyenne de la NLL sur les L tokens. L'exponentiation de cette NLL moyenne correspond à la perplexité (PPL), une mesure standard pour évaluer les modèles de langue (Jurafsky & Martin, 2026). Cependant, comme L dépend du tokeniseur, la PPL n'est généralement pas comparable entre les

3. Réalisées par des post-LLM sans exemples contextuels « *zero-shot* ».

4. Certains modèles (par exemple, Qwen3-8B) n'ont pas de token de début de séquence. Par conséquent, la NLL est calculée à partir de $i = 2$, la perplexité est normalisée par $L - 1$, et nous décomptons les caractères de t_1 pour BPC et BPEC.

5. En pratique, nous utilisons le même prompt que Lerner & Yvon (2026) : `"<src_lang>: <src_text>\n<tgt_lang>: <tgt_text>"` pour calculer la NLL conditionnelle. Par exemple : `"English: In Europe, we have freedom of the press.\nFrench: En Europe la presse est libre."`

Modèle	Version	URL	PPL conditionnelle		PPL inconditionnelle	
			ρ sBLEU	ρ COMET	ρ sBLEU	ρ COMET
Gemma-3-4B	pré	gemma-3-4b-pt	-0,690	-0,690	-0,262	-0,429
Gemma-3-4B	post	gemma-3-4b-it	-0,524	-0,690	-0,357	-0,548
Qwen3-4B	pré	Qwen3-4B-Base	-0,833	-0,952	-0,667	-0,881
Qwen3-4B	post	Qwen3-4B	-0,643	-0,905	-0,667	-0,881
Qwen3-8B	pré	Qwen3-8B-Base	-0,833	-0,857	-0,738	-0,810
Qwen3-8B	post	Qwen3-8B	-0,690	-0,833	-0,738	-0,810
Llama-3.1-8B	pré	Llama-3.1-8B	-0,571	-0,738	-0,405	-0,548
Llama-3.1-8B	post	Llama-3.1-8B-Instruct	-0,667	-0,833	-0,429	-0,619
Llama-3.1-70B	pré	Llama-3.1-70B	-0,671	-0,643	-0,527	-0,524
Llama-3.1-70B	post	Llama-3.1-70B-Instruct	-0,635	-0,714	-0,527	-0,524

TABLE 1 – Le ρ de Spearman présente une corrélation négative entre la PPL (↓) et les scores Borda basés sur sBLEU et COMET (↑) à travers les partis politiques, à la fois pour la PPL des pré-LLM et des post-LLM. A gauche, la PPL est conditionnée sur la source. À droite, la PPL est calculée sur le seul texte cible t, sans conditionnement par le texte source.

modèles (Jurafsky & Martin, 2026). Une manière de rendre ces mesures indépendantes du tokeniseur consiste à normaliser la NLL par le nombre de caractères $\mathrm{C}(t)$ du texte, correspondant au nombre de bits par caractère (BPC, Sutskever *et al.*, 2011). Cependant, comme le soulignent Cotterell *et al.* (2018), le BPC est sensible aux artefacts orthographiques (par exemple, le mot /putʃ/ s'écrit en 3 caractères « *puč* » en tchèque mais en 6 caractères « *putsch* » en allemand). Disposant d'un jeu de données parallèles, Cotterell *et al.* (2018) calculent le nombre de bits par caractère anglais (BPEC) en normalisant la NLL par le nombre de caractères $\mathrm{EC}(t)$ dans la traduction anglaise de t. Nous présenterons les résultats en utilisant ces trois métriques, implémentées avec 🤔 `pplm`[6].

4 Jeu de données parallèles

4.1 21-EuroParl

21-EuroParl (Lerner & Yvon, 2026) est un sous-ensemble multi-parallèle dérivé du jeu de données EuroParl en sélectionnant et alignant des débats du Parlement européen des années 2009-2011. Il comprend 72 234 exemples, où chaque exemple est une phrase traduite dans $M = 21$ langues : BUL, CES, DAN, DEU, ELL, ENG, EST, FIN, FRA, HUN, ITA, LAV, LIT, NLD, POL, POR, RON, SLK, SLV, SPA et SWE (ISO 639-3). Chaque exemple est annoté avec des métadonnées politiques détaillées, incluant l'affiliation partisane du locuteur, distribuée sur $K = 8$ groupes possibles (voir le tableau 2). Nous menons toutes les expériences sur l'ensemble de test (défini dans Lerner & Yvon, 2026) de $N = 23\,386$ exemples.

6. https://github.com/PaulLerner/pplm

Modèle	Version	NGL	S&D	EFA	ALDE	PPE	ECR	EFD	NI
Gemma-3-4B	pré	41,0	36,5	43,5	42,2	36,5	45,1	37,6	45,1
Gemma-3-4B	post	174,0	157,3	198,6	182,9	153,1	197,2	158,5	210,5
Qwen3-4B	pré	19,6	17,4	20,4	19,4	17,2	20,7	17,8	21,3
Qwen3-4B	post	35,1	31,0	36,9	35,0	30,5	37,5	31,8	39,2
Qwen3-8B	pré	16,0	14,2	16,6	15,9	14,2	16,9	14,6	17,2
Qwen3-8B	post	24,5	21,7	25,7	24,4	21,5	26,1	22,3	26,8
Llama-3.1-8B	pré	27,7	25,2	30,1	29,1	25,1	30,2	25,4	30,8
Llama-3.1-8B	post	32,8	30,0	36,0	34,6	29,8	35,9	30,1	36,9
Llama-3.1-70B	pré	19,7	17,7	21,0	20,6	17,8	21,6	17,9	21,7
Llama-3.1-70B	post	22,7	20,4	24,4	23,8	20,4	24,9	20,5	25,2

TABLE 2 – PPL (↓) des pré-LLM et post-LLM pour les huit partis politiques de 21-EuroParl. La PPL est calculée sur le texte cible t, sans contexte, notamment sans le texte source. La PPL n'est *pas* comparable entre les familles de modèles (par exemple, entre Qwen3 et Llama-3.1). Chaque ligne est colorée du magenta (PPL la plus haute, i.e. la pire) au vert (la plus basse, i.e. la meilleure). Les colonnes positionnent approximativement les partis politiques de gauche à droite. NGL : Groupe de la Gauche ; S&D : Alliance progressiste des socialistes et démocrates ; EFA : Groupe des Verts/Alliance libre européenne ; ALDE : Groupe de l'Alliance des démocrates et des libéraux pour l'Europe ; PPE : Groupe du Parti populaire européen ; ECR : Groupe des conservateurs et réformistes européens ; EFD : Groupe Europe de la liberté et de la démocratie ; NI : Non-inscrits.

