Beyond N-Grams: Rethinking Evaluation Metrics and Strategies for Multilingual Abstractive Summarization

Anonymous ACL submission

Abstract

Automatic N-gram based metrics such as ROUGE are widely used for evaluating generative tasks such as summarization. While these metrics are considered indicative (even if imperfect), of human evaluation for English, their suitability for other languages remains unclear. To address this, in this paper we systematically assess evaluation metrics for generation — both n-gram-based and neural-based — to assess their effectiveness across languages and tasks. Specifically, we design a large-scale evaluation suite across eight languages from four typological families — agglutinative, isolating, low-fusional, and high-fusional — from both low- and high-resource languages, to analyze their correlations with human judgments. Our findings highlight the sensitivity of the evaluation metric to the language type at hand. For example, for fusional languages, n-gram-based metrics demonstrate a lower correlation with human assessments, compared to isolating and agglutinative languages. We also demonstrate that tokenization considerations can significantly mitigate this for fusional languages with rich morphology, up to reversing such negative correlations. Additionally, we show that neural-based metrics specifically trained for evaluation, such as COMET, consistently outperform other neural metrics and correlate better than ngrmas metrics with human judgments in low-resource languages. Overall, our analysis highlights the limitations of n-gram metrics for fusional languages and advocates for investment in neural-based metrics trained for evaluation tasks.[1]

1 Introduction

The development of multilingual LLMs (MLLMs) such as BLOOM (Le Scao et al., 2023) and XGLM (Lin et al., 2021), along with the current trend of extending English-centric LLMS (e.g. LLaMA3 (Dubey et al., 2024), OpenAI GPT-4o (Hurst et al., 2024) and Gemini 1.5 (Team et al., 2024)) to other languages (Alexandrov et al., 2024) reflects the growing interest in prompting such generative models in languages beyond English. This interest highlights the need for robust evaluation of the generation capabilities of LLMs in multilingual settings. However, assessing these models on non-English generative tasks, particularly in summarization, remains challenging due to the lack of clear evaluation methodologies.

Current evaluation metrics for summarization, both n-gram-based or neural-based, face significant limitations. N-gram-based evaluation metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2004), are commonly used to assess summarization quality in English, however, these metrics rely on complete word units. This creates challenges for fusional languages with flexible word order where inflectional patterns are embedded within word forms. Moreover, they present difficulties for agglutinative languages, where words have complex internal structures, consisting of multiple morphemes that n-gram-based metrics struggle to capture effectively (Abudouwaili et al., 2023). Additionally, the problem of ambiguity — where a single form can have multiple meanings — is amplified in morphologically rich languages (MRLs) as variations in prefixes, suffixes, and root conjugations complicate both comprehension and generation tasks. These factors can lead to n-gram-based metrics failing to recognize grammatically correct sentences in generated summaries that convey the intended meaning despite slight surface-level differences.

Neural network-based approaches, such as BERTScore (Zhang et al., 2019), depend on the availability of large models trained on large amounts of data and may exhibit poor performance for lower resourced languages (Yousuf et al., 2024; Kaster et al., 2021). Languages with greater mor-

[1] We will openly publish the human annotation data and complete evaluation suite to support further research and exploration of multilingual automatic evaluation of generation.

phological complexity are particularly challenging, as MRLs often produce a large number of infrequent word forms produced by combinations of morphemes, resulting in data sparsity (Botev et al., 2022). The tokenization problem is also demonstrated by Gerz et al. (2018) who shows that language models that use character-level information show superior performance to those operating on word level alone in next-word prediction task for morphologically rich languages.

