Assessing the Political Fairness of Multilingual LLMs: A Case Study based on a 21-way Multiparallel EuroParl Dataset

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

Abstract

The political biases of Large Language Models (LLMs) are usually assessed by simulating their answers to English surveys. In this work, we propose an alternative framing of political biases, relying on principles of fairness in multilingual translation. We systematically compare the translation quality of speeches in the European Parliament (EP), observing systematic differences with majority parties from left and right being better translated than outsider parties. This study is made possible by a new, 21-way multiparallel version of EuroParl, the parliamentary proceedings of the EP, which includes the political affiliations of each speaker. The dataset consists of 1.5M sentences for a total of 40M words and 249M characters. It covers three years, 1000+ speakers, 7 countries, 12 EU parties, 25 EU committees, and hundreds of national parties.

Keywords: Parallel Dataset, Political Fairness, Political Biases, Large Language Models

1. Introduction

Parliamentary debates have long been used as a major source of parallel texts in the history of Machine Translation (Brown et al., 1990). Notably, the EuroParl dataset (Koehn, 2005) enabled the study of an increasing diversity of language pairs as the European Union (EU) was growing and played a key part in WMT's shared tasks (Koehn and Monz, 2006; Bojar et al., 2015; Kocmi et al., 2022). Given their content, parliamentary debates are also relevant for various fields, including political sciences (Høyland et al., 2014; Deligiaouri, 2016; Schwalbach, 2024; Sebők et al., 2025). Unfortunately, existing corpora derived from the European Parliament's debates (including the noteworthy EuroParl[1]) share two limitations (see Table 1):

- they are biparallel, i.e., they pair texts in two languages, although speeches were originally translated into n languages, which prevents any multilingual comparable analysis;
- the associated metadata is incomplete.

To fill this gap, we build on LinkedEP (Noy et al., 2017) to present 21-EuroParl, a 21-way multiparallel dataset aligned at the sentence level with extensive metadata, including information about the speaker's political affiliation (national and EU party), see example on Figure 1. The dataset contains a grand total of 72,234 instances and spans over three years, from 2009 to 2011 (included). It comprises speeches from 1000+ speakers coming from 27 countries (see Figure 2), 12 EU parties, 25 EU committees, and hundreds of national parties.

[1] Also available via OPUS (Tiedemann, 2012).

speech	2011-01-19-Speech-3-084-000
date	2011-01-19
speaker	Eumember_28251 (Jörg Leichtfried)
EU Committee	TRAN
EU Party	S&D
National party	SPÖ_Austria
Country	Austria
Bulgarian (bg)	В Европа имаме свобода на печата.
Czech (cs)	V Evropě máme svobodu tisku.
Danish (da)	I Europa har vi pressefrihed.
German (de)	**In Europa gibt es die Pressefreiheit.**
Modern Greek (1453-) (el)	Στην Ευρώπη, έχουμε ελευθερία του τύπου.
English (en)	In Europe, we have freedom of the press.
Spanish (es)	En Europa tenemos libertad de prensa.
Estonian (et)	Euroopas valitseb ajakirjandusvabadus.
Finnish (fi)	Meillä on Euroopassa lehdistönvapaus.
French (fr)	En Europe la presse est libre.
Hungarian (hu)	Európában sajtószabadság van.
Italian (it)	in Europa vige la libertà di stampa.
Lithuanian (lt)	Europoje naudojamės spaudos laisve.
Latvian (lv)	Eiropā pastāv preses brīvība.
Dutch (nl)	In Europa hebben we persvrijheid.
Polish (pl)	W Europie obowiązuje wolność prasy.
Portuguese (pt)	Na Europa, existe liberdade de imprensa.
Romanian (ro)	În Europa există libertatea presei.
Slovak (sk)	V Európe máme slobodu tlače.
Slovenian (sl)	V Evropi imamo svobodo tiska.
Swedish (sv)	I EU har vi tryckfrihet och pressfrihet.

Figure 1: A fully worked-out example of 21-EuroParl: a German (DE) source sentence is aligned to translations in 20 other languages. Metadata includes the speaker's political affiliation.

To showcase the potential of 21-EuroParl, we conduct a study on the *political fairness* of multilingual Large Language Models (LLMs). This question is critical as LLMs are used daily through chatbots by hundreds of millions of users (Milmo, 2023; Top Websites Ranking, 2025) and are increasingly embedded in democratic procedures and processes (Small et al., 2023; Tessler et al., 2024; Revel and Penigaud, 2025). However, their biases have so far mostly been assessed through

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 246–265
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Dataset	# Lang. Pairs	Sent. Align.	Multiparallel	EU Committee	EU Party	National Party	OL
LinkedEP	420	✗	✓	✓	✓	✓	✓
MPDE	42	✗	✓	✗	✗	✗	✓
DCEP	253	✓	✗	✗	✗	✗	✗
EuroParl-UdS	2	✓	✗	✓	✓	✓	✓
EuroParl v10	420	✓	✗	✗	✓	✗	✓
Europarl-ST	12	✓	✓	✓	✓	✓	✓
21-EuroParl	420	✓	✓	✓	✓	✓	✓

Table 1: 21-EuroParl compared to related parallel datasets. OL: Original Language.

simulated surveys, written in English (Feng et al., 2023; Rozado, 2023; Hartmann et al., 2023; Santurkar et al., 2023; Durmus et al., 2024; Motoki et al., 2024; Potter et al., 2024; Boelaert et al., 2025; Röttger et al., 2024; Ceron et al., 2024). We propose an alternative framing of political biases and ask the following question: *does LLM translation quality systematically vary depending on the political content of the source text?* The answer is not straightforward as existing Machine Translation metrics are not well calibrated across language pairs. We therefore introduce a novel way to aggregate translation scores across multiple language pairs and partisan speeches based on Borda counts. We find systematic differences in scores depending on the source languages, as well as the party affiliation, across several model sizes and families. In summary, our contributions are threefold:

- a new, 21-way multiparallel version of EuroParl, including the political affiliation of speakers (Section 3);
- a new Borda-based method to aggregate rankings across language pairs (Section 4);
- observations regarding the politically unfair translation of LLMs (Section 5).

We make our dataset, code, and translation outputs available for reproducibility[2].

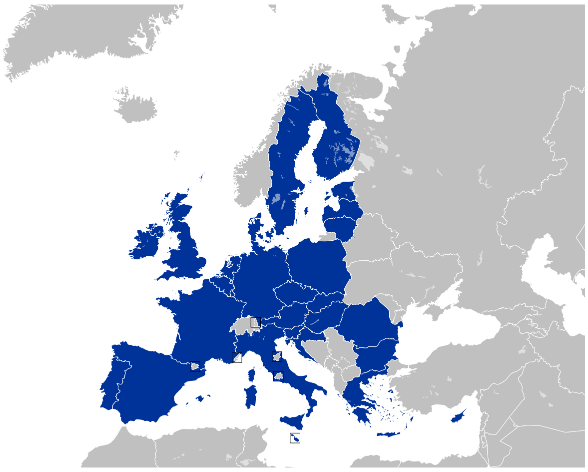

Figure 2: Countries covered in 21-EuroParl, members of the European Union in 2009-2011 (Kolja21, CC BY 3.0, via Wikimedia Commons)

2. Related Work

2.1. From Parliamentary Speeches into Multilingual Corpora

The preparation and release of EuroParl based on the processing of speeches at the EP (Koehn, 2005) has had a major impact on machine translation for European languages. Since then, multiple works have, like ours, extended EuroParl with additional information or metadata, such as the original source language or the political affiliation of speakers. For instance, CoStEP (Graën et al., 2014) includes multiple automatic linguistic annotations enabling various types of cross-lingual corpus analyses. However, they are often limited to a small number of languages, see Table 1. EuroParl-UdS (Karakanta et al., 2018) extends EuroParl with metadata about the original source language and the political affiliation of the speaker, focusing on only three languages. MPDE (Amponsah-Kaakyire et al., 2021) augments EuroParl with information about the original source language and about the use of pivot translation through a third language (e.g. FR-EN then EN-DE instead of FR-DE directly). As this information is not explicitly available, they approximate it by considering all translations from 2004 onward as performed through a pivot language. Moreover, they are limited to 7 languages and do not align texts at the sentence level. Iranzo-Sánchez et al. (2020) focus on speech translation and, like us, rely on LinkedEP to collect metadata. However, their study is also limited to a subset of four languages, dubbed Europarl-ST.

The undocumented EuroParl v10 now includes speaker metadata, notably their EU party affiliation.[3] However, it is lacking the mention of the relevant EU committee as well as the national party

[2] https://github.com/PaulLerner/21-EuroParl

[3] https://www.statmt.org/europarl/v10/

affiliation. Moreover, it is only biparallel, which prevents comparable analysis across languages.

Related to parliamentary debates, the DCEP dataset covers various documents from the European Parliament, such as press releases (Hajlaoui et al., 2014). However, to avoid overlap with EuroParl, it does not contain the verbatim reports of speeches made in the European Parliament's plenary. Moreover, it is only biparallel and does not include metadata.

Finally, Yang et al. (2023) propose Multi-EuP to assess the multilingual fairness of Information Retrieval methods, for which they consider topics of the EP as queries and relevant documents as any speech from this topic.

2.2. Political Biases of LLMs

English surveys To the best of our knowledge, this work is the first to leverage Machine Translation as a way to question the political biases of multilingual LLMs. Indeed, most existing works on political biases: (a) focus on monolingual, English-centric LLMs; (b) rely on surveys "administered" to LLMs (e.g., querying a model about a degree of agreement with statements such as "*Sex outside marriage is usually immoral*"). More precisely, we refer to the following works: (Feng et al., 2023; Rozado, 2023; Santurkar et al., 2023; Motoki et al., 2024; Potter et al., 2024).

Brittleness Following works have demonstrated the brittleness of the survey approach, as the computation of LLMs' answers is sensitive to the exact phrasing of the prompt, also yielding a lack of overall coherence (Röttger et al., 2024; Ceron et al., 2024; Boelaert et al., 2025).

Multilingual Durmus et al. (2024); Helwe et al. (2025) rely on MT to translate English surveys in multiple languages. However, this poses a confound since MT may (i) obviously contain errors, including stance-changing errors (Shafiabadi and Yvon, 2026); (ii) be biased towards a political party, as we will find out. Nevertheless, Durmus et al. (2024) find that surveying LLMs in English, Russian, Chinese, or Turkish always leads to occidental responses, i.e. the answers are consistent across languages. On the contrary, Helwe et al. (2025) finds that "language has a strong influence", although they do not control for the spurious prompt effects described above. Labat et al. (2026) have jointly assessed the variation caused by language and prompt phrasing. They found that the effect of language was question-dependent, i.e., for a subset of questions the answer is consistent across languages while for another subset the answer differs across languages.

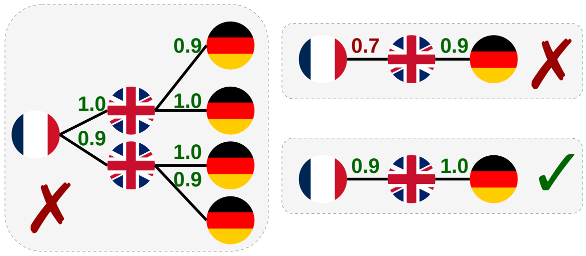

Figure 3: Illustration of various alignment scenarios: (left) a 1-FR-2-EN alignment may result in a 1-FR-4-DE alignment if each EN sentence is itself aligned to 2 DE sentences; (top, right) although EN-DE is well aligned, FR-EN is not, so we reject the multialignment; (bottom, right) valid multialignment: all bialignments are 1-1 and have a score above the threshold of 0.8

Other Methods In addition to their survey-based method, Potter et al. (2024) conduct a human-computer interaction experiment. They find that after debating with an LLM in English, Trump supporters increased their Biden-leaning. Ceron et al. (2025) study the political content of LLMs English pretraining and posttraining data. Using a left-right content classifier, they find a majority of left-leaning data, which concurs with the results discussed above. Röttger et al. (2025) assess the stance of LLMs when prompted to generate English texts about debated issues. They find that LLMs align closely with US Democrats, as opposed to Republicans.

Fairness We depart from these works and, instead of measuring political biases of LLMs based on the detected political leaning of the text they produce, we frame it in terms of a *group-fairness analysis*, based on a practical application (Machine Translation), and assess biases over a large set of 420 language pairs. This allows us to also cover a much more diverse set of texts, reflecting the fact that the political spectrum of the European Union is arguably more varied than that of the bipartisan US, which largely maps to a liberal-conservative axis (Ramaciotti et al., 2024; Vendeville et al., 2025). In the EU, other ideological dimensions emerge, e.g., Green-Alternative-Libertarian *versus* Traditional-Authoritarian-Nationalist (aka GAL-TAN), Urban-Rural, or Pro-Anti Immigration (Jolly et al., 2022).

Our methodology thus borrows from the *fairness* literature (Verma and Rubin, 2018; Czarnowska et al., 2021; Liang et al., 2023; Barocas et al., 2023; Gallegos et al., 2024), as we treat political affiliation as a *protected attribute* which models should not be sensitive to (instead of more commonly-used attributes such as gender (Savoldi et al., 2022)).

3. Preparing 21-EuroParl

3.1. LinkedEP

21-EuroParl heavily builds upon LinkedEP (Noy et al., 2017), which represents the European Parliament's debates as a knowledge graph in which speeches, speakers, and political parties are nodes linked by relations.