4.2 Corrélation entre la perplexité et les métriques de traduction

Le tableau 1 rapporte la corrélation ρ de Spearman entre la PPL conditionnée sur la source et les scores Borda basés sur sBLEU et COMET pour tous les partis politiques de 21-EuroParl. La corrélation est négative, à la fois pour les pré-LLM et leur équivalent post-LLM : lorsqu'un modèle donne une faible probabilité à la traduction de référence d'un texte donné, il est peu probable qu'il la génère exactement, ce qui conduit à de mauvais scores sBLEU et COMET. Le résultat est cohérent pour les deux métriques mais la corrélation est plus forte avec COMET. Le résultat vaut également avec le BPC et le BPEC conditionnés sur la source (voir l'annexe B).

Nous prolongeons cette analyse avec la corrélation ρ de Spearman entre la PPL (inconditionnelle) du texte cible et les scores Borda basés sur sBLEU et COMET, à travers les partis politiques (tableau 1). La corrélation, bien que plus faible, est également négative. Cela signifie que lorsqu'un modèle assigne une faible probabilité au texte d'un parti politique, même en dehors de tout contexte de traduction, il est peu probable qu'il le génère, y compris dans un contexte de traduction, conduisant à un mauvais score sBLEU (ou COMET). Le résultat est cohérent pour les deux métriques mais la corrélation est plus forte avec COMET. De même, cela vaut pour le BPC et le BPEC (voir l'annexe B).

4.3 Équité du pré-entraînement sur un jeu de données parallèles

Le tableau 2 rapporte enfin la PPL des pré-LLM et post-LLM pour les huit partis politiques de 21-EuroParl. Les classements des partis selon la PPL des pré-LLM sont cohérents avec ceux de leur équivalent post-LLM utilisé pour la TA : le PPE et le S&D obtiennent la meilleure PPL (i.e. la plus basse), tandis que les petits partis comme NI obtiennent la pire PPL (i.e. la plus haute). Ces

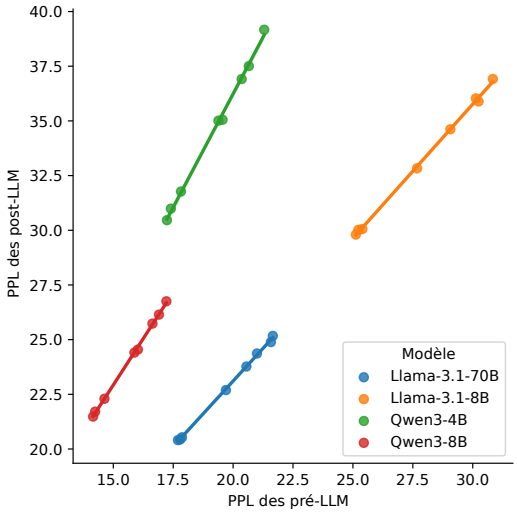

FIGURE 2 – PPL (↓) des pré-LLM contre celle des post-LLM. Pour chaque modèle, chaque point représente un parti politique de 21-EuroParl. La PPL est calculée sur le texte cible t, sans contexte, notamment sans le texte source. La PPL n'est *pas* comparable entre les familles de modèles (par exemple, entre Qwen3 et Llama-3.1).

observations sont stables à travers toutes les familles et tailles de modèles. La corrélation entre la PPL des pré-LLM et leur équivalent post-LLM est illustrée à la figure 2 (Gemma est exclu de la figure pour améliorer la lisibilité, en raison de sa haute PPL, comme indiqué dans le tableau 2). De façon générale, les pré-LLM ont une PPL plus faible que les post-LLM, car les premiers, contrairement aux seconds, sont explicitement optimisés en minimisant ce critère. L'annexe B.2 rapporte également les métriques BPC et le BPEC, qui montrent des résultats similaires.

Comme Lerner & Yvon (2026), nous ne trouvons pas de biais politique graduel (par exemple, de gauche à droite) mais plutôt que certains partis politiques, à savoir la gauche traditionnelle S&D et la droite traditionnelle PPE, sont favorisés par rapport aux petits partis, de la gauche radicale (NGL) et de l'extrême droite (NI) mais aussi du centre (EFA et ALDE).

5 Jeux de données comparables

5.1 Motivation

Lerner & Yvon (2026) ont utilisé la traduction comme moyen d'évaluer l'équité politique des LLM, mais sont limités à un seul jeu de données, 21-EuroParl, car les textes parallèles annotés avec des métadonnées politiques sont rares. Dans la section 4.2, nous avons constaté que les performances de traduction d'un LLM donné, mesurées par sBLEU et COMET, étaient négativement corrélées à la PPL des textes (ainsi qu'au BPC et au BPEC). Cette constatation nous permet d'étendre l'évaluation de l'équité des LLM aux corpus monolingues, plus abondants. Cette équité monolingue peut ensuite être extrapolée à l'équité de la traduction automatique, grâce à la corrélation trouvée.

