No Text Needed: Forecasting MT Quality and Inequity from Fertility and Metadata

Jessica M. Lundin
Institute for Disease Modeling
Gates Foundation

Ada Zhang
University of San Francisco

David Ifeoluwa Adelani
Mila, McGill University &
Canada CIFAR AI Chair

Cody Carroll
University of San Francisco

Abstract

We show that translation quality can be predicted with surprising accuracy from token-level statistics and linguistic metatdata alone, *without inspecting the translated text*. Using only a handful of features: token fertility ratios, token counts, and basic linguistic metadata (language family, script, and region), we can forecast ChrF scores for GPT-4o translations across languages in the FLORES-200 benchmark. Gradient boosting models achieve favorable performance ($R^2 = 0.66$ for XX→English and $R^2 = 0.72$ for English→XX). Feature importance analyses reveal that typological factors dominate predictions into English, while fertility plays a larger role for translations into diverse target languages, and the importance of fertility varies by model. We are not proposing methodology for quality estimation, rather these findings suggest explainability of translation quality shaped by both token-level fertility and broader linguistic typology, offering insight for multilingual evaluation and quality estimation. Our findings reveal systematic performance disparities across language families and regions, with implications for fairness and equity in multilingual NLP systems.

1 Introduction

Machine translation (MT) quality evaluation has evolved from early rule-based and statistical methods to large-scale neural models. Traditional evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) have been widely used but have been criticized for their limited sensitivity to linguistic diversity and their reliance on surface-level n-gram matches. More recent alternatives, such as ChrF (Popović, 2015) and ChrF++ (Popović, 2017), leverage character-level representations to better capture morphologically rich languages, and have shown strong correlation with human judgments in multilingual settings. The difference between quality estimation and this work is that we are not proposing a method for quality estimation, but rather, seek to understand systematic cross-linguistic patterns in model behavior.

One central factor in MT quality is fertility, originally formalized in IBM Models for statistical machine translation (Brown et al., 1993). Fertility refers to how many target tokens are generated per source word, and imbalances in fertility often lead to errors such as under-translation or over-translation. Figure 1 shows the language codes with highest and lowest fertility values for FLORES-200. While fertility has historically been studied in statistical MT, its role in neural MT evaluation remains underexplored. Recent work in large-scale multilingual evaluation, such as MEGA (Ahuja et al., 2023), has underscored the importance of tokenization and representation choices in shaping model performance across diverse languages, suggesting that fertility may still play a critical role.

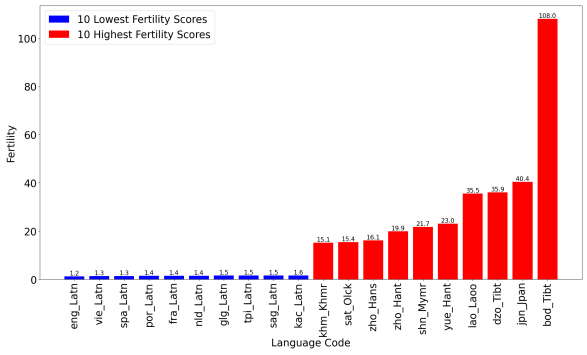


Figure 1: Languages with highest and lowest fertility for FLORES-200 dataset. Latin scripts tend to have lower fertility values.

At the same time, linguistic metadata such as language family, script type, and geographic region have been shown to influence translation performance. For example, scaling efforts such as NLLB (Costa-jussà et al., 2024), SIB-200 (Adelani et al., 2024) and mSTEB (Beyene et al., 2025)

demonstrate that typological and resource disparities across languages lead to systematic variation in model accuracy. Similarly, the WMT Quality Estimation shared tasks (Specia et al., 2020) have highlighted the value of incorporating language-level features into predictive models of translation quality.

These findings point to the need for a systematic investigation of how fertility and linguistic metadata interact with translation quality. Rather than building a runtime quality estimator, our research question: What factors explain quality variation across 200 languages? By modeling these features explicitly, we find interpretable patterns.