As LinkedEP is stored in RDF triples, we query it using SPARQL to retrieve, for each speech, all translations as well as their metadata. LinkedEP contains 191,921 speeches, partially translated into 23 languages. However, HR (Croatian) and MT (Maltese) are hardly covered[4] so we exclude them to arrive at the 21 languages of 21-EuroParl: BG, CS, DA, DE, EL, EN, ES, ET, FI, FR, HU, IT, LT, LV, NL, PL, PT, RO, SK, SL, and SV (ISO 639-1 codes).

Not all the 191,921 speeches are available in 21 languages: we found that many texts tagged as translations in LinkedEP were in fact duplicates of the source speech. After cleaning using fastText's language identification tool (Joulin et al., 2017), we were left with 37,799 speeches that were available in 21 languages. Less than a thousand were then filtered out because of missing metadata. The next important step was to compute a sentence alignment.

3.2. Sentence Alignment

Sentence alignment was performed with Bertalign (Liu and Zhu, 2023), which uses LaBSE sentence embeddings (Feng et al., 2022). After initial experiments, we replaced Bertalign's original sentence segmenter[5] by Trankit's[6] (Nguyen et al., 2021).

Bertalign computes alignments for pairs of languages (bialignments). We first performed English-centric bialignments, from English to the 20 other languages, yielding 20 bialignments per speech. To control the alignment quality and avoid one-to-many alignments propagating into one-to-many-many (see Figure 3), we only kept examples with one-to-one alignments in the 20 bialignments. Furthermore, we filtered out examples if any of their bialignment had a Bertalign score (LaBSE's cosine similarity between the source and target sentences) of less than 0.8.

The entire process yielded 72,234 examples (1,516,914 sentences) aligned in 21 languages, a reduction of about 76% from the original LinkedEP.

We will make alignments available so that the dataset can easily be extended with less conservative thresholds or one-to-many alignments.

[4]Croatian only acquired the status of an official language of the EU in 2013. Maltese and Irish were, until recently, subject to a derogatory status, likely due to a lack of trained interpreters/translators, and they were only partially supported. Ireland and Malta also have English as a co-official language.

[5]https://github.com/mediacloud/sentence-splitter

[6]github.com/nlp-uoregon/trankit

Metadata	Unique Values
Speech	27,873
Speaker	1,094
National party	272
EU party	12
EU committee	25
Country	27

Table 2: Summary of the available metadata in 21-EuroParl

Language	# Words	# Chars
BG	2,057,994	12,032,635
CS	1,778,026	10,668,881
DA	1,949,508	11,414,059
DE	2,027,728	13,216,642
EL	2,140,104	13,126,027
EN	2,088,541	11,541,038
ES	2,240,743	12,706,854
ET	1,524,824	10,738,616
FI	1,487,870	11,628,885
FR	2,213,235	13,027,959
HU	1,802,043	12,171,311
IT	2,027,404	12,546,078
LT	1,651,547	10,893,475
LV	1,682,997	10,807,953
NL	2,119,408	12,763,061
PL	1,781,796	11,921,243
PT	2,182,858	12,322,265
RO	2,102,942	12,336,509
SK	1,777,101	10,948,052
SL	1,792,083	10,570,973
SV	1,847,453	11,381,764
Total	**40,276,205**	**248,764,280**

Table 3: Total word and character counts in the 72,234 sentences for each 21-EuroParl language

3.3. Overview of the dataset

21-EuroParl spans over three years, from 2009 to 2011 (included). While we aligned the speeches at the sentence level, we kept the speech identifier of each example along with the date of the debate and the language in which the speech was originally uttered. Importantly, we also have a unique identifier for the 1000+ speakers in the dataset, along with their political affiliation: national party (and corresponding country), EU party, and EU committee (see Figure 1). The dataset covers 27 countries (see Figure 2), 12 EU parties, 25 EU committees,

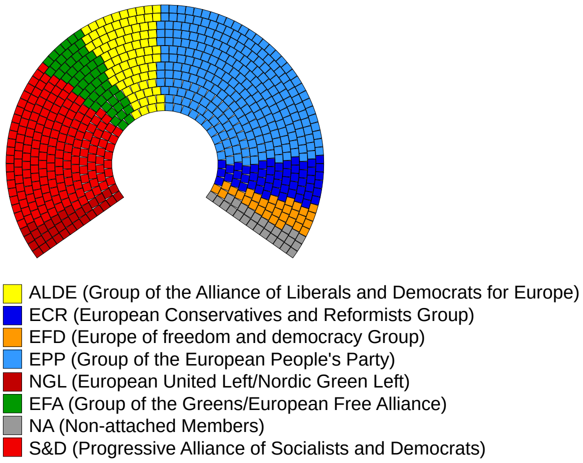

Figure 4: Composition of the 7th European Parliament, representing EU parties in 21-EuroParl, (Ssolbergj, CC BY-SA 3.0, via Wikimedia Commons)

and hundreds of national parties. These numbers are summarized in Table 2. The entire dataset contains 40,276,205 words[7] and 248,764,280 characters, a breakdown per language is in Table 3.

We split the dataset chronologically, reserving year 2009 for training, 2010 for validation, and 2011 for testing. All results in this study are computed on the test set of 23,386 examples; the split is provided to facilitate future research. The eight EU parties represented in the test set are displayed in Figure 4.

4. Experiments

4.1. Method

Translation fairness As a first study, we would like to assess whether the quality of translations provided by multilingual LLMs varies depending on the EU party, [8] which may indicate a political bias of the model. We thus generate automatic translations of each speech, considering all possible directions of the 21-way multiparallel dataset, that is $21 \times 20 = 420$ language pairs.

Metrics Translation quality is estimated with two standard reference-based metrics: sBLEU[9], i.e. per-sentence BLEU (Papineni et al., 2002; Chen and Cherry, 2014), and COMET[10] (Rei et al., 2022). It is however well-known that these metrics are not well calibrated across language pairs (Bugliarello et al., 2020; Zouhar et al., 2024) and that the scores cannot be easily compared. More precisely, COMET scores are not well calibrated across language pairs (Zouhar et al., 2024), likely because it inputs not only the hypothesis and reference, but also the source sentence. sBLEU, while also not well calibrated across target languages due e.g. to tokenizations issues, still enables to compare multiple source languages translating into the same target language (Section 5.2), as it only relies on a reference and a hypothesis in that target language.

Borda Count Therefore, to measure political fairness based on sBLEU and COMET, we compute a Borda Count of EU parties per language pair (Section 5.3). Borda Count is a simple, yet effective, way to aggregate rankings (McLean, 2019). For a given model, metric, and language pair, all EU parties are ranked according to the metric score. The EU party with the lowest score gets 0 points, the second-to-last gets 1 point, etc. until the best one gets 7 points (as there are 8 parties). While initially proposed as a voting system, Borda Counts has been used in Information Retrieval to aggregate rankings (Aslam and Montague, 2001; Bassani and Romelli, 2022). It has also been used more recently for benchmarking models across tasks with different metrics (Colombo et al., 2022) or meta-evaluating metrics across language pairs (Freitag et al., 2023). Our method differs from the usual fairness metrics, e.g. demographic parity or group calibration, that are designed for classification settings (Czarnowska et al., 2021; Barocas et al., 2023).

4.2. Implementation

We experiment with three families of posttrained LLMs: Gemma-3-4B (Team, 2025), Qwen3 (Yang et al., 2025), in two sizes (4B and 8B), and Llama-3.1 (Grattafiori et al., 2024), in two sizes as well (8B and 70B).

We originally carried out experiments with the smaller Llama-3.2 (in sizes 1B and 3B) but found the translation scores overall too poor to be reliably compared across parties (for most language pairs, sBLEU scores were below 10, see Appendix B.1). Llama-3.1 is a multilingual model trained with a very large share of English texts, accounting for 92% of the pretraining data (or 75% when excluding code) and 97% English posttraining data (or 82% when excluding code). Little is known about Gemma or Qwen3's data mix, except that Qwen3's covers 119 languages.

We use a standard prompt template for zero-shot translation: "<src_lang>: <src_text>\n<tgt_lang>:" where <src_lang> and <tgt_lang> are replaced by the English name of the source and target languages, respectively,

[7] Word tokenization is performed using spacy's multilingual tokenizer https://spacy.io/models/xx.

[8] We focus only on EU parties, given the exceedingly large number of national parties.

[9] nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.5.1 (Post, 2018)

[10] wmt22-comet-da

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	26.3	28.6	27.9	30.4	44.6	38.6	15.6	17.7	32.3	20.9	28.1	18.0	19.5	26.8	24.3	32.8	33.1	24.6	25.5	28.4
CS	27.8	–	27.2	27.0	27.6	41.2	36.3	15.4	17.4	30.8	20.5	26.7	16.7	18.1	26.4	23.9	30.7	30.5	31.2	8.2	27.3
DA	27.5	24.3	–	27.7	27.2	44.2	36.9	15.0	17.9	31.6	19.7	27.2	16.0	17.8	26.9	22.2	31.0	30.7	22.6	23.3	32.6
DE	25.1	25.0	28.7	–	26.6	43.2	37.2	15.3	17.7	31.5	20.0	27.1	16.3	17.5	28.6	22.4	30.9	31.1	23.4	23.9	28.5
EL	20.2	14.7	22.5	23.4	–	45.3	37.7	10.5	13.9	31.3	12.7	26.5	10.9	11.8	19.3	15.1	29.4	30.1	16.2	21.2	22.3
EN	37.0	32.2	37.5	36.2	38.8	–	50.1	19.0	22.2	40.4	25.2	34.6	20.4	23.9	32.3	28.7	41.7	44.3	30.3	31.2	37.5
ES	27.8	24.6	29.3	21.8	30.8	48.7	–	15.7	18.1	34.1	20.0	30.0	17.1	18.2	27.2	23.7	36.1	35.4	24.2	24.6	29.3
ET	23.9	21.7	23.9	23.7	23.6	34.9	32.5	–	17.4	27.3	18.9	23.6	15.5	16.7	23.4	20.3	26.9	26.4	20.0	19.9	24.1
FI	23.5	21.4	24.3	23.7	24.0	35.6	32.4	15.1	–	27.5	18.8	23.8	14.8	16.0	24.0	19.8	27.2	26.4	19.8	20.0	25.5
FR	25.0	23.4	27.1	26.7	28.3	42.8	38.8	13.0	16.2	–	19.6	29.0	14.9	16.0	26.8	22.2	33.0	32.0	21.8	22.8	27.2
HU	24.1	21.7	24.3	23.8	24.1	36.1	33.2	13.6	16.0	28.1	–	24.4	15.0	16.1	24.1	20.3	28.4	26.6	19.9	20.5	24.5
IT	22.7	21.9	25.4	22.0	25.6	39.9	36.2	13.6	13.5	31.5	18.1	–	14.3	15.9	25.8	21.1	31.1	29.9	21.0	20.9	25.6
LT	23.0	20.1	22.2	21.7	21.8	33.2	30.7	13.6	14.6	26.0	16.9	22.2	–	17.1	21.7	19.1	25.5	24.9	18.8	19.3	22.6
LV	24.0	21.7	24.1	23.4	23.6	35.9	32.7	14.9	15.8	27.2	18.5	23.4	17.0	–	23.1	20.5	27.2	26.6	20.0	20.5	24.0
NL	24.3	21.9	25.4	25.6	24.9	37.4	33.8	13.5	16.1	29.2	18.6	25.5	15.0	16.0	–	20.5	28.7	27.3	20.2	20.9	25.8
PL	26.2	24.0	25.3	21.7	26.1	38.1	34.6	14.4	16.2	29.4	19.3	25.5	16.2	17.3	24.7	–	29.4	28.6	22.5	22.7	25.7
PT	29.0	25.2	28.5	26.1	30.8	46.8	41.6	15.2	17.6	34.3	20.4	29.9	16.8	18.1	27.3	23.3	–	33.9	23.4	23.9	28.4
RO	29.6	25.8	29.6	25.5	32.3	48.7	42.1	15.5	17.5	34.8	21.2	30.0	17.1	19.1	27.7	23.5	35.7	–	24.3	24.7	29.7
SK	28.4	32.1	27.6	27.8	27.8	42.0	37.0	14.8	17.1	31.2	19.9	26.9	16.6	18.5	26.4	24.3	31.1	31.0	–	24.9	27.4
SL	27.7	25.7	27.1	27.2	26.9	41.2	36.6	14.5	16.6	30.7	19.4	26.4	16.8	18.0	25.9	23.3	30.4	30.4	24.7	–	27.0
SV	26.3	23.4	31.4	26.5	26.3	42.1	35.7	14.9	17.4	30.7	19.1	26.6	15.7	17.2	26.3	21.7	30.1	29.7	21.5	22.1	–

Table 4: sBLEU scores for Llama-3.1-70B: source languages in rows and target languages in columns. Scores are comparable within the same column (target language) but not across columns (within the same row, i.e. source language).

and `<src_text>` is replaced by the source sentence. For example: `"English: In Europe, we have freedom of the press.\nFrench:"`.[11] This prompt is then formatted as a user query using model-specific templates. We use the recommended decoding parameters, i.e. nucleus sampling with $p = 0.9$ and temperature $\tau = 0.6$ (Holtzman et al., 2020). Models are implemented using vLLM (Kwon et al., 2023).

A couple of thousands V100 and H100 GPU hours were used for the experiments.