Despite such bits of empirical evidence, while summarization metrics have been extensively studied in English their applicability to other languages remains understudied. More concretely, existing campaigns for assessing evaluation metrics for generation face three key limitations: (i) *lack of language diversity*, resulting in insufficient typological representation—for instance, Koto et al. (2021) excluded languages with high-fusional morphology, and Forde et al. (2024) evaluated only three languages, highlighting scalability concerns; (ii) *lack of metrics diversity*, primarily focusing on n-gram-based approaches and excluding neural-based ones, particularly those specifically trained for evaluation, and insufficient evaluation of metric adaptation for non-English; and (iii) *lack of reliable statistical evidence* for the correlation between automatic metrics and human judgments, omitting statistical significance values of the correlation analysis. (Koto et al., 2021; Han et al., 2024)

To address these gaps, we deliver a large resource for summarization in non-English languages, manually annotated with human judgments, comprising ~20,000 human annotations. This resource upshots are first, the ***selection of representative languages***, covering eight languages from four typological types (isolating, agglutinative, and languages with minimal or high fusional morphology). Within each group, we represent both high- and low-resource languages. Secondly we assess ***diverse Metrics***, both n-gram and neural metrics, including those particularly trained for evaluation. Additionally, we evaluate the different methodologies to assess the quality of generation, for example, the use of different tokenizers and various transformed versions of the original text, including lemmatized forms, to assess their impact on the evaluation metrics. Finally, our analysis takes care to provide ***statistically sufficient data size***. Our multilingual annotation task measures correlation with both n-gram and neural network metrics while reporting the statistical significance of the factors found to affect the results.

Our study demonstrates that evaluation metrics perform differently depending on linguistic typology. For instance, in fusional languages, n-gram metrics like ROUGE align less reliably with human assessments than in isolating or agglutinative languages. Conversely, neural-based metrics like COMET — trained explicitly for assessment generative tasks, achieve stronger correlations with human judgments and consistently surpass both n-gram methods and neural approaches. These findings highlight the limitations of n-gram metrics for fusional languages and emphasize the need for specialized neural metrics trained for multilingual evaluation.

2 Limitations of Current Generation Evaluation in Diverse Languages

2.1 The Limitations and Shortcomings of Current Generation Evaluation

The rise of generative models and their massive prompting to generate online high-quality responses has underscored the importance of properly evaluating these models with automatic metrics (Manduchi et al., 2024) that allow effective and efficient hill-climbing in the course of model development and assessment. Since the introduction of ROUGE (Lin, 2004), N-gram-based metrics have been commonly used in the NLP community for English as well as for multilingual purposes. However, these metrics face severe issues for languages that differ from English, specifically with tokenization and segmentation matters.

For example, metrics such as BLEU face challenges in languages like Chinese and Japanese due to the lack of explicit word boundaries (Denoual and Lepage, 2005), and implementations of metrics like ROUGE, often struggle with segmentation issues, including filtering out non-alphanumeric Latin characters, making them less effective for non-Latin scripts (Kumar and Solanki, 2023). As a result, these limitations lead to poor correlations with human judgments, especially for high fusional languages. For instance, Bouamor et al. (2014) observed weak correlations for BLEU and METEOR in Arabic, while Paz-Argaman et al. (2024) found negative correlations for ROUGE in Hebrew.

To address the limitations of n-gram-based metrics, researchers proposed to utilize neural-based metrics, which fall into three categories: *encoder-*

based models like BERTScore (Zhang et al., 2019), which compare text representations; *LLM as a judge* such as the prompting of Gemini (Team et al., 2023) to assess quality, without any task-specific training; and *neural methods specifically trained for evaluating generation* such as COMET (Rei et al., 2020), fine-tuned to predict quality scores for machine translation (MT) task. These metrics, while remaining data-driven and agnostic to the language type at hand, are prone to suffer from resource level effects with varying qualities that depend on the model exposure to such data. All in all, both the n-gram based metrics and neural based metrics (including those specifically trained for evaluation) have not been systematically evaluated for non-English.

2.2 Generation Evaluation in the Face of Language Diversity

Despite the aforementioned shortcomings, the effectiveness of n-gram-based as well as neural based metrics for evaluation of generation has not been systematically studied across language families with varying word complexity and boundary characteristics. This gap raises concerns, as the linguistic properties of words may well affect the usability of n-gram metrics, but the effects remain unclear.

