Register Sensitivity in Scalar MT Evaluation: Evidence from Spanish–Basque Informal Discourse

Nora Aranberri
HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
nora.aranberri@ehu.eus

Abstract

Automatic scalar metrics are widely used for machine translation (MT) evaluation, yet their behavior under sociolinguistic variation remains underexplored, particularly in under-resourced and minority-language contexts. We present a small, controlled empirical analysis of reference-based evaluation in Spanish–Basque informal discourse. Register is operationalized as indexical density, capturing dialectal forms, informal lexicon, code-switching and ortographic stylization. Across two MT systems and prompting conditions, sentence-level scores from chrF++, COMET-DA, and XCOMET-XL show a consistent negative association with indexical density under the original informal reference. In a reference-perturbation design that holds MT outputs constant while replacing the informal reference with a standardized Batua version, scores increase systematically, particularly for high-density items, and the density–score association weakens. These results provide controlled evidence that evaluation outcomes in this setting depend in part on reference register configuration. In minority-language and informal domains, reference design choices may influence how translation quality is measured and interpreted.

Keywords: under-resourced languages, machine translation evaluation, register variation

1. Introduction

Automatic MT evaluation is predominantly operationalized through scalar metrics that approximate translation quality via reference similarity or learned adequacy predictions. While such metrics often correlate with aggregate human judgments, they compress multiple quality dimensions into a single numerical signal. Sociolinguistic alignment—particularly register—is rarely modeled as an independent dimension and may instead be subsumed within adequacy-oriented scores.

This issue may be especially salient in informal and under-resourced or minority-language contexts. Informal discourse frequently exhibits dense sociolinguistic marking—dialectal forms, stylization, and code-switching—that encodes stance and identity alongside propositional content. In such settings, scalar evaluation may insufficiently distinguish between semantic adequacy and sociolinguistic alignment. While these challenges are not specific to under-resourced languages, their consequences are amplified in such contexts: data scarcity can reduce evaluation robustness, and the specific case of minority-language settings, misinterpretation of evaluation signals may have more serious consequences due to the sociolinguistic fragility of the language.

We explore this question in a small, controlled study of Spanish–Basque informal digital discourse. Using a stratified subset of Instagram direct messages (DMs), we operationalize indexical density as an aggregate measure of register-marked variation and analyse its relationship to sentence-level scores from widely used metrics (chrF++ and COMET-family models). We further introduce a reference-perturbation design in which MT outputs are held constant while the reference register is modified, and assess whether evaluation outcomes vary as a function of reference configuration.

Our analysis addresses three questions: (i) Does register density co-vary with scalar metric scores? (ii) Is any observed association consistent across systems and prompting conditions? (iii) Do evaluation outcomes shift when reference register is normalized and the translation output remains fixed?

The contributions of this study are threefold. First, we articulate a theoretical account of register as a structured quality dimension that may be compressed in scalar MT evaluation. Second, we introduce a controlled perturbation framework for examining register sensitivity in an under-resourced and minority-language setting. Third, we provide initial empirical evidence that, in this setting, reference-based scalar evaluation outcomes vary systematically with register density and reference configuration.

2. Register and Automatic MT Evaluation

This section outlines the theoretical framework motivating the empirical analysis. It situates scalar MT evaluation within discussions of translation quality and measurement, focusing on register as a structured dimension of sociolinguistic variation. We examine how scalar metrics aggregate mul-

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 19–32
11–12 May 2026.

tiple quality dimensions into a single signal and how register-sensitive phenomena surface in informal and under-resourced settings, particularly in minority-language contexts.

2.1. Scalar Evaluation and the Compression of Quality Dimensions

Automatic MT evaluation has increasingly converged on scalar metrics that summarize translation quality in a single score, typically through reference-based similarity or learned adequacy predictions (Papineni et al., 2002; Post, 2018; Rei et al., 2020). While such metrics can correlate well with aggregate human judgments in certain contexts, they necessarily summarize multiple dimensions of translation quality—such as semantic adequacy, fluency, stylistic alignment, and sociolinguistic appropriateness—into a single numerical outcome. This summarization entails a compression of quality dimensions: distinctions that are meaningful in human communication may be absorbed into an undifferentiated scalar signal. Register is rarely modeled as an explicit and separable dimension of translation quality and is typically subsumed within broader adequacy judgments.

Register as an unmodeled dimension. Despite methodological evolution, a core structural assumption persists: translation quality is approximated as a scalar quantity, within which dimensions such as register are not explicitly modeled. Meta-evaluation research shows that both classical and neural metrics correlate most strongly with aggregate human judgments of adequacy and fluency (Callison-Burch et al., 2007; Freitag et al., 2021), while also documenting sensitivity to evaluation design choices (Mathur et al., 2020; Post, 2018). Within this scalar paradigm, multiple quality dimensions are combined into a single evaluative signal. While recent work has explored generative or LLM-based evaluation frameworks that produce comparative or multi-dimensional feedback, scalar metrics remain central in MT benchmarking and meta-evaluation, and their behavior under sociolinguistic variation warrants careful examination.

In practice, sociolinguistic appropriateness—particularly register—is rarely operationalized independently Neural metrics may implicitly reflect stylistic mismatches insofar as annotators penalize them, yet register is not typically treated as a distinct construct in evaluation design. Metrics do not isolate sociolinguistic alignment nor provide interpretable signals about stylistic deviation. From a measurement-theoretic perspective, this configuration may give rise to construct underrepresentation (Messick et al., 1989). Translation quality, as conceptualized in translation studies (House, 2015; Baker, 2018), encompasses semantic adequacy, fluency, and pragmatic or stylistic equivalence. Human evaluation frameworks such as MQM explicitly distinguish among these dimensions (Lommel et al., 2014). Automatic metrics, by contrast, summarize them into a single scalar output. This dimensional compression may obscure trade-offs and limit diagnostic insight. A translation that preserves semantics but shifts register may receive a high score; a translation that aligns register but diverges lexically from a reference may receive a lower one.

