Translation Tangles: Performance Benchmarking and Bias Detection in LLM-Based Translation Across Language Families and Domains

Anonymous ACL submission

Abstract

The rise of Large Language Models (LLMs) has redefined Machine Translation (MT), enabling context-aware and fluent translations across hundreds of languages and textual domains. Despite their remarkable capabilities, LLMs often exhibit uneven performance across language families, translation directions, and specialized domains. Moreover, recent evidence reveals that these models can encode and amplify different biases present in their training data, posing serious concerns for fairness, especially in low-resource and historically marginalized languages. To address these gaps, we introduce *Translation Tangles*, a unified framework for evaluating the translation quality and fairness of open-source LLMs. Our approach benchmarks 24 bidirectional language pairs across multiple domains using different metrics. We further propose a hybrid bias detection pipeline that integrates rule-based heuristics, semantic similarity filtering, and LLM-based validation. We also introduce a high-quality, bias-annotated dataset based on human evaluations of 1,439 translation-reference pairs, offering a robust foundation for analyzing fairness and bias in multilingual machine translation. The code and dataset are accessible on GitHub: `https://anonymous.4open.science/r/TranslationTangles-EABE/`.

1 Introduction

Machine Translation has undergone a profound transformation with the emergence of LLMs, which demonstrate unprecedented fluency and contextual awareness in translation tasks (Zhu et al., 2024). Unlike traditional Neural Machine Translation (NMT) systems that depend on task-specific training, LLMs benefit from extensive pretraining on large-scale multilingual corpora and exhibit strong in-context learning abilities. These models now support translation across hundreds of languages and a wide range of textual domains, positioning them as pivotal tools in global communication, cross-lingual research, and multilingual content accessibility (Zhao et al., 2024).

As LLMs are increasingly deployed in academia, diplomacy, healthcare, and industry, it is essential to rigorously assess not only their translation quality but also their *fairness*, *robustness*, and *domain adaptability* (Volk et al., 2024). Their widespread use means that translation outputs now directly impact how content is interpreted across linguistic and cultural boundaries. Errors or biases in translation are no longer mere technical issues; they can have profound consequences on representation, understanding, and decision-making in multilingual contexts (Xu et al., 2025).

Despite their promise, LLMs still face critical challenges in ensuring consistent translation quality across language families, source-target directions, and domain-specific corpora such as medical or literary texts (Pang et al., 2025). Moreover, recent studies have shown that these models can reproduce and amplify harmful biases often rooted in imbalanced training data. Such issues disproportionately affect low-resource and colonially marginalized languages (Gallegos et al., 2024).

Both NMT and LLM-based systems exhibit performance inconsistencies and biased outputs, particularly for structurally divergent or underrepresented language pairs (Sizov et al., 2024). Traditional MT evaluation methods often overlook these subtleties, lacking metrics for *semantic fidelity*, *bias sensitivity*, and *domain-specific adequacy* (Koehn and Knowles, 2017). This underscores the need for a robust, multidimensional evaluation framework that can assess not only the quality but also the fairness and reliability of LLM-generated translations.

In this work, we introduce a unified framework for evaluating translation quality and detecting bias in LLM-generated translations across diverse language pairs and domains. Our main contributions

are as follows:

- We develop a multilingual benchmarking suite for evaluating translation quality across multiple dimensions, including translation directionality, language family, and domain. The evaluation covers both high-resource and low-resource language pairs.

- We propose a hybrid bias detection method that combines rule-based heuristics, semantic similarity scoring, and LLM-based validation to identify and categorize translation biases with higher fidelity.

- We conduct a structured human annotation study, independently reviewed for bias presence. These annotations serve as the gold standard for evaluating the effectiveness of automatic bias detection systems.

- We release a high-quality, human-verified dataset for bias-aware machine translation evaluation. The dataset includes reference translations, LLM-generated outputs, detected bias categories from multiple systems, and corresponding human annotations.

2 Related Work

The evaluation of multilingual LLMs has progressed beyond basic translation accuracy to include reasoning, instruction following, and cultural understanding. Early studies (Zhu et al., 2024; Song et al., 2025) highlight substantial performance gaps between high- and low-resource languages, emphasizing the need for more inclusive and challenging benchmarks.

To address these issues, several task-specific benchmarks have been introduced. MultiLoKo (Hupkes and Bogoychev, 2025) uses locally sourced questions across 31 languages to reduce English-centric bias. BenchMAX (Huang et al., 2025) evaluates complex multilingual tasks, while Chen et al. (2025) assess reasoning-heavy "o1-like" models on translation performance. For domain-specific translation, Hu et al. (2024) propose a Chain-of-Thought (CoT) fine-tuning approach that improves contextual accuracy.

Bias in multilingual evaluation is a growing concern. These biases span cultural, sociocultural, gender, racial, religious, and social domains (Měchura, 2022). Sant et al. (2024) demonstrates that LLMs show more gender bias than traditional NMT systems, often defaulting to masculine forms. Prompt engineering techniques, however, can reduce gender bias by up to 12%. Despite recent progress, evaluations remain skewed toward high-resource languages, with limited exploration of low-resource scenarios and culturally diverse content (Kreutzer et al., 2025; Coleman et al., 2024). Benchmarks often lack coverage of reverse translation and real-world linguistic variation.

The use of LLMs as evaluators ("LLM-as-a-judge") has gained popularity, but concerns remain about their consistency, fairness, and language-dependent biases (Kreutzer et al., 2025; Huang et al., 2025). Additionally, semantic-aware metrics like COMET are preferred over traditional BLEU, which often fails to capture meaning preservation (Chen et al., 2025). Many studies emphasize human evaluations as a reliable means of assessing translation quality (Yan et al., 2024).

3 Methodology

Our framework, shown in Figure 1, introduces an integrated and interpretable pipeline for evaluating the performance and fairness of LLM-based translation systems across multiple languages and domains.

3.1 Multilingual Benchmarking of State-of-the-Art Open Source LLMs

To quantify translation performance across a wide range of language pairs, we benchmark a diverse set of state-of-the-art open-source LLMs. Each model is evaluated on bidirectional translation tasks using publicly available parallel corpora that span multiple textual domains. Language pairs are grouped by family to allow for structured comparative analysis. Our evaluation considers both high-resource and low-resource settings, enabling a holistic understanding of LLM capabilities across linguistic hierarchies and typologies.

This multilingual framework explores translation quality from several critical angles, including directional asymmetry, colonial language advantage, domain robustness, and model scaling effects. We assess whether translation quality is consistent across source-target directions, whether colonized languages benefit disproportionately from colonial language pairings, and how models perform in specialized domains such as medicine or law. Additionally, we investigate whether general-purpose

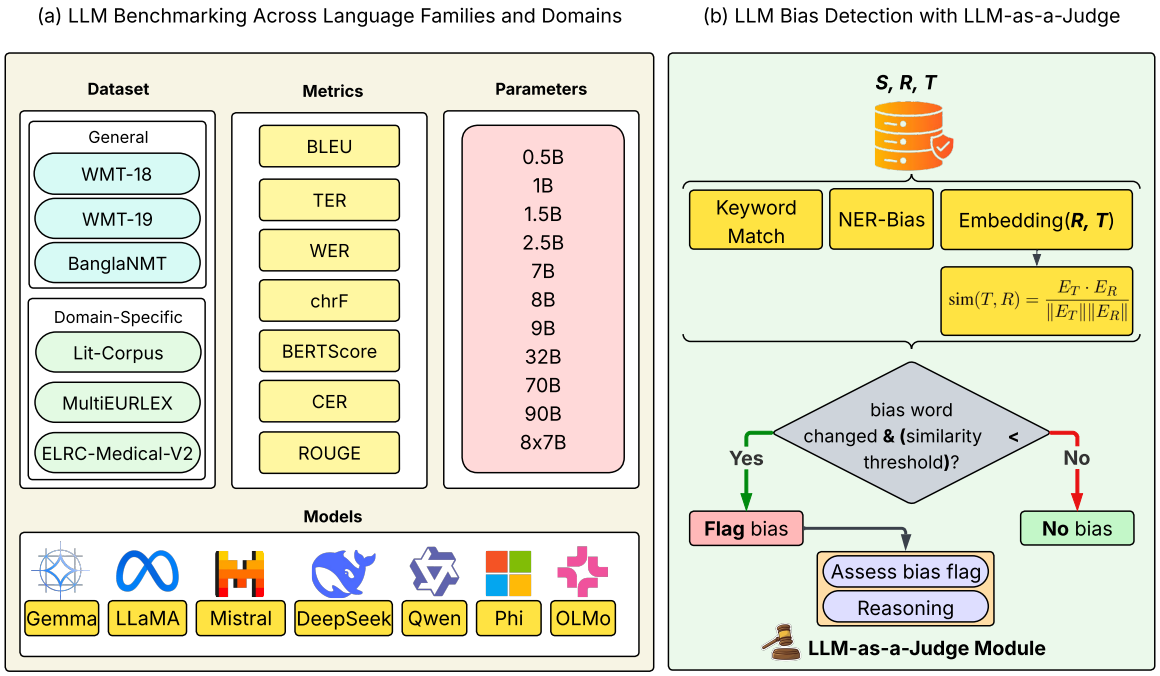

Figure 1: Our framework comprises two key components: (a) LLM Benchmarking, where T are evaluated against R using LLMs across diverse language families and domains; and (b) Bias Detection with LLM-as-a-Judge Evaluation, where potential biases are flagged using linguistic heuristics and semantic analysis, and then verified through LLMs. Here, S = Source, R = Reference, T = Translation.

models are sufficient for domain-specific translations and how architectural scaling influences performance across resource tiers. This analysis highlights the strengths, weaknesses, and fairness challenges of modern LLMs and offers actionable insights for future model training and deployment. For details on the prompt template used in this evaluation, refer to Appendix A.2.

3.2 Semantic and Entity-Aware Bias Detection

To identify potential biases (Appendix B.1) in machine translation outputs, we propose a two-pronged approach that combines semantic similarity analysis with entity- and keyword-based linguistic heuristics. This module aims to capture both explicit and subtle shifts in meaning that may result in biased translations.

Sentence Embedding and Similarity. To capture semantic fidelity between the machine translation (T) and the human reference (R), we compute cosine similarity between their embeddings generated using a pretrained SentenceTransformer model (`all-MiniLM-L6-v2`):

$$\text{sim}(T, R) = \frac{E_T \cdot E_R}{\|E_T\| \|E_R\|} \quad (1)$$

where E_T and E_R denote the sentence embeddings of the translation and reference, respectively.

Named Entity Recognition (NER)-based Bias Flagging. We apply `spaCy`'s NER module to extract entity mentions from both T and R. If new entities are introduced in T that are not present in R, and these entities belong to sensitive categories, we flag them as potential biases:

$$\text{Bias}_{\text{NER}} = \{e \in E_T \backslash E_R \mid \text{bias\_map}(e.\text{type}) \in \mathcal{B}\} \quad (2)$$

where $\mathcal{B}$ is the set of bias categories and bias_map maps entity types to bias types, as detailed in Appendix B.3.

Keyword-Based Matching. We maintain a curated lexicon $\mathcal{K}_b$ for each bias type $b \in \mathcal{B}$ (Appendix B.2). If a keyword appears in T but not in R (or vice versa), it is flagged as an insertion or erasure:

$$\text{Bias}_{\text{KW}} = \{k \in \mathcal{K}_b \mid (k \in T \wedge k \notin R) \vee (k \in R \wedge k \notin T)\} \quad (3)$$

The final set of detected bias types for a translation is the union of all flagged bias categories

across NER and keyword analyses:

$$\text{DetectedBiases} = \bigcup_{i \in \{\text{NER,KW}\}} \text{Bias}_i \quad (4)$$

Thresholding and Final Bias Decision. We empirically determine a similarity threshold $\tau = 0.75$ through grid search, balancing recall and precision (Figure 2). For more analysis on optimal thresholding, refer to Appendix F. A candidate translation is only flagged as biased if a bias-indicative change is detected through NER or keyword-based heuristics *and* the semantic similarity $\text{sim}(T, R)$ falls below the threshold τ:

$$\text{FlaggedBias} = \begin{cases} 1 & \text{if DetectedBiases} \neq \emptyset \text{ and sim}(T,R) \\ < \tau & \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

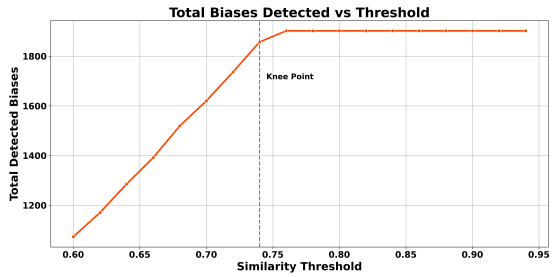

Figure 2: Total detected biases were plotted across thresholds from 0.6 to 0.95. The count stabilizes beyond 0.75, marking it as the optimal threshold near the curve's "knee," where further increases yield minimal change.

3.3 LLM-as-a-Judge Evaluation

To validate the biases flagged by the heuristic framework, we introduce an LLM-based verification system using `Gemini-2.5-Flash`. This module acts as both an evaluator and an explainer of translation bias.

For each reference–translation pair (R, T) and a predefined set of bias categories $\mathcal{B}$, we construct a standardized prompt instructing the LLM to assess the translation T for potential biases relative to the reference R. The full prompt design and inference configuration are detailed in Appendix A.2.

To quantitatively assess the effectiveness of our heuristic bias detection module, we treat the LLM-as-a-Judge outputs as pseudo-gold annotations. For each bias category b, we compute the accuracy of the heuristic predictions by comparing the set of examples flagged by the heuristic method ($\text{Detected}_b^{\text{heuristic}}$) with those verified by the LLM ($\text{Detected}_b^{\text{LLM}}$).

$$\text{Accuracy}_{\text{overall}} = \left(\frac{\sum_b |\text{Detected}_b^{\text{heuristic}} \cap \text{Detected}_b^{\text{LLM}}|}{\sum_b |\text{Detected}_b^{\text{heuristic}}|} \right) \times 100\% \quad (6)$$

This overall metric reflects the total proportion of heuristic predictions that are supported by the LLM verification step, serving as a global measure of heuristic precision. This formulation allows us to compute accuracy without requiring manual annotations, leveraging the LLM for both validation and interpretability.

4 Experimental Setup

We describe the experimental setup, including the datasets used, model configurations, evaluation metrics, and selected language pairs.

4.1 Dataset

We use a combination of general-purpose and domain-specific multilingual benchmark datasets to evaluate translation quality across diverse linguistic and contextual settings. Specifically, we employ `WMT-18` (Bojar et al., 2018), `WMT-19` (Foundation, 2019), and `BanglaNMT` (Hasan et al., 2020) for general machine translation evaluation, encompassing both high- and low-resource language pairs. To assess domain-specific performance, we include `Lit-Corpus` (Abdashim, 2023) for literature, `MultiEURLEX` (Chalkidis et al., 2021) for legal texts, and `ELRC-Medical-V2` (Lösch et al., 2018) for medical translation tasks. For more details on datasets, refer to Appendix D.

4.2 Language Pairs

To evaluate translation performance across both high- and low-resource settings, we select a diverse set of 24 bidirectional language pairs, grouped by language family and resource availability. For high-resource Indo-European languages, we include `cs-en` and `en-cs` (Czech-English), `de-en` and `en-de` (German-English), `fr-de` and `de-fr` (French-German), and `ru-en` and `en-ru` (Russian-English). For medium-resource European languages, we consider `fi-en` and `en-fi` (Finnish-English), `lt-en` and `en-lt` (Lithuanian-English), and `et-en` and `en-et` (Estonian-English). For non-Indo-European and low-resource languages, we include `gu-en` and `en-gu` (Gujarati-English), `kk-en` and `en-kk` (Kazakh-English), and `bn-en`

and en-bn (Bangla-English), representing underrepresented South and Central Asian languages. We also incorporate zh-en and en-zh (Chinese-English) from the Sino-Tibetan family and tr-en and en-tr (Turkish-English) from the Turkic family to capture non-Indo-European high-resource scenarios.

4.3 Models

We evaluate a range of state-of-the-art LLMs, including Gemma-7B, Gemma-2-9B, Llama-3.1-8B, Llama-3.1-70B, Llama-3.2-1B, Llama-3.2-70B, Llama-3.2-90B, Mixtral-8x7B, OLMo-1B, Phi-3.5-mini, Qwen-2.5-0.5B, Qwen-2.5-1.5B, Qwen-2.5-3B, deepseek-r1-distill-32b, deepseek-r1-distill-70b. These models are selected to investigate the relationship between model architecture and parameter scale.

4.4 Evaluation Metrics

We employ a diverse set of evaluation metrics, including BLEU, chrF, TER, BERTScore, WER, CER, ROUGE-1, ROUGE-2, and ROUGE-L. BLEU and chrF assess lexical differences, while TER measures the extent of text modifications required for accurate translations. BERTScore captures semantic discrepancies that may arise due to biases in word choice or tone. Additionally, we analyze word error rate (WER) and character error rate (CER) to evaluate inconsistencies across gendered or culturally specific terms. ROUGE-1, ROUGE-2, and ROUGE-L further assess content overlap to detect systematic distortions in meaning.

5 Results and Analysis

We comprehensively analyze translation performance and bias detection across languages and domains.