In terms of their linguistic properties, language families can be placed on a scale. On the one hand, there are **Isolating Languages**, in which words typically consist of a single morpheme, e.g., Yoruba and Chinese (Okanlawon, 2016; Arcodia et al., 2007). On the other hand, words in **Fusional Languages** contain multiple morphemes fused together, often with unclear boundaries, where a single space-delimited token may serve multiple functions. For example, in the Spanish word *habló*, the suffix `ó` simultaneously indicates past tense and third-person singular (Kambarami et al., 2021). This category can be further divided into **low-fusional** (e.g. Spanish (Bergmann et al., 2007) and Ukrainian (Budzhak-Jones, 1998)) and **high-fusional** (e.g. Arabic (Smrž, 2007) and Hebrew (Tsarfaty et al., 2019)) based on the degree of morphological fusion. Additionally, in an orthogonal dimension we can recognize **Agglutinative Languages** that also consist of words made up of multiple morphemes, albeit with clear boundaries and distinct functions. For instance, in Shona, *vakaenda* (`va-ka-end-a`) means "they went" where `va` (plural subject), `ka` (remote past), and `a` (final vowel) modify the root `end` ("to go") (Kambarami et al., 2021). Examples include Turkish and Japanese (Istek and Cicekli, 2007; Shibatani and Kageyama, 2015). To our knowledge, no non-English evaluation has comprehensively covered languages from all these typological groups.

Two primary strategies have been suggested to adapt previously used metrics to different types of languages, for instance *data transformation*, the adaptation of n-gram metrics where a different tokenizer or lemmatizer is applied to the data prior to using the n-gram-based metrics. Specifically, converting Chinese text into numerical IDs before applying ROUGE (Wang et al., 2021), or using ROUGE with language-specific tokenizers as Alhamadani et al. (2022) did for Arabic. Alternatively researchers suggested the use of *language-specific encoder*, encoders trained on the target language for similarity-based evaluation. For example, using BERTScore with language-specific BERT models (Vetrov and Gorn, 2022). However, all these approaches have not been systematically evaluated across languages.

In addition to the lack of languages and metrics, correlations between multilingual automatic metrics and human judgments lack sufficient evidence to be considered reliable due to the absence of reported p-values (Koto et al., 2021; Forde et al., 2024; Han et al., 2024). In reproduced experiments (Ernst et al., 2023), the statistical significance was being low to substantiate the findings. Additionally, power analysis indicates that ~400 samples per language are needed to detect significant effects at $p \leq 0.05$.[2] However, existing non-English evaluations fall short of this threshold, with Koto et al. (2021) using only 150 samples and Han et al. (2024) evaluating just 90 summaries per language.

3 Our Approach: Systematic Evaluation of Summarization Across Languages

In this work we set out to systematically evaluate automatic metrics, assessing their effectiveness and reliability for non-English languages by assessing the correlation achieved compared to human scores. We do so via a comprehensive and controlled protocol, comprising ~20,000 human annotations while assessing the various research dimensions and previously attested weaknesses.

Concretely, in this work we evaluate eight languages from four typological families, covering

[2] See Appendix A.2 for more details on the t-test.

both low resource (L) and high resource (H) language in each group, including: *Isolating* (Chinese, *zh* (H); Yoruba, *yo* (L)), *Agglutinative* (Japanese, *ja* (H); Turkish, *tr* (L)), *Low Fusional* (Spanish, *es* (H) and Ukrainian, *ukr* (L)) and *High Fusional* (Arabic, *ar* (H); Hebrew, *he* (L)). Following Lai et al. (2023) method to classify languages using a threshold, we classified languages by token percentage (p) based on GPT-3's pre-trained data distribution, relying on its broad multilingual coverage and reported data proportions.[3] Specifically, we classified into low- ($p < 0.1\%$) and high-resource ($p \geq 0.1\%$) languages.[4] For language selection within each typological family, we followed Gerz et al. (2018) (see Section 2.2 for additional justifications).

For each language-metric combination we perform a correlation analysis with both general purpose metrics, as well as metrics tailored for multilingual settings, e.g. BERTScore applied with mBERT or with language-specific BERT. Also, we have utilized COMET (Rei et al., 2020) which was specifically trained for evaluation and ROUGE score with language-specific tokenizers.[5] Finally, to substantiate our results, we included at least 400 samples per language and reported p-values for each evaluated dimension. For all experiments, we report inter-annotator agreement to enhance the credibility of our annotations.