Register as a structured dimension of translation. To clarify why this distinction matters, it is necessary to specify the nature of register as a structured form of sociolinguistic variation. In systemic-functional linguistics (Halliday and Hasan, 1976), register captures systematic variation conditioned by situational context. Corpus-based research shows that registers are realized through clusters of co-occurring linguistic features (Biber and Conrad, 2009). Sociolinguistic theory further emphasizes that such features carry indexical meanings, signaling stance, identity, solidarity, authority, and social positioning (Silverstein, 2003; Eckert, 2008). Crucially, register is not merely ornamental stylistic variation but an integral component of communicative meaning in interaction. Translation may preserve propositional content while altering these indexical signals; such shifts can modify pragmatic force (Levinson, 1983), even when truth conditions remain intact.

It is often noted that translation is "many-to-many." While multiple target realizations are possible for a given source utterance, sociolinguistic systems are structured. Linguistic features cluster in ways that index specific social meanings; not all stylistic substitutions are equivalent. Youth slang does not index the same identity as regionally marked speech, and casual professional tone differs from everyday informal interaction. Translation thus admits not unlimited variation but a bounded and structured set of sociopragmatically plausible realizations within the target language. Register equivalence can therefore be understood as constrained alignment within a sociolinguistic system.

Most widely used automatic MT metrics do not explicitly model register as a distinct aspect of translation quality, even if neural metrics may capture some stylistic effects indirectly through correlations with human judgments. What remains limited is an interpretable and separable register-sensitive component within the evaluation signal. Informal registers make this limitation more visible; under-resourced and minority-language settings may heighten its practical implications.

2.2. Register Sensitivity in Informal MT

Building on the structured view of register outlined above, informal discourse translation provides a particularly informative setting for examining register as a quality dimension in MT evaluation. Informal digital discourse is characterized by dense indexical marking, flexible orthographic practices, and high paraphrastic variability (Androutsopoulos, 2014). Dialectal features, code-switching, stylization, and expressive punctuation often contribute meaning beyond propositional content, shaping stance and social positioning. In such contexts, semantic adequacy alone may not fully capture communicative equivalence; alignment in register becomes central to how translations are interpreted.

Register-sensitive translation therefore involves more than selecting semantically equivalent lexical items. It requires alignment with the sociolinguistic constraints of the target system and with a projected speaker identity, including coherent combinations of dialectal features, levels of stylization, and patterns of code-switching. Register features tend to cluster in socially meaningful configurations and must remain internally consistent within an utterance or interaction. Incoherent mixtures can disrupt indexical alignment even when semantic content is preserved.

To illustrate the distinction between propositional content and register-marked variation, consider the following example from Spanish–Basque translation (with the English gloss for clarity):

> **Source (ES):** Tiia tenmos q qedarrrr
> **Reference (informal EU):** Tiia egon behar gea noizbaitt
> **Reference (standard EU):** Aizu noizbait egon behar gara
> *Gloss (EN):* "Girlll we gotta hanggg outttt"

The propositional content in both references is the same and properly aligned with the source. However, the informal reference includes register-marked features such as code-swithicng (*tía* instead of the standard *aizu*), dialectal use (*gea* instead of the standard *gara*) and stylization (double *i* and double *t*). These differences illustrate how indexical variation may affect evaluation outcomes independently of semantic adequacy.

Evaluation implications in informal MT. Because informal discourse admits multiple contextually valid realizations that may differ lexically from a single reference, reference-based metrics may assign lower scores to translations that diverge in surface form despite preserving pragmatic alignment (Callison-Burch et al., 2006; Mathur et al., 2020). Neural metrics trained on aggregate human judgments may likewise incorporate stylistic variation into broader adequacy signals without isolating it as a distinct evaluative dimension. Informal MT therefore provides a setting in which potential interactions between sociolinguistic alignment and scalar evaluation can be examined.

An additional asymmetry is relevant. MT systems often exhibit normalization tendencies, whereby stylistically marked or socially salient signals in the source may be attenuated in translation (Rabinovich et al., 2017). Research on controllable MT further indicates that non-propositional dimensions such as politeness and formality typically require explicit modeling and supervision (Sennrich et al., 2016; Briakou et al., 2021). Stylistic preservation is therefore not guaranteed by optimizing for semantic adequacy alone.

Related findings in affective translation suggest that non-propositional meaning can be weakened even when propositional content is largely preserved (Troiano et al., 2020; Kumari et al., 2021; Aranberri, 2026). These observations motivate closer examination of how scalar evaluation signals behave when register marking is dense and socially meaningful.

While informal MT provides a useful setting for studying register sensitivity, broader implications depend on context. In under-resourced and minority-language environments—where informal and non-standard varieties play an important communicative role—evaluation design choices may shape translation quality interpretation. Rather than presupposing systematic bias, the present study examines under controlled conditions whether scalar evaluation outcomes vary with register density and reference configuration in one such setting.

2.3. Register Sensitivity in Under-Resourced and Minority-Language Settings

In under-resourced and minority-language contexts, the interaction of data scarcity, single-reference evaluation, orthographic variability, and sociolinguistic stratification may amplify the practical consequences of register insensitivity. What may appear in high-resource settings as a relatively subtle evaluation limitation can become more consequential when linguistic diversity is unevenly represented in training and evaluation resources.

NLP resource distribution is highly unequal across languages (Joshi et al., 2020). For many under-resourced languages, parallel corpora are often dominated by formal or institutional registers, such as parliamentary proceedings, religious texts, or news. Domain skew is a well-documented MT challenge (Koehn and Knowles, 2017), and systems trained primarily on standardized varieties may reproduce institutional norms while underrepresenting informal and dialectal variation.