5.2 Manifesto et Parlamint

Chaque sous-corpus monolingue de Manifesto couvre les programmes, segmentés en phrases, de dizaines de partis nationaux (Merz *et al.*, 2016). Nous utilisons les textes de 2018 à aujourd'hui comme ensemble de test. Chaque sous-corpus monolingue de Parlamint contient des débats de parlements nationaux, annotés avec le parti du locuteur (Erjavec *et al.*, 2024). Nous utilisons les textes de 2022 à aujourd'hui comme ensemble de test.

Contrairement à 21-EuroParl, nous avons constaté que, pour chaque langue, les textes des partis avaient des distributions très différentes en nombre de caractères. Cela influencerait nos résultats car le BPC (section 3.2) normalise la NLL par le nombre de caractères $\mathrm{C}(t)$ d'un texte t. Bien que cette normalisation soit censée équilibrer la formulation de la NLL (section 3.2), nous observons une corrélation négative entre $\mathrm{BPC}(t)$ et $\mathrm{C}(t)$, à travers tous les modèles, toutes les langues et les deux jeux de données. Pour corriger ce problème, nous avons stratifié chaque sous-corpus monolingue de sorte que chaque parti ait une distribution similaire en nombre de caractères. La distribution cible est le nombre minimum de textes par intervalle sur tous les partis. Par conséquent, nous n'avons considéré que les partis ayant au moins 100 textes. Nous utilisons 100 intervalles à échelle logarithmique. L'algorithme de stratification est donné dans l'annexe A. Les langues suivantes sont donc exclues car elles comptent moins de deux partis avec au moins 100 textes chacun : GLG et LAV dans Manifesto ; FIN, ISL, RUS et SRP dans Parlamint. Le nombre final de partis et le nombre total de textes pour chaque langue des deux jeux de données sont rapportés dans l'annexe A.

5.3 Équité du préentraînement des partis nationaux

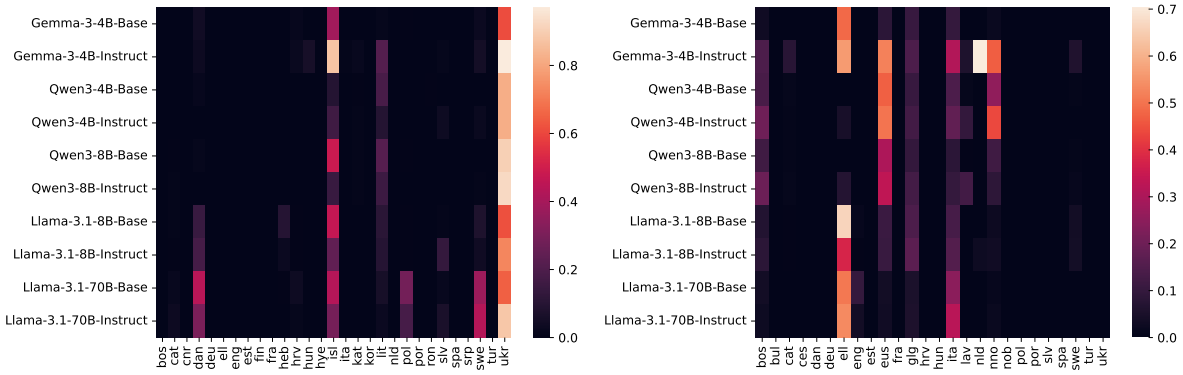

FIGURE 3 – Analyse de Kruskal-Wallis de la variation du BPC en fonction de la variable parti, montrant les p-valeurs à travers les langues (colonnes) et les modèles (lignes) sur le sous-ensemble stratifié par caractères de Manifesto (gauche) et Parlamint (droite). Les couleurs foncées indiquent une p-valeur faible, donc un test statistiquement significatif selon le seuil fixé.

Comme Manifesto et Parlamint ne sont pas des corpus parallèles, mais seulement des corpus comparables, chacun de leurs sous-corpus monolingues contient des textes de dizaines de partis politiques, dont les scores ne peuvent pas être agrégées (par exemple, *Partido Socialista* et *Partido Social Democrata* en POR et *La France insoumise* et *Renaissance* en FRA). Par conséquent, pour évaluer si, dans chaque langue, certains partis politiques sont modélisés de manière inéquitable, nous menons une analyse de Kruskal-Wallis sur le BPC des sous-ensembles stratifiés par caractères de Manifesto et Parlamint. Kruskal-Wallis est une version non paramétrique de l'ANOVA (Kruskal & Wallis, 1952).

Les p-valeurs des tests de Kruskal-Wallis sont représentées dans la figure 3. Nous constatons que, à travers les jeux de données et les modèles, la variance du BPC de la plupart des langues est significativement expliquée par la variable parti ($p < 0.01$). Cela signifie que les textes de certains partis politiques sont modélisés de manière inéquitable par les LLM. Comme précédemment, nous constatons que ce résultat est cohérent à travers les stades d'entraînement (préentraînement ou post-entraînement), les tailles et les familles de modèles.

Suite aux résultats de la section 4.2, nous nous attendons à ce que la traduction basée sur les LLM des textes de ces partis soit également systématiquement inéquitable. Les quelques langues pour lesquelles le test de Kruskal-Wallis n'est souvent pas statistiquement significatif ($p > 0.01$) à travers les modèles disposent de peu de données après stratification : seulement 63 textes par parti pour ISL et UKR dans Manifesto, et moins de 150 textes par parti pour ELL, EUS, ITA et NNO dans Parlamint (voir l'annexe A).

Savoir que les partis politiques sont modélisés de manière inéquitable dans la plupart des langues et dans les deux jeux de données n'est pas entièrement satisfaisant. Nous cherchons à savoir *quels* partis sont *bien* ou *mal* modélisés, de manière similaire à la section 4.3. C'est pourquoi nous consacrons la prochaine section au portugais (POR), bien couvert à la fois par Manifesto et Parlamint.