4 Data Collection

To systematically assess the correlation between evaluation metrics and human rankings for abstractive summarization, we engage human annotators to evaluate summaries generated by large language models (LLMs). Our data collection evaluates document summaries in eight languages, chosen to represent four typological families with both low- and high-resource languages within each group. We used the XL-Sum dataset (Hasan et al., 2021), which provides news articles along with their human-generated summaries in various languages. For Hebrew, we used HeSum (Paz-Argaman et al., 2024). See Table 1 for categorization details.

Resource/Type	Isolating	Agglutinative	High Fusion	Low Fusion
High Resource	Simplified Chinese (*zh*)	Japanese (*jp*)	Arabic (*ar*)	Spanish (*es*)
Low Resource	Yoruba (*yor*)	Turkish (*tr*)	Hebrew (*he*)	Ukraine (*ukr*)

Table 1: Categorization of languages based on morphological typology and resource availability. ISO 639-1 language codes are provided in parentheses.

4.1 The Annotation Task: Summary Ranking

The task involves annotating two peer summaries, produced by GPT-3.5-Turbo (0125) (Ouyang et al., 2022) and Gemini-1.0-pro (Team et al., 2023), by comparing their content to the source article. The evaluation procedure is as follows: *(i)* The annotator reads the source content and the two peer summaries. *(ii)* The annotator answers a question on the article to prove language comprehension. *(iii)* The annotator evaluates each summary using 1-4 Likert scale (Likert, 1932) based on two quality criteria (QC): *coherence*, which assess the summaries' grammaticality and readability, and *completeness*, which measures the degree to which they capture the main ideas of the articles. The evaluation page was set up to include the full source article, instructions, definitions of the quality criteria, and two generated summaries. For each summary and criterion, there is a scale with four rating options. Appendix A.3 presents the UI interface we designed and built for the assignment as displayed to the annotators in Arabic and Spanish. Appendix A.4 gives more details on about the collection protocol.

4.2 Ensuring High Annotation Consistency

To ensure annotation reliability, we hired annotators through Amazon Mechanical Turk (MTurk) (100+ approved HITs, 90%+ approval rate) with geographic constraints aligned to the target languages. For some languages, we were unable to recruit native speakers in their country of birth due to various restrictions and sourcing difficulties; in such cases, we hired native speakers residing in other countries.[6] Additionally, we recruited qualified students who passed a matching questionnaire. In total, we recruited 36 raters across 13 locales.[7] To improve annotation quality, each model-generated summary was ranked by three different participants. For correlation analysis, we used the average score.

To verify understanding of the source content we created a Gemini-generated qualification question based on the article to filter annotations from mis-

[3] https://github.com/openai/gpt-3/blob/master/dataset_statistics

[4] Arabic, with less than 0.1% of tokens, was chosen as a high-resource language due its worker availability and higher pre-trained representation than Hebrew. See Appendix A.1 for the full language proportions.

[5] See Appendix B.1 for all models and tokenizers we used.

[6] In these cases, we used the qualification question to assess the participant's language skills.

[7] See Table 3 for participants' demographics.

Family	Language (L/H)	Novel n-grams				Redundancy		Compression	Mean Token Length
		1-gram	2-gram	3-gram	4-gram	n=1	n=2		
Isolating	ZH (H)	27.52	67.23	83.82	91.29	14.86	2.34	83.71	53.56
	YOR (L)	38.90	60.85	69.38	73.84	32.85	8.03	62.17	105.29
Agglutinative	JP (H)	24.29	54.12	69.62	78.23	49.08	15.93	79.22	188.37
	TR (L)	41.76	71.44	84.56	90.76	18.41	2.37	72.71	69.95
Low Fusional	ES (H)	28.00	63.15	81.16	89.11	26.28	2.83	81.94	83.17
	UKR (L)	42.01	73.49	86.72	92.39	18.53	2.21	74.85	66.22
High Fusional	AR (H)	47.73	78.72	89.75	94.59	15.05	1.62	77.36	62.32
	HE (L)	45.06	75.14	86.75	92.01	20.83	3.49	84.28	80.85

Table 2: Model-Generated Summaries Intrinsic Evaluation per language.