Evaluation implications in under-resourced settings. Evaluation datasets often mirror this imbalance and frequently rely on a single reference translation. Single-reference evaluation increases sensitivity to surface variation and lexical divergence (Callison-Burch et al., 2006; Mathur et al., 2020). In languages with rich morphology or flexible orthography, relatively minor variation can substantially affect overlap-based scores. Informal and dialectal realizations may therefore be more likely to diverge from a canonical reference instantiation.

Sociolinguistic research emphasizes that standardization is socially embedded and often politically salient (Labov, 1972; Milroy and Milroy, 1999). Dialect and register encode regional belonging, generational identity, and stance. Work on dialects in NLP demonstrates that non-standard varieties pose persistent challenges for language technologies and are associated with performance disparities (Jørgensen et al., 2015; Blodgett et al., 2016; Blodgett and O'Connor, 2017). These disparities have been discussed not only as technical performance issues but also in relation to representational fairness and linguistic inequality in NLP systems (Blodgett et al., 2020).

The present work does not attempt to adjudicate broader questions of bias or fairness in MT evaluation. Rather, it examines whether, in a controlled Spanish–Basque informal setting, scalar evaluation outcomes vary systematically with register density and reference configuration. By situating the analysis in an under-resourced and minority-language context where standard and non-standard varieties coexist, we examine how evaluation design interacts with sociolinguistic structure under conditions of limited and unevenly distributed data. The next section operationalizes these considerations in a controlled empirical setting.

3. Empirical Study

Building on the preceding arguments, we conduct a small controlled analysis of Spanish–Basque informal discourse to test whether indexical density interacts with evaluation scores. We operationalize register as *indexical density* and examine its relationship to sentence-level metric scores across systems, prompts, and reference configurations. By manipulating only the reference while holding outputs constant, we isolate how evaluation configuration shapes scoring outcomes.

3.1. Data Selection and Annotation

The data derive from the GazteSare corpus (Elordui et al., 2020), which contains around 17,000 Instagram direct messages (DMs) characterized by dialectal variation, Spanish code-switching, orthographic stylization, and expressive punctuation. For a subset of DMs, Spanish parallel versions were previously constructed to enable controlled translation experiments. In the present study, these Spanish texts serve as source inputs, while the original Basque DMs function as references.

The Spanish versions were produced through translation and review by native bilingual youth speakers, including a primary translator with formal training in Basque philology and a second bilingual reviewer who refined idiomaticity and orthographic stylization. A final pass ensured consistency and the absence of unintended distortions. The process aimed to preserve propositional content, stance, and interactional tone, with orthographic stylization and affective cues mirrored using conventional informal Spanish practices where possible. As a result, Basque–Spanish code-switching in the originals was largely normalized into natural Spanish, whereas Basque–English switching was retained or pragmatically mirrored.

From the subset with available Spanish versions, we excluded non-linguistic messages (e.g., emoji-only or URL-only items) and utterances without Basque text. A feasibility-driven sample of 250 DMs was assembled for manual annotation (mean length 4.64 tokens, SD = 3.02; min = 1, max = 18). While most items were drawn from the eligible pool, additional low-density cases were included to ensure variation in register marking. Distributional statistics are reported for transparency rather than representativeness.

3.2. Operationalization of Indexical Density

DMs were classified according to *indexical density*, which we use to capture the extent to which an utterance encodes social and stylistic meaning beyond propositional content. Concretely, indexical density is defined as the number of distinct register-marked feature types present in the DMs. Four feature dimensions were annotated: (i) dialectal forms (including speech-based contractions and non-standard phonological realizations of dialectal morphology), (ii) informal lexical items, (iii) code-switching, and (iv) expressive orthographic stylization (e.g., repeated punctuation, character elongation, and laughter strings such as "jajaja"). Higher indexical density thus corresponds to a greater concentration of such features within an utterance.

Annotation was binary at the feature-type level (presence vs. absence within the DMs), irrespective of token frequency. Phonological spellings were treated as dialect to avoid double-counting, with stylization counted separately only when combined with independent expressive elongation or repetition. Isolated typographical errors or spelling devia-

tions were not treated as indexical features. Annotation was conducted by a bilingual Basque–Spanish speaker with expertise in Basque sociolinguistic variation.

Indexical density therefore reflects the accumulation of heterogeneous register-marked feature types—structural variation, cross-linguistic insertion, lexical informality, and expressive stylization—rather than token frequency or perceived sociolinguistic salience. Given the brevity of the DMs, this type-based operationalization provides a practical proxy for concentration of register marking while minimizing length-related bias.

For analysis, density is treated both as an ordinal variable (0–4) and, where appropriate, as a binary contrast (lower density: 0–1; higher density: 2+). Applying this operationalization to the 250-item working sample yields the following distribution: 0 (35), 1 (103), 2 (71), 3 (38), and 4 (3).

3.3. Systems and Prompting Conditions

For each of the 250 DMs, translations were generated using two systems: (i) ChatGPT (version 5.2) and (ii) Latxa (Etxaniz et al., 2024), an open-source Basque-specific large language model. Each system was evaluated under two prompting conditions (Appendix A): a baseline translation instruction and a context-rich instruction providing sociolinguistic cues, yielding 1,000 translation outputs (250 messages $\times$ 2 systems $\times$ 2 prompts).

ChatGPT-5.2 outputs were generated via the public ChatGPT web interface (February 2026 release) using default decoding settings; no sampling parameters were manually adjusted, and a single output was recorded per input. For Latxa, version Latxa-Llama-3.1-70B-Instruct-exp_2_101_v2-FP8 (temperature 0.7; top-p 0.9) was used.