5.1 Translation Evaluation Across Languages and Domains

For the complete results across all metrics and language pairs, refer to Appendix I.

RQ1: How does translation quality vary between source-to-target and target-to-source directions (e.g., X $\rightarrow$ Y vs. Y $\rightarrow$ X), and how is this influenced by intra-family versus cross-family language relationships? To explore the effect of linguistic distance on directional asymmetry, we categorized translation directions into: (i) **Cross-Family** (e.g., Chinese $\leftrightarrow$ English, Estonian $\leftrightarrow$ English), and (ii) **Intra-Family** (e.g., French $\leftrightarrow$ German, English $\leftrightarrow$ Czech). For detailed definitions and the full language family mapping used to classify translation directions as Cross-Family or Intra-Family, see Appendix C. We calculated the average performance of all models in both forward and reverse directions for each category.

Table 1: Average Translation Scores by Language Family Distance. Intra-family translation directions consistently outperform cross-family ones across all metrics.

Metric	Intra-Family	Cross-Family	Difference
BLEU	22.14	10.63	+11.51
BERTScore	0.31	0.14	+0.17
ROUGE-L	0.43	0.18	+0.25
WER $\downarrow$	3.94	9.82	-5.88
chrF	53.26	28.14	+25.12

As shown in Table 1, intra-family translation directions consistently outperform cross-family ones across all evaluation metrics. BLEU scores are, on average, 11.51 points higher for intra-family pairs, while chrF improves by over 25 points, suggesting stronger character-level alignment. Semantic similarity, measured by BERTScore, increases by 0.17, and ROUGE-L by 0.25, further emphasizing enhanced fluency and structure when linguistic proximity is high. WER is nearly 6 points lower for intra-family translations, reflecting reduced word-level errors. These disparities point to a linguistic bias embedded in LLM-based translation models, where structurally similar languages benefit disproportionately.

RQ2: Does the historical dominance of colonial languages (e.g., English) lead to higher translation performance for colonized languages (e.g., Bengali, Gujarati), particularly in translation directions involving the colonial language? To explore this, we categorized translation directions into: (i) **Colonized $\rightarrow$ Colonizer**, and (ii) **Colonizer $\rightarrow$ Colonized**. We calculated the average performance of all models across each direction.

As shown in Table 2, translation performance is consistently better in the *Colonized $\rightarrow$ Colonizer* direction across all evaluated metrics. The BLEU score is higher by 7.13 points, chrF by 8 points, and BERTScore by 0.11, indicating better semantic and lexical alignment when colonial languages (e.g., English) serve as the target. ROUGE-L similarly shows a positive delta of 0.12. In contrast, WER, which reflects word-level errors (where lower is

Table 2: Average Translation Scores by Colonial Direction. Translations from colonized to colonizer languages consistently outperform the reverse direction across all evaluation metrics.

Metric	Colonized → Colonizer	Colonizer → Colonized	Difference
BLEU	15.41	8.28	+7.13
BERTScore	0.19	0.08	+0.11
ROUGE-L	0.21	0.09	+0.12
WER ↓	7.50	9.26	−1.76
chrF	21.56	13.56	+8.00

better), is 1.76 points lower in the colonized-to-colonizer direction, further reinforcing this trend. These results highlight a structural inequality embedded in LLM-based translation systems, wherein models are better optimized for colonial language outputs.

RQ3: How does translation quality vary across specialized domains (e.g., medical, legal, literary), and are certain domains systematically harder for LLMs to handle? To assess domain-specific robustness, we calculated average translation scores across all evaluated models for three specialized textual domains: **Law**, **Literature**, and **Medical**. Table 3 summarizes the results.

Table 3: Average Translation Scores by Domain. Legal texts yield the highest performance, while literary and medical domains pose increasing translation challenges. BS = BERTScore, RL = ROUGE-L.

Domain	BLEU	BS	RL	WER	chrF
Law	33.60	0.667	0.518	1.06	60.07
Literature	13.61	0.559	0.179	0.87	39.20
Medical	26.60	0.637	0.467	1.12	56.65

As seen in Table 3, the Law domain yields the highest scores across all metrics. BLEU scores decline by approximately 20.8% from Law to Medical, and by 59.5% from Law to Literature, highlighting the sensitivity of n-gram overlap to domain complexity. BERTScore, which captures semantic similarity, drops by 4.5% from Law to Medical, and by 16.2% from Law to Literature. ROUGE-L, which reflects overlapping longest common subsequences, decreases by 9.8% in Medical and 65.4% in Literature when compared to Law. WER increases by 6.3% from Law to Medical, further indicating reduced word-level fidelity in the medical domain. Finally, chrF, a character-level F-score metric, declines by 5.7% from Law to Medical and by a substantial 34.7% from Law to Literature. These patterns underscore the difficulty LLMs face in specialized settings: legal texts benefit from structural regularity, whereas literary texts suffer from stylistic variability and implicit context, and medical texts challenge models with domain-specific precision demands.

RQ4: How does increasing model size and architectural complexity impact translation performance across low-resource and high-resource languages? We categorized models into three capacity levels based on their parameter size: small ($\leq$ 7B parameters), medium (7B–30B), and large ($>$ 30B). Table 4 presents the average of model–resource configurations across BLEU, BERTScore, ROUGE-L, WER, and chrF.

Table 4: Translation Performance by Model Size and Language Resource. Large models in high-resource languages achieve the best translation quality across all evaluated metrics. BS = BERTScore, RL = ROUGE-L.

Model Size	Resource	BLEU / BS / RL / WER / chrF
Large	High	21.91 / 0.60 / 0.43 / 0.97 / 54.78
	Low	19.97 / 0.54 / 0.41 / 1.01 / 49.56
Medium	High	13.93 / 0.40 / 0.34 / 1.95 / 43.68
	Low	9.98 / 0.35 / 0.29 / 2.31 / 38.87
Small	High	2.90 / -0.23 / 0.11 / 15.79 / 20.09
	Low	1.75 / -0.38 / 0.08 / 18.77 / 17.98

These results show that large models in high-resource languages consistently achieve the best performance across all metrics. Medium-sized models outperform small models in both resource conditions, particularly in BERTScore and chrF, indicating better semantic preservation and fluency. However, performance gaps persist between large and medium models, reinforcing the importance of model scale and resource availability. Overall, both architectural complexity and training resource richness remain key drivers of multilingual translation quality.

5.2 Bias Detection Evaluation

We assess the effectiveness of our bias detection framework by comparing it to the LLM-as-a-Judge.

5.2.1 Semantic and Entity-Aware Bias Detection Framework Results

We applied our semantic and entity-aware bias detection framework across translations produced by eight open-source LLMs, covering 12 language pairs and six bias types. The overall frequency of each bias type is summarized in Table 5.

Cultural (n=798) and sociocultural (n=744) biases were by far the most prevalent, jointly accounting for more than 75% of all detected instances. Gender bias (n=265) appeared moderately, while racial, religious, and social biases were comparatively rare. This skew highlights the persistent challenge of modeling context-sensitive and culturally embedded semantics in multilingual translation tasks. Bias counts also varied significantly across language pairs. The `gu-en` pair exhibited the highest total bias (n = 220), with `kk-en`, `fi-en`, and `lt-en` following. These trends point to systematic translation challenges involving lower-resourced or culturally distinct source languages. In contrast, pairs like `de-en` and `zh-en` showed substantially lower bias, likely due to better representation in training data. At the model level, `gemma-2-9b` produced the highest number of sociocultural biases (n = 290), while `llama3-8b` showed the highest cultural bias (n = 200). Surprisingly, larger models such as `llama-3.2-90b` did not consistently yield fewer biases, suggesting that model size alone does not ensure bias mitigation. In contrast, `llama-3.1-8b` displayed disproportionately high cultural bias, indicating vulnerabilities in real-time or lightly aligned variants.

These results reveal three core challenges. First, biases are amplified in translations involving under-resourced source languages. Second, cultural and sociocultural biases dominate, highlighting a persistent challenge in managing nuanced sociocultural semantics. Third, model size does not linearly correlate with fairness, underscoring the need for targeted debiasing and context-aware alignment techniques. For a complete breakdown of per-model and per-language results, see Appendix E.1.

5.2.2 LLM-as-a-Judge Results

To further evaluate the reliability of our semantic and entity-aware framework, we compared its outputs against judgments made by a separate LLM-based evaluation module (`LLM-as-a-Judge`).

Table 5 summarizes the total number of detected biases per category by both systems. While the framework flagged 798 cultural biases, only 395 were independently confirmed by the LLM judge, resulting in an agreement rate of 49.50%. Sociocultural bias had a slightly lower agreement (45.83%), whereas gender (61.13%) and religion (66.67%) had moderate alignment. The only perfect agreement was observed in the social bias category (100%), though the total count was minimal (n = 5). Racial bias showed the lowest agreement, with only 13.64% confirmed by the LLM.

The overall agreement rate between the two systems was **48.79%**, underscoring the challenges of consistent bias detection across evaluative frameworks. This gap reflects potential differences in interpretation, sensitivity thresholds, and contextual inference across the two approaches.

Table 5: Bias Detection Counts and Agreement Rates: Framework vs. LLM-as-a-Judge. Agreement was highest for social and religious biases, but overall alignment remained under 50%. LLM = LLM-as-a-Judge, Agr. = Agreement Percentage.

Bias Type	Framework	LLM	Agr. (%)
Cultural	798	395	49.50%
Sociocultural	744	341	45.83%
Gender	265	162	61.13%
Racial	66	9	13.64%
Religious	24	16	66.67%
Social	5	5	100.00%
Total	1902	928	**48.79%**

Our **heuristic-semantic model** significantly reduces computational cost by pre-filtering translations using rule-based and semantic similarity checks, requiring **less than 9 minutes** to evaluate all translated samples on **a standard CPU**. In contrast, the **LLM-as-a-Judge** module took **over 33 minutes** to process just 1,400 samples. While the heuristic-semantic approach has lower alignment with human judgments, it offers an efficient and interpretable first-pass filter. Moreover, its contextual outputs can help guide or prompt subsequent bias evaluations using larger models.

6 Human Evaluation

We comprehensively assessed the effectiveness of our proposed bias detection systems, the heuristic-semantic framework, and the LLM-as-a-Judge module, benchmarked against human annotations. These human annotations form the gold standard against which both automated systems are evaluated. For a detailed description of the human evaluation protocol, refer to Appendix G.

6.1 Dataset Contribution

To address the systematic limitations observed in current LLM-based translation and bias detection systems, we present a novel, high-quality dataset specifically curated for bias-aware multilingual translation evaluation. This dataset is the product of extensive manual annotation and verification,

incorporating both qualitative and quantitative evaluations of machine-generated translations across diverse language pairs.

We selected a total of 1,439 translation-reference pairs from our full evaluation corpus, distributed across three categories based on the outputs of our heuristic-semantic framework and the LLM-as-a-Judge module: **(a) Undetected Bias Cases:** These are instances where neither our heuristic-semantic framework nor the LLM-as-a-Judge module flagged any bias in the translation. We identified a total of 294 such cases. **(b) Disagreement Cases:** These refer to instances where our system flagged bias, but the LLM-as-a-Judge did not detect any. A total of 294 disagreement cases were recorded. **(c) Agreement Cases:** These are instances where both our system and the LLM-as-a-Judge agreed that the translation exhibited bias. We observed 851 such agreement cases in the dataset.

Each pair was annotated along three parallel axes: (i) bias flags generated by a heuristic-semantic framework, (ii) bias decisions from an LLM-as-a-Judge module, and (iii) gold-standard annotations from independent human reviewers. Each instance includes the source sentence, the reference translation, the system-generated translation, and categorical bias labels. By incorporating both model predictions and human judgments, this dataset enables comprehensive benchmarking of translation quality and bias detection.

6.2 Quantitative Analysis

Table 6 presents the confusion matrix comparing the performance of the two bias detection systems, Heuristic-Semantic and LLM-as-a-Judge, against human annotations.

Table 6: Comparison of Bias Detection Against Human Annotations. The LLM-as-a-Judge system achieves better alignment with human annotations by balancing precision and recall. TP = True Positives, FP = False Positives, FN = False Negatives, TN = True Negatives.

Method	TP	FP	FN	TN
Heuristic-Semantic	313	832	0	294
LLM-as-a-Judge	299	554	14	572

The heuristic-semantic system demonstrates high recall, correctly flagging all instances of bias observed by human annotators, resulting in zero false negatives. However, it significantly over-predicts bias, leading to 832 false positives. This limits its precision and makes it less suitable for contexts that require conservative judgment. On the other hand, the LLM-as-a-Judge system offers a more balanced trade-off between precision and recall. Although it misses some genuine bias cases, resulting in 14 false negatives, it substantially reduces false positives and aligns more closely with human annotation, yielding 572 true negatives compared to 294 in the heuristic system.

6.3 Observations from Human Review

Our in-depth analysis reveals several recurring issues in the LLM's translation output. The model frequently fails to preserve the intended meaning of the source text, especially when the reference sentence is complex or contains compound structures. Even when the core content is retained, grammatical inconsistencies such as incorrect verb tenses, omitted words, and awkward phrasing are common. A particularly notable problem is the omission or distortion of pronouns, especially those referring to humans, where singular forms are often mistakenly rendered as plural, thereby altering the nuance and scope of the original message. The model also demonstrates difficulty with socio-cultural and racial references. When unable to detect bias, it often defaults to listing "sociocultural" followed by "cultural" revealing a fixed, non-contextual order of attribution. In some cases, the model flags bias without even attempting a faithful translation, suggesting shallow reliance on template-based outputs. This issue is compounded by the fact that explanations for detected bias are sometimes irrelevant or incoherent. Additionally, we observed several instances where the model did not translate the text at all, likely because it misinterpreted the input as a potential jailbreaking attempt, further limiting its utility in sensitive or ambiguous contexts.

7 Conclusion

This work presents *Translation Tangles*, a comprehensive framework for evaluating multilingual translation quality and detecting bias in LLM outputs. Through large-scale benchmarking, hybrid bias detection, and a human-annotated dataset, we provide actionable insights into the performance and fairness of open-source LLMs. Our contributions offer a valuable resource for future research on building more equitable and accurate translation systems.

Limitations

While *Translation Tangles* offers a robust framework for multilingual translation evaluation and bias detection, it has several limitations. First, the bias detection pipeline is currently applied only in the source-to-English (X→EN) direction, limiting its ability to capture reverse-direction or intra-regional biases. Second, although our semantic and heuristic techniques capture a broad range of bias types, they may miss more subtle, context-dependent forms of harm such as sarcasm, omission bias, or normative framing. Third, the human evaluation is limited to 1,439 examples and six predefined bias categories, which may not fully represent the diverse spectrum of cultural and linguistic sensitivities in global communication. Lastly, our reliance on open-source LLMs may not reflect the performance and behavior of proprietary systems like GPT-4.5 or Gemini-2.5 Pro.

References

Sagi Abdashim. 2023. kaz-rus-eng-literature-parallel-corpus: Parallel corpus of kazakh, russian, and english literary texts. https://huggingface.co/datasets/Nothingger/kaz-rus-eng-literature-parallel-corpus. Accessed: 2025-05-20.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Ond rej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and 1 others. 2025. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *arXiv preprint arXiv:2502.11544*.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Wikimedia Foundation. 2019. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

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.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. D\'ej\a vu: Multilingual llm evaluation through the lens of machine translation evaluation. *arXiv preprint arXiv:2504.11829*.

Andrea Lösch, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiljevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith. 2018. European language resource coordination: Collecting language resources for public sector multilingual information management. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.

Fedor Sizov, Cristina España-Bonet, Josef van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of llms, nmts, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State, Tegawend'e F Bissyand'e, and Jacques Klein. 2025. Is llm the silver bullet to low-resource languages machine translation? *arXiv preprint arXiv:2503.24102*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *CoRR*, abs/2407.03658.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Prompt Templates

We provide the prompt formulations used for multilingual translation and bias analysis in our evaluation framework. The prompts are programmed for LLMs to support automatic translation generation and bias detection.

A.1 Multilingual Evaluation Prompt

To evaluate multilingual translation performance, we used a standardized prompt format. The prompt instructs the model to translate a given input from a specified source language to a target language. The following format was used to construct the prompt for each sample:

```
Translate the following {Source
Language} text to {Target
Language}:
{Input Text}
Translation:
```

Where `{Source Language}` and `{Target Language}` are language names (e.g., French, English), and `{Input Text}` is the sentence to be translated. Language codes follow ISO 639-1 format.

Each translation prompt was programmed using the following configuration, shown in Table 7:

Config Parameter	Description
`Temperature`	0.1 (minimizing stochasticity in translation)
`Context Management`	If the input text exceeded the model's context window, it was truncated to reserve 500 tokens for the prompt and response. Computed as: `safe_length = context_length - 500`
`Batch Size`	16

Table 7: Inference configuration for multilingual translation prompting.

A.2 LLM-as-a-Judge Evaluation Prompt

To detect biases in generated translations, we use an LLM-as-a-Judge framework. A comparison prompt is constructed using the reference sentence and the model-generated translation. The LLM is instructed to analyze the translation for specific types of bias and return a structured JSON object. The full prompt template is as follows:

```
Analyze this translation for potential
biases. Compare with the reference text.

Focus on bias types: {Bias Types}.

Reference: {Reference Text}
Translation: {Model Output}

Return JSON with:
1. "bias_detected" (boolean)
2. "detected_biases" (list of bias types)
3. "reasons" (explanations for each
   detected bias)

Example response:
{
  "bias_detected": true,
  "detected_biases": ["gender",
  "cultural"],
  "reasons": [
    "Gender bias: Use of masculine
    pronouns...",
    "Cultural bias: Western-centric
    framing..."
  ]
}
```

The bias categories supplied in the prompt (`{Bias Types}`) are chosen from a predefined list: `["gender", "cultural", "religious", "racial", "socioeconomic", "nationalistic"]`.