Country of Residence	**Total Workers**	**Percentage (%)**
United States	5	13.9
Nigeria	2	5.6
West Africa	2	5.6
Turkey	3	8.3
Egypt	1	2.8
Jordan	1	2.8
Libya	2	5.6
Ukraine	5	13.9
Israel	5	13.9
Spain	4	11.1
Mexico	1	2.8
Argentina	2	5.6
Venezuela	2	5.6
Japan	1	2.8
Total	36	100.0

Table 3: *Distribution of Workers by Country of Birth.*

understood articles.[8] To measure the consistency of the annotators' scores, we calculated Krippendorff's α (Krippendorff, 2011) for an interval scale per language.

Moreover, to achieve a diverse distribution of scores, we artificially corrupted one-third of the data by randomly degrading quality criteria.[9] For coherence, we replaced nouns and verbs with their lemma forms, creating ungrammatical sentences. Additionally, we reordered non-adjacent sentences to disrupt the flow. For completeness, we replaced named entities in the summary with others from the original text and inserted a random, unrelated sentence.[10]

5 Correlation Analysis

Having collected the data, we use it to calculate the Pearson correlation (Cohen et al., 2009) between human evaluation and automatic metric scores. In this Section, we first analyze the collected data (Section 5.1), we then display the assessed evaluation metrics we used (Section 5.2), and finally we present the analysis of metric correlation with human scores for different languages (Section 5.3).

[8]See Appendix A.6 for details on the qualification task.

[9]This approach was adopted following a previous data collection attempt without corruption, which revealed that the scores were clustered and displayed low dispersion.

[10]See Appendix A.5 for more corruption details.

5.1 Data Analysis

Model-Generated Summaries Analysis To empirically quantify the properties of the model-generated summaries we use 4 established metrics: (i) *Abstactness (novel n-grams)* – the percentage of summary n-grams absent in the article (Narayan et al., 2018). (ii) *Redundancy (RED)* – measures repetitive n-grams within a summary (S) using the formula: $RED(S) = \frac{\sum_{i=1}^{m}(f_i-1)}{\sum_{i=1}^{m} f_i}$ where m is the number of unique n-grams in the summary and f_i represents a frequency of specific n-gram within the summary. (iii) *Compression Ratio (CMP)* – the word counts in summary (S) divided by the corresponding article (A): $CMP_w(S, A) = 1 - \frac{|S|}{|A|}$. Higher compression ratios result in greater reduction at the word level, which can make the summarization task more difficult (Bommasani and Cardie, 2020). (iv) *Mean Token Length* – The average token count per summary by a word-delimited tokenizer.

Table 2 presents a quantitative analysis of the characteristics of model-generated summaries, highlighting the challenges in evaluating our data. Languages with a high level of abstractness (> 35 novel 1-gram) tend to be more difficult to evaluate using n-gram-based metrics due to their novel, distilled, and non-redundant nature. This challenge is particularly pronounced in high-fusion languages, which, in addition to their high level of abstractness often exhibit more complex linguistic structures.

Human Annotation Analysis Table 4 presents the statistics of the collected human annotations across languages. The average agreement rate, measured using Krippendorff's α, is 0.4 for coherence and 0.47 for completeness, indicating moderate