5. Results

5.1. Overall Translation Quality

For the 23,386 examples of the test set, we compute sBLEU and COMET scores for the 420 language pairs, giving 9,822,120 observations for each metric. We report the average sBLEU scores of Llama-3.1-70B per language direction (source languages in rows and target languages in columns, regardless of the party) in Table 4 (results for all models are reported in Appendix B). This model overall achieves good translation performance, especially for high-resource pairs such as ES-EN where sBLEU scores exceed 40. However, results are not uniform and even for the most favorable case (translating into English), sBLEU scores vary by a wide margin. We see that, overall, larger models outperform smaller models (e.g., Llama-3.1-70B vs. -8B, or Qwen3-8B, vs. -4B); Qwen3-8B slightly outperforms Llama-3.1-8B and Gemma-3-4B outperforms Qwen3-4B; but the tendencies across language pairs are largely similar: in fact, the sBLEU scores of Llama-3.1-8B and Llama-3.1-70B have a Pearson ρ correlation of .96. Even across families, the sBLEU scores of Qwen3-8B and Llama-3.1-8B have a Pearson ρ correlation of .98. We can also see a strong correlation of the two metrics: the COMET and sBLEU scores of Qwen3-4B have a Pearson ρ correlation of .91, although this number goes down to .77 for Qwen3-8B and Llama-3.1-8B. See (Kocmi et al., 2024) on how to interpret COMET scores.

5.2. Language Fairness

Before diving into the political fairness of LLMs, we first study their *language fairness*, trying to evaluate which language do they translate the worst, and which one do they translate the best. A hasty conclusion would rely on the average of sBLEU/COMET scores across language pairs, to find which language gets the highest score when translating into/out of it. However, COMET scores are not well calibrated across language pairs (Section 4.1). sBLEU, while not well calibrated across target languages, still enables to compare multiple source languages translating towards the same target language.[12] Therefore, we apply the Borda method described above as follows: for each target

[11] For Gemma, experiments with that prompt resulted in chatbot boilerplate answers, e.g.: `"Here are a few options for the English translation, with slightly different nuances:\n\n* **Most straightforward:** I believe that [...]"`. We therefore changed the prompt to explicitly ask for the translation output only: `"You are a professional <src_lang> to <tgt_lang> translator. Please translate the following <src_lang> text into <tgt_lang>:\n<src_text>\n\nProduce only the <tgt_lang> translation, without any additional explanations or commentary."`

[12] Assuming that the source texts are similar, as is the case in this section.

Model	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
Fair Model	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5	9.5
Gemma-3-4B	15.7	10.9	12.7	11.9	13.2	18.9	16.0	2.2	2.5	10.2	1.4	7.2	0.7	4.0	5.3	6.3	12.9	16.1	12.9	9.7	8.6
Qwen3-4B	14.9	11.6	12.9	12.9	6.8	18.9	17.1	1.1	3.6	13.1	3.2	8.8	1.2	4.3	6.3	9.1	12.5	14.9	10.5	7.5	8.8
Qwen3-8B	15.4	10.9	12.0	14.0	7.3	18.9	17.2	1.5	2.6	13.2	2.7	8.5	0.2	3.5	5.6	6.7	14.4	15.4	11.6	9.4	9.0
Llama-3.1-8B	15.0	10.9	9.1	13.9	13.8	18.9	17.2	1.3	2.5	13.5	3.1	9.2	0.6	2.6	5.7	7.4	14.4	13.2	11.9	8.0	7.9
Llama-3.1-70B	15.9	11.7	12.4	12.7	4.0	18.9	15.3	3.5	3.6	10.2	4.4	5.8	1.1	4.1	5.4	7.0	14.7	16.3	12.9	10.5	9.2

Table 5: Borda Count of source language rankings averaged over the 21 target languages (from 0-19, the higher the better), based on sBLEU rankings.

language, we rank all source languages and assign 0 points to the worst, up to 19 points to the best. For example in the BG column of Table 4, EL-BG has the worst sBLEU score of 20.2 (regardless of the party), so it gets 0 points, while EN-BG has the best score of 37.0 (regardless of the party) so it gets 19 points.

Results are reported in Table 5. As expected, we find that EN is the best translated language,[13] consistently across models, followed by ES. Unlike what Table 4 would suggest (the ET column has the lowest sBLEU scores), we find that LT is the worst translated language across models, especially for Qwen3-8B where it is almost always ranked last. ET is also among the worst translated languages. A surprising result is that DE and FR, which have a special status within the EP, have a lower rank than ES, but also than BG and RO. We also observe, for all models, that all of the 21 target languages have their performance variance significantly explained by the source language variable according to a Kruskal-Wallis test (a non-parametric version of ANOVA; Kruskal and Wallis, 1952) – confirming that some source languages are more difficult to process than others (Bugliarello et al., 2020).

An ideally language-fair translation model, on the contrary, would not have its performance variance explained by the source language. It would have near identical performance for all source languages, i.e. random rankings for each target language, which would result in the same Borda Count for all source languages: 9.5.

5.3. Political Fairness

We now turn to the political fairness of LLMs. We apply the same Borda method to rank political parties across language pairs. For a given model, metric, and language pair, all EU parties are ranked according to the metric score (Section 4.1). Therefore, the language fairness (previous section) does not influence political fairness, as each language pair equally contributes to the final Borda count. This can also be done with COMET, as COMET scores are comparable within the same language pair.

Results are reported in Table 6. We do not find a gradual bias from left to right. Rather, we find that EPP, S&D, and ALDE get the best translation quality and NA and NGL the worst, across all models and metrics. The preferences are very marked, with EPP nearing 7 points (i.e. always ranked first) and NGL nearing 0 points (i.e. always ranked last), especially with COMET. The preferences are especially consistent within the same model family: the Borda Count of Llama-3.1-8B and Llama-3.1-70B have a Pearson ρ correlation of .98, for both sBLEU- and COMET-based scores. Even across model families, results are very consistent, e.g. the Borda Count based on sBLEU of Qwen3-8B and Llama-3.1-8B have a Pearson ρ correlation of .96, up to .98 for COMET-based scores. For most language pairs, the party variable significantly explains the variance of performance according to a Kruskal-Wallis test ($p < 0.01$), especially with COMET. An ideally fair model, on the contrary, would not have its performance variance explained by the party variable. It would have near identical performance for all parties, i.e. random rankings for each language pair, which would result in the same Borda Count for all parties: 3.5.

While Table 6 reveals consistent political bias towards some parties across language pairs, we also report the per-target-language sBLEU scores of Qwen3-8B for every EU party in Table 7 (see other models in Appendix C). For every model, and target language, the performance variance is significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$). Although multilingual LLMs are very English-centric, we find that their political bias follows similar tendencies across all target languages, e.g. S&D, EPP, and ALDE are often the best translated parties, and NGL and NA often the worst (which is reflected in their Borda Count in Table 6). Computing a Borda ranking from Table 7 gives very close results to Table 6,[14] i.e. our results are consistent regardless if we aggregate sBLEU scores per language pair or only per target

[13] The only target languages where it is not ranked first, across all models, are CS and SK, where it is then ranked second (and SK – resp. CS – is ranked first, the two languages being closely related).

[14] For example, in the XX-CS row of Table 7, S&D gets the highest sBLEU score of 24.18 (averaged over all source languages: CS-BG, DA-BG, etc.) so it gets 7 points, and NGL gets the worst score of 22.30, so it gets 0 points.

Model	Ranking	$p < 0.01$	NGL	S&D	EFA	ALDE	EPP	ECR	EFD	NA
Fair Model	–	0	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.50
Gemma-3-4B	sBLEU	333	1.09	4.77	3.44	5.00	4.41	4.16	3.27	1.87
Qwen3-4B	sBLEU	360	1.41	4.99	3.70	4.98	5.17	3.23	3.06	1.45
Qwen3-8B	sBLEU	363	1.18	5.22	3.55	5.13	5.17	3.02	3.08	1.65
Llama-3.1-8B	sBLEU	304	1.08	5.11	3.63	4.83	4.56	3.73	2.84	2.20
Llama-3.1-70B	sBLEU	312	1.06	5.08	3.81	4.73	4.65	3.33	3.33	2.00
Gemma-3-4B	COMET	403	0.57	5.22	1.29	5.26	6.04	4.27	3.72	1.63
Qwen3-4B	COMET	418	0.94	5.72	1.63	4.51	6.72	3.20	4.49	0.79
Qwen3-8B	COMET	419	0.80	5.82	1.47	4.82	6.64	3.41	4.02	1.03
Llama-3.1-8B	COMET	419	0.50	5.61	1.13	4.94	6.37	4.15	3.58	1.72
Llama-3.1-70B	COMET	415	0.34	5.47	1.06	5.16	5.81	4.32	3.44	2.40

Table 6: Borda Count of political parties rankings averaged over the 420 language pairs (from 0-7, the higher the better), based on sBLEU or COMET rankings. Columns approximately position political parties from left to right. The $p < 0.01$ column shows how many language pairs (out of 420) have their performance variance significantly explained by the party variable according to a Kruskal-Wallis test.

tgt	NGL	S&D	EFA	ALDE	EPP	ECR	EFD	NA
BG	25.66	26.52	26.04	25.63	26.26	26.52	26.19	24.93
CS	22.30	24.18	23.52	23.74	24.01	23.63	23.90	23.39
DA	26.17	27.05	27.54	27.69	27.04	27.46	26.42	25.95
DE	24.22	25.97	24.93	25.75	25.58	26.01	25.73	22.32
EL	25.51	27.79	26.67	27.58	27.74	26.98	26.81	25.55
EN	39.76	41.59	39.90	41.12	41.31	40.07	41.54	39.78
ES	35.80	36.72	35.02	36.03	37.48	34.94	37.11	35.26
ET	14.02	14.47	15.29	15.21	14.63	15.16	14.42	14.72
FI	16.00	16.88	17.65	17.60	16.76	17.05	16.38	16.91
FR	30.38	31.20	29.89	30.18	31.63	29.54	31.07	28.76
HU	18.09	19.48	19.87	19.67	19.51	18.62	19.11	19.49
IT	26.44	27.20	26.56	26.61	27.20	25.41	26.57	25.49
LT	15.01	16.28	16.50	16.16	15.96	16.39	15.90	16.25
LV	16.03	17.45	17.94	17.71	17.60	17.05	17.42	17.01
NL	23.98	26.36	25.37	26.62	25.63	25.76	25.39	24.86
PL	21.01	22.28	21.71	21.91	22.28	21.10	21.69	21.38
PT	28.08	31.23	30.64	31.16	31.16	29.69	31.05	30.41
RO	29.80	30.70	30.64	30.51	30.41	30.07	30.47	30.39
SK	21.71	22.73	22.13	22.57	22.66	23.04	22.08	22.06
SL	21.11	22.37	21.76	22.40	22.20	22.05	21.58	20.69
SV	25.84	27.29	27.17	27.79	27.25	27.59	26.94	26.14

Table 7: sBLEU scores of Llama-3.1-70B per target language. Columns approximately position political parties from left to right. Each target language has its performance variance significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$).

language.

6. Discussion

Our first study on 21-EuroParl reveals significant differences in translation quality of political speeches across different EU parties. This result provides a different perspective on the political biases of multilingual LLMs, favoring the majority parties from left and right over outsiders such as *European United Left/Nordic Green Left* and *Non-attached Members*. Our results are consistent across model sizes and families. Further analyses are needed to better understand the origin of these differences, possibly originating either from the pretraining or posttraining data (Gehman et al., 2020; Matic et al., 2020; Hovy and Prabhumoye, 2021; Birhane et al., 2021; Ceron et al., 2025; Resnik, 2025). Another, complementary explanation, would relate these differences to differences in topic across EU parties. Indeed, we found that both the party and topic variables had independent effects (see Appendix A.1). In any case, these results call for extreme caution when using automatic translation systems in political processes, such as multilingual deliberations (Cabrera, 2024).[15] Besides political fairness, we also found, as expected, unfair translation quality across source languages, with English and Spanish being the best translated, while Estonian and Lithuanian are the worst translated. These differences may of course be due to differences in training data but also to the morphological complexity of these languages (Marco et al., 2022). A limit of our Borda metric is that it does not take into account the magnitude of the differences between the sBLEU or COMET scores. In future work, we could add ties to the Borda ranking, based on the significance of the difference (Deutsch et al., 2021; Freitag et al., 2023).

While we focused on EU parties, future work may leverage 21-EuroParl to study the biases regarding national parties or countries. A related question is the effect of the *original language* (in which the speech was uttered), including the well-known translationese effect (Kurokawa et al., 2009; Lembersky et al., 2012; Volansky et al., 2013; Läubli et al., 2018; Toral et al., 2018; Zhang and Toral, 2019). This is challenging as parties are not equally

[15]As was done during the "Conference on the Future of Europe". Available online: `https://wayback.archive-it.org/12090/20230418091815/https://futureu.europa.eu/`.

represented in all *original* languages, so we could not apply the same Borda method. We found that excluding the translationese subset (where the *target* language is the *original* language) did not affect the results of Section 5.3. Using a two-way ANOVA, we also found that both the party and original language had independent effects (see Appendix A.2).

Finally, while we focused on LLMs, 21-EuroParl may prove useful to study other Machine Translation algorithms, e.g. exemplar-based, statistical, learning or decoding methods.

7. Limitations

Given the scale of our experiments (23,386 examples times 420 language pairs times 5 models), we had to rely on automatic metrics to assess translation quality. However, our results should be confirmed using human judgments at a smaller scale before taking any actions.

Again, because of the scale of experiments, we could not afford to run the largest models, e.g. the 235B Qwen3 or 405B Llama-3.1. However, our experiment with Llama-3.1 shows that, although the performance overall improves from 8B to 70B, the "political bias" stays consistent.