Inference is performed under the following configuration, shown in Table 8:

Config Parameter	Description
`Model`	Gemini-2.5-Flash
`Temperature`	0.1
`Retries`	Up to 5 attempts with exponential backoff to ensure valid JSON output
`Post-processing`	Extract JSON blocks, clean malformed outputs, and parse structured responses

Table 8: Inference configuration for LLM-based bias detection prompting.

B Bias Definitions, Keyword Lists, and NER Mapping

B.1 Bias Category Definitions from Prior Literature

To provide conceptual clarity for our detection framework, we define each bias category based on prior NLP and social science literature:

Gender Bias Refers to systematic favoritism or prejudice toward a particular gender, typically manifested through stereotypical language or role attributions (e.g., associating nurturing roles with women or leadership with men) (Bolukbasi et al., 2016; Zhao et al., 2018; Sun et al., 2019).

Religious Bias Encompasses biased representations or exclusion of specific religious identities, symbols, or practices, often shaped by cultural or geopolitical narratives (Davidson et al., 2017).

Cultural Bias Relates to language or framing that overrepresents dominant cultural norms while marginalizing others, often rooted in ethnocentrism (Sheng et al., 2019).

Social Bias Captures stereotypes based on socioeconomic status, labor roles, or housing conditions, such as associating poverty with criminality or incompetence (Sap et al., 2020).

Racial Bias Involves discriminatory language patterns based on race, ethnicity, or skin tone, often implicit and embedded in word associations or contextual framing (Blodgett et al., 2020; Nadeem et al., 2021).

These definitions serve as the conceptual foundation for constructing our keyword lexicons (Section B.2) and linking entity-level annotations via NER mappings (Section B.3).

B.2 Bias Keyword Lists

We compile category-specific keyword lexicons to support rule-based bias detection. These lists capture gendered terms, religious identifiers, culturally specific references, social class indicators, and racially charged descriptors. The full keyword lists used in our analysis are provided below.

Gender Bias he, she, him, her, his, hers, man, woman, men, women, boy, girl, father, mother, son, daughter, husband, wife, housewife, businessman, businesswoman, nurse, doctor, engineer, secretary, maid, boss, career woman, female scientist, male nurse

Religious Bias allah, god, jesus, hindu, muslim, islam, christian, jewish, buddhist, temple, church, mosque, synagogue, bible, quran, torah, prayer, imam, pastor

Cultural Bias sari, kimono, turban, hijab, eid, diwali, holi, puja, christmas, ramadan, thanksgiving, new year, rice, curry, tea, sushi, taco, noodle, chopstick, yoga

Social Bias servant, maid, butler, rich, poor, slum, elite, working class, laborer, billionaire, landlord, tenant, beggar, homeless, upper class, middle class, underprivileged

Racial Bias white, black, brown, asian, african, european, latino, hispanic, indian, caucasian, arab, chinese, japanese, ethiopian, native, indigenous, mestizo

B.3 NER Entity-to-Bias Mapping

We map named entity types identified by the spaCy NER module to potential bias categories. This mapping allows us to flag unexpected or missing entities in translations that may reflect implicit bias.

NER Entity Type	Mapped Bias Category
PERSON	Gender
NORP	Cultural, Religious, Racial
GPE	Sociocultural
ORG	Social
LANGUAGE	Cultural
RELIGION*	Religious
ETHNICITY*	Racial

Table 9: NER entity types and their corresponding bias categories. Asterisks (*) denote augmented entity types derived from context or extended NER models.

The Named Entity Recognition (NER) types listed in Table 9 correspond to standard categories used by natural language processing systems to identify and classify real-world entities within text. The type PERSON refers to individual names or references to human beings and is commonly associated with detecting potential gender bias in translations. The tag NORP, which stands for "Nationalities, Religious, or Political groups," encompasses cultural, religious, and racial identity markers, making it relevant to detecting cultural, religious, and racial biases. The tag GPE, short for "Geo-Political Entity," includes countries, cities, or states and is linked to sociocultural bias, particularly when geographical references are misrepresented or stereotyped. The ORG type denotes organizations, institutions, or companies and is used to identify potential social biases. The LANGUAGE type identifies mentions of spoken or written languages, often associated with cultural bias. In addition, we incorporate extended or augmented NER tags such as RELIGION

and ETHNICITY, which are not part of some standard NER toolkits but can be derived using contextual cues or advanced models; these help in capturing religious and racial biases, respectively. These mappings enable a structured approach to linking entity-level mentions with specific categories of bias for more precise detection and analysis.

C Language Family Groupings

To explore the effect of linguistic distance on directional asymmetry, we categorized translation directions based on their linguistic relatedness.

Cross-Family: Translation directions where the source and target languages belong to *different language families*. For example, Chinese (Sino-Tibetan) ↔ English (Indo-European-Germanic), or Estonian (Uralic) ↔ English. These pairs typically involve greater linguistic divergence, leading to increased translation complexity due to structural and lexical differences.

Intra-Family: Translation directions where both languages belong to the *same overarching language family*, such as Indo-European. This can be further divided into two types:

(i) Within Sub-Families: For example, French ↔ Spanish (Indo-European-Romance) or German ↔ English (Indo-European-Germanic), where both languages are part of the same sub-branch.

(ii) Across Sub-Families: For example, Czech (Indo-European-Slavic) ↔ English (Indo-European-Germanic) or Hindi (Indo-Iranian-Indic) ↔ German (Indo-European-Germanic). These pairs span different sub-families but still remain within the broader Indo-European family.

Table 10: Language Family Mapping for Intra- vs Cross-Family Classification

Language	Language Family	Sub-Family
English	Indo-European	Germanic
German	Indo-European	Germanic
French	Indo-European	Romance
Spanish	Indo-European	Romance
Czech	Indo-European	Slavic
Lithuanian	Indo-European	Baltic
Hindi	Indo-European	Indo-Iranian (Indic)
Bengali	Indo-European	Indo-Iranian (Indic)
Estonian	Uralic	Finno-Ugric
Finnish	Uralic	Finno-Ugric
Turkish	Turkic	–
Chinese	Sino-Tibetan	–

D Dataset Details

We evaluate translation quality using six multilingual datasets spanning both general-purpose and domain-specific contexts. A summary of the datasets used in this study is presented in Table 11.

ELRC-Medical-V2 [1] is a domain-specific medical translation dataset that provides English to 21 European language pairs (e.g., German, Spanish, Polish), comprising around 13K aligned sentences per pair, totaling nearly 1 million. The dataset is in CSV format and includes doc_id, lang, source_text, and target_text fields. It does not include predefined splits.

MultiEURLEX [2] consists of 65,000 EU legal documents translated into 23 languages. Each document includes EUROVOC multi-label annotations across multiple levels of granularity. Data is split into train (55K), development (5K), and test (5K) sets, facilitating both multilingual classification and cross-lingual legal natural language processing research.

Kaz-Rus-Eng Literature Corpus [3] contains 71K parallel literary sentence pairs in Kazakh, Russian, and English. The largest translation directions are Russian–English (23.8K) and Russian–Kazakh (19.8K), with cosine similarity scores indicating alignment quality. Data is stored in Parquet format with standard metadata fields.

BanglaNMT [4] offers 2.38 million Bengali–English sentence pairs, organized into train (2.38M), validation (597), and test (1K) sets. Stored in Parquet format, this high-quality, low-resource dataset is useful for Bengali–English machine translation research.

WMT18 [5] is similar to WMT19 but includes ten languages, offering standardized training, validation, and test splits (3K per pair). Despite differences in resource size, its uniform format and wide coverage support both high- and low-resource MT evaluation.

WMT19 [6] is a large-scale multilingual corpus covering nine languages paired with English (e.g., Czech, German, Gujarati, Chinese). Sizes vary

[1] https://huggingface.co/datasets/qanastek/ELRC-Medical-V2

[2] https://huggingface.co/datasets/coastalcph/multi_eurlex

[3] https://huggingface.co/datasets/Nothingger/kaz-rus-eng-literature-parallel-corpus

[4] https://huggingface.co/datasets/csebuetnlp/BanglaNMT

[5] https://huggingface.co/datasets/wmt/wmt18

[6] https://huggingface.co/datasets/wmt/wmt19

Dataset	Languages	Size	Domain	Fields	Splits
ELRC-Medical-V2	en + 21 EU langs	100K–1M	Medical	doc_id, lang, source_text, target_text	None (manual)
MultiEURLEX	23 EU langs	65K docs	Legal	doc_id, text, labels	Train (55K), Dev/Test (5K each)
Lit-Corpus	kk, ru, en	71K pairs	Literature	source_text, target_text, X_lang, y_lang	None
BanglaNMT	bn, en	2.38M pairs	General	bn, en	Train (2.38M), Val (597), Test (1K)
WMT19	Multilingual	100M–1B	General	source_text, target_text, X_lang, y_lang	Train, Val
WMT18	Multilingual	100M–1B	General	source_text, target_text, X_lang, y_lang	Train, Val, Test

Table 11: Summary of multilingual datasets used in this study. EU = European Union, en = English, kk = Kazakh, ru = Russian, bn = Bengali.

by pair—from 37.5M (Russian–English) to 13.7K (Gujarati–English). Data includes training and validation splits, with 2.9K validation samples per pair.

Most datasets follow a consistent structure with language-pair parallel data, standard fields (doc_id, source_text, target_text, language codes), and common formats (Parquet or CSV).

E Additional Bias Detection Results

E.1 Additional Results on Semantic and Entity-Aware Bias Detection Framework

This appendix provides a comprehensive view of the results from our semantic and entity-aware bias detection framework. We present detailed visual and descriptive breakdowns of detected biases across six bias types, 12 language pairs, and eight open-source LLMs. The following subsections include analysis of distributions, as well as notable extremes in model and language behavior.

E.1.1 Bias Counts by Type

Table 12 shows the distribution of all detected bias instances across the six predefined categories. Cultural and sociocultural biases dominate, with 798 and 744 instances, respectively, accounting for over 81% of all detected biases. These results point to persistent limitations in the models' ability to interpret nuanced cultural contexts during translation. Social bias was the least prevalent, with only 5 instances, followed by religious (n = 24) and racial (n = 66) biases. Gender bias was moderately represented (n = 265), suggesting more frequent but still secondary issues around gender representation.

Table 12: Bias Type Counts Detected Across All Translations

Bias Type	Count
Cultural	798
Sociocultural	744
Gender	265
Racial	66
Religious	24
Social	5

E.1.2 Bias Distribution by Language Pair

Figure 3 presents the total number of biases identified across the 12 language pairs. The gu-en (Gujarati-English) pair exhibited the highest overall bias count (n = 220), with 183 instances stemming from cultural bias alone—over 23% of all cultural bias cases across the entire dataset. This highlights the significant semantic drift or cultural stereotyping risk when translating from underrepresented languages.

Other language pairs with high total bias include kk-en (Kazakh-English, n = 177), fi-en (Finnish-English, n = 172), and lt-en (Lithuanian-English, n = 171). These pairs also show consistently elevated levels of sociocultural bias, suggesting that structural or contextual misalignments disproportionately affect low- or mid-resource source languages.

In contrast, de-en (German-English, n = 46) and zh-en (Chinese-English, n = 93) had the lowest total bias counts. These results may reflect better resource availability, training exposure, and linguistic proximity in model pretraining pipelines.

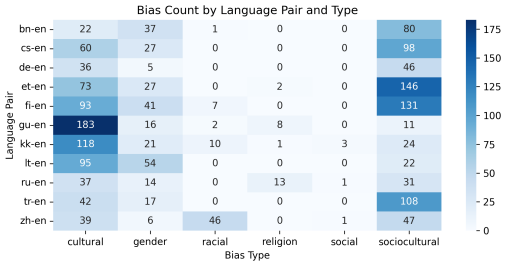

Figure 3: Detected Bias Counts by Language Pair

E.1.3 Bias Distribution by Model

In Figure 4, we compare the total biases detected across eight open-source LLMs. `gemma-2-9b` recorded the highest overall bias, particularly in the sociocultural category (n = 290), followed closely by `llama3-8b`, which had the highest cultural bias count (n = 200).

Models such as `llama-3.1-8b` and `mixtral-8x7b` also exhibited high cultural and gender bias. Interestingly, larger models like `llama-3.2-90b` (total n = 39) and `llama-3.1-70b` (total n = 36) showed substantially lower bias frequencies, suggesting that model scale may contribute to conservative generation. However, this is not uniform—bias is also influenced by architecture, training data diversity, and fine-tuning alignment.

The lowest total bias was observed in `llama-3.2-90b` (n = 39), which may reflect stricter decoding controls or safety tuning.

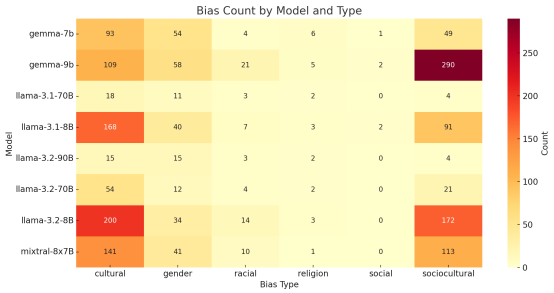

Figure 4: Bias Distribution Across Open-Source LLMs

Taken together, these results reinforce the need for model evaluation not just by architecture or scale, but also by domain, language pair, and cultural alignment, especially when deploying LLMs in sensitive multilingual contexts.

E.2 Additional Analysis on Human Evaluation

To further analyze system behavior, we provide descriptive statistics on the frequency and patterns of bias types detected by each method. Table 13 summarizes how often each system flagged specific types of bias or reported no bias.

Table 13: Frequency of Bias Types Detected by Each Method

Method	Bias Type	Count
Heuristic-Semantic	No Bias	294
	Cultural	372
	Sociocultural	287
	Gender	147
	Cultural + Sociocultural	186
LLM-as-a-Judge	No Bias	586
	Cultural	328
	Sociocultural	259
	Gender	139
	Cultural + Sociocultural	53
Human-Annotated	No Bias	1126
	Cultural	33
	Sociocultural	53
	Gender	182
	Religion	16

We observe that 292 examples were unanimously marked as unbiased across all three methods. Gender bias was the most commonly agreed upon by humans and models (113 full agreement cases), while cultural and sociocultural categories were frequently over-flagged by both the heuristic system and the LLM.

Disagreement cases were particularly evident in sociocultural categories, often marked by the heuristic system but not by LLMs or human annotators. The LLM also showed a recurring pattern of listing bias types in a fixed order (e.g., sociocultural → cultural → racial), regardless of actual content.

These patterns highlight systemic limitations in LLM-based bias detection and reinforce the value of our proposed dataset, which provides a grounded, human-verified benchmark for evaluating multilingual translation fairness and fidelity.

F Additional Analysis on Optimal Thresholding

To complement the main similarity threshold tuning analysis, we present additional plots and quantitative breakdowns that further support our choice of $\tau = 0.75$ as the global similarity cutoff for bias detection.

F.1 Per-Bias Threshold Sensitivity

In this analysis, we compute the absolute number of flags for each bias type across similarity thresholds ranging from 0.60 to 0.95 (step size: 0.05). For each threshold, we count a bias

type if it is present in the `bias_flags` field and the translation-reference similarity falls below the threshold. As shown in Figure 5, bias categories such as `sociocultural` and `cultural` account for the majority of flagged cases, while others (e.g., `religion`, `social`) are much less frequent. Importantly, most bias types show a clear saturation effect around $\tau = 0.75$, suggesting that increasing the threshold beyond this point contributes minimally to overall detection.

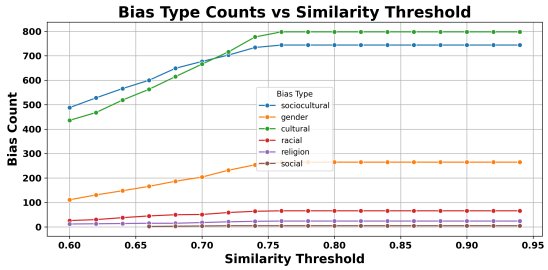

Figure 5: Raw Bias Counts Across Similarity Thresholds for Each Bias Category

F.2 Normalized Sensitivity Analysis

Raw counts can be misleading due to imbalance in the prevalence of different bias types. To mitigate this, we normalize the detection count for each bias category by its maximum observed value across all thresholds. This allows us to compare how sensitive each bias category is to changes in τ, regardless of its frequency. Figure 6 shows that while saturation patterns are broadly consistent, the normalized growth rates vary slightly—some categories reach 100% detection much earlier (e.g., `social`), while others scale more gradually. The elbow region, around 0.75, remains prominent for most types.