3.4. Evaluation Metrics and Statistical Analysis

We compute automated quality scores using three reference-based MT metrics and one reference-free quality estimation metric to examine how scalar evaluation interacts with register-marked variation.

Reference-based metrics. We include chrF++ (Popović, 2017), a character n-gram F-score metric sensitive to morphological and orthographic variation (computed with `sacrebleu`); COMET-DA (Rei et al., 2022), a neural metric trained to predict human direct assessment (DA) scores (using the `wmt22-comet-da` checkpoint); and XCOMET-XL (Guerreiro et al., 2024), an extension incorporating MQM-style annotations (using the `xcomet-xl` checkpoint). Together, these metrics contrast surface overlap with adequacy-oriented and error-aware neural evaluation signals.

All reference-based metrics are computed at the sentence level under two reference configurations: (i) the original informal Instagram DM (INSTA), and (ii) a standardized Basque (Batua) version (BATUA). The Batua references were produced through manual normalization by a bilingual speaker trained in Basque philology, with the aim of preserving propositional content while removing dialectal and informal register marking. The normalized versions were reviewed to ensure semantic equivalence with the original DMs.

Reference-free metric. To assess whether density-related effects are specific to reference anchoring, we additionally compute sentence-level scores using CometKiwi-XL (Rei et al., 2023), a reference-free quality estimation (QE) model using the `wmt23-cometkiwi-da-xl` checkpoint. Because QE models predict translation quality without direct comparison to a reference, they can provide a control condition for disentangling translation difficulty from reference-dependent evaluation effects.

Statistical analyses. Our analyses focus on three complementary comparisons. All hypothesis tests are two-sided unless otherwise noted.

1. **Ordinal association between indexical density and metric scores.** We compute Spearman rank correlations (ρ) between indexical density (0–4) and sentence-level metric scores. Spearman correlation is used due to the ordinal nature of density and non-normal score distributions.

   Binary contrast (low vs. high density). To complement the ordinal analysis, we compress density into two groups: low (0–1 features) and high (2+ features). We compare sentence-level metric scores between these groups using the non-parametric Mann–Whitney U test. Effect sizes are reported as rank-biserial correlations.

2. **Reference perturbation effects.** For reference-based metrics, we compute per-item reference deltas:

$$\Delta_i^{(m,c)} = \text{score}_{\text{BATUA},i}^{(m,c)} - \text{score}_{\text{INSTA},i}^{(m,c)}, \quad (1)$$

   where m indexes the metric and c the system–prompt condition. Positive Δ values indicate higher scores under the standardized reference. Paired Wilcoxon signed-rank tests assess whether median Δ differs from zero. We further examine the association between density and Δ using Spearman correlation. Bootstrap confidence intervals for differences in correlations ($\Delta\rho$) are estimated by resampling sentence-level observations with replacement

while preserving pairing across reference conditions.

Robustness checks. Because indexical density correlates moderately with DM length, we compute partial rank correlations controlling for token length. These analyses assess whether density-related effects are reducible to length variation rather than register marking per se. Token length was computed as whitespace-separated tokens in the Basque output.

4. Metric Sensitivity to Indexical Density

We first analyse whether indexical density co-varies with scalar evaluation scores under the original informal reference configuration.

4.1. Ordinal Association Between Density and Metric Scores

To examine whether scalar evaluation metrics are sensitive to register-marked variation, we first assess the ordinal association between indexical density (0–4) and sentence-level metric scores. If scalar metrics do not explicitly separate register from other dimensions of translation quality, density and metric scores would be expected to show weak or unstable association. Conversely, consistent correlations would indicate that register-marked variation co-varies with the scalar evaluation signal.

Using the original informal (INSTA) reference, indexical density exhibits a consistent negative association with sentence-level scores for all reference-based metrics across the four output conditions (two systems × two prompting conditions) (see Figure 1). Spearman correlations were computed separately for each system–prompt condition and then averaged. Across conditions, mean correlations are $\rho = -0.36$ for chrF++, $\rho = -0.38$ for COMET-DA, and $\rho = -0.36$ for XCOMET-XL. In each condition, correlations are statistically significant ($p < 10^{-8}$). By contrast, the reference-free quality estimation metric CometKiwi-XL shows no meaningful association with density (mean $\rho \approx 0.00$; all $p > .75$) (Appendix B).

These results suggest that, when evaluation is anchored to the informal reference, higher concentrations of register-marked features are associated with lower scores under overlap-based and reference-based neural metrics. Importantly, indexical density is not itself a quality label: higher density reflects a greater accumulation of dialectal forms, informal lexicon, stylization, or code-switching, rather than an a priori reduction in semantic adequacy. The absence of a comparable association in the reference-free QE metric is consistent with the interpretation that the density–score relationship is linked to reference anchoring rather than reflecting a uniform neural adequacy signal.

4.2. Binary Contrast: Low vs. High Density

To complement the ordinal analysis, we compare low-density items (0–1 features) with high-density items (2+ features). Across systems and prompting conditions, low-density items receive higher scores than high-density items under chrF++, COMET-DA, and XCOMET-XL (Mann–Whitney $p < .001$; moderate effect sizes). CometKiwi-XL shows no significant differences between density groups. Summary statistics are reported in Table 1. The binary contrast corroborates the ordinal analysis: reference-based metrics assign lower scores to high-density messages, whereas the reference-free QE metric shows no comparable separation.

4.3. Interpreting the Density Effect

An alternative explanation is that higher-density messages may be intrinsically more challenging to translate, leading to lower-quality outputs and consequently lower metric scores. The present study does not include independent human adequacy annotation and therefore does not exclude the possibility that translation difficulty contributes to the observed pattern.