2 Methods

We developed 5 regression models (Linear, Lasso, MLP, Random Forest, XGBoost) to predict ChrF (Character n-gram F-score) translation quality scores for GPT-4o translations in the FLORES-200 benchmark dataset (NLLB Team et al., 2024).

For data, we utilize the LLM text translations of the FLORES-200 benchmark and annotated features released by mSTEB (Beyene et al., 2025). Our analysis covered two translation directions: multilingual-to-English (XX→English) and English-to-multilingual (English→XX) across 200 languages.

Table 1 shows features used to fit the ChrF regression. We extracted both linguistic and text-level features from the translation pairs. Text-level features included token counts for both source and target texts using the "`o200kbase`" tokenizer, as well as fertility ratios (tokens per word) for source and target texts. Language-level metadata include language (ISO language codes, script (29 scripts), Joshi class (0-5), language family, and geographic region. We compared the fit of multiple regression approaches: **Linear (OLS, Lasso)**, **Tree ensembles (Random Forest, XGBoost)**, and **multi-layer perceptron (MLP)**, using metrics R^2, RMSE, MAE. For reproducibility, we used 20% hold out and values from hyperparameter grid search are shown in Table 2.

Feature importances were extracted from the trained Random Forest and XGBoost models to identify which variables most strongly influence translation quality predictions. For Random Forest models, importances are calculated using mean decrease in impurity (Gini importance). During tree construction, each feature's contribution to reducing variance at split nodes is tracked and averaged across all 300 trees in the forest. Features that consistently produce larger reductions in the residual sum of squares when used for splitting receive higher importance scores. For XGBoost models, we use the default gain-based importance metric, which measures the average improvement in prediction accuracy (reduction in loss function) contributed by each feature across all trees.

Marginal averages are calculated by first training Random Forest and XGBoost regression models on the training set using all features (categorical language features and numeric tokenization features). After training, the models generate predictions on the held-out test set. For each categorical feature (Region, Family, Script, Joshi Class, Language Code), we group the test set observations by their categorical values and compute the mean predicted score within each group. For example, to determine the marginal average score for "Africa," we average all predicted translation quality scores for test samples where the region is Africa, marginalizing over all other feature values (different language families, scripts, fertility rates, etc.). Both predicted means and actual means are calculated for each category to assess model fit.

3 Results and Discussion

3.1 Model Performance Comparison

Table 3 demonstrates a clear performance hierarchy across all model types and translation directions. The substantial performance gap between linear models ($R^2 \approx 0.25$-0.31) and tree-based models ($R^2 \approx 0.66$-0.72) indicates strong non-linear relationships in the data that simple linear combinations cannot capture effectively. XGBoost achieves the highest performance in both directions: English-to-XX and XX-to-English respectively had R^2 values of 0.719 and 0.663, while Random Forest achieved values of 0.701 and 0.588.

Neural networks show moderate performance (0.586 train R^2 and 0.684 test R^2) but remain substantially below ensemble methods, while linear approaches (Linear and Lasso Regression) perform poorly across both directions with nearly identical results. The consistent superiority of XGBoost over Random Forest suggests that gradient boosting is particularly well-suited for capturing the complex interactions between language-level metadata and translation quality, with performance gaps of 0.018 R^2 for English-to-XX and 0.075 R^2 for XX-

Feature	Example	Description
Joshi Class	0-5	Joshi class labels (Joshi et al., 2021), categorizing languages by resource availability and computational support. These are imputed for the FLORES-200 dataset.
Region	Africa, Europe, Americas	9 geographic regions where the language is primarily spoken, used to capture regional linguistic patterns
Family	Austronesian, Afro-Asiatic, Indo-European	Linguistic family classification, grouping languages by common ancestral origins and structural similarities
Script	Arab, Latn, Cyrl, Deva	29 scripts used (Arabic, Latin, Cyrillic, Devanagari, etc.)
Code	ace, afr, amh, ara	ISO language code identifier, uniquely identifying each language
Reference Fertility	2.5, 3.2, 4.1	Average number of tokens per word in reference (human) translations, measuring morphological complexity
Candidate Fertility	2.3, 3.0, 3.8	Average number of tokens per word in LLM-generated translations, measuring model's tokenization efficiency
Reference Tokens	54, 101, 160	Total number of tokens in the reference translation for a given text
Candidate Tokens	48, 78, 200	Total number of tokens in the LLM-generated translation for a given text

Table 1: Features used in the ChrF predictive analysis.