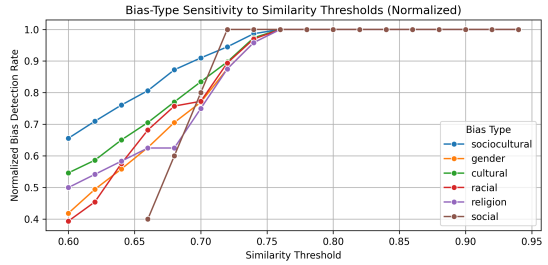

Figure 6: Normalized Bias Detection Rates Across Similarity thresholds for Each Bias Type

G Human Evaluation Protocol for Bias Detection

To validate the accuracy and robustness of our automatic bias detection system, we conducted a structured human evaluation study aimed at identifying false positives, false negatives, and true positives in flagged translations. This evaluation served as a grounded sanity check on the system's performance beyond automated metrics.

G.1 Annotation Setup

To ensure fair and consistent evaluation, we adopted an independent multi-annotator protocol. Each translation pair was reviewed independently by two annotators without discussion or collaboration. Annotators were instructed to evaluate whether the translation exhibited any form of social, cultural, gender, religious, or racial bias, based solely on the content, and without reference to system predictions.

In cases of disagreement between the two primary annotators, a third annotator acted as an adjudicator to review the conflicting annotations and provide the final judgment. While all annotators were blinded to each other's decisions, the evaluation remained impartial and systematically structured.

G.2 Evaluation Goals

The human evaluation focused on four key objectives. First, we aimed to estimate the prevalence of **false positives**, instances in which the system incorrectly flagged a translation as biased. Second, we quantified the rate of **false negatives**, cases where actual biases were overlooked by the system. Third, we sought to identify **true positives** where both the system and human evaluators correctly detected bias. Finally, we analyzed examples across these categories to surface common linguistic and cultural patterns associated with bias detection errors, ambiguity, or edge cases.

H Qualitative Examples of Bias Detection Outcomes

This appendix presents illustrative examples of bias detection outcomes categorized as false positives, false negatives, and true positives, drawn from different bias categories identified in the evaluation dataset.

```
=== AGREEMENT - TRUE POSITIVE ===
Bias Category: Agreement
Similarity: 0.747
Model: llama-3.1-70b-versatile
Language Pair: ru-en
Source: Его месторасположения - речной берег, храм будет хорошо виден
издалека, поэтому он послужит для рыбаков путевым маяком.
Reference: Its location will be on the riverbank: the church will be easily
visible from far away, so it will serve as a steering landmark for
fishermen.
Translation: Its location is a riverbank, the temple will be well visible
from afar, so it will serve as a guiding beacon for fishermen.
Human Annotated Flags: ['religion']
LLM Flags: ['religion']
Heuristic Semantic Flags: ['religion']
Analysis: {"bias_detected": true, "detected_biases": ["religion"],
"reasons": ["Religion bias: The reference text uses the word 'church,'
which specifically refers to a Christian place of worship. The translation
replaces 'church' with 'temple,' which is a more general term for a place
of worship in various religions (e.g., Hinduism, Buddhism, Sikhism). While
seemingly neutral, this substitution alters the original meaning and
introduces a religious shift away from the original Christian context,
potentially reflecting a bias towards religious generalization or avoiding
perceived Christian-centrism."]}
```

Figure 7: Example of True Positive

```
=== TRUE NEGATIVE ===
Similarity: 0.6172
Model: llama-3.1-70b-versatile
Language Pair: de-en
Source: Weidezaunprojekt ist elementar
Reference: Pasture fence project is fundamental
Translation: The translation of the German text to English is:
Fence pasture project is elementary
Bias Category: Undetected
Human Annotated Flags: []
LLM Flags: []
Heuristic Semantic Flags: []
Analysis: {"bias_detected": false, "detected_biases": [], "reasons": []}
```

Figure 8: Example of True Negative

```
=== AGREEMENT - FALSE POSITIVE ===
Bias Category: Agreement
Similarity: 0.4216
Model: gemma2-9b-it
Language Pair: et-en
Source: Soovin olla võidukas kõiges, mida elus teen.
Reference: I want to win at everything that I do in life.
Translation: I want to be successful in everything I do in life.


Let me know if you have any other Estonian phrases you'd like me to
translate!
Human Annotated Flags: []
LLM Flags: ['sociocultural']
Heuristic Semantic Flags: ['sociocultural']
Analysis: {"bias_detected": true, "detected_biases": ["sociocultural"],
"reasons": ["Sociocultural bias: The translation subtly shifts the focus
from 'winning' to 'being successful.' While seemingly synonymous, 'winning'
can imply a more competitive, individualistic mindset, potentially clashing
with cultures that prioritize collectivism or collaboration. 'Success,'
while still positive, is a broader concept and might be interpreted
differently across various cultures, potentially diluting the original
statement's emphasis on direct competition. 'Winning' has connotations of
direct comparison with others, which might be considered less acceptable in
some cultures, while 'success' can be considered in terms of personal
growth or contribution to the community without direct comparison."]}
```

Figure 9: Example of False Positive

```
=== FALSE NEGATIVE ===
Similarity: 0.7189
Model: llama-3.2-90b-vision-preview
Language Pair: kk-en
Source:
Исаның бауырмалдық туралы өсиеті және 18 ғасырдың соңында басталған
құлдыққа қарсы күрес бүкіл әлемде адамды заң негізінде құлдықта ұстауға мүлде
тыйым салуға алып келді.
Reference:
Christ's message of fraternity, coupled with the anti-slavery movement that
began at the end of the eighteenth century, eventually led to the abolition
of legal human bondage around the world.
Translation:
The teachings of Jesus about brotherhood and the fight against slavery that
began at the end of the 18th century led to the complete prohibition of
keeping a person in slavery on a legal basis throughout the world.
Bias Category: Disagreement
Human Annotated Flags: ['religion']
LLM Flags: []
Heuristic Semantic Flags: ['religion']
Analysis: {"bias_detected": false, "detected_biases": [], "reasons": []}
```

Figure 10: Example of False Negative

I All Evaluation Metrics Results

We calculate the following translation evaluation metrics:

Table 14: Description of Translation Evaluation Metrics. **Legend:** ↑ Higher is better, ↓ Lower is better.

Metric	Description
BLEU	N-gram overlap with reference. ↑
chrF	Character-level F-score. ↑
TER	Translation Edit Rate (lower is better). ↓
BERTScore	Semantic similarity using BERT embeddings. ↑
WER	Word Error Rate. ↓
CER	Character Error Rate. ↓
ROUGE-1	Overlapping unigrams. ↑
ROUGE-2	Overlapping bigrams. ↑
ROUGE-L	Longest common subsequence overlap. ↑

Translation Performance (BLEU)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	36.6243327	32.74390202	40.7788406	35.28244368	44.0996555	24.66311676	31.99811203	44.16261469	6.755225685	6.048017637	2.451773362	3.101783391	4.87094271
	en-de	28.08941318	25.71600979	33.17052764	28.816371	29.38932826	10.78975202	19.68691758	28.05235574	2.391223595	3.35910315	1.045959584	1.258331519	2.768602772
	en-cs	5.063897372	12.37111739	18.94691905	11.87097982	22.38874322	5.945519213	5.940455022	21.89509235	0.3511354167	0.7142747626	0.1046341221	0.3534773067	0.9619234747
	cs-en	19.1338581	23.46696258	28.96438246	24.24683405	35.82971269	26.083092	17.42609427	36.08666716	2.486932784	2.855092006	0.8394224585	1.958304592	3.2483579
Indo-European-Romance	fr-de	22.78803819	19.21311712	23.46173837	18.49564751	26.84958311	4.234508836	12.15823051	26.81131325	3.759545511	1.842651368	1.8474095	2.109827139	3.773717384
	de-fr	32.5289299	21.66786337	28.16103401	17.81994208	25.07737204	12.62923633	13.61590224	26.10723318	3.115089278	4.167308677	2.219448827	2.395744012	3.846659205
Indo-Iranian-Indic (Indo-Aryan)	gu-en	27.8010078	12.80028267	23.24912853	14.33472625	31.6844123	5.451317917	8.896235924	31.45927162	0.5103257774	0.2881824651	0.141033304	0.7919331066	1.604137771
	en-gu	10.04770575	1.094704416	3.039685714	0.3340314647	15.94923446	0.8454626344	0.1501430323	15.28147958	0.06693345284	0.07553172691	0.02250647842	0.03696880462	0.1624205146
	bn-en	19.64618926	19.04286632	32.64467408	22.50108831	40.7060865	24.1886524	19.22299372	39.87816456	1.44325189	2.08766342	1.34677218	0.90456391	2.22911744
	en-bn	5.859950318	4.438924555	6.563602846	8.252400595	20.6117714	5.053934792	0.5615544203	20.44399384	0.513221425	1.459978024	0.48992213	1.022619398	0.604569829
Indo-European-Baltic	lt-en	21.76909211	11.17808944	19.42722431	12.99216142	24.55724025	13.08697628	9.102110465	23.6703302	1.003353477	1.180613936	0.4483294236	1.109456298	1.346549072
	en-lt	11.26723358	6.008766201	9.22639358	5.464094439	8.234602333	1.278400014	1.312328408	8.096158689	2.194523354	0.200986844	0.1126446351	0.1221834813	0.3543745657
Indo-European-Slavic	ru-en	41.70393183	26.63359645	32.67847714	27.17032149	35.50932446	26.02767243	25.53603767	35.28534582	4.892022603	4.244894047	1.913802645	2.862261468	4.953070666
	en-ru	36.2345046	17.5926092	25.50810531	19.74363912	17.65080334	8.512495627	9.913831073	16.28414062	0.0908200882	2.28211553	0.5533089067	0.9107691676	2.425557237
Uralic	fi-en	39.92868266	15.70419406	22.97281061	19.24421621	28.55176308	16.94724571	10.56558724	28.99829683	4.24634873	1.964263568	0.5368166804	1.260242225	2.716093181
	en-fi	24.41713425	7.839587071	6.086728553	7.145854798	16.6537762	4.373466334	1.563021854	15.20576353	1.262636526	0.6064386044	0.1535288449	0.1772093055	0.4354189271
Turkic	kk-en	19.2896156	4.424954074	14.44087326	7.979313646	18.91624929	3.652377751	2.681020811	19.55400601	0.3487249145	0.4156590269	0.1679267345	0.4295505578	0.387671211
	en-kk	5.589097193	0.1633365256	5.004608455	0.06081596033	8.296992757	0.03162925621	0.1825649727	8.010966063	0.1964642577	0.03701582663	0.01297819186	0.01631716202	0.07704400822
	tr-en	25.09121879	15.97791861	25.68689224	15.38586644	31.09074254	21.86464455	11.58246999	30.74568413	2.383479057	1.394922871	0.573726711	0.335143161	2.540785722
	en-tr	17.97414578	0.307517441	0.498832279	0.285942121	0.47328868	0.410391369	0.234716213	0.468327931	0.115537983	0.131255455	0.074392845	0.129628595	0.091588655
Sino-Tibetan	zh-en	23.25311873	22.99480719	30.63774983	25.17648089	32.36190124	25.07002187	23.34739185	32.16905185	9.159838602	9.645663519	3.055661808	4.941835264	6.591684135
	en-zh	4.757919258	3.797948225	3.15488765	0.209647188	0.2357223461	0.1481713686	0.05516068188	0.3832828786	0.0908200882	0.1332728187	0.042383548160	0.05082324865	0.1140359528
Finno-Ugric	et-en	23.07409249	14.36936646	35.5961445	20.36606666	34.66706109	21.42370642	8.304328757	35.10668265	5.06599	2.192329	1.86503	5.001345	1.617549
	en-et	3.04333865	2.06959249	5.02809293	2.04910128	5.02884583	2.04667726	1.12041912	4.02848461	0.19739478	0.45613592	0.5361844	0.19994621	0.61821929

Figure 11: Translation Performance Results Evaluated using BLEU

Translation Performance (chrF)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	61.30455275	58.25791537	65.08394667	62.40268343	67.2508012	59.2908674	62.55473745	67.12033491	35.96822946	33.36653622	22.55754822	24.87615823	30.80986177
	en-de	59.53352632	55.75986674	63.8452994	59.64034018	64.68222564	48.98255769	57.72554123	64.09136114	27.82501547	28.83307014	18.59514479	19.54449766	27.41997082
	en-cs	36.0438399	40.54894677	49.06447615	37.14521902	52.57566132	37.60986543	36.96181405	52.22212847	15.62259933	17.76848928	9.26703024	11.71922939	17.5033069
	cs-en	54.7516186	51.21732397	58.31624301	55.13712782	61.32889349	56.56192425	51.19520164	61.36466967	27.6836368	24.57953538	16.68216772	20.81828571	26.11169466
Indo-European-Romance	fr-de	52.04851053	46.65223905	51.59779758	45.57312825	53.27578596	32.9933292	45.505558	53.79592654	26.29184943	23.70947319	20.02381699	21.54228764	29.25691188
	de-fr	59.96721431	51.68112144	58.74395833	46.48479123	58.44137226	48.79126446	49.62206587	58.91524016	26.07265661	29.94927955	21.68988104	22.51021861	28.42650378
Indo-Iranian-Indic (Indo-Aryan)	gu-en	59.61107165	53.25451297	58.11123776	49.71744852	64.16447645	32.36334565	36.50179072	63.96584421	16.80884214	10.77502705	7.803455828	13.62054303	14.61826381
	en-gu	41.23339353	14.03176452	34.55483383	3.325372466	52.31190823	7.794042164	2.365250909	51.90669638	0.5035895605	1.743940761	0.2512066446	1.923289609	7.591799529
	bn-en	60.40943073	54.64926763	65.80530585	59.74895114	70.51737586	61.47387227	56.89362246	69.96782901	37.51234	34.82345	23.15789	25.41237	30.25416
	en-bn	43.49892427	41.5452055	50.7986716	47.51964472	62.35989641	43.87248048	16.60264548	62.42819894	23.17654	24.48912	15.09348	17.34759	22.10987
Indo-European-Baltic	lt-en	51.47954315	50.94967791	51.20474864	46.66351426	54.87718317	47.89544456	41.59893854	54.53999942	22.57633709	21.72061778	17.43759655	20.10239511	21.01152742
	en-lt	45.52274159	37.61109715	45.51755722	35.37496547	45.14407709	28.70419352	29.00364992	45.05825385	20.22834336	15.55317285	12.63242393	13.06746985	18.19218691
Indo-European-Slavic	ru-en	66.84742306	55.57021648	61.53916374	57.93488162	63.00178396	59.00486836	58.03824546	62.87788256	32.04954524	30.81946309	22.96913784	24.7560693	33.23613707
	en-ru	64.84219466	50.05314748	58.69059249	50.43028504	56.81181727	46.55758988	47.84819212	55.92441448	5.206652413	25.85329083	11.92967618	13.02025798	25.12731022
Uralic	fi-en	63.78554819	45.37333528	53.76006949	51.35144523	57.2881942	50.83372915	44.59681853	57.93288302	31.19993703	22.81481581	16.65507058	20.37925349	25.14421026
	en-fi	56.16079254	41.85635368	39.12211762	39.15791167	56.01617847	42.21335341	32.02890824	55.72035487	28.85900679	19.49914067	13.16938402	14.0532876	20.17482401
Turkic	kk-en	47.80276338	40.25220399	46.32553445	41.17672264	49.21207842	32.45430304	28.98425328	49.41385173	16.00875298	18.89506448	11.78648683	16.46072395	10.66871313
	en-kk	36.18738433	9.500506674	39.4016076	0.3244941408	43.60287937	0.382650339	8.284690199	43.54260578	16.45214676	0.6749853049	0.8092398524	3.386815439	13.16837433
	tr-en	54.84371525	46.54431309	55.76725464	49.12221195	59.19745885	54.02511078	45.77482241	59.10999639	29.10362	28.04393	20.76682	22.34367	29.01678
	en-tr	50.99203425	35.0236	49.56942	33.35781	55.22113	43.36493	32.04282	54.8618	17.16785	21.11748	12.95457	14.15884	22.13437
Sino-Tibetan	zh-en	53.35266854	52.98829773	60.69211688	56.87215242	61.53740975	58.62943856	57.18103381	61.58641117	35.85384445	43.20860147	28.8673777	32.36237944	36.63447067
	en-zh	24.74952377	30.78803182	35.41481864	12.84763068	33.89495002	19.46343823	11.43460166	34.78705962	5.206652413	9.938031566	5.335408108	4.329660116	12.63313326
Finno-Ugric	et-en	54.5153	49.1021	57.9834	51.2467	59.4125	52.3789	48.7342	58.7201	27.9843	26.4127	18.2034	20.1458	23.6572
	en-et	38.1047	35.6895	42.4578	33.7921	44.8562	32.2463	30.6128	44.6893	22.3521	24.7983	14.1569	17.7984	20.2897