21-EuroParl covers three years, from 2009 (as Bulgaria and Romania only joined EU in 2007, along with their respective languages) to 2011 (included). However, it could easily be extended, provided an update of LinkedEP. One concern with evaluating LLMs on data from 2011 is that the data may have leaked into their training set. In fact, we *did not* experiment with translation-savvy models such as Tower (Alves et al., 2024) nor EuroLLM (Martins et al., 2024) because they were trained using WMT's EuroParl. However, we find that only a very small portion of the automatic translations exactly match the reference (i.e. have an sBLEU of 100), e.g. 0.2% for Qwen3-8B, which suggests very little contamination.

More generally, we emphasize that our study is limited to a domain, proceedings of parliamentary debates, and a context, the European Union. Our findings may differ given a different domain (e.g. social media), political context, or different languages.

8. Ethical considerations

We study the variation of translation quality according to the political affiliation of European parliamentarians in order to assess the political biases of LLMs. Note that the political affiliation of European parliamentarians is of public interest. Therefore, it is available and can be redistributed, as licensed by existing EU regulations (GDPR).

9. Acknowledgements

We thank the reviewers for their helpful feedback.

This research was funded by BPI-France under the project AI For Democracy - Democratic Commons, one of seven winners of BPI-France's "Digital Commons for Generative AI" call for projects, conducted as part of the France 2030 investment plan. We thank all the members of the Democratic Commons program at Sorbonne Université, Sciences Po, and make.org for the discussion and feedback on earlier versions of this work, in particular Nazanin Shafiabadi, Jean-Philippe Cointet, Salim Hafid, Léo Labat and David Mas. We also acknowledge the contribution of Hal Daumé III who visited our group in the spring of 2025. Finally, we thank Ziqian Peng for assistance with sentence alignment. This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-AD010615901).

10. Bibliographical References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F.T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*.

Javed A. Aslam and Mark H. Montague. 2001. Models for metasearch. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 275–284. ACM.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Elias Bassani and Luca Romelli. 2022. ranx.fuse: A Python Library for Metasearch. In *Proc. CIKM*, pages 4808–4812. ACM.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine Bias. How Do Generative Language Models Answer Opinion Polls? *Sociological Methods & Research*, 54(3):1156–1196.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Luis Cabrera. 2024. Babel Fish Democracy? Prospects for Addressing Democratic Language Barriers through Machine Translation and Interpretation. *American Journal of Political Science*, 68(2):767–782.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Tanise Ceron, Dmitry Nikolaev, Dominik Stammbach, and Debora Nozza. 2025. What is the political content in llms' pre-and post-training data? *arXiv preprint arXiv:2509.22367*.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. 2022. What are the best Systems? New Perspectives on NLP Benchmarking. *Advances in Neural Information Processing Systems*, 35:26915–26932.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267. Place: Cambridge, MA Publisher: MIT Press.

Anastasia Deligiaouri. 2016. Analysis of the political discourse of European parliament political groups during economic crisis (2008–2014). In *24th World Congress of Political Science*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics:*

EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Available at SSRN 4316084*.

Chadi Helwe, Oana Balalau, and Davide Ceolin. 2025. Navigating the Political Compass: Evaluating Multilingual LLMs across Languages and Nationalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17179–17204, Vienna, Austria. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Bjørn Høyland, Jean-François Godbout, Emanuele Lapponi, and Erik Velldal. 2014. Predicting Party Affiliations from European Parliament Debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 56–60, Baltimore, MD, USA. Association for Computational Linguistics.

Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75:102420.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of Machine Translation Summit XII: Papers*, Ottawa, Canada.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Léo Labat, Etienne Ollion, and François Yvon. 2026. Polyglots or multitudes? multilingual llm answers to value-laden multiple-choice questions.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Marion Weller-Di Marco, Matthias Huck, and Alexander Fraser. 2022. Modeling target-side morphology in neural machine translation: A comparison of strategies. *arXiv preprint arXiv:2203.13550*.

Pedro Henrique Martins, Patrick Fernandes, J Alves, Nuno M Guerreiro, Ricardo Rei, Duarte Miguel Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F.T Martins. 2024. EuroLLM: Multilingual language models for Europe. arxiv preprint 2409.16235.

Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. 2020. Identifying Sensitive URLs at Web-Scale. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 619–633, New York, NY, USA. Association for Computing Machinery.

I McLean. 2019. *Voting*, page 121–140. Oxford University Press.

Dan Milmo. 2023. ChatGPT reaches 100 million users two months after launch. *The Guardian*, 3:1017–1054.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Ramaciotti, Duncan Cassells, Zografoula Vagena, Jean-Philippe Cointet, and Michael Bailey. 2024. American politics in 3D: Measuring multidimensional issue alignment in social media using social graphs and text data. *Applied Network Science*, 9(1):2.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philip Resnik. 2025. Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics*, 51(3):885–906.

Manon Revel and Theophile Penigaud. 2025. Ai-facilitated collective judgements. *Available at SSRN 5167340*.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. Issuebench: Millions of realistic prompts for measuring issue bias in LLM writing assistance. *arXiv preprint arXiv:2502.08395*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3):148.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9(0):845–874.

Jan Schwalbach. 2024. Mind the context! The role of theoretical concepts for analyzing legislative text data. *Research & Politics*, 11(3):20531680241277466.

Miklós Sebők, Sven-Oliver Proksch, Christian Rauh, Péter Visnovitz, Gergő Balázs, and Jan Schwalbach. 2025. Comparative european legislative research in the age of large-scale computational text analysis: A review article. *International Political Science Review*, 46(1):18–39.

Nazanin Shafiabadi and François Yvon. 2026. Biases in translation: Assessing opinion distortion in machine translated texts. In *Proceedings of the Fifteenth International Conference on Language Resources and Evaluation (LREC'26)*.

Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. 2023. Opportunities and risks of llms for scalable deliberation with polis. *CoRR*, abs/2306.11932.

Gemma Team. 2025. Gemma 3 Technical Report.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

The Top Websites Ranking. 2025. The top 50 most visited websites for september 2025. Similarweb. Accessed on October 13th 2025. URL archived from October 10th 2025.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Antoine Vendeville, Jimena Royo-Letelier, Duncan Cassells, Jean-Philippe Cointet, Maxime Crépel, Tim Faverjon, Théophile Lenoir, Béatrice Mazoyer, Benjamin Ooghe-Tabanou, Armin Pournaki, Hiroki Yamashita, and Pedro Ramaciotti. 2025. Mapping the political landscape from data traces: multidimensional opinions of users, politicians and media outlets on X. Working paper or preprint.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

11. Language Resource References

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not Rely on Relay Translations: Multilingual Parallel Direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Johannes Graën, Dolores Batinić, and Martin Volk. 2014. Cleaning the Europarl corpus for linguistic applications.

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP -Digital Corpus of the European Parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Alina Karakanta, Mihaela Vela, and Elke Teich. 2018. Europarl-uds: Preserving and extending metadata in parliamentary debates. *ParlaCLARIN: Creating and Using Parliamentary Corpora*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Natasha Noy, Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2):271–281.

Jinrui Yang, Timothy Baldwin, and Trevor Cohn. 2023. Multi-EuP: The Multilingual European Parliament Dataset for Analysis of Bias in Information Retrieval. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 282–291, Singapore. Association for Computational Linguistics.

Model	Metric	Party	Topic	Party × Topic
Gemma-3-4B	sBLEU	307	329	340
Qwen3-4B	sBLEU	331	265	305
Qwen3-8B	sBLEU	318	275	335
Llama-3.1-8B	sBLEU	255	292	326
Llama-3.1-70B	sBLEU	288	331	375
Gemma-3-4B	COMET	405	406	398
Qwen3-4B	COMET	417	404	402
Qwen3-8B	COMET	419	412	415
Llama-3.1-8B	COMET	419	409	403
Llama-3.1-70B	COMET	411	417	407

Table 8: Number of language pairs, out of 420, for which a two-way ANOVA for the Party and Topic variables gives a statistically significant result (Fisher's test, $p < 0.01$), for sBLEU and COMET scores (the dependent variables)

A. Confounding Variables

A.1. Topic

To understand the results of Section 5.3, we wish to disentangle here the effect of the topic of the text, e.g., if NGL has worse sBLEU scores than EPP, is it because of a political bias of the LLM, or because NGL discusses different topics than EPP that are more complex or otherwise more difficult to translate? To avoid relying on a topic model, we use the EU Committee metadata, each committee being dedicated to a topic. We then conduct a two-way ANOVA for the Party and Topic variables. The results are reported in Table 8. We find that, while the topic has a significant effect, the party also has an independent effect for most language pairs. The effect is consistent for more language pairs with COMET than sBLEU, for both variables.

A.2. Original Language

We conduct the same analysis to study the effect of the *original language* (in which the speech was uttered). We assumed that the original language might have an effect because of translationese (Kurokawa et al., 2009; Lembersky et al., 2012; Volansky et al., 2013; Läubli et al., 2018; Toral et al., 2018; Zhang and Toral, 2019). However, we found that excluding the translationese subset (where the *target* language is the *original* language) did not affect the results of Section 5.3.

Results of the two-way ANOVA are reported in Table 9. As previously, we find that, while the Original Language has an effect, the party has an independent effect for most language pairs. Again, the effect is consistent across more language pairs with COMET than with sBLEU.

Model	Metric	Party	OL	Party × OL
Gemma-3-4B	sBLEU	254	308	406
Qwen3-4B	sBLEU	179	331	408
Qwen3-8B	sBLEU	238	318	412
Llama-3.1-8B	sBLEU	256	414	207
Llama-3.1-70B	sBLEU	287	414	242
Gemma-3-4B	COMET	405	409	367
Qwen3-4B	COMET	417	418	364
Qwen3-8B	COMET	419	419	387
Llama-3.1-8B	COMET	419	419	337
Llama-3.1-70B	COMET	411	420	365

Table 9: Number of language pairs, out of 420, for which a two-way ANOVA for the Party and Original Language (OL) variables gives a statistically significant result (Fisher's test, $p < 0.01$), for sBLEU and COMET scores (the dependent variables)

B. Results per Language Pair

B.1. sBLEU

We report average sBLEU results per language pair (source languages in rows and target languages in columns) for each model in Tables 10, 11, 12, 13, 14 and 15. Because the results of Llama-3.2-1B (Table 13) and Llama-3.2-3B (Table 14) are quite poor, we excluded them from the rest of the analysis.

B.2. COMET

We report COMET results per language pair (source languages in rows and target languages in columns) for each model in Tables 16, 17, 18, 19, and 20.

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	22.8	27.8	25.2	10.8	40.7	36.2	11.9	16.1	30.8	14.8	25.9	14.0	15.7	23.6	21.9	29.3	30.9	22.4	20.5	24.8
CS	28.9	–	25.8	23.9	10.3	35.9	32.8	11.3	15.3	28.3	14.3	24.0	13.1	14.4	22.5	21.2	25.7	27.4	29.6	20.2	22.8
DA	29.0	21.2	–	25.4	12.4	40.0	34.3	11.5	16.1	29.3	14.0	25.0	12.8	14.5	23.4	19.9	26.7	28.1	20.5	18.6	29.0
DE	28.7	21.7	27.1	–	11.7	38.4	34.0	11.4	15.6	28.9	14.2	24.2	12.7	14.0	24.6	19.9	26.7	28.1	21.5	19.1	24.7
EL	31.0	21.8	26.9	24.5	–	41.1	36.6	11.2	15.9	30.4	13.9	25.9	12.9	14.4	23.5	20.6	29.8	31.0	21.2	19.6	24.2
EN	38.2	27.2	35.3	31.4	13.7	–	47.3	13.5	19.2	38.5	16.5	31.9	15.4	17.8	27.7	24.9	37.4	39.8	26.5	23.8	31.8
ES	31.2	22.2	27.9	25.5	11.0	43.7	–	11.7	16.0	33.2	14.3	28.3	13.1	14.9	23.9	21.3	33.7	32.8	21.7	19.6	25.3
ET	24.2	17.5	21.4	19.7	9.0	29.3	28.2	–	15.1	23.9	12.8	20.3	11.8	13.0	19.1	17.3	22.2	22.8	17.0	15.5	19.8
FI	24.1	17.5	22.1	20.0	8.4	30.0	28.6	10.8	–	24.7	12.8	21.0	11.5	12.7	19.8	16.9	22.0	23.3	16.9	15.4	21.4
FR	29.0	20.6	26.0	23.8	9.8	38.7	37.1	10.7	14.7	–	13.8	28.2	12.5	14.0	23.4	20.1	30.3	30.2	20.0	18.3	23.4
HU	23.9	17.3	21.4	19.7	9.2	29.7	28.9	9.8	13.4	24.0	–	20.7	10.9	12.1	19.7	16.9	22.9	22.5	16.6	15.4	19.4
IT	27.2	19.6	24.3	22.0	10.8	35.7	34.6	10.4	14.0	30.9	13.1	–	11.8	13.1	22.3	19.1	28.4	27.8	18.8	16.9	22.2
LT	24.2	17.0	20.7	18.6	8.7	28.3	27.2	10.1	13.6	23.3	12.1	19.5	–	13.8	18.3	16.8	21.5	21.7	16.5	15.4	18.7
LV	25.7	18.2	22.2	20.0	8.6	31.0	29.2	10.7	14.4	24.8	12.9	20.7	13.0	–	19.5	17.8	23.1	23.3	17.3	16.1	20.0
NL	25.8	19.0	24.1	23.9	10.4	33.3	31.1	10.2	14.3	27.2	13.3	23.0	11.6	12.9	–	18.3	24.9	25.1	18.1	17.2	22.4
PL	27.4	20.9	23.8	21.6	10.5	33.3	31.5	10.8	14.1	26.9	13.4	22.8	12.5	13.6	21.2	–	25.3	26.0	20.4	18.1	21.6
PT	30.5	21.4	26.7	24.4	11.3	42.1	40.2	11.1	15.4	32.9	14.0	27.7	12.6	14.2	23.4	20.8	–	31.4	20.7	18.6	24.3
RO	32.1	22.4	28.3	26.1	10.3	43.7	39.7	11.5	16.0	33.2	14.5	27.9	13.3	14.8	24.1	21.6	32.5	–	21.7	19.8	25.4
SK	29.9	29.6	26.3	24.5	10.3	37.0	33.8	11.4	15.7	28.8	14.5	24.6	13.2	14.6	22.7	21.5	26.7	28.1	–	21.1	23.6
SL	28.7	22.8	25.5	23.6	11.1	35.8	33.0	11.0	15.1	28.0	14.0	23.5	12.9	14.2	22.2	20.2	26.1	27.3	22.8	–	22.7
SV	27.6	19.9	30.6	24.0	9.2	37.3	32.2	11.2	16.2	28.2	13.3	23.9	12.0	13.9	22.5	19.1	25.2	27.0	19.1	17.8	–