5.4 Étude de cas du portugais

Manifesto et Parlamint comprennent des centaines de textes pour chacun des neuf partis politiques portugais (tableau 3). Étant donné que chaque jeu de données a été stratifié de sorte que les textes de ses partis politiques aient une distribution comparable en nombre de caractères, leur NLL est comparable et nous la rapportons dans le tableau 3 (car elle est plus lisible que le BPC).

Nous constatons que les discours de BE (gauche radicale) et CH (extrême droite) ont la NLL la plus élevée (i.e. la pire), sur les sous-ensembles stratifiés par caractères de Manifesto et Parlamint. À l'opposé, les discours de PS (gauche) ont la NLL la plus basse (i.e. la meilleure). Ces résultats sont cohérents avec ceux rapportés dans la section 4.3 sur 21-EuroParl et ceux de Lerner & Yvon (2026) pour la traduction automatique ; ils sont également cohérents à travers les stades d'entraînement des modèles (préentraînement ou post-entraînement), les tailles et les familles de modèles. Nous nous attendons donc à ce que les discours du PS seront mieux traduits que ceux de BE et CH.

5.5 Équité du préentraînement des familles de partis

Tandis que la section précédente se concentrait sur les partis nationaux d'un seul pays, nous élargissons ici notre étude en agrégeant les métriques des partis nationaux appartenant à la même famille (de gauche à nationaliste, voir le tableau 4). Cependant, contrairement à 21-EuroParl, Manifesto n'est pas un corpus parallèle, seulement un corpus comparable. Par conséquent, nous ne pouvons pas calculer directement le BPC d'une famille de partis donnée, disons CON (conservateurs), en concaténant ses manifestes dans chaque langue, car les familles de partis ne sont pas également représentées dans chaque langue. À la place, nous utilisons la même agrégation par classement de Borda que décrite dans la section 3.1 car chaque langue fournit un classement des K familles de partis basé sur leur BPC. En cherchant à maximiser K, le nombre de familles de partis considérées, ainsi que le nombre de langues de vote, nous nous limitons aux 10 langues suivantes qui couvrent chacune des $K = 5$ plus grandes familles de partis dans Manifesto (listées dans le tableau 4) après stratification par caractères :

Modèle	Version	Données	BE	PCP	PS	SDP	L	PAN	CDS	IL	CH
Gemma-3-4B	pré	Parlamint	667	633	636	658	687	660	686	667	684
Gemma-3-4B	pré	Manifesto	136	135	128	134	133	126	133	134	155
Gemma-3-4B	post	Parlamint	901	869	902	914	923	894	924	905	940
Gemma-3-4B	post	Manifesto	184	184	172	181	175	172	180	179	201
Qwen3-4B	pré	Parlamint	687	657	656	681	705	680	702	691	707
Qwen3-4B	pré	Manifesto	127	125	118	126	119	116	123	126	144
Qwen3-4B	post	Parlamint	797	760	770	794	820	791	811	801	823
Qwen3-4B	post	Manifesto	148	144	137	148	136	136	141	145	163
Qwen3-8B	pré	Parlamint	647	617	612	637	668	642	664	650	666
Qwen3-8B	pré	Manifesto	121	119	113	120	114	111	118	121	138
Qwen3-8B	post	Parlamint	724	687	695	717	749	721	744	727	748
Qwen3-8B	post	Manifesto	135	130	124	134	125	124	129	134	150
Llama-3.1-8B	pré	Parlamint	656	622	615	643	678	649	673	660	675
Llama-3.1-8B	pré	Manifesto	136	134	128	137	130	127	134	137	152
Llama-3.1-8B	post	Parlamint	688	651	647	674	709	680	706	690	706
Llama-3.1-8B	post	Manifesto	143	142	135	144	135	132	140	143	161
Llama-3.1-70B	pré	Parlamint	574	539	521	556	598	570	596	579	592
Llama-3.1-70B	pré	Manifesto	125	120	117	126	122	113	124	127	141
Llama-3.1-70B	post	Parlamint	600	564	547	581	626	596	622	605	619
Llama-3.1-70B	post	Manifesto	130	124	121	130	125	119	128	132	145

TABLE 3 – NLL (↓) des textes portugais, moyennées sur Manifesto et Parlamint, en bits. Chaque jeu de données est stratifié séparément pour la distribution du nombre de caractères dans les textes. Par conséquent, les chiffres ne sont pas directement comparables entre les lignes adjacentes. Chaque ligne est colorée du magenta (NLL la plus haute, i.e. la pire) au vert (la plus basse, i.e. la meilleure). Les colonnes positionnent approximativement les partis politiques de gauche à droite. BE : Bloco de Esquerda ; PCP : Partido Comunista Português ; PS : Partido Socialista ; SDP : Partido Social Democrata ; L : Livre ; PAN : Pessoas-Animais-Natureza ; CDS : Centro Democrático Social - Partido Popular ; IL : Iniciativa Liberal ; CH : Chega.

DAN, DEU, ENG, FIN, HEB, ITA, POR, SLV, SPA et SWE.

Les résultats sur le sous-ensemble stratifié par caractères de Manifesto sont rapportés dans le tableau 4. Comme précédemment, nous constatons que les partis NAT (nationalistes et d'extrême droite) ont le plus souvent le pire BPC tandis que les partis SOC (social-démocrates) ont souvent le meilleur BPC. La cohérence entre langues est frappante : par exemple, pour Llama-3.1-70B-pré, NAT a un score de 0,4 (c'est-à-dire presque toujours classé dernier à travers les langues) tandis que SOC a un score de 3,2 (c'est-à-dire presque toujours classé premier à travers les langues). Nous ne pouvons pas pousser l'analyse plus loin faute d'avoir accès aux données d'entraînement de ces modèles, ni même la proportion des langues sur lesquelles ils ont été entraînés. De nouveau, les résultats sont largement cohérents à travers les tailles de modèles, les stades d'entraînement et les familles de modèles.