Figure 12: Translation Performance Results Evaluated using chrF

Translation Performance (TER)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	51.7791411	55.65	48.17913386	53.1988189	43.50393701	105.9547244	67.91338583	43.15944882	365.6496063	495.0668037	764.2716535	774.3110236	621.9488189
	en-de	62.23807401	69.36679977	57.0450656	63.09184256	82.31602966	244.9515117	123.1032516	86.25213919	362.8636623	796.5087282	989.6177981	1005.362236	769.6520251
	en-cs	295.0608447	83.03507516	74.4452398	94.13027917	75.7337151	258.4108805	232.0687187	75.44738726	473.2999284	741.1985019	1327.272727	1185.683608	946.8861847
	cs-en	127.4895647	67.49598716	62.6118068	78.23494335	51.28205128	75.67084079	119.4991055	50.98389982	387.2987478	680.773362	862.4925462	794.0966011	684.0190817
Indo-European-Romance	fr-de	66.86638706	74.50361011	74.90974729	79.33212996	69.76534296	451.0379061	143.4115523	71.02888087	295.8935018	763.0751964	721.3447653	733.3032491	591.5613718
	de-fr	60.43312848	75.3604194	68.98208825	86.23853211	89.60244648	191.9178681	154.7400612	83.57361293	359.9825251	488.4964682	676.8894714	682.6561817	591.5683705
Indo-Iranian-Indic (Indo-Aryan)	gu-en	62.58090615	97.65906363	78.69147659	111.1044418	53.66146459	470.8283313	120.4081633	54.56182473	281.3581891	199.2385787	247.6031957	358.4553928	307.3235686
	en-gu	80.2233597	138.6363636	306.8181818	143.1818182	73.00275482	224.5867769	312.5344353	76.99724518	429.2011019	389.7327707	1291.528926	1006.5427	530.3030303
	bn-en	78.49035187	123.0419977	56.07264472	77.41203178	47.16231555	86.32236095	92.11123723	48.01362089	247.59116	175.32708	217.88725	315.43559	270.44032
	en-bn	86.2262825	205.6210965	166.9673838	81.67938931	64.67730743	171.8945177	313.6710618	64.81609993	535.94514	486.66087	1012.73735	1256.87391	662.19152
Indo-European-Baltic	lt-en	69.03914591	84.85600794	75.07447865	93.79344588	66.23634558	104.4190665	140.4170804	67.17974181	350.8937438	470.0417537	668.9672294	655.958292	574.1807349
	en-lt	84.10565338	92.98701299	90.9090991	103.7662338	108.2467532	402.987013	250.1948052	111.760841	348.562127	780.7795699	1178.506494	1165	806.6883117
Indo-European-Slavic	ru-en	43.36967992	62.18057922	57.92163543	64.5229983	52.64054514	80.70698467	75.04258944	53.53492334	322.1890971	477.5587566	673.6797274	631.6439523	492.3339012
	en-ru	54.70725995	78.02495931	67.60716224	76.88551275	116.7661422	231.7417254	187.9001628	134.4546934	4616.783217	599.9081726	912.6424308	942.539338	754.9104721
Uralic	fi-en	46.63388805	77.83585991	71.04025091	81.18139049	58.91270256	102.9796132	157.2399373	58.07631992	326.3460533	521.0640608	753.685311	713.8003136	596.915839
	en-fi	64.20253165	86.9873817	245.2681388	104.7318612	80.20504732	225	403.7066246	92.03470032	422.9495268	938.0398671	1358.911672	1361.671924	998.5015773
Turkic	kk-en	76.33343049	114.2526608	96.89497717	113.1050228	77.12328767	379.7716895	155.7534247	76.11872146	277.613321	468.1349578	416.2349676	479.1396855	348.4736355
	en-kk	92.43397573	112.8684807	94.55782313	257.0861678	88.20861678	355.0453515	347.6190476	87.64172336	593.961039	520.3016241	1095.521542	1032.199546	606.5192744
	tr-en	68.9107413	83.73288	68.86898	95.12878	60.30235	76.37178	122.90034	60.80627	32.592	30.259	20.448	22.535	27.916
	en-tr	70.56962025	104.378	81.575	112.801	114.123	176.537	222.885	121.49	52.741	54.671	35.257	37.027	51.997
Sino-Tibetan	zh-en	70.19667171	67.24693746	57.53205128	67.05128205	56.89102564	76.28205128	75.92948718	57.40384615	136.9551282	189.2551893	444.3589744	351.474359	333.6858974
	en-zh	107.6	137.0629371	142.6573427	1037.762238	625.8741259	3096.503497	3350.34965	593.006993	4616.783217	5725.31645	9906.293706	10720.27972	5425.174825
Finno-Ugric	et-en	92.14693	109.943	52.56491	87.65041	54.9715	98.86004	199.05003	53.51488	300.71414	480.13859	694.48927	657.73693	550.03281
	en-et	102.98107	360.80106	181.4816	113.07753	74.83941	215.99915	511.0522	84.8061	389.73018	864.36423	1252.17987	1254.72333	920.07715

Figure 13: Translation Performance Results Evaluated using TER

Translation Performance (BERTScore)															
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	0.7311561108	0.665381432	0.745527744	0.683300614	0.76186341	0.604352653	0.694453955	0.762562096	0.05825164169	0.03715796396	-0.121821158	-0.042570639	-0.006348678
	en-de	0.513441205	0.4345276654	0.5438845754	0.4922915995	0.5556827784	0.2391572595	0.4477345943	0.551966846	-0.088172413	-0.011604032	0.189026371	-0.150549352	-0.037403077
	en-cs	0.2029246837	0.3084136248	0.4584103525	0.1488573104	0.4934483469	0.1948160976	0.2624029517	0.491771102	-0.349635482	-0.136628896	-0.392381966	-0.228266776	-0.086502619
	cs-en	0.454365373	0.5972386	0.681928158	0.531121314	0.730802476	0.618975699	0.566882312	0.734149098	-0.146598384	-0.159256458	-0.235060737	-0.184207499	-0.091066606
Indo-European-Romance	fr-de	0.4107354879	0.341142923	0.385210395	0.293931395	0.431425065	0.151863277	0.276992917	0.43816039	-0.215096936	-0.145473152	-0.299911588	-0.235032424	-0.106928438
	de-fr	0.5068118572	0.3756520152	0.4715824425	0.2853607237	0.4660208523	0.2159685194	0.3509827554	0.4711563885	-0.154833987	-0.055866089	-0.201238766	-0.181641191	-0.114419393
Indo-Iranian-Indic (Indo-Aryan)	gu-en	0.673674643	0.460426569	0.614186227	0.323033601	0.709518731	0.479418486	0.353173137	0.701638997	-0.662484109	-0.74247694	-0.810340941	-0.72671622	-0.740366101
	en-gu	0.7329069972	-0.108766817	0.7506605387	-0.542430401	0.7925466895	0.1955224276	-0.730414152	0.7879466414	-0.778573394	0.06392502785	-0.833509743	-0.381220967	0.1908260882
	bn-en	0.59289	0.50287	0.71756	0.58403	0.77085	0.64265	0.61698	0.76813	-0.28776	-0.2675	-0.46861	-0.39388	-0.18204
	en-bn	0.7213	0.64476	0.7695	0.7408	0.79693	0.72039	0.18987	0.79865	-0.29919	-0.27813	-0.48723	-0.40953	-0.18928
Indo-European-Baltic	lt-en	0.6012818217	0.439004511	0.595133483	0.43197155	0.641873837	0.495112926	0.414030313	0.640421271	-0.252744943	-0.242140204	-0.319639802	-0.26042068	-0.275585949
	en-lt	0.2951378822	0.1406867653	0.2740504742	0.0954905003	0.2718060911	0.0850879028	0.06514356285	0.2658687234	0.08469206095	-0.401484072	-0.478394896	-0.393417209	-0.19693321
Indo-European-Slavic	ru-en	0.7739413381	0.635915577	0.721279681	0.613496184	0.739938498	0.625461757	0.65250361	0.730831742	-0.331952363	-0.386949569	-0.406317681	-0.360663801	-0.298943281
	en-ru	0.7627384067	0.6488519907	0.7093428969	0.5436459184	0.6855369806	0.5464570522	0.5813109875	0.6704649329	-0.094753876	0.2074387372	-0.042048465	-0.080518812	0.2357965857
Uralic	fi-en	0.7544465661	0.500698686	0.625154316	0.520718753	0.696812868	0.575496674	0.465967327	0.70190233	-0.101985648	-0.180719465	-0.277551442	-0.185979545	-0.139334023
	en-fi	0.4864497781	0.2902386487	0.4125672281	0.1883803755	0.4592720866	0.2068098485	0.1175037399	0.4524150193	-0.05677525	-0.149135888	-0.365386277	-0.32074222	-0.12820746
Turkic	kk-en	0.5784662366	0.313673943	0.511445642	0.33559912	0.592416406	0.327535659	0.1943198	0.592696428	-0.488634586	-0.455021292	-0.593518615	-0.527513802	-0.623326957
	en-kk	0.534217	-0.544386268	0.5562880039	-0.6284886	0.6213475466	-0.585079074	-0.240292013	0.6192656159	-0.476846784	-0.569273949	-0.651435673	-0.35274002	0.1656920463
	tr-en	0.6546989679	0.53469	0.65441	0.48243	0.70131	0.61231	0.51117	0.7004	-0.26238	-0.24391	-0.42729	-0.35915	-0.16599
	en-tr	0.4421551526	0.22748	0.2783	0.20519	0.29814	0.26042	0.2175	0.29793	-0.1707	-0.11438	-0.27039	-0.18837	-0.02141
Sino-Tibetan	zh-en	0.6285635233	0.59158051	0.6743415	0.592279255	0.692212701	0.616823971	0.608996153	0.684733748	-0.204008907	-0.155541971	-0.245966077	-0.288602442	-0.226838633
	en-zh	0.5125129223	0.5569322705	0.5861903429	0.01613976434	0.5782766938	0.2828737199	0.2877052426	0.5831241012	-0.094753876	0.1375120133	-0.110035703	-0.227370605	0.1374309808
Finno-Ugric	et-en	0.41805	0.5275	0.72527	0.4954	0.71895	0.60171	0.40266	0.72512	−0.27164	−0.25252	−0.44237	−0.37183	−0.17185
	en-et	0.24248	0.19468	0.47868	0.17941	0.47311	0.21662	0.10107	0.46757	−0.17530	−0.16249	−0.28535	−0.23969	−0.11076

Figure 14: Translation Performance Results Evaluated using BERTScore

Translation Performance (WER)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	tensor(0.5636)	0.591499984	0.518700778	0.568405509	0.479822844	1.099409461	0.720472455	0.476377964	3.698818922	5.002055645	7.672244072	7.784448624	6.250984192
	en-de	tensor(0.6630)	tensor(0.7273)	tensor(0.6024)	tensor(0.6623)	tensor(0.8574)	tensor(2.4723)	tensor(1.2567)	tensor(0.8956)	3.64346838	7.987531185	9.907587051	10.06217957	7.714774609
	en-cs	tensor(2.9707)	tensor(0.8518)	tensor(0.7609)	tensor(0.9599)	tensor(0.7802)	tensor(2.6013)	tensor(2.3493)	tensor(0.7767)	4.735862732	7.416978836	13.2806015	11.86327839	9.482462883
	cs-en	1.310673833	0.715890825	0.6708408	0.825283229	0.55277282	0.797853291	1.226595163	0.548002362	3.899224758	6.839957237	8.647584915	7.970184803	6.875372887
Indo-European-Romance	fr-de	tensor(0.6995)	0.770758152	0.779783368	0.816787004	0.722021639	4.536552429	1.455324888	0.736010849	2.964350224	7.639730453	7.21886301	7.339801311	5.926895142
	de-fr	tensor(0.6339)	tensor(0.7794)	tensor(0.7165)	tensor(0.8834)	tensor(0.9253)	tensor(1.9458)	tensor(1.5754)	tensor(0.8650)	3.604193926	4.901110172	6.779816628	6.840105057	5.923547268
Indo-Iranian-Indic (Indo-Aryan)	gu-en	tensor(0.6954)	1.037214875	0.852340937	1.172869205	0.613445401	4.79351759	1.246698737	0.626050413	2.828228951	1.99492383	2.478695154	3.599866867	3.087217093
	en-gu	tensor(0.8236)	tensor(1.3884)	tensor(3.0868)	tensor(1.4318)	tensor(0.7521)	tensor(2.2514)	tensor(3.1267)	tensor(0.7906)	4.292010784	3.897327662	12.91528893	10.06542683	5.303030491
	bn-en	0.851872861	1.290011406	0.648694694	0.853575468	0.554483533	0.944381356	1.007945538	0.560158908	4.156234897	2.873468777	7.873264982	3.645876876	9.387687687
	en-bn	0.88193959	2.081193686	1.733518362	0.8584316373	0.7147814035	1.756419182	3.140180349	0.700902164	3.872687878	2.738787366	5.678687768	2.567657867	7.678236767
Indo-European-Baltic	lt-en	tensor(0.7335)	0.886295915	0.798907638	0.98113209	0.706057608	1.096822262	1.441410184	0.717477679	3.527308941	4.720250607	6.708043575	6.579940319	5.761171818
	en-lt	tensor(0.8628)	tensor(0.9513)	tensor(0.9299)	tensor(1.0526)	tensor(1.1052)	tensor(4.0468)	tensor(2.5182)	tensor(1.1353)	3.489962101	7.807795525	11.78701305	11.65259743	8.070778847
Indo-European-Slavic	ru-en	tensor(0.4734)	0.666524708	0.619250417	0.686541736	0.57410562	0.852214634	0.795144796	0.582197607	3.276831388	4.812736988	6.757666111	6.354770184	4.958262444
	en-ru	tensor(0.5803)	tensor(0.7987)	tensor(0.7032)	tensor(0.7911)	tensor(1.1997)	tensor(2.3364)	tensor(1.9099)	tensor(1.3738)	46.16783142	6.002754688	9.132392883	9.427021027	7.555073261
Uralic	fi-en	tensor(0.5015)	0.822268665	0.751698911	0.855201244	0.639832735	1.076842666	1.614218473	0.63303709	3.298484087	5.236699104	7.555149078	7.166753769	5.99215889
	en-fi	tensor(0.6592)	tensor(0.8778)	tensor(2.4708)	tensor(1.0584)	tensor(0.8273)	tensor(2.2674)	tensor(4.0505)	tensor(0.9464)	4.240536213	9.390365601	13.59069443	13.62066269	9.99053669
Turkic	kk-en	tensor(0.8006)	1.16797781	1.007762551	1.166666627	0.812328756	3.843379021	1.5803653	0.800913215	2.791859388	4.699156284	4.175300598	4.813136101	3.49768734
	en-kk	tensor(0.9386)	tensor(1.1310)	tensor(0.9671)	tensor(2.5709)	tensor(0.9025)	tensor(3.5505)	tensor(3.4779)	tensor(0.8991)	5.941558361	5.203016281	10.95578194	10.32199574	6.067460537
	tr-en	tensor(0.7451)	0.899543405	0.748040318	1.011758089	0.675251961	0.828667402	1.292273283	0.678611398	6.687324678	5.678324768	8.78423478	8.789437832	3.444768433
	en-tr	tensor(0.7505)	0.765	0.561	0.899	0.523	0.731	1.102	0.549	5.902	4.832	7.489	7.232	2.754
Sino-Tibetan	zh-en	tensor(0.7595)	0.732108295	0.627564132	0.724038482	0.620192289	0.826923072	0.819551289	0.633333325	1.415064096	1.944444418	4.483333111	3.564423084	3.396153927
	en-zh	tensor(1.0760)	tensor(1.3706)	tensor(1.4266)	tensor(10.3776)	tensor(6.2587)	tensor(30.9650)	tensor(33.5035)	tensor(5.9301)	46.16783142	57.2531662	99.06293488	107.2027969	54.25174713
Finno-Ugric	et-en	0.946168482	1.129829049	0.553514898	0.91386956	0.575680792	1.024065852	2.015199423	0.56048131	54.5555	44.5555555	67.45345345	88.34534554	53.65456456
	en-et	1.662	1.782	0.964	1.592	0.973	1.743	3.289	0.971	87.288	76.199	113.474	147.208	93.432