Table 10: sBLEU scores of Gemma-3-4B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	14.2	16.8	18.9	13.8	37.5	29.7	6.0	8.6	25.1	9.3	19.3	8.9	8.4	16.5	13.9	22.1	21.1	10.4	11.8	14.8
CS	17.1	–	16.0	18.6	13.0	34.0	26.7	5.8	7.8	22.1	9.3	17.7	8.1	7.7	15.5	13.8	19.6	19.3	19.7	11.1	14.6
DA	16.1	13.1	–	19.0	11.4	36.4	26.9	6.0	8.8	22.7	9.1	18.5	8.2	8.0	16.8	12.0	20.6	19.4	12.3	10.6	20.9
DE	16.5	13.9	17.7	–	11.4	37.1	28.7	6.1	8.7	23.9	9.4	18.8	8.0	7.8	18.7	12.6	21.2	20.0	4.5	11.3	16.6
EL	16.9	11.8	12.1	17.3	–	34.5	28.7	5.4	7.6	23.9	7.2	18.6	7.5	7.3	15.4	7.2	19.1	19.3	4.1	10.4	13.2
EN	22.3	17.4	23.9	25.7	18.6	–	41.6	7.2	11.1	33.2	11.2	25.6	10.0	9.9	21.1	16.7	31.4	30.0	16.0	13.5	21.9
ES	18.5	14.2	18.2	20.1	15.8	42.2	–	6.2	8.5	29.1	9.6	23.2	8.6	8.0	16.9	13.9	29.9	25.2	12.0	11.4	16.5
ET	13.1	9.6	11.8	12.8	10.1	24.4	18.2	–	6.5	16.0	7.8	12.0	7.3	6.3	11.3	10.1	14.2	13.7	7.3	6.7	11.0
FI	13.7	10.5	13.3	14.6	11.3	27.7	22.1	5.0	–	18.6	8.7	14.3	7.1	6.3	13.3	10.4	16.9	15.4	8.2	7.6	12.7
FR	17.0	13.2	16.9	19.0	12.7	37.1	32.6	6.0	7.9	–	9.3	23.1	8.1	7.7	17.0	12.9	26.0	23.0	4.5	10.6	15.6
HU	13.6	10.4	12.3	15.0	10.7	27.7	23.2	4.6	7.3	19.5	–	15.2	7.1	6.3	12.0	10.5	17.6	15.5	7.4	8.8	11.7
IT	15.8	12.4	15.6	17.1	12.7	34.2	30.3	5.7	7.2	26.3	8.7	–	7.7	7.3	16.4	12.0	23.9	21.1	4.7	10.0	14.1
LT	14.0	9.9	10.9	12.2	10.0	24.6	20.1	4.5	7.0	17.0	7.3	13.2	–	5.2	11.8	9.9	14.8	13.8	6.6	7.5	11.1
LV	14.8	10.8	12.9	14.4	11.2	27.9	22.3	4.7	7.6	19.1	8.2	14.1	8.3	–	12.7	10.7	16.9	15.7	8.0	7.0	11.7
NL	14.9	11.8	15.8	15.5	10.9	31.8	25.4	5.7	7.9	21.5	8.6	17.4	7.7	7.2	–	10.7	19.5	17.7	4.9	9.5	14.8
PL	17.0	6.5	15.3	16.7	13.0	32.0	26.7	5.6	8.3	22.6	9.1	17.9	8.1	6.0	15.4	–	20.5	19.1	13.1	10.7	14.4
PT	17.3	12.9	16.5	18.1	14.9	40.1	36.6	5.7	8.1	28.9	7.8	22.8	8.1	7.4	16.7	13.0	–	23.7	10.3	10.5	15.1
RO	18.2	13.9	17.1	19.0	16.1	41.4	33.9	6.0	8.5	28.2	8.9	21.7	8.5	7.6	16.6	13.7	25.6	–	8.9	11.4	15.7
SK	17.1	17.9	14.6	18.2	12.9	34.1	27.0	5.6	7.5	22.1	8.8	18.2	8.4	7.9	15.3	13.8	20.1	19.3	–	10.4	14.3
SL	16.6	14.7	13.4	16.6	12.2	32.2	23.2	4.8	8.1	17.7	8.8	16.4	7.9	7.5	14.4	12.9	18.9	18.4	10.3	–	13.3
SV	15.6	12.4	20.4	18.0	10.1	33.7	24.6	5.9	8.8	21.8	8.8	17.6	7.9	7.4	16.3	11.7	19.1	18.7	11.4	10.0	–

Table 11: sBLEU scores of Qwen3-4B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	18.7	21.7	22.2	20.2	40.0	33.4	8.9	12.0	27.8	13.2	22.4	13.0	13.7	20.3	17.8	26.0	26.3	18.5	16.5	20.2
CS	22.0	–	20.3	21.7	18.8	36.3	30.3	8.6	11.3	25.7	12.4	21.0	11.7	12.4	19.2	17.4	23.3	23.5	26.2	15.6	18.5
DA	21.4	17.1	–	22.9	18.4	39.7	31.5	8.8	11.9	26.9	12.7	21.4	11.4	12.5	20.1	15.8	24.0	23.4	16.7	14.6	25.4
DE	21.9	18.3	22.4	–	19.0	39.2	32.6	9.2	12.1	27.4	12.9	22.2	11.9	12.3	22.5	16.6	24.8	24.8	18.0	15.8	21.2
EL	22.6	15.8	20.0	19.7	–	35.8	32.1	6.3	10.7	26.5	11.5	21.1	9.9	11.1	19.2	15.1	22.5	22.8	16.3	14.6	18.1
EN	29.3	22.5	29.0	29.5	25.8	–	45.1	10.6	14.6	35.9	14.8	28.7	14.3	15.6	24.6	20.7	34.3	35.3	21.9	19.3	26.8
ES	24.6	18.6	22.6	23.8	22.0	44.3	–	9.3	12.3	31.7	13.0	26.1	12.6	12.8	21.2	17.5	32.4	29.4	18.5	16.5	21.0
ET	17.8	14.1	16.4	17.2	14.8	28.7	25.3	–	10.9	21.5	10.9	16.5	10.3	9.3	16.1	13.4	19.7	18.7	13.4	11.9	15.7
FI	18.0	14.2	17.5	18.0	15.6	30.2	26.7	6.9	–	22.8	11.6	17.1	9.6	9.7	17.1	13.2	20.5	19.5	13.7	12.1	17.1
FR	22.5	17.4	20.9	22.3	20.2	39.2	35.5	8.4	11.0	–	12.6	26.0	11.8	12.5	20.8	16.6	28.7	27.1	16.9	14.9	19.6
HU	18.5	14.1	17.1	18.2	15.8	30.6	27.2	7.1	10.2	22.7	–	18.0	9.8	10.1	16.8	13.9	20.2	19.6	13.3	12.4	16.1
IT	20.7	16.2	19.4	18.9	19.2	36.0	33.1	8.0	10.5	28.6	11.9	–	10.9	11.2	19.6	15.7	26.9	24.9	15.8	13.6	18.0
LT	18.4	13.7	15.9	16.1	14.7	28.6	24.8	6.8	10.0	21.2	10.2	16.1	–	7.9	15.6	13.2	18.8	18.0	13.2	11.5	15.2
LV	19.6	15.0	17.5	17.5	15.6	31.4	26.9	6.4	10.9	22.9	11.4	17.8	10.8	–	17.0	14.3	20.4	19.9	14.1	12.4	16.2
NL	19.3	15.4	19.3	21.9	16.8	33.7	28.5	8.0	10.5	25.1	11.6	20.1	10.7	11.0	–	14.7	21.9	21.2	14.6	13.2	18.8
PL	21.4	17.2	19.1	19.5	17.4	30.5	28.9	8.4	10.7	24.7	11.7	19.9	11.2	11.8	18.5	–	22.4	22.4	17.1	14.5	17.8
PT	23.8	17.6	21.4	22.5	21.2	42.8	39.4	8.8	11.8	31.6	12.5	25.9	12.0	12.1	20.6	17.4	–	28.4	17.3	15.5	20.0
RO	24.3	18.4	22.0	22.2	22.5	44.1	37.3	8.4	12.0	31.2	12.9	24.5	11.9	12.5	21.0	17.8	29.4	–	18.0	15.7	20.6
SK	22.7	24.8	20.6	22.1	19.2	37.2	31.0	8.2	11.4	26.1	12.5	21.4	11.7	12.1	19.2	17.5	23.7	24.0	–	14.6	18.6
SL	21.7	19.3	20.0	21.2	18.6	36.2	30.5	7.9	11.4	25.5	12.5	20.6	11.5	11.4	18.9	16.5	23.4	23.2	18.7	–	18.4
SV	20.6	15.9	25.7	21.7	17.2	37.5	29.6	8.8	12.1	25.7	11.9	20.7	10.9	11.6	19.6	15.3	22.5	22.6	15.3	13.8	–

Table 12: sBLEU scores of Qwen3-8B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	5.8	6.8	10.0	3.1	22.8	16.4	2.5	3.5	13.2	4.6	9.6	2.8	2.8	9.1	6.8	12.4	8.7	3.8	4.0	8.4
CS	2.0	–	5.0	11.4	1.9	24.3	16.1	2.1	3.5	13.1	4.2	8.5	2.6	2.5	8.0	6.0	11.0	6.9	5.6	3.8	7.4
DA	2.1	4.6	–	12.1	3.2	18.7	11.7	2.4	3.6	11.7	4.3	6.9	2.4	2.4	7.8	4.8	9.7	6.5	2.6	2.9	12.5
DE	2.2	4.7	5.4	–	4.3	26.5	17.1	2.2	3.9	14.2	4.9	9.2	2.5	2.1	12.7	5.6	13.8	7.4	2.5	3.1	9.9
EL	1.9	3.5	3.1	8.5	–	21.8	14.3	0.8	1.9	12.1	2.2	8.4	1.5	1.3	6.0	3.0	6.8	5.4	1.2	1.8	4.2
EN	3.4	6.8	9.8	16.4	7.0	–	24.9	2.5	5.1	22.9	6.4	16.7	2.6	2.1	12.2	8.5	19.6	15.6	1.7	4.1	15.3
ES	3.1	3.8	4.2	5.0	5.1	22.1	–	1.9	3.3	18.0	4.1	11.6	2.4	2.1	7.1	4.6	16.0	11.1	2.3	2.7	6.8
ET	1.8	2.4	2.1	3.5	1.7	6.2	3.9	–	2.7	4.2	2.7	2.1	2.2	2.3	3.5	2.2	3.4	2.3	2.1	2.2	2.4
FI	1.8	3.6	3.9	7.3	2.2	14.6	10.3	2.4	–	8.4	3.5	5.6	2.3	2.0	5.9	3.5	7.5	3.9	2.3	2.3	5.9
FR	3.1	3.7	3.6	8.0	4.9	17.9	18.7	2.0	3.3	–	4.1	14.2	2.4	2.2	8.2	5.3	14.7	11.8	2.5	2.7	7.0
HU	2.0	3.4	2.8	7.2	2.5	16.6	10.9	1.5	3.0	9.2	–	5.4	2.1	1.8	6.2	4.2	7.3	4.6	2.1	2.6	4.8
IT	2.5	4.0	4.0	7.8	4.2	22.5	19.7	1.7	2.4	17.7	4.1	–	2.3	2.0	7.5	5.7	15.0	9.4	2.1	2.6	5.7
LT	2.1	2.7	3.0	3.7	2.1	6.6	5.1	2.1	2.6	4.7	3.2	3.4	–	2.5	3.9	3.0	4.1	3.2	2.2	2.3	3.6
LV	2.0	2.6	2.4	3.7	1.9	7.7	4.2	2.1	2.4	4.8	3.0	2.9	2.3	–	3.5	2.5	3.9	3.0	2.1	2.2	3.3
NL	3.0	3.7	5.0	14.7	3.7	25.6	15.5	2.0	3.6	13.9	4.5	9.4	2.5	2.2	–	5.0	12.6	7.3	2.3	2.6	8.1
PL	2.4	3.7	4.9	10.6	2.7	23.7	17.1	2.2	3.7	14.5	4.5	10.5	2.4	2.3	8.6	–	12.6	8.1	3.2	3.2	7.9
PT	3.1	4.6	6.0	10.5	4.4	29.7	29.0	2.0	3.4	21.7	4.5	16.4	2.5	2.3	9.0	6.5	–	12.6	2.9	3.0	8.6
RO	1.9	3.6	2.7	7.4	2.4	26.4	19.2	1.1	3.2	17.9	3.6	11.5	2.0	1.7	6.6	5.3	11.8	–	1.9	2.3	5.8
SK	1.6	4.0	2.9	6.1	1.3	18.3	11.7	1.5	2.5	8.2	2.7	5.1	1.5	1.7	4.5	3.1	7.5	3.5	–	2.5	5.0
SL	2.0	3.0	2.9	5.9	2.4	12.4	7.5	1.3	2.7	6.9	3.2	5.5	2.1	2.1	4.9	3.7	5.8	3.4	2.5	–	5.3
SV	2.2	4.4	8.6	10.2	3.1	24.0	11.4	1.7	3.4	12.1	3.8	7.7	2.3	2.1	7.8	3.9	10.6	7.7	2.0	2.9	–