Model	Direction	Hyperparameters
Linear Regression	both	tol=1e-06
LassoCV	En→XX	cv=5, α=0.0347
	XX→En	cv=5, α=0.0191
Random Forest	both	n_est=300, depth=None, feat='sqrt', split=2, leaf=5, seed=42
XGBoost	En→XX	n_est=300, depth=7, lr=0.1, sub=0.6, colsamp=1.0, child_wt=5,
	XX→En	n_est=300, depth=10, lr=0.05, sub=0.8, colsamp=0.8, child_wt=1,
MLP	En→XX	layers=(128,64), act=tanh, ep=50, batch=64
	XX→En	layers=(256,128,64), act=relu, ep=50, batch=32

Table 2: Sklearn (Pedregosa et al., 2011) hyperparameters for regression models predicting translation quality

to-English translation.

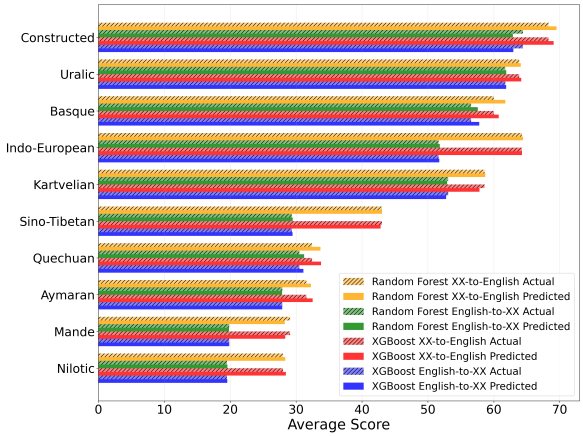
Figure 2 presents a comparison of marginal average translation quality scores across three key categorical features: geographic region, language family, and script. The analysis reveals systematic performance patterns that are remarkably consistent between XGBoost and Random Forest models across both translation directions, indicating that these linguistic biases are inherent to the underlying data rather than model-specific artifacts. Language family effects (2a) show dramatic variation between resource levels, where constructed languages like Esperanto and high-resource families like Indo-European score 15-20 points higher than low-resource families such as Niger-Congo and Austronesian. Regional disparities (2b) demonstrate clear geographic variation, with European languages achieving scores of 55-65 compared to 35-45 for African languages. Effects from the script type (2c) reveal the advantage of the top 5 scripts: Armn, Hebr, Thai, Grek, and Cyrl. In a twist, Latin (Latn) is not in the top nor bottom 5 scripts in terms of ChrF. The nearly identical performance patterns between XGBoost and Random Forest, combined with close alignment between predicted and actual scores, suggest that both models capture the same underlying linguistic regularities with high fidelity, making the observed categorical biases robust findings rather than model-dependent effects.

3.2 Feature Importance Analysis

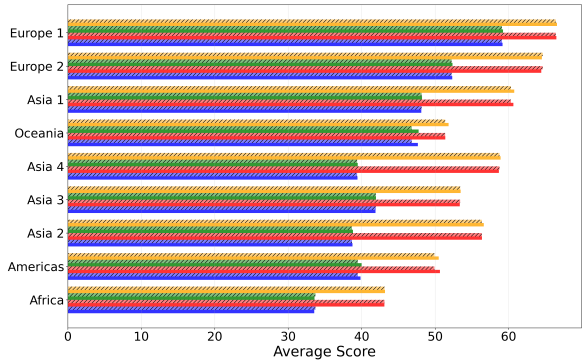
Table 4 reveals distinct patterns in how XGBoost and Random Forest prioritize language-level metadata across translation directions. For English-to-XX translation, XGBoost places overwhelming emphasis on Joshi Class (0.365 importance), indicating that resource-level categorization is the most critical factor when translating into diverse target languages. This is followed by regional patterns (0.206) and language family relationships (0.133), creating a clear hierarchical structure focused on language categorizations.