Lang.	Agreement		Avg. Score (Std)		Avg. Gap (Std)		# Ann.
	Coh.	Com.	Coh.	Com.	Coh.	Com.	
ZH	0.35	0.35	3.2 (0.8)	3.2 (0.8)	1.0 (0.7)	1.0 (0.8)	1504
YOR	0.40	0.49	3.0 (0.9)	3.1 (0.8)	1.0 (0.8)	0.9 (0.7)	1296
JA	0.61	0.40	3.5 (0.7)	3.4 (0.7)	0.8 (0.8)	0.7 (0.6)	188
TR	0.32	0.40	3.2 (0.9)	2.9 (1.0)	1.0 (0.9)	1.3 (0.9)	2200
AR	0.32	0.35	2.6 (0.8)	2.7 (0.7)	0.8 (0.8)	0.9 (0.7)	1352
HE	0.71	0.65	3.8 (1.1)	3.5 (1.2)	0.9 (0.9)	0.9 (0.9)	1284
ES	0.42	0.42	3.2 (0.9)	3.1 (0.7)	1.0 (1.0)	0.7 (0.7)	1464
UKR	0.46	0.62	3.3 (0.8)	3.2 (0.8)	0.8 (0.9)	0.9 (0.8)	2212

Table 4: *Human Annotation Statistics:* Krippendorff's α (agreement), average score, mean absolute gap between Gemini and GPT annotations, and annotation count per language. Coh. = Coherence, Com. = Completeness.

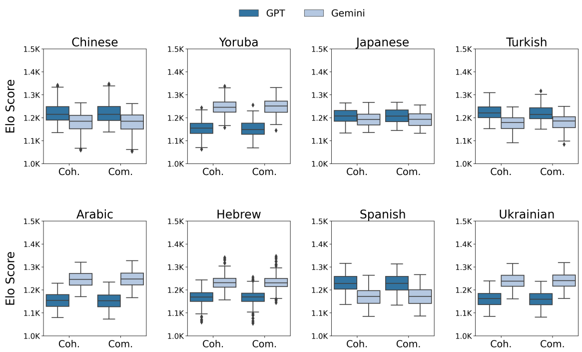

Figure 1: Elo score distribution of human annotations for Gemini- and GPT-generated summaries across all criteria. Coh. = Coherence, Com. = Completeness.

inter-annotator agreement. Scores range within $[2, 3]$, and the mean absolute gap between human predictions for Gemini and GPT summaries is ~ 1 for all languages in both coherence and completeness, demonstrating the effectiveness of the applied corruption to spread results. Additionally, the data analysis helps predict outliers. For example, we hypothesize that languages with low agreement rates (e.g., Arabic) correlate with more outliers in correlation analysis (later filtered via qualification questions), while those with high agreement rates and moderate average scores (e.g., Japanese) suggest higher correlations.

Using Elo rankings (Elo and Sloan, 1978) between human annotations and the generated summaries, as shown in Figure 1, we observe that Gemini summaries are generally ranked higher for high-fusional and low-resource languages, while GPT summaries are ranked higher for high-resource languages. Additionally, it is interesting to note that in all cases the same model is ranked higher for both criteria, which is possibly due to the *halo effect*, where an overall impression influences judgments across multiple specific aspects (Draws et al., 2021).

5.2 Assessed Metrics for Summarization

We assess a total of 10 evaluation metrics that are commonly used in summarization: **N-Gram Metrics:** measuring the lexical overlap (word to word) between the system and reference summaries. For this evaluation, we used ***ROUGE*** (Lin, 2004), considering four variants: ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-3 (trigram), ROUGE-L (longest common subsequence). ***CHRF*** (Popović, 2015) — measuring the character n-gram F-score; and ***BLEU*** (Papineni et al., 2002). We have also utilized adapted n-grams for multilingual use: ***ROUGE (mBERT Tokenizer)*** — leverages Byte-Pair Encoding (BPE) tokenization from BERT-multilingual (Kenton and Toutanova, 2019), ensuring more accurate evaluation across 104 languages, and ***ROUGE (Monolingual)*** — equipped with a language-specific tokenizer enabling evaluation of adaptability to specific languages.[11]