However, two observations suggest that evaluation configuration may also play a role. First, the reference-free QE metric does not exhibit a comparable association between density and score. If high-density items systematically resulted in lower-quality translations, a similar trend might be expected in the QE scores. Second, as shown in Section 5, modifying only the reference register—without altering the MT output—systematically shifts metric scores upward, with larger shifts for high-density items. This pattern indicates that evaluation outcomes are sensitive to reference configuration in ways that interact with register marking.

Taken together, these findings suggest that the negative association between density and reference-based metric scores may reflect an interaction between register-marked variation and scalar evaluation design. While translation difficulty may contribute, the results are consistent with the view that evaluation signals anchored to a fixed reference are not fully register-neutral.

4.4. System-Level Density Effects

To assess whether the density penalty observed is system-specific, we compute density–score correlations separately for Latxa and GPT outputs (pooling prompting conditions) under the INSTA refer-

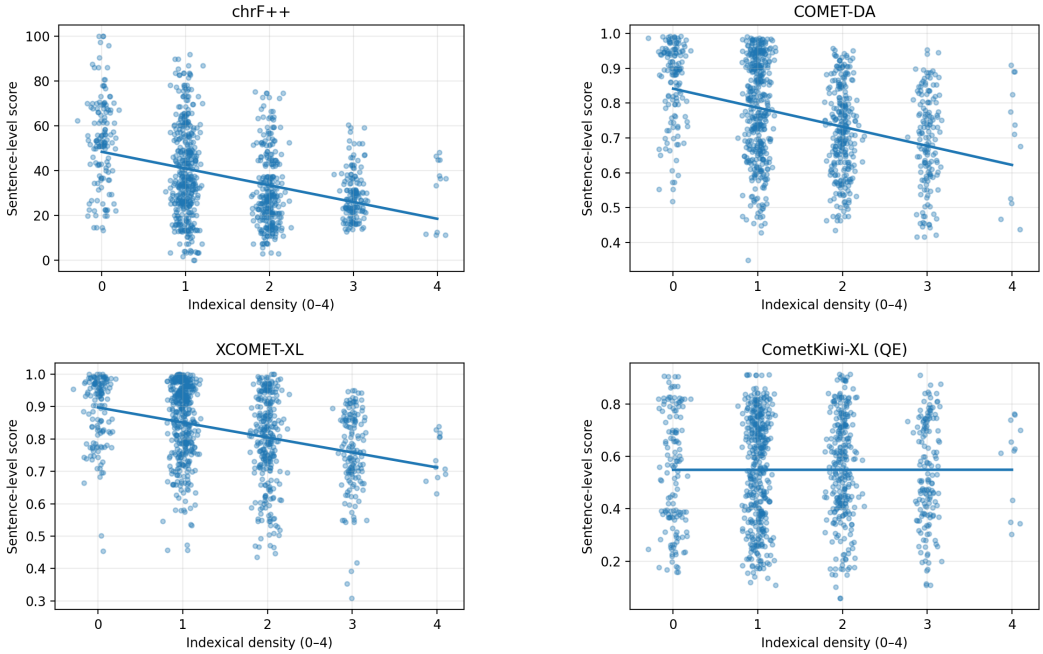

Figure 1: Sentence-level metric scores as a function of indexical density (0–4). Each point represents one DM under one output condition (2 systems × 2 prompts; $N = 1000$ per panel). Points are jittered for readability; fitted lines indicate monotonic trends. Results are aggregated across MT systems and prompting conditions; trends were consistent across conditions.

Metric	**Mean ρ**	**p-range**	**Low vs. High (dir.)**	**Mean r_{rb} / p-range**
chrF++	-0.362	$< 10^{-8}$	Low $>$ High	0.401; $p < .001$
COMET-DA	-0.378	$< 10^{-8}$	Low $>$ High	0.417; $p < .001$
XCOMET-XL	-0.358	$< 10^{-8}$	Low $>$ High	0.389; $p < .001$
CometKiwi-XL	-0.004	.75–.97	(n.s.)	(n.s.)

Table 1: Spearman correlations relate indexical density (0–4) to sentence-level metric scores, averaged across output conditions. Low vs. high density contrasts compare density 0–1 vs. 2+ using Mann–Whitney tests; effect sizes are rank-biserial correlations (r_{rb}).

ence. Across metrics, both systems exhibit comparable negative associations between indexical density and metric scores: for chrF++, $\rho = -0.359$ (Latxa) and $\rho = -0.354$ (GPT); for COMET-DA, $\rho = -0.394$ and $\rho = -0.357$; and for XCOMET-XL, $\rho = -0.352$ and $\rho = -0.374$, respectively. Bootstrap confidence intervals for the difference in correlations ($\Delta\rho$) include zero for all metrics, indicating no strong evidence of systematic system-level divergence. These results suggest that the density-related score reduction under reference-based evaluation is not driven by a single model, but appears comparably across systems.

Partial correlations controlling for token length are reported in Appendix C, complete correlation estimates and confidence intervals in Appendix D, and condition-wise correlations in Appendix B.

5. Reference Register Sensitivity

We next examine whether scalar evaluation outcomes shift when the register configuration of the reference is altered while holding MT outputs constant.

5.1. Setup and Computation of Reference Deltas

To assess the extent to which scalar evaluation depends on reference register configuration, we evaluate each fixed MT output against two Basque references: the original informal message (INSTA) and a standardized Batua version (BATUA). This design holds the MT output constant and perturbs only the reference, allowing us to isolate how reference choice shapes evaluation outcomes in register-rich informal discourse.