Model	XX→English			English→XX		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Linear Regression	0.25	16.58	13.50	0.31	16.19	13.01
Lasso Regression	0.25	16.58	13.51	0.31	16.19	13.01
MLP	0.59	12.34	9.71	0.68	10.97	8.33
Random Forest	0.59	12.31	9.73	0.70	10.68	8.11
XGBoost	0.66	11.14	8.71	0.72	10.36	7.88

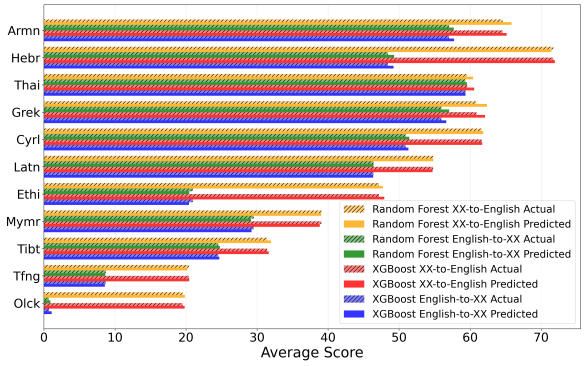
Table 3: Model Performance Comparison across Translation Directions



(a) Language Family



(b) Region



(c) Script

Figure 2: Marginal Average Translation Quality Scores by Categorical Features. Each panel shows XGBoost vs Random Forest performance across both English-to-XX and XX-to-English translation directions, displaying both predicted and actual scores. Panel (a) examines the top and bottom 5 language families, demonstrating the substantial performance gap between high-resource families like Constructed and Indo-European languages versus low-resource families such as Mande and Nilotic. Panel (b) compares performance across geographic regions, revealing systematic differences where European languages consistently have the highest performance. Panel (c) shows performance by script, with the top 6 (Latn is 6th) and bottom 5 scripts. Across panels, predicted scores (solid bars) align closely with actual scores (hatched bars), indicating good model calibration, and XGBoost and Random Forest show remarkably similar performance patterns, suggesting that the underlying linguistic patterns are robust across different ensemble methods.

English-to-XX		XX-to-English	
XGBoost	**Random Forest**	**XGBoost**	**Random Forest**
Joshi Class (0.365)	Joshi Class (0.205)	Region (0.278)	Region (0.198)
Region (0.206)	Language Id Code (0.183)	Family (0.208)	Family (0.163)
Family (0.133)	Region (0.161)	Joshi Class (0.178)	Language Id Code (0.156)
Script (0.127)	Reference Fertility (0.115)	Script (0.103)	Joshi Class (0.116)
Language Id Code (0.092)	Candidate Fertility (0.081)	Language Id Code (0.092)	Candidate Fertility (0.089)
Reference Fertility (0.025)	Family (0.073)	Reference Fertility (0.040)	Reference Fertility (0.082)
Candidate Fertility (0.021)	Reference Tokens (0.072)	Reference Tokens (0.041)	Script (0.077)
Reference Tokens (0.018)	Script (0.066)	Candidate Fertility (0.029)	Reference Tokens (0.061)
Candidate Tokens (0.013)	Candidate Tokens (0.044)	Candidate Tokens (0.031)	Candidate Tokens (0.059)

Table 4: Feature Importance Rankings by Model and Translation Direction

Random Forest shows a more distributed approach for English-to-XX translation, with Joshi Class leading (0.205) but at much lower intensity than XGBoost, followed closely by individual language codes (0.183) and regional groupings (0.161). This more balanced distribution suggests Random Forest's ensemble approach captures a broader range of linguistic patterns rather than focusing heavily on a single dominant feature.