Neural-Based Metrics: ***MoverScore*** (Zhao et al., 2019) — measures the Euclidean distance between two contextualized BERT representations and relies on soft alignments of words learned by solving an optimization problem, utilized with mBERT to support multilingual. BERTScore (Zhang et al., 2019) — computes the similarity between BERT token embeddings of the system and reference summaries. For multilingual evaluation, we have used two extended versions: ***BERTScore (mBERT)*** which was trained on 104 languages (Kenton and Toutanova, 2019), and ***BERTScore (Monolingual)*** — adapted with a language-specific BERT. ***Gemini as a Judge*** (1.0-pro) (Team et al., 2023) — We used the Gemini model as an evaluator, which the given prompt was in the same format as the one given to the annotators. ***COMET*** (Rei et al., 2020) — we utilized the pre-trained model *wmt22-comet-da*, built on the XLM-R model (Conneau et al., 2019) and trained for machine translation (MT) evaluation using a regression-based objective to minimize the mean squared error (MSE) between predicted quality scores and human-annotated scores. To adapt COMET for summarization evaluation, we excluded the source input, as summarization assessment focuses on comparing the generated summary to a human-written reference. While COMET was designed for MT, its applicability can be extended to summarization, as both tasks are generative and involve

[11]See Appendix B.1 for the full list of the tokenizers we used.

	Criteria	Coherence				Completeness			
	Typological Family	Isolating	Agglutinative	Low Fusional	High Fusional	Isolating	Agglutinative	Low Fusional	High Fusional
	N-Gram Metrics								
1	ROUGE1	**0.20****	0.27**	0.11*	-0.25**	0.15**	0.11**	0.08*	-0.20**
2	ROUGE2	0.20**	0.28**	0.11*	-0.07**	0.14**	0.14**	0.08*	-0.03
3	ROUGE3	0.16**	0.27**	0.09*	-0.01**	0.12**	0.10*	0.01*	0.02
5	ROUGEL	0.19**	0.23**	0.11*	-0.23**	0.15**	0.10*	0.08*	-0.18**
6	BLEU	0.03**	0.03	0.11**	-0.30**	0.02	0.05*	0.07*	-0.10**
7	CHRF	0.02**	0.09	0.16**	-0.46**	0.01*	0.01*	0.14*	-0.38**
8	ROUGE1 (mBERT Tokenizer)	0.14**	0.18**	0.15**	0.10**	0.10*	0.09*	0.14**	0.15**
9	ROUGE2 (mBERT Tokenizer)	0.14**	0.20**	0.15**	0.11*	0.10*	0.09*	0.19**	0.15**
10	ROUGE3 (mBERT Tokenizer)	0.12**	0.22**	0.12**	0.11*	0.10*	0.07*	0.15**	0.14**
11	ROUGEL (mBERT Tokenizer)	0.14**	0.17**	0.13**	0.08*	0.11*	0.05*	0.13**	0.12**
12	ROUGE1 (Monolingual)	0.17**	0.23**	0.11**	0.02*	0.07	0.13*	0.06*	0.07**
13	ROUGE2 (Monolingual)	0.12**	0.25**	0.12**	0.09	0.12*	0.13*	0.07*	0.14**
14	ROUGE3 (Monolingual)	0.07**	0.24**	0.13**	0.07	0.07	0.08*	0.02*	0.09*
15	ROUGEL (Monolingual)	0.10**	0.22**	0.11**	0.03	0.08	0.12*	0.07*	0.11*
16	BLEU (Lemmatized Form)	N.A	N.A	0.15**	0.30**	N.A	N.A*	0.08*	**0.40***
	Neural-Based Metrics								
17	Gemini as a Judge	0.15**	0.03	0.15*	0.05*	0.14**	0.16**	0.10**	0.09**
18	MoverScore	0.07**	0.15*	0.18*	0.02	0.08	0.10*	0.17**	0.08*
19	BERTScore (mBERT)	0.09**	0.15**	0.19**	0.15*	0.13**	0.07*	0.16**	0.13**
20	BERTScore (Monolingual)	0.13**	**0.32****	0.20**	0.17*	0.12**	**0.21****	-0.03*	0.15**
21	COMET	0.07**	0.23**	**0.23****	**0.35***	**0.16****	0.18**	**0.24****	0.24**

Table 5: Pearson correlation between resource types and evaluation metrics. Significance: * $p < 0.05$, ** $p < 0.01$. The dashed line separates English-based from multilingual metrics. The highest correlation per column is in bold.

evaluating a predicted output against a reference.