For each DM and output condition (2 systems × 2 prompts), we compute a per-item reference

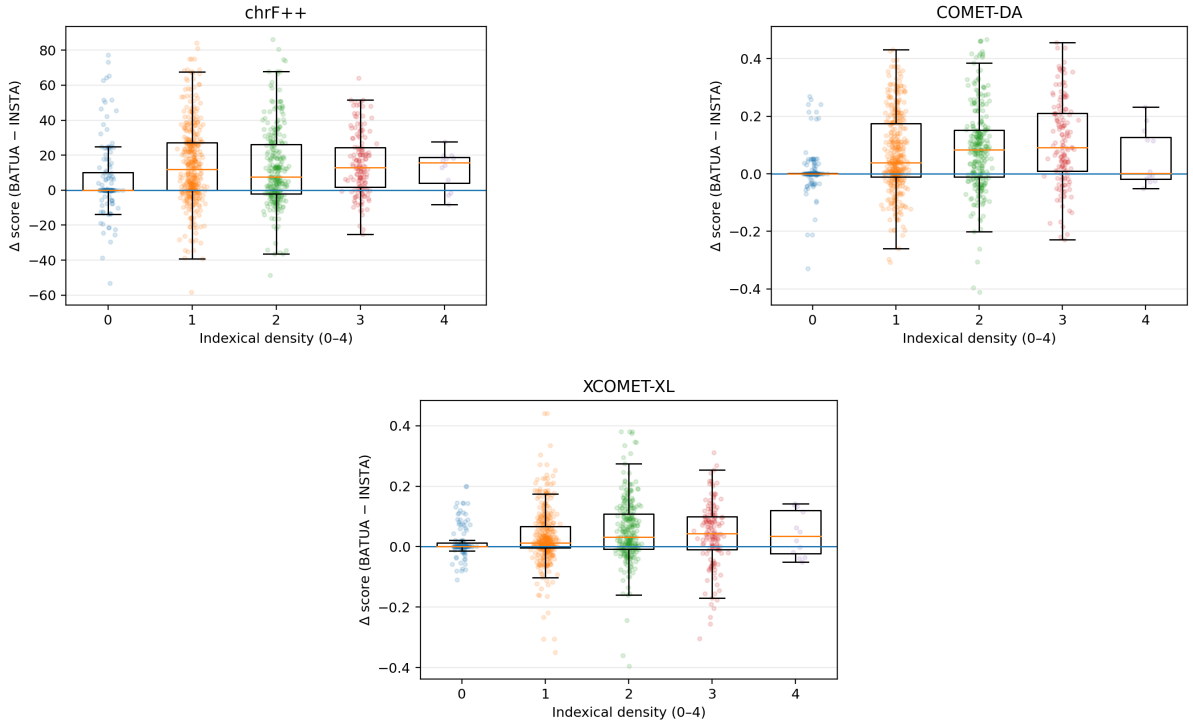

Figure 2: Reference register sensitivity as a function of indexical density. Each panel shows $\Delta = score_{\text{BATUA}} - score_{\text{INSTA}}$; positive values indicate higher scores under the standardized reference. Boxplots summarize Δ across density levels (0–4) with jittered observations. Because metrics differ in scale, magnitudes are not directly comparable; interpretation focuses on direction and density-dependent trends.

delta for each reference-based metric, as defined in Section 3.4. This measure captures the score difference between the standardized (BATUA) and informal (INSTA) reference configurations. Positive values indicate higher scores under BATUA.

5.2. Overall Effect of Reference Register

Across systems and prompting conditions (1,000 paired observations per metric), all reference-based metrics exhibit positive median Δ values, indicating higher scores under BATUA. Median Δ values are 8.30 for chrF++, 0.043 for COMET-DA, and 0.016 for XCOMET-XL. In each case, approximately two-thirds of items receive higher scores under the standardized reference (63–67%). Paired Wilcoxon signed-rank tests strongly reject the null hypothesis of no difference (all $p < 10^{-38}$).

These results show that evaluation outcomes systematically shift when the reference is normalized to standard Batua, even though the MT output remains unchanged.

5.3. Density Dependence of Reference Sensitivity

We next check whether the magnitude of reference sensitivity varies with indexical density. Spearman correlations between density and Δ are positive for all three metrics: $\rho = 0.097$ for chrF++, $\rho = 0.207$ for COMET-DA, and $\rho = 0.167$ for XCOMET-XL (all $p < .01$), indicating that the score increase associated with standardization is larger for messages with higher concentrations of register-marked features. Figure 2 visualizes this pattern, showing upward shifts in the distribution of per-item score differences (Δ = BATUA − INSTA) as density increases.

To assess whether the overall density–score association depends on reference configuration, we compare density–score correlations under INSTA and BATUA. Across metrics, the negative association weakens under standardization; for example, for COMET-DA, $\rho = -0.376$ (INSTA) and $\rho = -0.274$ (BATUA), yielding $\Delta\rho = 0.102$ (95% CI $[0.049, 0.156]$). Comparable attenuation is observed for chrF++ and XCOMET-XL, indicating that the strength of the density penalty is partially contingent on reference register configuration. Full correlation values and bootstrap confidence intervals are provided in AppendixE, with condition-wise breakdowns in Appendix F.

5.4. Interpreting Reference Sensitivity

Because MT outputs are identical across reference conditions, the observed score shifts cannot be attributed to differences in translation behavior. In-

stead, they suggest sensitivity of scalar evaluation metrics to the register configuration of references.

While we do not claim that standardization necessarily distorts quality assessment, the results demonstrate that evaluation signals are not invariant to reference register choice. Moreover, the positive association between density and Δ indicates that this sensitivity is structured rather than uniform: DMs containing register-marked features experience larger score increases under reference normalization. Taken together, these findings are consistent with the view that reference-anchored scalar evaluation interacts systematically with sociolinguistic variation in under-resourced settings.