Table 13: sBLEU scores of Llama-3.2-1B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	9.2	9.2	14.6	4.7	29.6	23.5	3.9	5.1	19.3	6.1	15.6	3.9	3.7	13.8	8.7	13.0	8.1	4.6	6.1	11.9
CS	9.2	–	11.1	17.0	5.3	32.1	24.5	3.7	6.2	19.9	6.3	16.5	3.9	4.0	15.4	10.6	17.7	14.3	9.3	6.8	14.7
DA	7.9	8.1	–	13.3	3.8	12.1	20.3	3.5	5.6	14.6	4.8	12.1	3.9	3.8	13.8	7.7	10.6	9.2	4.4	5.7	17.3
DE	8.8	9.2	12.9	–	5.5	30.1	23.4	4.1	6.2	18.9	6.3	12.9	4.0	3.7	17.4	7.4	16.0	12.0	4.8	6.6	15.1
EL	5.6	7.9	5.9	10.3	–	29.2	21.4	2.3	3.1	12.2	1.8	12.7	1.9	2.1	9.7	5.0	11.6	3.4	2.9	2.5	6.6
EN	14.6	15.9	20.5	25.8	14.4	–	41.5	5.5	9.7	32.7	12.4	25.9	5.7	6.1	21.9	16.4	31.2	28.1	9.7	10.1	24.1
ES	9.5	9.6	9.4	14.2	7.1	26.2	–	4.2	6.3	22.4	6.9	18.0	4.1	3.8	15.6	9.4	23.3	14.7	5.3	6.3	12.6
ET	4.2	4.9	5.3	6.9	2.9	13.6	9.9	–	4.1	8.1	3.1	5.4	3.0	2.3	7.4	5.5	5.9	4.7	2.8	3.4	6.4
FI	6.1	5.7	7.8	8.9	3.6	21.7	15.5	2.6	–	12.8	4.1	7.3	2.7	2.1	10.6	5.7	9.8	6.4	3.5	3.6	11.0
FR	8.8	8.4	9.9	16.0	6.5	23.3	27.9	3.7	5.6	–	6.2	19.2	4.0	3.6	15.2	7.9	20.5	14.5	5.0	6.2	13.5
HU	5.9	6.3	5.8	11.8	2.8	20.0	18.3	3.0	4.6	14.9	–	10.2	3.1	3.2	11.8	6.7	12.5	8.4	3.0	4.5	9.6
IT	8.5	8.9	8.4	12.6	6.2	21.4	26.1	3.8	4.6	21.3	5.6	–	4.0	3.7	15.2	8.8	20.4	12.8	5.2	6.0	12.1
LT	5.4	5.0	4.3	7.5	3.7	13.7	12.8	2.8	4.1	9.5	3.2	6.6	–	2.5	8.4	5.6	7.5	5.4	2.7	3.2	6.6
LV	6.1	6.1	6.9	9.5	4.1	18.4	14.9	3.0	4.9	12.4	4.6	8.3	2.5	–	9.5	6.5	8.6	5.5	2.7	3.6	8.9
NL	8.2	6.8	8.7	15.5	5.3	21.3	20.8	3.7	5.4	17.9	6.1	11.5	3.5	3.7	–	7.6	13.9	10.0	5.0	5.3	12.5
PL	9.0	8.1	9.3	13.3	5.8	24.9	23.0	3.7	5.5	19.2	5.0	14.2	3.8	3.7	13.5	–	15.9	12.4	4.4	5.5	12.6
PT	8.9	8.8	8.7	13.9	6.2	25.8	31.0	3.7	5.5	22.1	4.9	17.7	3.8	3.5	14.8	8.9	–	11.2	4.3	5.2	10.3
RO	10.0	9.2	8.8	16.8	6.9	29.4	30.8	3.5	5.6	25.0	6.7	16.9	3.8	3.6	15.6	10.3	19.0	–	4.5	4.5	13.4
SK	8.9	9.8	9.0	13.8	5.0	28.1	22.4	3.6	5.3	17.7	5.0	13.0	3.1	3.4	13.9	8.8	14.9	11.7	–	4.7	12.4
SL	8.0	7.2	6.0	11.7	4.1	22.7	18.1	3.5	3.6	14.8	4.3	10.4	3.3	2.7	11.9	7.7	10.6	6.9	3.1	–	9.6
SV	7.4	8.4	8.2	10.4	3.0	19.5	18.5	3.6	5.9	14.6	4.3	9.1	3.7	3.2	13.8	6.2	11.0	7.6	3.3	5.1	–

Table 14: sBLEU scores of Llama-3.2-3B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	19.6	20.4	21.3	19.1	39.3	32.2	8.5	10.5	27.4	13.5	22.1	10.0	10.4	20.9	18.4	25.2	24.8	15.7	16.6	21.3
CS	19.5	–	19.4	20.9	17.6	35.7	29.8	8.2	10.0	25.2	12.5	20.7	9.0	9.2	20.0	16.5	23.0	21.5	22.3	15.0	20.1
DA	17.2	17.5	–	21.5	16.6	38.3	28.3	7.8	10.5	25.3	12.3	20.0	8.3	9.0	20.2	15.3	21.0	19.8	12.7	12.9	26.2
DE	19.0	18.5	20.7	–	18.2	38.6	31.3	8.5	10.9	26.4	13.2	21.0	9.3	9.1	23.2	16.4	24.2	23.2	14.7	15.1	22.3
EL	19.7	18.4	20.3	21.4	–	40.5	33.4	8.2	10.6	27.8	12.6	22.3	9.2	9.4	20.9	17.2	26.9	25.6	14.1	14.9	21.2
EN	26.2	24.1	28.2	29.0	27.2	–	45.3	10.6	13.7	35.9	16.2	29.6	11.7	12.5	26.1	21.9	35.2	35.6	19.0	19.9	29.7
ES	20.5	18.8	21.2	22.7	22.0	44.3	–	8.7	11.2	30.8	13.4	25.8	9.7	9.8	21.8	17.8	31.9	27.7	14.9	15.4	22.0
ET	15.0	13.8	15.0	16.1	14.0	27.2	23.6	–	9.8	20.0	11.4	15.6	7.1	6.9	15.7	13.1	18.4	15.5	10.4	10.2	16.2
FI	15.1	14.0	16.2	16.6	14.6	29.1	24.4	7.4	–	21.1	11.6	16.6	6.5	6.0	16.6	12.9	19.1	16.3	10.6	10.0	18.0
FR	19.2	17.6	19.7	21.5	19.8	38.8	35.0	7.9	10.4	–	12.8	25.7	9.3	9.6	21.3	17.0	28.5	25.9	13.8	14.8	21.2
HU	15.5	14.5	15.5	17.6	14.9	30.4	26.3	6.4	8.9	22.2	–	17.8	6.9	8.0	17.1	13.6	20.9	17.1	10.1	11.1	17.2
IT	17.5	16.9	18.4	19.8	18.2	35.8	32.3	7.7	9.9	28.5	11.9	–	8.7	8.9	20.4	16.1	26.5	22.7	13.0	13.4	19.7
LT	15.5	13.3	14.0	15.0	13.3	26.1	22.1	6.5	8.8	19.3	10.6	15.0	–	8.2	14.5	13.0	17.1	16.0	9.7	10.3	15.3
LV	16.2	14.3	15.7	16.2	14.1	28.6	23.5	6.9	9.5	20.5	11.5	16.1	7.5	–	15.9	13.5	18.5	17.2	10.7	11.1	16.4
NL	16.7	16.0	18.4	21.5	15.9	33.3	27.6	7.2	9.7	24.3	11.8	19.4	8.2	8.3	–	14.6	21.4	18.6	11.9	11.9	19.7
PL	18.7	17.7	17.9	18.7	16.7	33.5	28.7	7.6	9.7	25.1	11.9	19.9	8.5	8.9	18.9	–	22.7	20.8	13.9	14.0	19.1
PT	20.1	18.1	20.2	21.9	21.0	42.5	38.0	8.1	10.5	30.8	12.7	25.8	9.3	9.2	21.2	17.4	–	25.6	13.7	14.4	21.6
RO	20.7	18.1	20.2	22.4	22.2	43.2	36.3	7.3	10.1	30.6	12.9	24.3	8.5	8.9	20.8	17.0	28.1	–	13.3	13.7	21.6
SK	19.7	28.5	19.3	21.1	18.1	35.9	30.1	8.3	10.2	25.1	12.8	20.6	9.0	9.3	19.9	17.7	23.0	22.4	–	15.3	20.2
SL	19.2	19.6	18.5	20.2	17.3	34.5	29.0	7.0	9.9	24.4	12.7	19.2	8.1	7.9	19.3	16.3	22.1	20.7	15.5	–	19.3
SV	16.8	16.3	24.1	20.6	16.3	36.8	27.1	7.9	10.6	24.7	11.9	19.8	8.4	8.7	19.7	15.2	20.4	20.9	12.4	12.1	–

Table 15: sBLEU scores of Llama-3.1-8B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	91.9	90.3	87.2	70.9	89.5	88.7	82.0	90.1	88.0	84.7	90.0	84.6	84.0	87.9	91.7	88.5	91.8	90.0	87.7	89.9
CS	91.5	–	90.1	88.4	70.7	89.0	88.1	81.3	90.0	87.6	84.7	89.7	83.2	82.4	88.6	92.3	87.5	91.1	93.2	87.3	90.1
DA	91.6	91.6	–	87.9	74.7	90.0	88.4	81.7	90.6	87.4	84.3	89.7	83.5	82.8	88.2	91.6	88.1	91.3	89.1	86.2	91.6
DE	91.1	92.0	90.2	–	72.7	89.2	87.5	81.3	90.0	87.2	84.1	89.3	82.7	82.2	88.7	91.7	87.4	90.9	90.0	86.4	90.0
EL	91.5	91.4	90.0	86.7	–	89.9	88.7	81.0	89.9	88.2	83.6	90.0	83.9	82.8	87.4	91.4	88.4	91.5	89.7	86.6	89.8
EN	93.0	92.5	91.8	89.2	73.6	–	90.9	84.8	92.6	90.3	86.4	91.5	84.7	84.6	89.7	92.2	90.3	93.4	90.7	87.9	91.6
ES	91.7	91.2	89.9	86.6	70.4	89.9	–	81.6	90.3	89.0	84.5	90.9	83.0	82.2	87.5	91.7	90.1	91.9	89.4	86.3	89.6
ET	90.2	90.3	89.0	86.3	69.6	89.2	87.0	–	89.4	86.8	82.7	88.7	82.2	81.1	87.2	91.0	87.0	89.8	87.2	84.4	88.9
FI	91.1	91.3	90.2	87.1	70.2	90.3	88.4	80.3	–	88.1	83.1	89.7	82.9	82.0	88.1	91.5	88.0	90.6	88.5	85.2	90.4
FR	91.4	91.0	89.4	86.7	68.9	89.4	89.4	80.7	90.0	–	83.9	90.9	82.2	82.1	87.8	91.4	89.0	91.6	89.0	85.8	89.0
HU	90.5	90.4	89.0	86.4	69.3	88.8	87.5	79.3	88.5	86.8	–	89.0	81.2	80.8	87.1	90.8	87.4	90.0	87.7	84.3	88.6
IT	91.3	91.1	89.4	86.4	71.4	88.8	89.1	80.3	89.5	88.7	83.4	–	82.5	81.8	86.9	91.5	88.9	91.4	88.8	85.8	89.2
LT	90.2	89.3	87.9	85.2	69.4	87.3	86.1	78.2	88.1	84.9	82.0	87.8	–	82.2	85.9	90.4	85.6	88.2	86.1	83.7	87.7
LV	90.8	89.6	88.6	85.9	67.2	88.2	86.4	79.2	88.5	86.0	82.7	88.2	83.4	–	86.6	90.4	86.1	89.3	86.6	83.7	88.3
NL	90.7	91.0	89.4	87.7	71.1	88.6	87.2	79.9	89.4	87.0	83.4	88.9	82.1	81.5	–	91.0	86.9	90.4	88.7	85.3	89.2
PL	91.4	92.1	89.5	87.3	71.7	87.8	88.0	80.9	89.8	87.3	84.0	89.6	83.1	82.2	87.8	–	87.6	91.0	90.1	87.3	89.4
PT	91.6	91.1	89.8	86.6	71.1	89.6	90.3	80.9	90.0	89.0	84.1	90.8	82.4	81.7	87.4	91.4	–	91.6	88.9	85.8	89.6
RO	92.0	91.6	90.3	87.5	67.5	90.5	89.6	80.1	89.8	88.9	83.7	90.8	82.7	82.0	88.3	91.8	89.3	–	89.4	86.4	89.9
SK	91.7	94.5	90.2	88.7	69.4	89.2	88.4	81.0	89.9	87.8	84.9	89.9	82.7	82.0	88.8	92.2	87.8	91.3	–	87.5	90.1
SL	91.4	91.5	89.8	87.6	71.3	88.9	87.9	79.7	89.2	86.8	83.6	89.3	82.5	81.7	87.7	91.7	87.4	90.8	89.4	–	89.6
SV	91.5	91.7	91.8	87.9	69.6	89.8	88.3	82.0	90.7	87.2	84.1	89.7	83.3	82.6	88.2	91.7	87.9	91.1	89.0	86.3	–