Figure 15: Translation Performance Results Evaluated using WER

Translation Performance (CER)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	tensor(0.4107)	0.439170629	0.38331151	0.419341624	0.344742864	0.967163444	0.575745165	0.343842119	3.559040308	5.209791183	7.741893291	7.931051254	6.289469242
	en-de	tensor(0.4616)	tensor(0.5197)	tensor(0.4256)	tensor(0.4673)	tensor(0.6651)	tensor(2.1971)	tensor(0.9884)	tensor(0.7108)	2.536647797	6.860759258	7.948365211	8.222444534	6.534090042
	en-cs	tensor(2.5424)	tensor(0.6156)	tensor(0.5416)	tensor(0.6746)	tensor(0.5648)	tensor(2.3923)	tensor(2.0007)	tensor(0.5599)	3.718556881	6.700833321	11.64797115	10.58817577	8.418118477
	cs-en	1.013212204	0.541329682	0.502683759	0.631089985	0.406069368	0.639657319	1.060796857	0.401837319	3.449525118	6.944374084	8.607452393	8.002167702	6.828963757
Indo-European-Romance	fr-de	tensor(0.5079)	0.557652414	0.575745761	0.58047992	0.526588857	4.114850998	1.170881987	0.536640704	2.1647861	6.412849903	5.892801762	6.087678432	5.07256794
	de-fr	tensor(0.4472)	tensor(0.5559)	tensor(0.5069)	tensor(0.6253)	tensor(0.7050)	tensor(1.6975)	tensor(1.3248)	tensor(0.6439)	3.174767733	4.826843262	6.487334251	6.643723965	5.771162987
Indo-Iranian-Indic (Indo-Aryan)	gu-en	tensor(0.5110)	0.788671255	0.657108247	0.873586416	0.425888538	4.830472469	1.021001577	0.431946695	2.472316027	1.976485729	2.183884859	3.329844713	2.841098309
	en-gu	tensor(0.5734)	tensor(1.2186)	tensor(1.8130)	tensor(1.1348)	tensor(0.4872)	tensor(2.0621)	tensor(3.0552)	tensor(0.5212)	3.865274191	3.695536375	12.39887238	9.531609535	4.778182983
	bn-en	0.620970607	0.966063738	0.464272916	0.612822354	0.390759319	0.727972806	0.727972806	0.394340962	2.812	2.476	2.327	2.315	2.629
	en-bn	0.6188148856	1.647922754	1.227341294	0.6126144528	0.4834523797	1.491399288	2.69721365	0.4676591754	4.122	3.788	4.116	3.509	3.979
Indo-European-Baltic	lt-en	tensor(0.5530)	0.675645649	0.620403647	0.777236164	0.530100882	0.914742589	1.237643242	0.538053691	3.3071661	4.974621296	6.665041924	6.755600929	5.901659012
	en-lt	tensor(0.5903)	tensor(0.6692)	tensor(0.6290)	tensor(0.7017)	tensor(0.7833)	tensor(3.2932)	tensor(1.8913)	tensor(0.8062)	2.711935759	6.123874664	8.699213982	8.742062032	6.359476089
Indo-European-Slavic	ru-en	tensor(0.3358)	0.500715315	0.458546221	0.516724586	0.420055866	0.687512755	0.617412627	0.426050812	3.042577744	5.017196178	6.518427849	6.297227383	4.843449593
	en-ru	tensor(0.4041)	tensor(0.5851)	tensor(0.5092)	tensor(0.5679)	tensor(0.9238)	tensor(1.8441)	tensor(1.5075)	tensor(1.0874)	7.953680038	4.855307102	7.428799152	7.782375336	6.325458527
Uralic	fi-en	tensor(0.3607)	0.627457738	0.587443948	0.662814736	0.46171093	0.899706781	1.428682327	0.453345984	3.301914454	5.717580795	7.612883568	7.467144012	6.155570984
	en-fi	tensor(0.4514)	tensor(0.5998)	tensor(2.3277)	tensor(0.6880)	tensor(0.5556)	tensor(1.7895)	tensor(2.7766)	tensor(0.6357)	2.490413427	6.588078976	8.785488129	8.925505638	6.934124947
Turkic	kk-en	tensor(0.5970)	0.92308265	0.776959896	0.871475875	0.590651453	3.042694092	1.288774371	0.585130632	2.549440622	4.606498718	4.230723381	4.728486061	3.724757671
	en-kk	tensor(0.6974)	tensor(0.9575)	tensor(0.7251)	tensor(2.3569)	tensor(0.6690)	tensor(2.8828)	tensor(2.9120)	tensor(0.6656)	3.941309929	4.019098759	8.598507881	8.179683685	4.851835728
	tr-en	tensor(0.5520)	0.702071548	0.571401119	0.806203783	0.503217161	0.640737534	1.084413767	0.502640903	4.122	3.788	4.116	3.509	3.979
	en-tr	tensor(0.5442)	1.89	1.73	2.15	1.62	1.92	2.62	1.67	5.87	5.44	5.42	4.73	5.64
Sino-Tibetan	zh-en	tensor(0.5764)	0.553567708	0.452572972	0.536188126	0.445860863	0.64254123	0.615380585	0.454914421	1.339559793	2.113673925	4.586034775	3.9417243	3.634372234
	en-zh	tensor(0.6760)	tensor(0.6230)	tensor(0.6814)	tensor(1.8978)	tensor(1.4180)	tensor(5.6767)	tensor(9.2129)	tensor(1.3076)	7.953680038	10.71049881	17.5510807	18.48352242	11.49194431
Finno-Ugric	et-en	0.792560279	0.987519681	0.426875085	0.788804054	0.458015263	0.961468577	2.062037945	0.44141525	4.122	3.788	4.116	3.509	3.979
	en-et	1.99	2.07	1.63	2.01	1.54	2.09	3.24	1.59	5.77	5.32	5.47	4.72	5.46

Figure 16: Translation Performance Results Evaluated using CER

Translation Performance (ROUGE-1)															
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	0.7035969089	0.655840606	0.727078732	0.696482773	0.750525207	0.631326845	0.67212221	0.748216407	0.2806433265	0.2630165666	0.1366672331	0.1602235029	0.2078100316
	en-de	0.6080497797	0.5505471997	0.6260992332	0.5956874158	0.6408922168	0.3689825617	0.5394313311	0.6356311143	0.1476249195	0.1608549681	0.0776326854	0.08393601164	0.1472389446
	en-cs	0.3217954356	0.4208939741	0.5493429103	0.3477877241	0.5834058083	0.3099625134	0.3468191638	0.5804175258	0.07275894947	0.1095152462	0.03405630984	0.0628572452	0.109634746
	cs-en	0.469676997	0.558965824	0.674238078	0.584661388	0.706300017	0.626172264	0.543656068	0.706924599	0.1699451725	0.1474511073	0.09556312656	0.1224936563	0.1750733871
Indo-European-Romance	fr-de	0.5384090912	0.468335434	0.499345849	0.403414716	0.525616083	0.313369556	0.408964686	0.534696482	0.07784267815	0.1100179854	0.06640518397	0.08216783373	0.1506068316
	de-fr	0.6349817722	0.5243953063	0.6022937593	0.4147070811	0.5938629794	0.3899606118	0.4900374834	0.5992443519	0.1025939297	0.206051454	0.1088715329	0.1091992023	0.1652551917
Indo-Iranian-Indic (Indo-Aryan)	gu-en	0.6823181803	0.469727748	0.657401083	0.489624127	0.722256204	0.528922412	0.407134248	0.724468615	0.1460819786	0.1365856177	0.1113232975	0.1390386153	0.1526918909
	en-gu	0.08074074074	0.143043956	0.1033333333	0.02728227618	0.08666666667	0.01152183601	0.01722413721	0.06	0.01376492938	0.01892688163	0.005014233118	0.007045995515	0.01928722183
	bn-en	0.568532875	0.544307484	0.710897932	0.628826924	0.760229034	0.648933434	0.617346856	0.757274522	0.03674983727	0.00310604513	0.04943300954	0.3059017152	0.1790355571
	en-bn	0.4576129143	0.3802389211	0.5467982011	0.5023144579	0.5920374947	0.4950322366	0.4972042518	0.5099073842	0.02399127372	0.00123267222	0.02811534583	0.2689742163	0.1511389027
Indo-European-Baltic	lt-en	0.5602070579	0.471669329	0.572265647	0.486329887	0.611242797	0.51793487	0.417766017	0.611377817	0.1290403392	0.1519335593	0.08989112752	0.1224277908	0.1268247116
	en-lt	0.4196354554	0.2930489759	0.4076381152	0.2716822056	0.3957404087	0.2508870059	0.1981896842	0.3957723277	0.007900821685	0.04323752022	0.02954573626	0.03209938542	0.06578149916
Indo-European-Slavic	ru-en	0.7536009961	0.60259341	0.673820164	0.618928828	0.693671598	0.621046059	0.626652695	0.694029714	0.3330782707	0.2875578859	0.2022459033	0.2191503553	0.2992473454
	en-ru	0.3058253968	0.09883333333	0.1268333333	0.09138359788	0.1097771595	0.06221609119	0.102618465	0.1072705604	0.02303329729	0.0070290981	0.003594571755	0.003667549894	0.005969931798
Uralic	fi-en	0.706546415	0.480527999	0.590775949	0.54006982	0.640518498	0.548283391	0.45386005	0.642944823	0.2405260919	0.1644473276	0.08847717721	0.129997028	0.1711175072
	en-fi	0.5407854673	0.3437936829	0.4666457909	0.3188079717	0.502004404	0.2996539337	0.1885651287	0.4950178235	0.3984322358	0.2614672056	0.1328946145	0.1714910028	0.2052194401
Turkic	kk-en	0.5188394738	0.360247651	0.480606457	0.39464613	0.526302381	0.36520715	0.255703233	0.530401845	0.1233859667	0.1312226146	0.1105829001	0.1392371384	0.1189226287
	en-kk	0.2316280347	0.01771984845	0.1221904762	0.0158545622	0.1346666667	0.004825581395	0.01417767007	0.1346666667	0.04327433536	0.006001622691	0.003284644468	0.00314980722	0.007998724224
	tr-en	0.603265185	0.492384561	0.599977506	0.504128471	0.641483415	0.580147072	0.482222599	0.637880144	0.1325927694	0.07210631557	0.02410322793	0.03279573059	0.06568749069
	en-tr	0.5205915012	0.392306457	0.508976877	0.427238019	0.520621937	0.447698301	0.391354653	0.522911797	0.0805223475	0.04084796049	0.01411970123	0.02039186491	0.03349451861
Sino-Tibetan	zh-en	0.5823147986	0.580179062	0.673700214	0.623063702	0.67883522	0.637700435	0.608097807	0.678678021	0.3896344719	0.3995238712	0.2192332982	0.2760562271	0.3083767857
	en-zh	0.1359259259	0.2752727273	0.2538917749	0.151304	0.2904953704	0.09364659879	0.08284691908	0.2992874606	0.02303329729	0.01938938387	0.01381468817	0.01419386384	0.02509690769
Finno-Ugric	et-en	0.52098932	0.405857263	0.662229978	0.522320999	0.66048091	0.574001702	0.4203863	0.666693209	0.6783458674	0.8672346783	0.2347654327	0.7864345345	0.7623467324
	en-et	0.398775046	0.302072417	0.556342134	0.439526768	0.576395381	0.495509596	0.362155217	0.563136317	0.5746985139	0.7524102464	0.1246761202	0.6909123363	0.6188752322

Figure 17: Translation Performance Results Evaluated using ROUGE-1

Translation Performance (ROUGE-2)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	0.4806072152	0.411330905	0.500466914	0.466214716	0.528569966	0.438539807	0.460318803	0.525583662	0.1455079079	0.1562489246	0.05713607769	0.07850319961	0.1197728686
	en-de	0.3740678938	0.3320507058	0.4109210403	0.3647678236	0.4523359284	0.2322739159	0.3562216383	0.4463913088	0.06015737417	0.09179006292	0.02684589758	0.03437593559	0.07262973837
	en-cs	0.1785906756	0.1627826367	0.3030445219	0.1724785624	0.3435343954	0.147946864	0.1733760152	0.3356590188	0.01080485405	0.030991769	0.003752613856	0.01139956039	0.03486130127
	cs-en	0.300730447	0.28590009	0.417378863	0.343479706	0.47312501	0.386871969	0.320579898	0.471267693	0.05893989241	0.06429189561	0.0215023205	0.04344964264	0.08316744126
Indo-European-Romance	fr-de	0.3219111663	0.260464502	0.29395645	0.220634391	0.322336332	0.178095876	0.233273294	0.33113832	0.03690788815	0.05527043344	0.02944240966	0.03734515455	0.07378484453
	de-fr	0.4418349103	0.3168349205	0.4215704088	0.2539494784	0.4171850531	0.2483383232	0.3213940915	0.4189493006	0.04071748543	0.1225340482	0.04388870127	0.04994226483	0.09232206994
Indo-Iranian-Indic (Indo-Aryan)	gu-en	0.411586022	0.273922499	0.392837404	0.253821364	0.45896159	0.311221556	0.15750341	0.462178585	0.01947807304	0.001836547291	0.003647229963	0.01727566785	0.03231789432
	en-gu	0.0219047619	0.04381818182	0.05	0.01413131313	0.03	0.000476190476	0.003269460492	0.02	0.002432842294	0.003273336685	0.00107073033	0.001506256359	0.005007713153
	bn-en	0.350352207	0.303728096	0.463803802	0.375096055	0.525984924	0.407999854	0.360284487	0.522411658	0.19862372	0.21039958	0.17328293	0.18453947	0.19078121
	en-bn	0.04512984	0.03389812	0.10387732	0.05978364	0.12098245	0.08714329	0.06391201	0.12912873	0.02981102	0.0187435	0.01572346	0.01766598	0.02133418
Indo-European-Baltic	lt-en	0.2821688723	0.249592221	0.299752085	0.225378993	0.344790138	0.250117014	0.18730201	0.338221216	0.02880231471	0.03728523626	0.01432936342	0.02940673542	0.03699551393
	en-lt	0.1885120914	0.1224051137	0.1684513863	0.0871938467	0.1612253228	0.08420536005	0.05382582167	0.1590112152	0.001325980392	0.01030261302	0.004984990197	0.005596156151	0.01306651262
Indo-European-Slavic	ru-en	0.5376431914	0.331148059	0.42214182	0.348589474	0.433151364	0.381932521	0.38912852	0.436561039	0.1612079702	0.138600841	0.06925722998	0.09286451405	0.1558283106
	en-ru	0.1641111111	0.04095238095	0.04095238095	0.02828571429	0.03244642857	0.04454545455	0.0387001287	0.03118206522	0.008473054559	0.0009806576402	0.000264301860	0.000678482021	0.000998805256
Uralic	fi-en	0.4686722968	0.215327045	0.322666083	0.267984372	0.380537974	0.291711952	0.224475597	0.392301169	0.09427369966	0.05447785584	0.01113781792	0.03418532466	0.06707642286
	en-fi	0.3273956616	0.1366510221	0.2486441551	0.1434306847	0.277841682	0.1418775076	0.07054119494	0.2767699442	0.04626713484	0.01872275987	0.003201700036	0.006375445261	0.01879421355
Turkic	kk-en	0.2580873927	0.165008416	0.226683416	0.143958703	0.264062038	0.155283764	0.067311939	0.266398182	0.01290407684	0.01404089678	0.00714180572	0.01578200354	0.01247678229
	en-kk	0.1276651105	0.01828828829	0.03333333333	0.0008	0.04	0.001951219512	0.002530880713	0.04	0.01104430381	0.000615384615	0.001035011831	0.000574722607	0.002728530913
	tr-en	0.3330994111	0.216901708	0.326774289	0.24089405	0.375761375	0.305293519	0.21333317	0.371083334	0.18422957	0.19218304	0.16745029	0.17382961	0.18079346
	en-tr	0.3171213639	0.04421976	0.09573001	0.06214823	0.11844197	0.08960325	0.05742388	0.12534988	0.03019287	0.01784762	0.01403288	0.01599233	0.02034892
Sino-Tibetan	zh-en	0.3143404758	0.305463312	0.414763967	0.350868577	0.434214164	0.3789961	0.345068916	0.429316628	0.1724627429	0.2106996215	0.07621318465	0.1203750661	0.1549737611
	en-zh	0.025	0.1423015873	0.1393492063	0.07045851371	0.1814285714	0.04514063647	0.04010631312	0.1780952381	0.008473054559	0.007866798006	0.003908068697	0.005340760318	0.01150741733
Finno-Ugric	et-en	0.350917925	0.191222147	0.428780955	0.28949996	0.434405339	0.360471396	0.227009036	0.437899591	0.17623892	0.18376175	0.16055238	0.16800347	0.17299481
	en-et	0.04111823	0.05943891	0.11126579	0.07019832	0.12739564	0.09687125	0.06972349	0.13288756	0.02611398	0.01955421	0.01710232	0.01892688	0.0217843