6. Conclusions

This paper examined how scalar MT evaluation behaves in the presence of register-marked variation in an under-resourced informal Basque setting. Building on the argument in Section 2 that scalar metrics compress multiple dimensions of translation quality into a single adequacy-oriented signal, we operationalized register as a structured dimension of sociolinguistic variation and assessed whether it remains independent under reference-anchored evaluation.

Across systems and prompting conditions, we found a consistent negative association between indexical density and reference-based metric scores. This pattern persisted under length controls and was broadly comparable across systems. When we manipulated only the reference register—holding MT outputs constant—scores increased under standardization, with larger gains for high-density items. The density–score association also weakened under reference normalization, suggesting that part of the apparent density penalty may be related to reference configuration rather than translation behavior alone.

Taken together, these findings suggest that scalar evaluation may not fully separate sociolinguistic alignment from adequacy. When multiple quality dimensions are compressed into a single scalar output, register-sensitive variation may be incorporated into adequacy signals rather than represented independently.

In under-resourced and minority-language contexts—where informal varieties are central and language normalization processes are ongoing—such dynamics may influence how translation quality is interpreted. Our results do not imply that current metrics are flawed; rather, they indicate that evaluation design choices, particularly reference selection, can shape how sociolinguistic variation enters scoring. Future work could explore evaluation frameworks that more explicitly distinguish adequacy from register-sensitive variation and examine how reference construction shapes metric behavior across diverse language settings. As evaluation practices increasingly incorporate LLM-based judge paradigms, understanding how generative evaluators handle register-sensitive variation remains an open direction for research.

7. Ethical Considerations and Limitations

Ethical considerations. This study uses data from the GazteSare corpus, specifically, Instagram direct messages (Elordui et al., 2020), which was collected under prior ethical approval as documented in the original corpus publication. Although the corpus is not publicly downloadable, it is shared with researchers for scholarly use. We do not introduce additional user data beyond this resource. DMs are analyzed in their original form (with emojis removed for automated evaluation), and no identifying metadata are reported. Our analysis focuses on evaluation behavior rather than user profiling or system deployment.

We do not treat dialectal forms, informal lexicon, stylization, or code-switching as deviations from a normative standard. Indexical density is operationalized as a descriptive measure of register marking, not as an indicator of linguistic quality. The Batua references were created specifically for this study by a trained bilingual linguist to enable controlled comparison of reference configurations. Their use is methodological rather than prescriptive. Given the sociolinguistic and historical significance of language normalization in Basque, the introduction of standardized references should not be interpreted as privileging Batua over informal or non-standard varieties.

Limitations. Several considerations constrain the interpretation of our findings.

First, we do not include independent human adequacy judgments. Although the reference-perturbation design isolates evaluation configuration from output variation, we cannot fully disentangle translation difficulty from register density without complementary human assessment. The results therefore concern evaluation sensitivity under controlled conditions rather than definitive statements about translation quality.

Second, indexical density is a simplified proxy for register. It aggregates heterogeneous feature types (dialectal morphology, informal lexicon, stylization, and code-switching) into a type-based measure annotated by an expert. The measure should therefore be interpreted as an operational approximation rather than a comprehensive sociolinguistic account.

Third, the BATUA references were manually con-

structed for this study to enable controlled comparison. Although semantic equivalence was reviewed, normalization may introduce subtle shifts beyond register, and reference construction necessarily reflects analytic decisions.

Fourth, the empirical setting is limited to one language pair (Spanish–Basque) and an informal digital discourse domain. While theoretically relevant given the under-resourced context and active normalization processes, the extent to which the findings generalize to other languages or domains remains open.

Finally, we examine a subset of widely used scalar MT metrics. Neural metrics are primarily trained on high-resource evaluation data, and alternative paradigms—including multi-dimensional human evaluation or LLM-as-judge approaches—may behave differently with respect to register variation. The present results therefore characterize scalar reference-anchored evaluation in this setting, rather than metric behavior across evaluation frameworks more broadly.

8. Acknowledgements

This work was partially supported by the MOLVI project (PID2024-157855OB-C32), funded by MICIU/AEI/10.13039/501100011033 and FEDER, EU, and by the project Desarrollo de Modelos ALIA, Resol. SEDIA 19.08.2024, within the framework of the National Language Technologies Plan (ENIA 2024), funded by MTDFP, PRTR, and the European Union–NextGenerationEU. The author thanks the volunteer participants for their contributions to the data creation and crowd-based evaluation.

9. Bibliographical References

Jannis Androutsopoulos. 2014. Mediatization and sociolinguistic change. In Nikolas Coupland, editor, *Sociolinguistics: Theoretical Debates*, pages 74–101. Cambridge University Press, Cambridge.

Nora Aranberri. 2026. Crowd-based evaluation of emotion intensity preservation in spanish–basque tweet machine translation. In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 123–133.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

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.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1119–1130.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Eleftheria Briakou, Shuyan Lu, Lesly Miculicich, and Graham Neubig. 2021. Searching for controllable machine translation in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6547–6564, Online. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Agurtzane Elordui, Jokin Aiestaran, Garbiñe Bereziartua, Irantzu Epelde, Ibarra Orreaga, Oroitz Jauregi, Libe Mimenza, Beñat Muguruza, and Ane Odria. 2020. Gaztesare corpusa eta datu-basea. *https://gaztesare.eus*.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of ACL Workshop on Noisy User-generated Text*, pages 9–18.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

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.

Divya Kumari, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021. Sentiment preservation in review ttranslation using curriculum-based re-inforcement framework. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 150–162.

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Samuel Messick, Robert L Linn, et al. 1989. Educational measurement. *Validity*, pages 13–103.

James Milroy and Lesley Milroy. 1999. *Authority in Language: Investigating Standard English*. Routledge.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084.