Table 16: COMET scores of Gemma-3-4B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	83.7	81.5	83.1	75.6	88.2	86.4	59.3	75.7	85.2	76.5	86.4	70.9	64.7	81.5	84.4	84.4	85.8	73.3	74.7	80.6
CS	81.0	–	79.8	84.4	75.3	88.0	85.3	56.9	72.1	82.4	75.8	85.0	67.4	61.1	81.0	84.4	83.0	84.6	87.4	74.3	79.9
DA	79.9	81.5	–	82.6	71.1	88.3	83.3	57.7	74.6	81.7	75.9	84.9	68.5	61.9	81.1	81.4	83.4	83.2	75.6	70.2	84.4
DE	79.8	82.3	80.5	–	70.8	88.7	84.5	57.6	73.8	83.0	75.4	85.0	68.8	61.5	82.3	82.3	82.4	82.9	64.6	70.3	80.9
EL	79.7	78.1	71.6	81.3	–	86.4	84.6	57.9	73.3	84.4	71.4	86.0	69.5	62.0	79.1	68.1	79.5	82.6	63.8	70.1	77.6
EN	84.0	85.7	84.8	86.6	81.0	–	89.7	64.0	81.0	88.8	81.4	89.4	74.1	66.9	85.5	85.8	88.7	89.8	80.1	75.3	84.7
ES	82.3	82.7	80.8	82.2	78.1	89.5	–	59.6	74.0	86.9	77.5	88.3	70.3	61.6	81.1	84.4	88.3	87.3	75.0	71.8	79.8
ET	77.8	76.8	73.7	77.5	73.2	86.3	77.8	–	69.9	78.7	71.3	79.5	65.3	60.7	74.7	80.8	78.9	80.0	67.4	63.8	73.3
FI	80.8	81.1	78.6	81.7	77.3	89.1	85.0	64.9	–	83.9	75.0	84.4	66.6	59.9	80.0	83.1	84.6	83.7	73.0	66.4	78.8
FR	80.9	81.6	79.6	81.8	72.3	88.8	86.6	59.3	73.8	–	75.7	87.8	68.0	61.4	81.0	82.7	86.0	85.2	64.5	70.5	79.9
HU	79.6	79.0	75.1	82.2	74.7	87.8	84.6	50.8	69.7	83.8	–	85.2	67.6	59.4	76.9	81.8	83.4	82.8	69.3	67.0	76.9
IT	80.4	81.2	78.7	80.6	73.8	88.0	86.3	57.6	69.8	84.8	74.6	–	68.8	60.9	80.0	81.4	84.8	84.8	64.6	70.2	76.5
LT	79.0	76.2	70.8	77.0	73.8	84.9	81.5	57.6	68.5	79.4	70.1	81.8	–	68.5	75.8	80.1	79.4	79.9	64.1	65.3	74.8
LV	80.1	78.4	74.4	79.9	74.8	86.7	82.5	59.6	70.4	82.0	73.2	82.7	69.0	–	76.8	80.8	81.8	82.0	66.5	65.4	74.4
NL	78.5	80.3	78.1	76.3	71.0	87.8	83.4	57.1	72.8	81.6	73.8	84.2	66.7	60.9	–	78.6	82.3	82.9	62.9	66.9	77.8
PL	82.3	85.4	79.3	83.4	76.9	87.3	86.1	58.1	74.0	84.9	76.4	87.0	71.1	64.7	80.8	–	85.4	86.0	80.3	74.8	80.2
PT	81.1	80.4	78.8	80.5	77.4	88.7	88.7	58.0	73.5	86.7	72.1	87.8	68.6	61.1	80.6	82.3	–	86.3	70.7	70.2	79.1
RO	81.0	81.7	79.2	81.0	77.5	89.6	86.5	57.0	72.9	85.6	74.4	86.7	67.6	60.3	80.8	83.4	85.2	–	70.8	71.2	79.2
SK	80.3	89.5	75.6	83.5	75.4	87.7	85.0	56.5	70.4	82.2	73.9	85.5	67.0	61.7	80.6	84.7	83.4	84.0	–	76.3	79.5
SL	79.8	81.6	72.5	80.2	73.8	86.9	80.1	55.8	70.1	74.5	73.6	83.2	66.3	60.6	77.4	82.1	81.9	82.7	73.2	–	76.6
SV	80.5	82.0	83.2	82.2	71.4	87.7	82.1	58.2	74.7	81.4	75.4	85.1	68.3	61.8	81.0	81.4	83.0	83.2	75.7	69.9	–

Table 17: COMET scores of Qwen3-4B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	89.6	86.7	86.3	85.8	89.3	88.0	75.1	85.1	87.2	85.4	88.7	83.7	82.2	86.3	89.9	87.3	90.3	86.5	84.3	87.2
CS	87.5	–	85.8	87.6	85.4	88.9	87.4	74.4	83.6	86.3	84.2	88.1	81.4	79.2	86.8	89.9	86.4	89.2	91.9	82.8	86.7
DA	86.9	88.4	–	86.9	85.6	89.8	87.4	74.8	84.3	86.5	85.3	87.5	82.6	80.1	85.8	88.8	86.5	87.8	84.6	81.3	89.7
DE	87.1	89.8	86.5	–	85.2	89.4	87.2	75.7	84.6	86.9	85.4	88.5	82.7	79.9	87.5	89.8	86.5	89.5	86.4	82.8	87.5
EL	86.3	82.4	84.4	81.9	–	84.9	85.5	64.2	82.0	85.6	82.4	86.4	77.1	75.7	84.2	86.0	80.6	84.4	82.0	79.4	83.8
EN	89.4	90.4	88.6	88.7	88.3	–	90.5	79.4	88.3	89.9	87.5	90.7	84.6	83.4	88.3	90.3	89.8	92.2	87.1	84.3	89.0
ES	88.3	89.2	86.0	86.0	86.6	90.0	–	76.3	85.1	88.7	85.2	89.9	82.7	79.9	85.9	89.7	89.7	90.6	86.1	82.9	86.6
ET	85.3	86.9	83.2	84.6	83.8	88.6	85.3	–	80.6	85.5	82.2	85.5	78.1	74.6	84.3	88.2	85.6	86.9	81.8	77.4	84.1
FI	86.6	88.1	85.4	86.2	85.8	90.2	87.8	72.6	–	87.5	83.5	86.9	76.6	75.2	86.0	89.0	87.4	88.1	83.3	78.9	86.9
FR	87.6	88.5	85.3	86.2	85.9	89.6	88.9	74.2	84.1	–	85.0	90.0	81.9	80.8	86.3	89.2	88.4	90.3	85.3	81.8	85.9
HU	86.6	86.6	84.4	85.8	84.8	89.0	86.9	68.8	80.9	86.3	–	87.5	78.6	77.2	83.6	88.6	85.2	87.8	81.9	78.6	85.0
IT	87.1	88.0	85.1	83.6	85.8	88.8	88.6	73.7	82.8	87.8	84.3	–	81.4	79.3	85.2	89.1	88.0	89.7	84.8	81.1	85.5
LT	86.1	85.0	82.4	83.0	84.1	87.2	85.2	68.4	80.6	84.1	80.3	85.8	–	76.7	83.2	87.5	84.2	85.4	79.3	75.0	83.4
LV	87.0	86.8	83.8	84.5	84.2	88.3	85.9	71.8	81.9	85.5	83.2	87.0	79.8	–	84.7	88.1	85.1	87.4	80.1	77.6	84.3
NL	86.2	87.5	84.2	86.9	84.0	88.7	85.9	73.2	82.8	85.9	83.3	87.1	81.2	79.1	–	88.2	85.2	87.1	82.8	79.2	85.8
PL	87.6	89.1	85.2	86.5	85.0	86.8	87.3	74.7	83.3	86.6	83.8	88.1	80.7	79.3	85.8	–	86.4	89.1	86.8	82.9	86.3
PT	88.1	87.7	85.7	85.6	86.2	89.8	90.2	75.6	84.1	88.8	84.8	90.2	81.9	79.3	85.5	89.4	–	90.3	84.8	81.9	86.4
RO	87.7	88.8	85.6	85.1	86.7	90.6	88.9	71.3	83.3	88.4	84.2	89.0	80.0	78.3	86.1	89.5	88.0	–	85.5	80.8	86.6
SK	87.9	93.4	85.8	87.7	85.5	89.1	87.5	72.3	82.8	86.2	84.1	88.2	79.9	78.6	86.6	90.0	86.4	89.1	–	82.2	86.6
SL	87.1	89.3	85.2	86.6	85.3	88.9	87.2	69.4	82.5	85.9	83.8	88.1	79.1	76.1	85.2	89.0	86.4	88.7	84.8	–	86.0
SV	87.0	88.6	89.2	87.0	85.4	89.8	87.3	75.8	84.7	86.3	84.7	88.0	82.0	79.1	86.0	89.2	86.7	87.9	83.7	80.9	–

Table 18: COMET scores of Qwen3-8B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	90.6	87.4	85.8	83.9	89.0	87.7	76.1	84.5	86.8	86.2	88.7	78.0	75.4	86.8	90.4	87.2	90.0	84.5	84.9	88.7
CS	85.7	–	87.0	87.3	84.3	88.8	87.1	74.1	83.0	85.5	84.5	88.3	74.1	71.0	87.4	89.5	86.2	88.1	90.3	78.5	89.1
DA	83.1	89.3	–	86.1	82.3	89.2	85.3	73.6	84.2	85.3	85.1	87.0	75.2	72.2	86.5	88.5	84.7	86.8	79.7	79.2	90.5
DE	85.2	90.4	86.7	–	83.7	89.1	86.5	75.2	84.7	85.9	85.9	87.8	76.3	71.8	87.9	89.8	86.1	89.1	82.9	82.7	89.1
EL	85.5	89.7	87.1	85.2	–	89.5	87.5	75.4	84.9	87.1	84.6	88.6	77.3	74.1	86.2	89.8	87.2	90.0	83.5	82.6	88.8
EN	88.3	91.8	89.5	88.4	88.5	–	90.3	82.1	89.0	89.6	89.0	90.8	80.4	77.9	89.2	91.1	89.7	92.5	85.3	85.5	91.1
ES	86.2	89.7	86.7	85.3	86.1	89.7	–	75.9	85.6	88.2	86.2	90.0	76.5	72.5	86.7	90.0	89.6	89.8	82.3	82.0	88.5
ET	83.3	87.3	83.9	84.0	82.6	87.9	84.8	–	80.4	84.4	83.7	85.4	70.0	67.5	84.7	87.5	84.8	84.9	76.5	73.7	86.6
FI	84.6	88.6	86.4	85.4	84.7	89.7	86.7	71.4	–	86.4	84.8	87.4	70.2	67.2	86.3	88.4	86.6	86.3	77.9	75.2	88.8
FR	86.0	89.1	86.3	85.5	85.2	89.2	88.6	74.1	85.3	–	86.0	90.1	76.0	73.6	86.9	89.8	88.2	89.9	81.6	82.2	88.1
HU	84.1	87.7	84.1	85.4	83.4	88.7	86.6	66.7	79.8	85.8	–	87.7	72.2	69.6	84.7	88.3	86.4	86.0	76.8	77.6	87.4
IT	84.6	89.3	85.7	84.9	84.1	88.4	88.1	73.9	84.1	87.4	85.1	–	75.4	72.0	86.2	89.7	87.8	88.5	81.5	80.8	87.9
LT	82.9	85.6	82.4	82.3	81.2	85.8	83.3	66.9	80.5	82.0	82.5	84.6	–	68.6	81.8	87.1	82.5	84.4	73.1	73.5	85.5
LV	84.0	86.9	84.2	83.6	81.6	86.9	83.8	68.3	82.0	83.6	83.8	85.7	72.3	–	84.0	87.1	83.6	85.9	75.6	77.0	86.6
NL	83.9	88.5	85.1	86.3	82.2	88.2	85.3	71.9	83.2	85.4	84.3	86.8	74.3	69.9	–	88.0	84.7	86.0	79.2	77.3	87.5
PL	86.0	89.9	86.0	85.9	84.1	87.6	87.1	71.8	83.4	86.2	84.9	88.4	75.0	71.2	86.3	–	86.5	88.5	82.6	81.2	88.5
PT	86.4	89.2	86.4	85.3	85.6	89.5	89.8	74.7	84.8	88.3	85.2	89.9	75.4	71.5	86.4	89.4	–	88.1	80.8	80.2	88.6
RO	86.4	89.1	86.0	86.1	85.7	90.2	88.6	69.3	82.8	87.8	84.8	88.9	72.4	69.1	86.2	88.7	87.1	–	79.7	77.8	88.4
SK	85.7	94.2	86.8	87.3	84.3	88.6	87.1	74.3	83.8	85.5	85.7	88.2	72.6	70.9	87.3	90.3	86.3	88.8	–	81.4	89.0
SL	85.2	90.3	86.0	86.0	83.5	88.2	86.4	68.4	82.6	84.9	84.8	87.3	71.5	66.4	86.1	89.0	85.4	87.3	82.1	–	88.1
SV	83.6	89.1	88.9	86.3	83.3	89.4	85.7	74.6	84.5	85.3	85.0	87.5	75.4	71.9	86.4	89.2	84.9	87.6	79.2	78.5	–