6 Discussion

Nous avons constaté que la PPL, le BPC et le BPEC d'un pré-LLM donné étaient négativement corrélés aux métriques de traduction en aval de son équivalent post-LLM, sur des textes de partis

Modèle	Version	LEF	SOC	LIB	CON	NAT
Modèle équitable	–	2,0	2,0	2,0	2,0	2,0
Gemma-3-4B	pré	2,0	3,3	1,7	2,5	0,5
Gemma-3-4B	post	1,6	3,0	1,8	3,1	0,5
Qwen3-4B	pré	1,9	3,0	1,8	2,6	0,7
Qwen3-4B	post	1,5	3,3	1,9	2,7	0,6
Qwen3-8B	pré	1,8	3,0	1,9	2,6	0,7
Qwen3-8B	post	1,9	3,2	2,0	2,5	0,4
Llama-3.1-8B	pré	2,2	3,0	1,6	2,5	0,7
Llama-3.1-8B	post	2,0	3,0	1,5	2,7	0,8
Llama-3.1-70B	pré	2,1	3,3	1,8	2,4	0,4
Llama-3.1-70B	post	2,1	3,1	1,7	2,5	0,6

TABLE 4 – Scores Borda des familles de partis politiques basé sur le BPC (de 0 à 4, plus c'est élevé mieux c'est), moyenné sur 10 langues dans Manifesto. LEF : Partis socialistes ou autres partis de gauche, SOC : Partis social-démocrates, LIB : Partis libéraux, CON : Partis conservateurs, NAT : Partis nationalistes et d'extrême droite.

politiques (section 4.2). C'est-à-dire que si un pré-LLM prédit mal le texte d'un parti politique donné, il sera peu susceptible de le générer, de sorte que la traduction sera de mauvaise qualité, même après post-entraînement. Ces résultats suggèrent que les biais politiques des LLM proviennent de leur préentraînement, et que ces biais sont peu affectés par le post-entraînement. Cette constatation ouvre des pistes pour de futurs travaux sur le post-entraînement.

Forts de cette constatation, nous avons ensuite évalué l'équité des pré-LLM et des post-LLM à travers les mêmes métriques issues de la théorie de l'information, d'abord sur un jeu de données parallèles (section 4.3), puis sur des jeux de données comparables couvrant davantage de langues (section 5). Dans les deux cas, nous avons constaté, contrairement aux travaux connexes basés sur des questionnaires qui détectent une tendance à pencher « à gauche » (section 2), que le biais politique des LLM n'est pas graduel. Nous constatons plutôt que les partis sociaux-démocrates sont favorisés par rapport aux partis d'extrême droite et nationalistes. Cependant, la gauche radicale et le centre sont également mal modélisés. Notre analyse politique est donc plus précise que les travaux antérieurs et cohérente avec l'étude sur l'équité de la traduction automatique de Lerner & Yvon (2026).

Enfin, le multilinguisme est au cœur de notre travail, notre analyse ayant été menée sur 420 paires de langues (§4.2), 21 langues (§4.3), et enfin 36 langues (§5). Comprendre comment les langues interagissent dans un seul modèle est une question fondamentale pour le TAL multilingue. Plusieurs travaux soulignent l'influence de l'anglais (comme langue majoritaire dans les données d'entraînement) et l'existence de neurones indépendants de la langue dans les couches intermédiaires du modèle (Tang *et al.*, 2024; Wendler *et al.*, 2024; Guo *et al.*, 2025; Wang *et al.*, 2025). Nous avons constaté que le biais politique d'un LLM donné était très cohérent à travers les langues. Cela suggère que les traits politiques des textes sont partagés entre toutes les langues au sein du modèle, plutôt que dans des sous-espaces spécifiques à chaque langue. Cela plaide en faveur d'une vision des LLM multilingues comme des polyglottes ayant des convictions cohérentes à travers les langues qu'ils modélisent. Cette constatation offre de nombreuses pistes pour de futurs travaux.

Remerciements

Nous remercions les membres du comité de programme pour leurs précieux commentaires.

Cette recherche a été financée par Bpifrance dans le cadre du projet « AI For Democracy - Democratic Commons », l'un des sept lauréats de l'appel à projets « Digital Commons for Generative AI » lancé par Bpifrance dans le cadre du plan d'investissement France 2030.

Nous remercions tous nos collègues du projet « AI For Democracy - Democratic Commons ».

Références

BLODGETT S. L., BAROCAS S., DAUMÉ III H. & WALLACH H. (2020). Language (Technology) is Power : A Critical Survey of "Bias" in NLP. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5454–5476, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.485.

BOELAERT J., COAVOUX S., ÉTIENNE OLLION, PETEV I. & PRÄG P. (2025). Machine bias. how do generative language models answer opinion polls ?1. *Sociological Methods & Research*, **54**(3), 1156–1196. DOI : 10.1177/00491241251330582.

CERON T., FALK N., BARIĆ A., NIKOLAEV D. & PADÓ S. (2024). Beyond Prompt Brittleness : Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, **12**, 1378–1400. DOI : 10.1162/tacl_a_00710.

CERON T., NIKOLAEV D., STAMMBACH D. & NOZZA D. (2025). What is the political content in LLMs' pre- and post-training data ? In *EurIPS 2025 Workshop on Private AI Governance*.

CHEN B. & CHERRY C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In O. BOJAR, C. BUCK, C. FEDERMANN, B. HADDOW, P. KOEHN, C. MONZ, M. POST & L. SPECIA, Éds., *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 362–367, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3346.

COTTERELL R., MIELKE S. J., EISNER J. & ROARK B. (2018). Are All Languages Equally Hard to Language-Model ? In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 536–541, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2085.

DUCEL F., NÉVÉOL A. & FORT K. (2024). La recherche sur les biais dans les modèles de langue est biaisée : état de l'art en abyme. *Revue TAL : traitement automatique des langues*, **64**(3), 119.