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

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 35–40.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & communication*, 23(3-4):193–229.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4340–4354.

A. Full Prompt Texts

The following prompts were used for both systems under the baseline and context-rich conditions. Prompts are reproduced verbatim in Basque, followed by an English gloss for transparency.

A.1. Baseline Prompt

Basque (original):
Itzuli hurrengo testua euskarara.

English gloss:
Translate the following text into Basque.

A.2. Context-Rich Prompt

Basque (original):
Itzuli hurrengo Instagrameko mezu zuzenak euskarara. Itzulpenean erabili gazteen arteko hizkera informala, herri hizkera, ahoskeran oinarritutako idazkera eta hizkuntza-alternantzia naturala, testuinguruari egokituta.

English gloss:
Translate the following Instagram direct messages into Basque. In the translation, use informal youth language, vernacular speech, pronunciation-based spelling, and natural code-switching, adapted to the context.

B. Condition-Wise Density Effects

To provide a full condition-wise breakdown, the following table reports density–score correlations for each system–prompt combination under the informal reference.

Metric	Latxa Base	Latxa Rich	GPT Base	GPT Rich
chrF++	-0.364	-0.353	-0.349	-0.358
COMET-DA	-0.401	-0.387	-0.352	-0.362
XCOMET-XL	-0.344	-0.360	-0.372	-0.355

Table 2: Condition-wise Spearman correlations between indexical density and metric scores under the INSTA reference (two systems $\times$ two prompting conditions). All correlations are statistically significant ($p < 10^{-8}$). The negative density–score association is consistent across systems and prompting conditions.

C. Robustness to DM Length

To assess robustness to DM length, the following table reports density–score correlations under the informal (INSTA) reference together with partial rank correlations controlling for token length.

Metric	ρ(density, score)	Partial ρ (control: length)
chrF++	-0.36***	-0.29***
COMET-DA	-0.38***	-0.31***
XCOMET-XL	-0.36***	-0.28***
CometKiwi-XL	-0.00	0.02

Table 3: Spearman correlations between indexical density and sentence-level metric scores (original informal reference), with and without controlling for DM length. Asterisks indicate significance levels: *** $p < .001$. Values represent averages across system and prompting conditions.

D. System-Level Density Effects

To evaluate system-level differences in density effects, the following table reports correlations computed separately for Latxa and GPT outputs.

Metric	ρ (Latxa)	ρ (GPT)	$\Delta\rho$ (GPT$-$Latxa)	95% CI
chrF++	-0.359	-0.354	0.006	$[-0.028,\ 0.039]$
COMET-DA	-0.394	-0.357	0.037	$[-0.018,\ 0.091]$
XCOMET-XL	-0.352	-0.374	-0.022	$[-0.078,\ 0.038]$

Table 4: System-level density effects under the INSTA reference. Spearman correlations between indexical density and metric scores are computed separately for Latxa and GPT outputs (pooling prompting conditions). $\Delta\rho$ denotes the difference in correlation strength between systems. Bootstrap 95% confidence intervals include zero for all metrics, indicating no strong evidence of systematic system-level divergence in density penalties.

E. Reference Configuration Diagnostics

To examine reference-configuration effects in detail, the following table compares density–score correlations under informal (INSTA) and standardized (BATUA) references.

Metric	ρ (INSTA)	ρ (BATUA)	$\Delta\rho$ (BATUA$-$INSTA)	95% CI
chrF++	-0.356	-0.243	0.113	$[0.059,\ 0.173]$
COMET-DA	-0.376	-0.274	0.102	$[0.049,\ 0.156]$
XCOMET-XL	-0.363	-0.310	0.053	$[0.013,\ 0.091]$

Table 5: Density–score correlations under informal (INSTA) and standardized (BATUA) references. $\Delta\rho$ denotes the difference in Spearman correlation between reference conditions. Positive $\Delta\rho$ indicates attenuation of the negative density–score association under reference standardization. Confidence intervals are bootstrap 95% intervals over paired observations (all exclude 0).

F. Condition-Wise Reference Sensitivity

To detail reference sensitivity across output conditions, the following table summarizes condition-wise reference deltas and their association with density.

Condition	Metric	Median Δ	% $\Delta > 0$	Wilcoxon p	ρ(density,Δ)	p
Latxa Base	chrF++	11.800	70.0	1.23×10^{-22}	0.119	$.060$
Latxa Base	COMET-DA	0.053	70.4	9.53×10^{-19}	0.257	$< .001$
Latxa Base	XCOMET-XL	0.018	67.6	1.73×10^{-13}	0.189	$.003$
Latxa Rich	chrF++	9.550	70.0	8.42×10^{-22}	0.134	$.034$
Latxa Rich	COMET-DA	0.050	68.0	1.53×10^{-16}	0.218	$< .001$
Latxa Rich	XCOMET-XL	0.015	63.6	1.62×10^{-11}	0.142	$.025$
GPT Base	chrF++	6.500	64.4	3.83×10^{-15}	0.150	$.018$
GPT Base	COMET-DA	0.042	61.6	1.21×10^{-12}	0.213	$< .001$
GPT Base	XCOMET-XL	0.012	60.8	1.49×10^{-10}	0.167	$.008$
GPT Rich	chrF++	9.150	65.2	3.37×10^{-12}	0.004	$.952$
GPT Rich	COMET-DA	0.029	62.0	2.82×10^{-9}	0.144	$.023$
GPT Rich	XCOMET-XL	0.014	61.6	1.82×10^{-7}	0.165	$.009$

Table 6: Condition-wise reference sensitivity (BATUA INSTA). Median Δ values indicate systematic score increases under the standardized reference across systems and prompting conditions. Positive correlations between density and Δ indicate larger reference-induced gains for high-density items.