Table 19: COMET of Llama-3.1-8B for each language pair (source language in rows and target language in columns)

src/tgt	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	HU	IT	LT	LV	NL	PL	PT	RO	SK	SL	SV
BG	–	93.2	90.7	88.0	90.7	89.8	89.1	89.7	90.5	88.5	90.3	90.3	90.0	88.8	88.7	92.6	89.1	92.3	91.2	91.3	90.9
CS	90.9	–	90.9	89.4	90.3	89.7	89.0	90.6	91.6	88.5	90.8	90.3	89.2	87.4	89.9	93.3	88.7	92.1	94.0	88.1	91.5
DA	91.0	93.2	–	88.6	89.7	90.3	88.9	90.6	92.1	88.2	90.6	90.2	89.5	87.9	89.5	92.1	89.0	92.2	91.0	90.7	92.5
DE	85.2	92.8	90.8	–	87.0	89.8	88.3	89.9	90.9	88.1	89.5	89.9	89.0	86.6	89.8	92.6	88.2	91.7	90.9	90.7	91.2
EL	73.1	67.5	79.6	79.9	–	89.7	86.6	72.1	80.5	86.3	70.5	87.4	73.7	68.6	74.6	75.0	82.2	85.9	71.2	83.6	81.9
EN	92.8	94.2	92.4	90.2	92.7	–	91.1	92.9	93.9	90.7	92.5	91.8	91.3	91.1	90.7	93.3	90.8	94.1	92.7	92.5	92.8
ES	87.1	89.9	90.5	74.4	89.3	90.3	–	90.1	91.4	88.6	87.8	90.6	89.3	86.5	88.0	92.3	90.4	92.4	90.5	90.0	90.8
ET	90.2	93.0	90.5	88.2	89.1	90.0	88.4	–	91.6	88.2	90.2	89.9	89.0	88.1	89.1	92.6	88.4	91.4	90.7	89.7	90.9
FI	90.3	93.3	91.3	88.6	90.8	90.9	89.4	91.0	–	89.2	90.5	90.4	88.9	86.8	89.6	92.9	89.3	91.8	90.9	90.1	92.0
FR	86.7	90.9	89.9	87.1	89.0	89.8	89.7	84.9	90.2	–	89.8	91.0	85.9	83.6	88.7	92.1	89.4	92.1	89.4	89.7	90.3
HU	90.4	92.4	90.6	87.9	89.1	89.7	88.8	89.5	90.8	88.3	–	90.0	88.6	87.6	88.8	91.9	88.8	91.1	90.5	89.6	90.6
IT	84.7	89.6	89.7	81.8	87.9	89.2	89.6	89.0	83.4	89.3	88.2	–	86.1	85.1	88.3	91.7	89.4	91.8	90.2	88.3	90.2
LT	89.0	91.0	89.4	86.5	87.3	88.0	87.3	89.3	88.4	86.2	88.1	88.8	–	88.4	87.0	91.6	86.6	90.1	89.2	88.9	89.4
LV	88.8	91.9	90.0	87.6	88.7	88.8	87.6	90.5	89.0	87.1	89.7	89.3	90.7	–	88.3	92.0	87.4	90.8	90.2	89.7	90.4
NL	89.4	92.4	90.2	87.6	88.7	89.1	88.0	89.7	91.2	87.9	89.6	89.4	88.5	86.5	–	91.7	87.8	91.2	90.4	89.5	90.7
PL	90.4	93.2	90.3	84.3	90.2	88.4	88.7	90.2	91.4	88.0	90.3	90.1	89.9	88.2	88.7	–	88.5	91.7	91.7	91.2	90.9
PT	91.2	92.8	90.8	84.6	91.0	90.1	90.6	90.2	91.6	89.5	90.5	91.2	89.3	87.3	88.7	92.5	–	92.4	90.9	90.3	91.1
RO	90.7	92.5	91.0	83.5	91.1	91.0	90.1	89.9	89.6	89.4	90.4	91.1	88.4	88.0	89.1	91.7	89.8	–	90.7	90.0	91.4
SK	90.9	94.6	90.9	89.7	89.2	89.8	89.1	88.2	89.6	88.6	88.6	90.3	87.8	87.5	89.4	93.4	88.8	91.9	–	90.2	90.6
SL	90.4	92.2	90.8	88.8	88.3	89.6	88.9	88.3	88.8	88.1	88.9	90.3	89.2	86.6	88.1	93.1	88.5	91.8	90.5	–	90.9
SV	91.0	93.0	92.3	88.4	90.1	90.3	89.0	90.6	91.3	88.1	90.1	90.3	89.6	88.2	89.5	92.6	89.0	92.0	91.0	90.6	–

Table 20: COMET scores of Llama-3.1-70B for each language pair (source language in rows and target language in columns)

C. Political Fairness per Target Language

We report the per-target-language sBLEU scores of every EU party for all models in Tables 21, 22, 23, and 24.

tgt	ALDE	ECR	EFA	EFD	EPP	NA	NGL	S&D
BG	27.67	28.70	27.98	28.61	28.67	26.25	27.10	28.72
CS	21.18	21.23	20.63	21.14	21.11	20.47	19.75	21.30
DA	26.50	26.12	25.72	25.20	25.76	24.46	24.64	25.88
DE	24.02	24.58	22.57	23.46	23.47	20.16	22.29	23.78
EL	10.46	10.02	10.64	10.57	10.28	9.92	10.50	10.52
EN	36.32	35.88	34.76	36.39	36.42	35.16	35.16	36.71
ES	33.16	32.53	31.93	34.04	34.50	32.76	32.96	33.80
ET	11.56	11.67	11.72	10.84	11.04	11.04	10.32	11.12
FI	16.10	15.73	16.17	14.67	15.23	15.30	14.66	15.28
FR	28.20	28.17	27.44	28.89	29.38	26.83	28.12	28.92
HU	14.06	13.40	14.43	13.46	14.01	13.80	12.84	13.77
IT	24.42	23.89	23.96	24.34	24.76	22.86	23.96	24.54
LT	13.15	13.00	12.94	12.61	12.57	12.78	11.65	12.84
LV	14.43	13.84	14.75	14.17	14.17	13.65	13.16	14.19
NL	23.42	22.85	22.02	21.92	22.20	21.52	20.86	22.94
PL	20.01	19.27	19.19	19.36	19.95	19.42	18.94	20.09
PT	27.33	26.58	26.55	27.26	27.21	26.80	24.78	27.23
RO	27.81	27.77	27.73	27.51	27.72	27.43	26.93	28.06
SK	20.53	21.11	20.18	20.21	20.59	19.80	19.44	20.59
SL	18.47	18.76	18.20	17.67	18.50	17.57	17.55	18.53
SV	24.17	24.21	22.89	23.32	23.39	22.72	22.09	23.43

Table 21: sBLEU scores of Gemma-3-4B per target language. Each language has its performance variance significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$).

tgt	ALDE	ECR	EFA	EFD	EPP	NA	NGL	S&D
BG	15.90	16.75	15.89	16.19	16.44	14.75	15.85	16.65
CS	12.60	12.46	12.44	12.39	12.69	12.04	11.56	12.76
DA	15.96	15.78	15.50	15.09	15.74	14.37	15.09	15.61
DE	17.59	17.77	16.39	17.36	17.59	14.67	16.26	17.60
EL	12.51	12.31	12.25	12.34	12.93	11.41	11.57	12.92
EN	33.19	32.89	31.86	33.34	34.01	31.88	32.21	33.93
ES	26.73	25.83	25.66	27.33	28.22	25.99	26.49	27.46
ET	5.90	5.81	6.09	5.42	5.56	5.63	5.28	5.63
FI	8.50	8.11	8.43	7.69	8.11	7.73	7.76	8.07
FR	22.26	21.90	21.82	22.77	23.61	21.05	22.40	23.05
HU	8.78	8.28	9.09	8.60	8.99	8.57	8.10	8.72
IT	18.23	17.15	17.85	17.95	18.60	17.08	17.91	18.42
LT	8.35	8.01	8.44	7.99	8.08	7.96	7.62	8.05
LV	7.57	7.30	7.88	7.21	7.40	7.13	6.85	7.30
NL	15.98	15.61	15.03	15.36	15.51	14.78	14.65	15.76
PL	12.10	11.61	11.87	11.85	12.29	11.78	11.68	12.19
PT	20.93	20.02	20.40	21.01	21.19	20.34	19.09	21.12
RO	19.34	18.80	19.16	19.52	19.69	18.72	18.56	19.70
SK	9.29	9.18	9.24	8.97	9.31	8.74	8.68	9.38
SL	10.26	9.98	10.29	9.61	10.08	9.68	9.92	10.09
SV	15.22	14.70	14.29	14.52	14.78	13.59	14.20	14.84

Table 22: sBLEU scores of Qwen3-4B per target language. Each language has its performance variance significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$).

tgt	ALDE	ECR	EFA	EFD	EPP	NA	NGL	S&D
BG	20.83	21.56	21.28	21.59	21.81	19.80	20.88	21.93
CS	17.21	16.90	16.80	17.29	17.37	16.47	15.99	17.46
DA	20.91	20.35	20.60	19.88	20.55	19.27	19.93	20.52
DE	21.35	21.73	20.38	21.00	21.10	18.10	19.70	21.34
EL	18.70	18.19	18.21	18.16	18.89	17.29	17.27	19.08
EN	36.04	35.65	34.67	36.00	36.35	34.65	35.01	36.60
ES	30.81	30.05	29.85	31.39	32.20	30.22	30.66	31.52
ET	8.62	8.27	8.67	7.91	8.16	8.14	7.66	8.19
FI	12.01	11.42	11.73	10.92	11.45	10.77	10.80	11.47
FR	25.99	25.45	25.35	26.27	27.19	24.74	25.68	26.78
HU	12.39	11.73	12.82	11.97	12.37	12.19	11.55	12.19
IT	21.27	20.15	20.70	21.23	21.78	19.94	20.77	21.50
LT	11.86	11.10	11.77	11.40	11.41	11.26	10.55	11.43
LV	11.91	11.09	11.88	11.63	11.87	11.24	10.75	11.59
NL	20.03	19.54	19.05	19.03	19.34	18.89	18.41	19.79
PL	16.12	15.39	15.79	15.59	16.20	15.76	15.22	16.21
PT	24.34	23.47	23.64	24.41	24.60	23.76	22.26	24.55
RO	23.91	23.07	23.58	23.40	23.86	23.09	22.81	24.13
SK	17.03	16.99	16.46	16.50	17.00	16.15	16.08	17.07
SL	14.57	14.30	14.28	13.93	14.64	13.88	13.96	14.50
SV	19.85	19.57	18.89	18.83	19.18	18.43	18.43	19.27

Table 23: sBLEU scores of Qwen3-8B per target language. Each target language has its performance variance significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$).

tgt	ALDE	ECR	EFA	EFD	EPP	NA	NGL	S&D
BG	18.13	19.00	18.10	18.60	18.39	17.51	17.94	18.73
CS	17.75	17.78	17.58	17.70	17.90	17.29	16.89	18.00
DA	19.80	19.61	19.54	18.45	19.17	18.50	18.53	19.22
DE	20.86	21.23	20.12	20.33	20.34	18.11	19.22	20.55
EL	17.90	17.73	17.11	17.54	18.10	16.55	16.49	18.20
EN	35.52	35.23	34.52	35.94	35.80	34.27	34.39	36.06
ES	29.52	28.84	28.80	30.26	30.87	29.15	29.48	30.21
ET	8.16	8.02	8.30	7.53	7.75	7.99	7.39	7.86
FI	10.70	10.61	10.84	9.78	10.23	10.17	9.78	10.27
FR	25.23	24.64	24.86	25.58	26.33	24.04	25.27	25.89
HU	12.63	12.52	13.12	12.13	12.56	12.68	11.84	12.51
IT	20.92	19.92	20.50	20.50	21.10	19.80	20.59	21.00
LT	8.91	8.57	9.05	8.53	8.64	8.71	8.01	8.73
LV	9.04	8.81	9.12	8.80	8.98	8.81	8.40	8.96
NL	20.50	20.16	19.72	19.38	19.63	19.19	18.54	20.02
PL	15.95	15.66	15.65	15.71	16.22	15.70	15.21	16.34
PT	23.83	23.16	23.27	23.79	23.97	23.48	21.95	23.84
RO	21.75	21.46	21.68	21.74	22.04	21.55	21.09	22.15
SK	13.70	13.62	13.42	13.36	13.65	13.21	13.00	13.84
SL	13.57	13.65	13.56	13.16	13.70	13.10	13.12	13.80
SV	21.08	20.89	20.02	20.09	20.40	20.00	19.50	20.56

Table 24: sBLEU scores of Llama-3.1-8B per target language. Each language has its performance variance significantly explained by the party variable according to a Kruskal-Wallis test ($p < 0.01$).