DURMUS E., NGUYEN K., LIAO T., SCHIEFER N., ASKELL A., BAKHTIN A., CHEN C., HATFIELD-DODDS Z., HERNANDEZ D., JOSEPH N., LOVITT L., MCCANDLISH S., SIKDER O., TAMKIN A., THAMKUL J., KAPLAN J., CLARK J. & GANGULI D. (2024). Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

ERJAVEC T., KOPP M., LJUBEŠIĆ N., KUZMAN T., RAYSON P., OSENOVA P., OGRODNICZUK M., ÇÖLTEKIN Ç., KORŽINEK D., MEDEN K., SKUBIC J., RUPNIK P., AGNOLONI T., AIRES J., BARKARSON S., BARTOLINI R., BEL N., CALZADA PÉREZ M., DARGIS R., DIWERSY S.,

GAVRIILIDOU M., VAN HEUSDEN R., IRUSKIETA M., KAHUSK N., KRYVENKO A., LIGETI-NAGY N., MAGARIÑOS C., MÖLDER M., NAVARRETTA C., SIMOV K., TUNGLAND L. M., TUOMINEN J., VIDLER J., VLADU A. I., WISSIK T., YRJÄNÄINEN V. & FIŠER D. (2024). ParlaMint II : Advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*. DOI : 10.1007/s10579-024-09798-w.

FENG S., PARK C. Y., LIU Y. & TSVETKOV Y. (2023). From pretraining data to language models to downstream tasks : Tracking the trails of political biases leading to unfair NLP models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11737–11762, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.656.

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

GONZALEZ-AGIRRE A., PÀMIES M., LLOP J., BAUCELLS I., DALT S. D., TAMAYO D., SAIZ J. J., ESPUÑA F., PRATS J., AULA-BLASCO J., MINA M., PIKABEA I., RUBIO A., SHVETS A., SALLÉS A., LACUNZA I., PALOMAR J., FALCÃO J., TORMO L., VASQUEZ-REINA L., MARIMON M., PARERAS O., RUIZ-FERNÁNDEZ V. & VILLEGAS M. (2025). Salamandra Technical Report. DOI : 10.48550/arXiv.2502.08489.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

GUO Y., CONIA S., ZHOU Z., LI M., POTDAR S. & XIAO H. (2025). Do large language models have an English accent ? evaluating and improving the naturalness of multilingual LLMs. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3823–3838, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.193.

HELWE C., BALALAU O. & CEOLIN D. (2025). Navigating the Political Compass : Evaluating Multilingual LLMs across Languages and Nationalities. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 17179–17204, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.883.

JURAFSKY D. & MARTIN J. H. (2026). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd édition. Online manuscript released January 6, 2026.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X : Papers*, p. 79–86, Phuket, Thailand.

KRUSKAL W. H. & WALLIS W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, **47**(260), 583–621.

LABAT L., OLLION E. & YVON F. (2026). Polyglots or multitudes ? Multilingual LLM answers to value-laden multiple-choice questions. In V. DEMBERG, K. INUI & L. MARQUEZ, Éds., *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3382–3398, Rabat, Morocco : Association for Computational Linguistics. DOI : 10.18653/v1/2026.eacl-long.156.

LERNER P. & YVON F. (2026). Assessing the Political Fairness of Multilingual LLMs : A Case Study Based on a 21-Way Multiparallel EuroParl Dataset. In S. PIPERIDIS, N. BEL, H. VAN DEN HEUVEL, N. IDE, S. KREK & A. TORAL, Éds., *Proceedings of the Fifteenth Language Resources*

and Evaluation Conference (LREC 2026), p. 246–265, Palma, Mallorca, Spain : European Language Resources Association (ELRA). DOI : 10.63317/3wwi6bzcsd86.

MARTINS P. H., FERNANDES P., ALVES J., GUERREIRO N. M., REI R., ALVES D. M., POMBAL J., FARAJIAN A., FAYSSE M., KLIMASZEWSKI M., COLOMBO P., HADDOW B., DE SOUZA J. G. C., BIRCH A. & MARTINS A. F. (2024). EuroLLM : Multilingual language models for Europe. arxiv preprint 2409.16235.

MCLEAN I. (2019). *Voting*, In *The Mathematical World of Charles L. Dodgson (Lewis Carroll)*, p. 121–140. Oxford University Press.

MERZ N., REGEL S. & LEWANDOWSKI J. (2016). The Manifesto Corpus : A new resource for research on political parties and quantitative text analysis. *Research & Politics*, **3**(2), 2053168016643346. DOI : 10.1177/2053168016643346.

MILMO D. *et al.* (2023). ChatGPT reaches 100 million users two months after launch. *The Guardian*, **3**, 1017–1054.

MOTOKI F., PINHO NETO V. & RODRIGUES V. (2024). More human than human : Measuring ChatGPT political bias. *Public Choice*, **198**(1), 3–23. DOI : 10.1007/s11127-023-01097-2.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318.

POTTER Y., LAI S., KIM J., EVANS J. & SONG D. (2024). Hidden Persuaders : LLMs' Political Leaning and Their Influence on Voters. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 4244–4275, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.244.

REI R., C. DE SOUZA J. G., ALVES D., ZERVA C., FARINHA A. C., GLUSHKOVA T., LAVIE A., COHEUR L. & MARTINS A. F. T. (2022). COMET-22 : Unbabel-IST 2022 Submission for the Metrics Shared Task. In P. KOEHN, L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, M. FREITAG, Y. GRAHAM, R. GRUNDKIEWICZ, P. GUZMAN, B. HADDOW, M. HUCK, A. JIMENO YEPES, T. KOCMI, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POPEL, M. TURCHI & M. ZAMPIERI, Éds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 578–585, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

RESNIK P. (2025). Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics*, **51**(3), 885–906. DOI : 10.1162/coli_a_00558.