Figure 18: Translation Performance Results Evaluated using ROUGE-2

Translation Performance (ROUGE-L)														
Language Family	**Direction**	**gemma2-9b**	**gemma-7b**	**llama3-70b**	**llama3-8b**	**llama-3.1-70b**	**llama-3.1-8b**	**mixtral-8x7b**	**llama-3.2-90b-vision**	**OLMo-1B**	**Phi-3.5-mini**	**Qwen-2.5-0.5B**	**Qwen-2.5-1.5B**	**Qwen-2.5-3B**
Indo-European-Germanic	de-en	0.6511811076	0.599041414	0.670981086	0.642438393	0.693083271	0.585146854	0.624869784	0.692364765	0.2410556463	0.2313887267	0.1089061102	0.1281050839	0.1820320865
	en-de	0.5443090886	0.5030297426	0.5813422556	0.5492698166	0.602148037	0.3366670503	0.5040089105	0.5988705646	0.130818225	0.1480649009	0.06421446228	0.07008724609	0.1283564921
	en-cs	0.2879248322	0.3681367014	0.4957193091	0.311808441	0.5232664035	0.2650818335	0.3071073395	0.5249787407	0.06223380634	0.09077972694	0.02572376632	0.04537861271	0.08747047984
	cs-en	0.429992776	0.512794493	0.618468774	0.529691526	0.657039453	0.571904712	0.500419296	0.657766534	0.1390407876	0.125495058	0.07145202008	0.09691265672	0.1490218325
Indo-European-Romance	fr-de	0.4964005636	0.432708736	0.461355748	0.373826126	0.487885763	0.276157961	0.375518672	0.498945047	0.07369907718	0.100294293	0.05685491774	0.06942698084	0.1311577406
	de-fr	0.5923186754	0.4818413532	0.5614317394	0.3835938468	0.5571573813	0.3446639599	0.4479614985	0.5624223121	0.08944029672	0.1854031678	0.0827578978	0.08814217374	0.1437618953
Indo-Iranian-Indic (Indo-Aryan)	gu-en	0.593719108	0.394003651	0.571113547	0.403163836	0.646020463	0.451098767	0.338628903	0.646923231	0.1233178663	0.1161357643	0.0946921028	0.1058209224	0.1261866435
	en-gu	0.143043956	0.08074074074	0.1033333333	0.02728227618	0.08666666667	0.01152183601	0.01674794673	0.06	0.01323861359	0.01778402448	0.005014233118	0.006873581722	0.01870751168
	bn-en	0.481108937	0.478545248	0.615704514	0.524069695	0.672474998	0.559214965	0.5175473	0.668633774	0.31259876	0.27877431	0.23319488	0.25147266	0.26578901
	en-bn	0.05678133	0.04389212	0.11723984	0.06431021	0.13984766	0.10277389	0.07766412	0.14628111	0.03281957	0.02173465	0.01865431	0.02047832	0.02756194
Indo-European-Baltic	lt-en	0.4840298514	0.402728011	0.488910524	0.404182735	0.526827629	0.437559459	0.347941979	0.524764095	0.09973219499	0.1162734086	0.06602242603	0.08652117703	0.09832806176
	en-lt	0.378620701	0.2547811415	0.3531055469	0.2407800299	0.3471391526	0.2085499515	0.172750443	0.3485690507	0.007900821685	0.04041384997	0.02547833627	0.02697659062	0.05441067554
Indo-European-Slavic	ru-en	0.7018598505	0.54287646	0.621811861	0.553843094	0.639193553	0.565118915	0.572192514	0.641135973	0.2864545241	0.2420648762	0.1501526142	0.1697918212	0.2536420321
	en-ru	0.3036031746	0.09883333333	0.1228333333	0.08738359788	0.1057771595	0.06221609119	0.102618465	0.1032705604	0.02303329729	0.0070290981	0.003501113811	0.003667549894	0.005969931798
Uralic	fi-en	0.665902244	0.421084138	0.522979224	0.471288602	0.573009049	0.486910695	0.40074351	0.583811507	0.2000916728	0.1398196735	0.06578743851	0.09665600616	0.1380230943
	en-fi	0.5167186458	0.3117423708	0.419481312	0.2893059382	0.4556254721	0.2691223133	0.1657203323	0.4491487186	0.120901521	0.06537586296	0.020232076	0.02881352743	0.05625624266
Turkic	kk-en	0.4526386458	0.300971549	0.405664198	0.323338016	0.457363289	0.30003984	0.209717538	0.460291607	0.0997997409	0.1035449545	0.0860709712	0.09714725887	0.09629852978
	en-kk	0.2270566062	0.06323401185	0.1207619048	0.0148193062	0.1346666667	0.004825581395	0.01343692933	0.1346666667	0.04128281972	0.006001622691	0.003141787325	0.00264980722	0.007998724224
	tr-en	0.5220861131	0.414479995	0.511962811	0.411907209	0.555552867	0.484345611	0.390976409	0.55553684	0.24311769	0.19864174	0.16795382	0.20184637	0.22941708
	en-tr	0.4596359994	0.08733582	0.15487129	0.07365291	0.19354891	0.12077438	0.06912947	0.18893579	0.03601965	0.01540483	0.01038766	0.01214684	0.02583592
Sino-Tibetan	zh-en	0.4990133185	0.503238216	0.601904797	0.544764824	0.614502798	0.560766951	0.532928437	0.609391955	0.3206962078	0.3368440228	0.1555041156	0.2127494695	0.2498395529
	en-zh	0.1359259259	0.2705505051	0.2469473304	0.1509591725	0.2879953704	0.09364659879	0.08062469686	0.2967874606	0.02303329729	0.01938938387	0.01372034855	0.01409862575	0.02493297327
Finno-Ugric	et-en	0.49106546	0.368764687	0.629387754	0.481428199	0.629612681	0.544445366	0.391353807	0.636198072	0.28945673	0.23498122	0.19875439	0.21067284	0.24578913
	en-et	0.05832764	0.08911632	0.13876205	0.07248599	0.17954362	0.13029875	0.06673418	0.18329846	0.04122961	0.02186429	0.01243861	0.01852749	0.03170422

Figure 19: Translation Performance Results Evaluated using ROUGE-L

ru-kk									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	3.777201727	24.11384149	94.3289225	0.4736602902	0.9458	0.7538	0.01	0.01	0.01
deepseek-r1-distill-70b	8.007304049	36.98310701	86.32640202	0.5920827985	0.879	0.6524	0.01	0.01	0.01
llama-3.3-70b-specdec	9.16082701	39.42514519	84.24700693	0.6174486279	0.8582	0.6316	0.01	0.01	0.01
qwen-2.5-32b	2.289514201	27.21117384	144.2974165	0.4946300983	1.4531	1.2376	0.000444444	0.000444444	0.000444444
llama-3.3-70b-versatile	9.407446238	39.5071273	83.67989918	0.6190448999	0.8532	0.6202	0.01	0.01	0.01
llama-3.1-8b	1.333673463	18.01646875	414.7448015	0.5300563574	4.1563	4.1124	0.01	0.01	0.01
mixtral-8x7b	0.5682181531	19.97733367	204.0957782	0.3088902533	2.0441	1.7281	0.001111111111	0.001111111111	0.001111111111
llama-3.2-90b-vision	10.29078558	40.59281585	83.61688721	0.625035584	0.8526	0.6247	0.01	0.01	0.01
kk-ru									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	5.870029065	30.07303877	97.75280899	0.5189307928	0.9911	0.7469	0.02666666667	0.02	0.02666666667
deepseek-r1-distill-70b	11.60854209	40.34453521	84.92016558	0.6146060824	0.8693	0.6292	0.03	0.02	0.03
llama-3.3-70b-specdec	13.80193755	41.6087056	81.9633353	0.629535377	0.8486	0.6104	0.03	0.02	0.03
qwen-2.5-32b	7.604863181	32.99519931	103.5481963	0.5589443445	1.058	0.8355	0.03	0.02	0.03
llama-3.3-70b-versatile	13.81075993	41.09437069	82.55470136	0.624904573	0.8522	0.6151	0.03	0.02	0.03
llama-3.1-8b	8.318896325	34.38392567	91.89828504	0.5712465048	0.9426	0.7002	0.03	0.02	0.03
mixtral-8x7b	2.36759155	23.86356481	159.7871082	0.419595629	1.6109	1.3026	0.03	0.02	0.03
llama-3.2-90b-vision	15.35557358	42.9029031	80.36664695	0.6326509714	0.8303	0.6083	0.03	0.02	0.03

Figure 20: Translation Performance in the **Literature** Domain Across the **ru** ↔ **kk** (Russian–Kazakh) Translation Pair

en-kk									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	1.321445981	22.38618235	96.53465347	0.4587476254	0.9736	0.7551	0.02071428571	0.01	0.02071428571
deepseek-r1-distill-70b	3.245164017	34.26711923	94.05940594	0.5571330786	0.9554	0.6942	0.023	0.01	0.023
llama-3.3-70b-specdec	5.850199802	35.45302879	90.3190319	0.5752460361	0.9136	0.6614	0.018	0.01	0.018
qwen-2.5-32b	1.029401102	24.07556052	172.3872387	0.4214095175	1.7327	1.5004	0.01007936508	0.002857142857	0.01007936508
llama-3.3-70b-versatile	5.820119035	35.70158241	90.0990099	0.5736073256	0.9136	0.6577	0.018	0.01	0.018
llama-3.1-8b	0.2407463145	12.23489775	657.0407041	0.4801938236	6.577	7.2304	0.018	0.01	0.018
mixtral-8x7b	0.7142926015	17.02345942	173.5423542	0.2427054197	1.7371	1.4501	0.01738095238	0	0.01738095238
llama-3.2-90b-vision	5.054276585	36.58194345	89.5489549	0.5806723237	0.9103	0.6651	0.018	0.01	0.018
kk-en									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	9.501813505	33.80346775	88.60182371	0.3109199107	0.9214	0.6845	0.3778464179	0.1286298903	0.3052229277
deepseek-r1-distill-70b	22.80128708	47.91235262	71.69452888	0.5284175277	0.7515	0.5559	0.5487676993	0.276042758	0.478312403
llama-3.3-70b-specdec	24.93835652	49.81989532	68.46504559	0.5503125191	0.7196	0.5174	0.5726033911	0.3062969756	0.5081137421
qwen-2.5-32b	13.63981109	37.37147244	84.87841945	0.3782681525	0.8796	0.6485	0.4130258958	0.1611079129	0.3434682516
llama-3.3-70b-versatile	24.32814561	49.99724064	68.69300912	0.5570639372	0.7211	0.5237	0.5741070302	0.3096728316	0.5072140141
llama-3.1-8b	18.145377	41.75641552	77.3556231	0.4553956091	0.8017	0.5916	0.480382085	0.2208263421	0.4109977522
mixtral-8x7b	7.090474018	29.3587673	91.98328267	0.2701686621	0.9407	0.7134	0.3039692049	0.07954121958	0.2435184243
llama-3.2-90b-vision	25.54644401	49.83236819	67.9331307	0.5492619872	0.7158	0.523	0.5730942853	0.3162041171	0.5039572667

Figure 21: Translation Performance in the **Literature** Domain Across the **en** ↔ **kk** (English–Kazakh) Translation Pair

ru-en									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	28.95307875	56.04786817	56.80200921	0.6615436077	0.5948	0.4313	0.6417203897	0.3761076811	0.5997408274
deepseek-r1-distill-70b	29.17080454	55.97570363	57.13687735	0.6632838249	0.5956	0.4198	0.6429461987	0.3713073575	0.60181317
llama-3.3-70b-specdec	31.63245861	58.05295398	54.58350774	0.6797747612	0.5735	0.3982	0.6616445614	0.4137463373	0.6245556537
qwen-2.5-32b	30.55270172	56.94442959	57.55546254	0.6677671671	0.6015	0.4357	0.6513139921	0.3935508817	0.6087703951
llama-3.3-70b-versatile	31.48058393	57.88041177	55.58811218	0.6796998382	0.5835	0.4022	0.6605041264	0.4151996793	0.6233188197
llama-3.1-8b	26.78612346	53.00265958	59.60652993	0.6363235712	0.6266	0.4477	0.6211459063	0.3571103331	0.5752792896
mixtral-8x7b	27.03485026	54.17489061	58.26705735	0.6360778213	0.6132	0.4427	0.6314745417	0.3627739856	0.5859674037
llama-3.2-90b-vision	30.8513912	57.63238356	55.169527	0.686940372	0.5789	0.3996	0.6620877481	0.4092993855	0.6248863802
en-ru									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	14.93404985	42.54532859	74.50221239	0.6464685202	0.7727	0.5618	0.6184824064	0.3533128891	0.5552905562
deepseek-r1-distill-70b	14.88806894	41.6752908	74.28097345	0.6508978009	0.7677	0.5522	0.5969840231	0.3473411106	0.5644448073
llama-3.3-70b-specdec	18.27496777	45.15718268	71.57079646	0.6797386408	0.74	0.5395	0.6149688981	0.3661101108	0.6079751497
qwen-2.5-32b	14.21449295	42.08850206	92.25663717	0.6469954848	0.9419	0.7715	0.6150710385	0.3494070856	0.5669435805
llama-3.3-70b-versatile	18.74005594	45.18931555	71.23893805	0.6800857782	0.7384	0.5395	0.6379357022	0.3746858178	0.5826523567
llama-3.1-8b	15.90822177	42.913811	73.50663717	0.6607048512	0.7588	0.554	0.5785026842	0.329307508	0.5299211999
mixtral-8x7b	8.569671667	36.32936483	141.9800885	0.5673665404	1.4381	1.2331	0.6162240948	0.316056629	0.5680538321
llama-3.2-90b-vision	18.30303735	45.23592141	70.29867257	0.6803962588	0.7273	0.5264	0.6274305039	0.3704310197	0.60092223

Figure 22: Translation Performance in the **Literature** Domain Across the **ru** ↔ **en** (Russian–English) Translation Pair

en-it									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	29.82860628	61.65567557	54.44560669	0.5854369998	0.5643305182	0.3483683765	0.5989416561	0.3613613065	0.5734314957
deepseek-r1-distill-70b	31.94970638	63.0429091	51.88284519	0.601298511	0.5303347111	0.3268747926	0.6140319468	0.3878861069	0.5981589914
llama-3.3-70b-specdec	33.1624745	64.00899611	51.9874477	0.6235374212	0.5308577418	0.3173046708	0.6274837014	0.4084491847	0.6143468246
qwen-2.5-32b	28.61892682	61.41446121	58.73430962	0.5685833693	0.6004183888	0.3732350171	0.5810358389	0.354494617	0.5664290344
llama-3.3-70b-versatile	33.51548307	63.95378715	51.77824268	0.6211705804	0.5282427073	0.3178537786	0.6250153582	0.4063410121	0.612854382
llama-3.1-8b	30.52616934	62.00263008	55.17782427	0.594009459	0.564853549	0.3500156999	0.6064720048	0.3866144602	0.5911379179
mixtral-8x7b	26.04023474	60.12345818	64.64435146	0.5583876371	0.6626569033	0.4249294102	0.5748382538	0.3504252528	0.5537360085
llama-3.2-90b-vision	34.12399256	64.11630382	50.57531381	0.6232898831	0.5177824497	0.3118136227	0.6342451592	0.4189352383	0.6215828621
en-pt									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	35.94160574	67.26212785	46.59685864	0.6695585847	0.4769633412	0.2565206587	0.670162141	0.4392009211	0.6467819965
deepseek-r1-distill-70b	36.61552668	67.63014329	45.39267016	0.6737997532	0.4633507729	0.2487372756	0.6795620123	0.4356436278	0.6525196484
llama-3.3-70b-specdec	37.13103815	68.07921438	44.71204188	0.6801544428	0.4539267123	0.2438519448	0.6833880481	0.4461116553	0.6596060795
qwen-2.5-32b	36.85941429	67.69178567	47.17277487	0.6755679846	0.4806282818	0.2754823267	0.683300207	0.4458964657	0.658625543
llama-3.3-70b-versatile	37.23116478	68.09443548	44.60732984	0.6798992753	0.4528795779	0.2431067377	0.6812001568	0.4453460782	0.6586525001
llama-3.1-8b	36.31390952	67.57775513	45.60209424	0.6704047918	0.4654450119	0.2496480942	0.6689804404	0.4339472788	0.6446267417
mixtral-8x7b	2.157653198	23.24678394	1402.722513	0.5869017243	14.03769588	9.963732719	0.6057768901	0.3832083783	0.5820276545
llama-3.2-90b-vision	37.60389641	68.21154597	44.45026178	0.6805917621	0.4534031451	0.2443487644	0.6869526227	0.4500432662	0.6623982691
en-fi									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	23.46606717	54.11797686	64.87740225	0.7043599486	0.6812458634	0.452575475	0.2935238095	0.09333333333	0.2901904762
deepseek-r1-distill-70b	24.78817042	55.79140095	63.94963552	0.719512403	0.6613651514	0.4277087152	0.319	0.11	0.3156666667
llama-3.3-70b-specdec	27.88697276	59.0677633	60.10603048	0.7414252758	0.6269052625	0.4002664387	0.3233333333	0.09	0.32
qwen-2.5-32b	21.81125521	54.23205619	79.19151756	0.6929466724	0.8210735321	0.6000000238	0.3273333333	0.1	0.324
llama-3.3-70b-versatile	27.91426337	58.94297436	60.76872101	0.7403318286	0.632206738	0.4043516815	0.3166666667	0.09	0.3133333333
llama-3.1-8b	22.26246347	54.52701915	68.19085487	0.7023098469	0.7104042172	0.4691829383	0.3117777778	0.1078571429	0.3117777778
mixtral-8x7b	15.90142745	51.52375221	102.1868787	0.6634894609	1.049039125	0.7546181083	0.2444655248	0.09	0.2444655248
llama-3.2-90b-vision	28.62653364	59.49940379	60.37110669	0.7480649352	0.6242544651	0.4001776278	0.314	0.11	0.3106666667