The XX-to-English direction shows different priorities for both models. XGBoost emphasizes regional patterns (0.278) followed by language family relationships (0.208) and Joshi Class (0.178), suggesting that when translating into English, geographic and phylogenetic groupings become more predictive than individual resource levels. Random Forest maintains region as the top feature (0.198) but shows more balanced importance across language family (0.163) and individual language codes (0.156).

The two model types diverge on fertility features. In Random Forest, combined fertility (reference+candidate) is 0.196 for English -> XX, nearly matching Joshi Class (0.205), the most dominant feature. XGBoost assigns fertility features lower than Random Forest, which could reflect difference in optimization strategies, where XGBoost gradient boosting prioritizes the single most discriminative splits (categorical features), while Random Forest bagging captures patterns that categorical features cannot fully represent.

4 Conclusion

This work shows that much of machine translation quality can be anticipated without ever looking at the translated words themselves. Using high-level linguistic metadata, fertility ratios, and token counts, tree-based models predict ChrF scores across 200 languages with striking accuracy. While the act of predicting ChrF is not inherently remarkable, the fact that such predictions are possible without examining the text itself underscores the systematic role of fertility and typology in shaping translation quality.

Equally important, tree ensembling methods such as XGBoost not only deliver the strongest predictive performance but also provide interpretable insights into which features matter most. The relative weights assigned to fertility, resource levels, and typological categories reveal consistent cross-linguistic patterns: target-side fertility explains predictability into diverse languages, while source-side typology dominates when translating into English. These interpretable rankings allow us to see translation quality through the lens of linguistic structure rather than opaque model behavior.

Looking forward, this perspective supports the role for lightweight quality estimation as a diagnostic tool for understanding multilingual systems and the typological factors that drive their performance. By showing that translation quality can be explained largely from fertility and metadata alone, our results highlight a path toward more efficient, interpretable, and linguistically grounded approaches toward improving translation.

5 Limitations

Our approach to predicting translation quality from fertility and metadata, while demonstrating strong performance, has limitations. Our analysis is restricted to GPT-4o translations on the FLORES-200 benchmark, which may not generalize to other

LLMs, traditional MT systems, or specialized domains. The exclusive reliance on ChrF scores, despite their correlation with human judgments, cannot capture nuanced aspects of translation quality such as cultural appropriateness or contextual accuracy. Our linguistic categorizations (family, script, region) represent coarse-grained groupings that may obscure important within-category variations, for example, treating all Niger-Congo languages uniformly ignores variation within this family. The use of a single tokenizer (o200kbase) limits our understanding of how different tokenization schemes might affect fertility-quality relationships.

6 Broader Impact

This work contributes to trustworthy multilingual NLP by making translation quality disparities visible and interpretable. By identifying which linguistic factors drive performance gaps, our findings can guide targeted investment in low-resource language development and inform fairer evaluation practices. In terms of risks, our findings reveal and potentially amplify existing biases in MT evaluation: the strong predictive power of resource-level indicators (Joshi Class) and regional groupings risks perpetuating a cycle where low-resource languages receive less attention due to anticipated lower quality scores. This could inadvertently discourage investment in improving translation for underserved languages, as stakeholders might view poor performance as inherent to these languages rather than a consequence of limited training data and research attention. Using our models for pre-deployment quality estimation could lead to discriminatory practices, such as withholding MT services from speakers of languages predicted to have lower quality, thereby exacerbating digital language divides. We therefore caution against using these predictions as gatekeeping mechanisms and instead recommend them solely as diagnostic tools for understanding systemic challenges in multilingual NLP.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *Preprint*, arXiv:2303.12528.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. msteb: Massively multilingual evaluation of llms on speech and text tasks. *arXiv preprint arXiv:2506.08400*.

Peter F Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world. *Preprint*, arXiv:2004.09095.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2015. Chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Lucia Specia, Andre FT Martins, Carolina Scarton, and 1 others. 2020. Findings of the wmt 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764.