REVEL M. & PENIGAUD T. (2025). AI-facilitated collective judgements. *Available at SSRN 5167340*.

RÖTTGER P., HOFMANN V., PYATKIN V., HINCK M., KIRK H., SCHUETZE H. & HOVY D. (2024). Political compass or spinning arrow ? Towards more meaningful evaluations for values and opinions in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15295–15311, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.816.

ROZADO D. (2023). The Political Biases of ChatGPT. *Social Sciences*, **12**(3), 148. DOI : 10.3390/socsci12030148.

RÖTTGER P., HINCK M., HOFMANN V., HACKENBURG K., PYATKIN V., BRAHMAN F. & HOVY D. (2026). Issuebench : Millions of realistic prompts for measuring issue bias in LLM writing

assistance. *Transactions of the Association for Computational Linguistics*, **14**, 318–340. DOI : 10.1162/TACL.a.626.

SANTURKAR S., DURMUS E., LADHAK F., LEE C., LIANG P. & HASHIMOTO T. (2023). Whose Opinions Do Language Models Reflect ? In *Proceedings of the 40th International Conference on Machine Learning*, p. 29971–30004 : PMLR.

SHAFIABADI N. & YVON F. (2026). Biases in translation : Assessing opinion distortion in machine translated texts. In S. PIPERIDIS, N. BEL, H. VAN DEN HEUVEL, N. IDE, S. KREK & A. TORAL, Éds., *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, p. 8596–8614, Palma, Mallorca, Spain : European Language Resources Association (ELRA). DOI : 10.63317/2pjio9ho8rxg.

SMALL C. T., VENDROV I., DURMUS E., HOMAEI H., BARRY E., CORNEBISE J., SUZMAN T., GANGULI D. & MEGILL C. (2023). Opportunities and risks of LLMs for scalable deliberation with polis. *CoRR*, **abs/2306.11932**.

SUTSKEVER I., MARTENS J. & HINTON G. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, p. 1017–1024, Madison, WI, USA : Omnipress.

TANG T., LUO W., HUANG H., ZHANG D., WANG X., ZHAO X., WEI F. & WEN J.-R. (2024). Language-specific neurons : The key to multilingual capabilities in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5701–5715, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.309.

TEAM G. (2025). Gemma 3 Technical Report. DOI : 10.48550/arXiv.2503.19786.

TESSLER M. H., BAKKER M. A., JARRETT D., SHEAHAN H., CHADWICK M. J., KOSTER R., EVANS G., CAMPBELL-GILLINGHAM L., COLLINS T., PARKES D. C., BOTVINICK M. & SUMMERFIELD C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, **386**(6719), eadq2852. DOI : 10.1126/science.adq2852.

WANG M., ADEL H., LANGE L., LIU Y., NIE E., STRÖTGEN J. & SCHUETZE H. (2025). Lost in multilinguality : Dissecting cross-lingual factual inconsistency in transformer language models. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5075–5094, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.253.

WENDLER C., VESELOVSKY V., MONEA G. & WEST R. (2024). Do Llamas Work in English ? On the Latent Language of Multilingual Transformers. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15366–15394, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.820.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C. *et al.* (2025). Qwen3 technical report. *arXiv preprint arXiv :2505.09388*.

A Manifesto et Parlamint

Le nombre de partis et le nombre total de textes pour chaque langue de Manifesto et Parlamint après stratification par caractères sont rapportés dans le tableau 5.

L'algorithme de stratification par caractères d'un sous-corpus monolingue donné est présenté dans l'Algorithme 1.

Algorithm 1 Algorithme de stratification par caractères d'un sous-corpus monolingue donné

```
all_chars ← nombre de caractères pour chaque texte
bins ← 100 intervalles à espacement logarithmique dans all_chars
min_counts ← nombre minimum de textes par intervalle de bins sur tous les partis
for all party do
    party_counts ← compter les textes par intervalle de bins
    upscale ← min(party_counts / min_counts)
    for all bin do
        n ← min_counts[bin] × upscale
        échantillonner n textes de party[bin]
    end for
end for
```

B Autres métriques

B.1 Corrélation entre les métriques de vraisemblance et de traduction

Le tableau 6 rapporte la corrélation ρ de Spearman entre le BPC et le BPEC conditionnés sur la source et les scores Borda basés sur sBLEU et COMET à travers les partis politiques.

Le tableau 7 rapporte la corrélation ρ de Spearman entre le BPC et le BPEC (sans contexte) et les scores Borda basés sur sBLEU et COMET à travers les partis politiques.

B.2 Équité du pré-entraînement

Les tableaux 8 et 9 rapportent respectivement le BPC et le BPEC des pré-LLM et post-LLM pour les huit partis politiques de 21-EuroParl.

Langue	Parlamint		Manifesto	
	# Partis	# Textes	# Partis	# Textes
BOS	4	348	6	1245
BUL	7	6477	–	–
CAT	9	348	4	672
CES	7	8382	–	–
CNR	–	–	8	348
DAN	12	4393	9	2426
DEU	5	1265	22	9384
ELL	6	819	7	683
ENG	13	8884	49	21387
EST	5	3710	5	1525
EUS	2	164	–	–
FIN	–	–	8	2745
FRA	17	1850	7	4407
GLG	3	1080	–	–
HEB	–	–	14	2644
HRV	14	1376	10	1953
HUN	10	2406	7	2991
HYE	–	–	6	1313
ISL	–	–	7	441
ITA	8	964	9	2705
KAT	–	–	7	802
KOR	–	–	4	2216
LAV	6	1284	–	–
LIT	–	–	8	1339
NLD	29	5541	24	18822
NNO	6	180	–	–
NOB	7	2070	–	–
POL	5	4266	7	1157
POR	10	16677	15	4728
RON	–	–	4	797
SLV	9	272	9	1138
SPA	17	631	48	14908
SRP	–	–	9	528
SWE	8	2009	8	1696
TUR	5	13541	6	3438
UKR	10	1820	2	126

TABLE 5 – Nombre de partis nationaux et nombre total de textes pour chaque sous-corpus monolingue stratifié par caractères de Parlamint et Manifesto.