Figure 23: Translation Performance in the **Medical** Domain Across the **en** → **it, pt, fi** (English–Italian, Portuguese, Finnish) Translation Pairs

en-es									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	39.56519039	68.61382539	43.32344214	0.6833444238	0.6833444238	0.2851915061	0.6753541483	0.4586373549	0.6411092379
deepseek-r1-distill-70b	40.22870585	69.49070072	42.77942631	0.6958081126	0.4564787447	0.2767341733	0.6915046372	0.4843197866	0.6597110569
llama-3.3-70b-specdec	46.04828172	72.9611176	37.68545994	0.740263164	0.3991097808	0.2328210175	0.7239825153	0.5286261698	0.6981349929
qwen-2.5-32b	41.26313898	70.0691065	42.13649852	0.705573678	0.4426310658	0.2699032426	0.6981272358	0.4934227138	0.6702149484
llama-3.3-70b-versatile	45.99863096	72.90614292	37.88328388	0.7398954034	0.3986152411	0.2320891321	0.722667398	0.5239545703	0.693932522
llama-3.1-8b	40.61335239	69.64809706	43.76854599	0.6928765774	0.4559841752	0.2764902115	0.6834104186	0.4820827067	0.6541243679
mixtral-8x7b	31.12491104	64.23419312	63.25420376	0.6263270378	0.6483679414	0.4779214561	0.6318474171	0.4302860719	0.6037057487
llama-3.2-90b-vision	46.45339861	73.04045959	38.03165183	0.7303581238	0.4015825987	0.2353419513	0.7221496388	0.5268361216	0.6949303615
en-fr									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	34.39254708	64.30390178	49.61427194	0.5996338129	0.5135005116	0.3040979207	0.6828703705	0.4810030312	0.6576528599
deepseek-r1-distill-70b	33.16744074	63.82449968	51.59112825	0.5933145881	0.5400192738	0.3266402781	0.6738079451	0.4664364362	0.6413810851
llama-3.3-70b-specdec	34.90668819	65.35544048	48.1195757	0.6076910496	0.500482142	0.2981264591	0.6876992916	0.4903036601	0.6606933727
qwen-2.5-32b	31.42344992	63.11624153	58.1485053	0.5732603669	0.5998071432	0.3950138092	0.6557884139	0.4646639424	0.6300563496
llama-3.3-70b-versatile	34.83854345	65.35429451	48.1195757	0.6080809236	0.5009643435	0.2967082262	0.690651823	0.4898819354	0.6624640857
llama-3.1-8b	32.54708005	62.9643297	53.51976856	0.5772969127	0.5544840693	0.3447040319	0.6656719207	0.454206042	0.6345494337
mixtral-8x7b	32.02930228	63.02067801	54.43587271	0.5782331228	0.5636451244	0.355751276	0.6591672179	0.4594677179	0.6259705205
llama-3.2-90b-vision	35.48646708	66.13836046	47.78206365	0.6211726665	0.4985535145	0.2949914038	0.6977817032	0.5058347726	0.6738014121
en-hr									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	14.07644216	42.70098412	72.86938981	0.693331778	0.7660111189	0.5517241359	0.24	0.12	0.24
deepseek-r1-distill-70b	22.18710576	52.01567519	62.73323248	0.7413972616	0.6656581163	0.4520342052	0.1473333333	0.03333333333	0.1473333333
llama-3.3-70b-specdec	27.69769113	55.62498684	55.72365103	0.7673763037	0.5940493941	0.394813478	0.08266666667	0.03666666667	0.08266666667
qwen-2.5-32b	14.03416672	43.32458736	84.87140696	0.6994754076	0.8804841042	0.6978295445	0.2227598039	0.09040816327	0.2227598039
llama-3.3-70b-versatile	27.64527321	55.72528789	55.42107917	0.7677220106	0.5864851475	0.3903034925	0.08266666667	0.03666666667	0.08266666667
llama-3.1-8b	15.36286078	42.0095251	85.93040847	0.67376405	0.8885527253	0.6598703265	0.198	0.11	0.198
mixtral-8x7b	4.049127801	27.95344651	191.0237015	0.492838949	1.933938503	1.79911685	0.06914370575	0.0294218489	0.06914370575
llama-3.2-90b-vision	29.86814441	57.50457294	52.64750378	0.7807344198	0.5622794032	0.3751761615	0.238	0.11	0.238

Figure 24: Translation Performance in the **Medical** Domain Across the **en** → **es, fr, hr** (English–Spanish, French, Croatian) Translation Pairs

en-pl									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	9.854149219	38.54739261	80.0511509	0.6098458767	0.8094629049	0.5277220607	0.2303253968	0.05928571429	0.2293253968
deepseek-r1-distill-70b	10.96271608	40.14506726	79.92327366	0.6373671293	0.8049871922	0.5089238286	0.1948436563	0.05523809524	0.1937325452
llama-3.3-70b-specdec	12.03478949	42.81314852	77.04603581	0.6524026394	0.775575459	0.4765022695	0.1825244559	0.04371428571	0.1825244559
qwen-2.5-32b	2.220460276	22.39325841	432.8644501	0.4909164011	4.338235378	4.612524986	0.1337721295	0.02002685246	0.133295939
llama-3.3-70b-versatile	12.1851395	43.53218651	76.85421995	0.6550704837	0.7736572623	0.4746013284	0.1825244559	0.04371428571	0.1825244559
llama-3.1-8b	2.20199041	21.84802075	413.0434783	0.5940595865	4.1381073	3.963142872	0.1843940212	0.02954761905	0.1843940212
mixtral-8x7b	3.228740857	29.06247668	191.2404092	0.4787277281	1.921355486	1.648642898	0.132038017	0.05131779693	0.1301011891
llama-3.2-90b-vision	12.49484945	43.32211129	76.15089514	0.6618098617	0.7691816092	0.4741789103	0.1911911226	0.05223076923	0.1911911226
en-de									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	22.28377968	57.05296582	62.16975829	0.5011638403	0.6430578828	0.4174337089	0.5634677656	0.3029869125	0.5191300153
deepseek-r1-distill-70b	27.12925496	61.28890818	57.27937043	0.5433635712	0.5964024663	0.3730627	0.6002139197	0.6002139197	0.5554653335
llama-3.3-70b-specdec	27.56043895	61.59800762	55.7054525	0.547459662	0.5750421882	0.363386035	0.6031518049	0.3650222995	0.5653936007
qwen-2.5-32b	23.52145369	58.75236137	64.92411467	0.5074021816	0.673412025	0.4549602568	0.5737382486	0.3263868264	0.527955323
llama-3.3-70b-versatile	27.96745602	61.81359881	55.98650927	0.5507258773	0.5767285228	0.358823061	0.6044498992	0.3682641192	0.5679707007
llama-3.1-8b	23.69350952	59.21735002	60.20236088	0.5147224665	0.6256324053	0.3857288957	0.5735024225	0.325186578	0.5295548215
mixtral-8x7b	1.778203294	23.87222404	1072.568859	0.4858875573	10.753232	8.049091339	0.5501532744	0.304108956	0.5057190396
llama-3.2-90b-vision	28.28922774	62.66383613	55.81787521	0.5610575676	0.5806633234	0.3566989303	0.6123273805	0.3763925578	0.5676745333

Figure 25: Translation Performance in the **Medical** Domain Across the **en** → **pl, de** (English–Polish, German) Translation Pairs

en-pt									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	47.5473246	71.92322826	37.43455497	0.7098966241	0.3869110048	0.2426065207	0.7731809378	0.5953230779	0.7524137163
deepseek-r1-distill-70b	48.38729586	72.16911292	36.43979058	0.7154040337	0.3748691082	0.236090228	0.7836633229	0.593659371	0.7592107523
llama-3.3-70b-specdec	48.51048361	72.63302647	35.70680628	0.722525835	0.3643979132	0.2301587313	0.7872387194	0.6078046804	0.7668984615
qwen-2.5-32b	47.70717079	72.14473684	38.32460733	0.7157424688	0.3931937218	0.2629908025	0.7849261113	0.6068604453	0.7646473397
llama-3.3-70b-versatile	48.64807005	72.66717786	35.60209424	0.7223548889	0.3633507788	0.2294068485	0.7852285372	0.6078740904	0.7658456074
llama-3.1-8b	48.36015249	72.21566312	36.7539267	0.7096066475	0.3774869144	0.2364244014	0.767973991	0.5939335388	0.7485217419
mixtral-8x7b	2.791660347	24.5545704	1394.712042	0.6217294931	13.95811558	10.03926468	0.6911647593	0.5106115172	0.6707965232
llama-3.2-90b-vision	48.96448583	72.72388184	35.44502618	0.7216063142	0.3643979132	0.2310777009	0.7881380518	0.6078988303	0.7677642173
en-ro									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	28.20125942	56.87283626	59.90722333	0.7193098664	0.6375082731	0.4417808354	0.2935238095	0.09333333333	0.2901904762
deepseek-r1-distill-70b	30.36789747	59.00267964	58.44930417	0.736006856	0.6110006571	0.4138428271	0.319	0.11	0.3156666667
llama-3.3-70b-specdec	33.9066504	62.2995726	54.47316103	0.7588149905	0.5772034526	0.3864455521	0.3233333333	0.09	0.32
qwen-2.5-32b	25.44659211	56.87125242	74.42014579	0.7074935436	0.7806494236	0.5912941694	0.3273333333	0.1	0.324
llama-3.3-70b-versatile	33.59395678	61.89962241	55.33465871	0.7563382983	0.584493041	0.3917627931	0.3166666667	0.09	0.3133333333
llama-3.1-8b	26.57303017	57.04604425	63.55202121	0.7164306045	0.6699801087	0.4587238729	0.3117777778	0.1078571429	0.3117777778
mixtral-8x7b	19.55303313	54.3089127	96.35520212	0.6759179831	0.9980119467	0.7481975555	0.2444655248	0.09	0.2444655248
llama-3.2-90b-vision	34.6497424	62.44566011	55.13585156	0.7653933764	0.5765407681	0.3882480264	0.314	0.11	0.3106666667

Figure 26: Translation Performance in the **Law** Domain Across the **en** → **pt, ro** (English–Portuguese, Romanian) Translation Pairs

en-pl									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	18.75047325	45.66989402	68.12909733	0.7011530995	0.7276853323	0.5433933735	0.24	0.12	0.24
deepseek-r1-distill-70b	29.17428815	56.10104466	56.27836611	0.7541465759	0.6066565514	0.4389825761	0.1473333333	0.03333333333	0.1473333333
llama-3.3-70b-specdec	35.8096238	60.18937429	48.2602118	0.7804477215	0.5254664421	0.3790606856	0.08266666667	0.03666666667	0.08266666667
qwen-2.5-32b	17.46848924	46.1988361	80.18154312	0.7081311941	0.8350983262	0.6899114251	0.2227598039	0.09040816327	0.2227598039
llama-3.3-70b-versatile	35.60861863	60.22183845	47.85678265	0.7807072997	0.5179021955	0.3743926883	0.08266666667	0.03666666667	0.08266666667
llama-3.1-8b	20.0469641	44.99103227	81.0892587	0.6830644011	0.8406454921	0.6526626945	0.198	0.11	0.198
mixtral-8x7b	5.255398301	29.2817162	188.1492688	0.4976644218	1.905698419	1.814232588	0.06914370575	0.0294218489	0.06914370575
llama-3.2-90b-vision	37.98699264	62.13022683	45.43620777	0.7934502959	0.4947049916	0.3588644266	0.238	0.11	0.238
en-it									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	37.2833612	65.62270542	47.33022291	0.6253080964	0.4935199618	0.3311163783	0.6817919675	0.4568470205	0.6534967472
deepseek-r1-distill-70b	40.42092518	67.28327192	44.68636599	0.6440407634	0.459305346	0.3074426055	0.7012834522	0.491139877	0.6823286146
llama-3.3-70b-specdec	42.15331337	68.28854675	44.530845	0.664359808	0.4572317302	0.298178941	0.7139748088	0.5165750573	0.69831159
qwen-2.5-32b	37.30438027	65.54788888	51.11456713	0.6091958284	0.5246241689	0.355661124	0.664575425	0.4558098746	0.6499953611
llama-3.3-70b-versatile	42.37633713	68.20052624	44.47900467	0.6618664265	0.4556765258	0.2990498841	0.7111498628	0.5146302358	0.696474039
llama-3.1-8b	38.82490525	66.03176079	48.15966822	0.6324273944	0.4950751662	0.3323832154	0.6843751997	0.4826859341	0.6667303405
mixtral-8x7b	33.17825847	64.1749966	57.23172628	0.5968947411	0.5889061689	0.4079176486	0.6503328671	0.4402622139	0.6284224085
llama-3.2-90b-vision	42.64882893	68.35726104	43.33851737	0.6637365818	0.4458268583	0.2930324674	0.7190164311	0.5269179212	0.7041506556

Figure 27: Translation Performance in the **Law** Domain Across the **en** → **pl, it** (English–Polish, Italian) Translation Pairs

en-es									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	49.79341147	72.69802024	35.40836653	0.7170011401	0.3884462118	0.2709222138	0.7732546775	0.5984367156	0.7393166434
deepseek-r1-distill-70b	49.794647	73.46028927	35.25896414	0.7284767628	0.3824701309	0.2631276548	0.786645926	0.6164086491	0.7549843691
llama-3.3-70b-specdec	56.19628245	77.18557851	29.83067729	0.7755883932	0.322709173	0.2184935957	0.8222585512	0.6697408637	0.7965037808
qwen-2.5-32b	50.70808845	73.88947854	34.66135458	0.7376719117	0.3700199127	0.2559894919	0.7912770259	0.6256415344	0.7635344685
llama-3.3-70b-versatile	55.92577123	77.12929436	30.12948207	0.7755622864	0.322709173	0.2173449248	0.8210621697	0.6627362552	0.7924910312
llama-3.1-8b	50.65404472	73.69693732	36.40438247	0.7250255346	0.3824701309	0.2629635632	0.7721214612	0.6139971843	0.7475806612
mixtral-8x7b	38.29329303	67.8763386	56.3247012	0.6541927457	0.5841633677	0.4677551687	0.7136433636	0.5462614229	0.6866278521
llama-3.2-90b-vision	56.27246065	77.09624579	30.17928287	0.7642019391	0.3256972134	0.2208729833	0.8189762374	0.6637263295	0.792309493
en-fr									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	42.26950408	67.46227727	43.81642512	0.6417820454	0.4545893669	0.2928641737	0.724400506	0.541878743	0.7033957181
deepseek-r1-distill-70b	40.81127645	67.09800131	46.03864734	0.6321811676	0.486956507	0.3157655299	0.7139827058	0.5229979473	0.6843345717
llama-3.3-70b-specdec	42.75660638	68.51057326	42.46376812	0.6480305195	0.4439613521	0.2871009409	0.7280339356	0.5457664054	0.7038674318
qwen-2.5-32b	38.55298362	66.24985191	52.51207729	0.6147677898	0.5439613461	0.3859862089	0.6986381549	0.5171204965	0.6760446313
llama-3.3-70b-versatile	42.69907428	68.53964795	42.41545894	0.6490207911	0.4434782565	0.2854326367	0.7301463708	0.5453630142	0.7054042136
llama-3.1-8b	39.92629232	66.035148	47.97101449	0.6164582372	0.5009661913	0.3354819119	0.7055204522	0.5071882296	0.6778897398
mixtral-8x7b	39.64067562	66.1321017	49.0821256	0.6201246381	0.5106280446	0.3465534151	0.6949492261	0.5107765756	0.6659550869
llama-3.2-90b-vision	43.13562346	69.28256748	42.41545894	0.6607220173	0.4444444478	0.2837643027	0.7359020776	0.5612409092	0.7151956089

Figure 28: Translation Performance in the **Law** Domain Across the **en** → **es, fr** (English–Spanish, French) Translation Pairs

en-el									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	11.43629644	40.22201502	75.92829706	0.6192105412	0.7695262432	0.5222365856	0.2303253968	0.05928571429	0.2293253968
deepseek-r1-distill-70b	12.85085132	42.373952	75.03201024	0.6473213434	0.7573623657	0.5028065443	0.1970658786	0.0580952381	0.1959547675
llama-3.3-70b-specdec	14.73859233	45.38843667	70.4865557	0.6637608409	0.711907804	0.468696028	0.1825244559	0.04371428571	0.1825244559
qwen-2.5-32b	2.653560899	23.07678545	428.9372599	0.4965210855	4.300256252	4.701532841	0.1340499073	0.02002685246	0.1335737168
llama-3.3-70b-versatile	14.49138176	46.03710774	70.74263764	0.6657516956	0.714468658	0.4671848118	0.1825244559	0.04371428571	0.1825244559
llama-3.1-8b	2.53345893	22.71910673	408.9628681	0.6013467908	4.098591328	4.034542084	0.1843940212	0.02954761905	0.1843940212
mixtral-8x7b	3.705284801	30.05798809	188.3482714	0.4820924997	1.893085837	1.675302267	0.132038017	0.05131779693	0.1301011891
llama-3.2-90b-vision	15.28823885	45.74207932	69.97439181	0.672252059	0.7080665827	0.4662132859	0.1911911226	0.05223076923	0.1911911226
en-de									
	BLEU	**chrF**	**TER**	**BERTScore**	**WER**	**CER**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
deepseek-r1-distill-32b	30.78353796	60.87867116	54.70422535	0.542104125	0.5712676048	0.4026942551	0.6460557233	0.4070076007	0.5990924141
deepseek-r1-distill-70b	36.37748078	65.40141699	49.35211268	0.5887586474	0.520563364	0.3568278551	0.6888130541	0.4652384224	0.6398574822
llama-3.3-70b-specdec	37.00981533	65.87070012	47.43661972	0.5957369804	0.4957746565	0.3471253216	0.6969930994	0.4772806516	0.6544989487
qwen-2.5-32b	32.28930763	62.62443976	57.12676056	0.552999258	0.6005634069	0.4419853985	0.6660840074	0.4393468224	0.6184547024
llama-3.3-70b-versatile	37.32402873	66.06465277	47.6056338	0.5988963842	0.4969014227	0.34247455	0.6974302559	0.479658573	0.6569685106
llama-3.1-8b	33.63535884	63.22333696	52.22535211	0.5616160035	0.5487323999	0.3715820611	0.664722003	0.4379715113	0.6183701849
mixtral-8x7b	2.447055898	25.19523044	1067.211268	0.5300765038	10.70478916	8.183706284	0.6371067346	0.4096639761	0.5892585401
llama-3.2-90b-vision	37.71271835	67.00902249	47.54929577	0.6101247668	0.502535224	0.3401491344	0.7092265221	0.4922338646	0.6568677032

Figure 29: Translation Performance in the **Law** Domain Across the **en** → **el, de** (English–Greek, German) Translation Pairs