Modèle	Version	sBLEU		COMET	
		BPC$(t\|s)$	**BPEC**$(t\|s)$	**BPC**$(t\|s)$	**BPEC**$(t\|s)$
Gemma-3-4B	pré	-0.619	-0.476	-0.786	-0.714
Gemma-3-4B	post	-0.524	-0.310	-0.690	-0.571
Qwen3-4B	pré	-0.714	-0.643	-0.952	-0.905
Qwen3-4B	post	-0.643	-0.452	-0.905	-0.762
Qwen3-8B	pré	-0.738	-0.690	-0.905	-0.833
Qwen3-8B	post	-0.690	-0.619	-0.833	-0.810
Llama-3.1-8B	pré	-0.571	-0.524	-0.738	-0.762
Llama-3.1-8B	post	-0.667	-0.524	-0.833	-0.762
Llama-3.1-70B	pré	-0.635	-0.479	-0.714	-0.714
Llama-3.1-70B	post	-0.695	-0.515	-0.810	-0.762

TABLE 6 – Le ρ de Spearman montre une corrélation négative entre le BPC conditionné sur la source (↓) et le BPEC conditionné sur la source (↓) et les scores Borda basés sur sBLEU et COMET (↑) à travers les partis politiques, à la fois pour le BPC et le BPEC conditionnés sur la source des pré-LLM et des post-LLM.

Modèle	Version	sBLEU		COMET	
		BPC(t)	**BPEC**(t)	**BPC**(t)	**BPEC**(t)
Gemma-3-4B	pré	-0.357	-0.238	-0.548	-0.524
Gemma-3-4B	post	-0.262	-0.310	-0.500	-0.571
Qwen3-4B	pré	-0.643	-0.452	-0.905	-0.762
Qwen3-4B	post	-0.524	-0.452	-0.810	-0.762
Qwen3-8B	pré	-0.690	-0.476	-0.833	-0.643
Qwen3-8B	post	-0.571	-0.476	-0.714	-0.643
Llama-3.1-8B	pré	-0.429	-0.310	-0.619	-0.595
Llama-3.1-8B	post	-0.310	-0.310	-0.548	-0.595
Llama-3.1-70B	pré	-0.527	-0.335	-0.524	-0.595
Llama-3.1-70B	post	-0.491	-0.335	-0.595	-0.595

TABLE 7 – Le ρ de Spearman montre une corrélation négative entre le BPC (↓) et le BPEC (↓) et les scores Borda basés sur sBLEU et COMET (↑) à travers les partis politiques, à la fois pour le BPC et le BPEC des pré-LLM et des post-LLM. Le BPC et le BPEC sont calculés sur le texte cible t, sans contexte, notamment sans le texte source.

Modèle	Version	NGL	S&D	EFA	ALDE	EPP	ECR	EFD	NA
Gemma-3-4B	Base	1,03	0,99	1,06	1,04	0,99	1,05	0,99	1,06
Gemma-3-4B	Instruct	1,43	1,39	1,48	1,44	1,38	1,46	1,39	1,49
Qwen3-4B	Base	1,09	1,04	1,11	1,09	1,04	1,11	1,05	1,12
Qwen3-4B	Instruct	1,30	1,25	1,33	1,30	1,25	1,32	1,26	1,35
Qwen3-8B	Base	1,02	0,97	1,04	1,01	0,97	1,03	0,98	1,05
Qwen3-8B	Instruct	1,17	1,12	1,20	1,17	1,12	1,19	1,13	1,21
Llama-3,1-8B	Base	1,10	1,06	1,13	1,11	1,06	1,12	1,06	1,14
Llama-3.1-8B	Instruct	1,15	1,12	1,19	1,17	1,12	1,18	1,12	1,20
Llama-3.1-70B	Base	0,98	0,94	1,01	1,00	0,95	1,01	0,95	1,02
Llama-3.1-70B	Instruct	1,03	0,99	1,06	1,05	0,99	1,06	0,99	1,07

TABLE 8 – BPC (↓) des pré-LLM et post-LLM pour les huit partis politiques de 21-EuroParl. Le BPC est calculé sur le texte cible t, sans contexte, notamment sans le texte source.

Modèle	Version	NGL	SD	EFA	ALDE	EPP	ECR	EFD	NA
Gemma-3-4B	Base	1,05	1,02	1,09	1,07	1,01	1,08	1,03	1,08
Gemma-3-4B	Instruct	1,45	1,44	1,53	1,49	1,41	1,50	1,44	1,52
Qwen3-4B	Base	1,12	1,09	1,16	1,13	1,07	1,15	1,10	1,15
Qwen3-4B	Instruct	1,34	1,31	1,39	1,36	1,29	1,38	1,32	1,38
Qwen3-8B	Base	1,04	1,01	1,08	1,06	1,00	1,08	1,02	1,07
Qwen3-8B	Instruct	1,20	1,17	1,25	1,22	1,15	1,24	1,18	1,24
Llama-3.1-8B	Base	1,12	1,10	1,17	1,15	1,08	1,16	1,10	1,16
Llama-3.1-8B	Instruct	1,17	1,16	1,23	1,21	1,14	1,22	1,16	1,22
Llama-3.1-70B	Base	1,00	0,98	1,05	1,03	0,97	1,05	0,98	1,04
Llama-3.1-70B	Instruct	1,05	1,03	1,10	1,08	1,01	1,09	1,03	1,09

TABLE 9 – BPEC (↓) des pré-LLM et post-LLM pour les huit partis politiques de 21-EuroParl. Le BPEC est calculé sur le texte cible t, sans contexte, notamment sans le texte source.