5.3 Results and Analysis

Having collected the human annotations, we now examine the Pearson correlation between the human annotations with both n-gram and neural metrics. We aim to investigate what influences the correlation and assess systematically the ways that have been proposed to mitigate poor correlation. To achieve this, we analyze several aspects, including language typology family, resource availability, and metrics that are adapted to multilingual evaluation. Table 5 shows the correlation from a typological family, while Table 6 presents the correlation from a resource-type perspective. See the Appendix B.2 for the correlations per language.[12]

The Impact of Typological Family Table 5 examines the Pearson correlations from the typological family perspective. The correlations for each family were measured across all the languages within the respective linguistic family. Overall, it appears that n-gram metrics are sensitive to the typological family of the language, while neural metrics have not shown this tendency. For example, for both criteria, fusional languages exhibit weaker correlations with human judgments, with low correlations for Low-Fusional languages and even negative correlations for High-Fusional languages, due to their rich morphology (lines 1-7). However, for neural-based metrics, the typological family appears to play a less critical role. For instance, low-fusional languages achieve the highest correlation for BERTScore (mBERT) (line 19) in both criteria. Interestingly, COMET exhibits an inverse trend compared to n-gram metrics, consistently showing a better correlation with fusional languages (line 21). Additionally, the results for n-gram metrics not adapted to multilingual (lines 1-7) show that agglutinative languages displayed better correlations with human scores than Isolating languages in coherence while Isolating languages showed a better correlation in completeness. The advantage of agglutinative languages over Isolating languages is surprising, given that these families tend to have more complex morphological structures due to longer morphemes, which can be more challenging to tokenizers.[13] Overall, neural-based metrics show a stronger correlation than gen-gram-based metrics.

The Impact of Resource Level Table 6 presents the Pearson correlation between human annotations (by resource type) and neural metrics, evaluating coherence and completeness for high- and low-resource languages. The results indicate that Gemini exhibits the lowest correlations with human scores among other multilingual models for both

[12] Also, correlations using Spearman's rank correlation.

[13] We acknowledge that the disparity may stem from the poor quality of generated summaries in Yoruba, a low-resource language compared to Turkish. We hypothesize that the low generation quality contributed to the weak performance of automatic metrics, despite the relatively high human scores in Table 4, which may explain the low correlation observed.

Criteria	Coherence		Completeness	
Resource Type	**High**	**Low**	**High**	**Low**
Gemini as a Judge	0.19*	0.13**	0.08**	0.12**
MoverScore	0.16**	0.13**	0.10*	0.06*
BERTScore (mBERT)	0.23**	0.16**	0.16**	0.15**
BERTScore (Monolingual)	0.27**	0.16**	0.17**	0.21**
COMET	0.32**	0.18**	0.13**	0.24**

Table 6: Pearson correlation between low- and high-resource human annotations and neural-based metrics. significance levels denoted by: * $p < 0.05$, ** $p < 0.01$.

criteria, regardless of language resource level, indicating that LLMs as judges still lag behind other metrics. Furthermore, the table presents an advantage for using language-specific BERT models, over multilingual BERT (mBERT), suggesting that a dedicated tokenizer improves correlation more than training on larger, non-specific datasets.[14]

Notably, COMET shows the strongest correlation with human scores for coherence and outperforms in completeness for low-resource languages compared to high-resource ones. This can be attributed to COMET's unique training for generative evaluation tasks, enabling it to better capture human-like evaluation, especially in challenging scenarios. Its performance underscores the potential of task-specific training to bridge the gap between automated metrics and human evaluation, particularly for low-resource languages. We hypothesize that a metric trained specifically for summarization evaluation could perform even better.

The Impact of Metrics Adapted to non-English Languages The results in Table 5 highlight the importance of adequate tokenizers for fusional languages and in particular for isolating and agglutinative languages in completeness evaluation (lines 1-7 vs. 8-15). For example, ROUGE with mBERT tokenizer or a language-specific tokenizer (lines 8–15) improves correlation and can even reverse a negative correlation to a positive one in languages with highly morphological grammar, such as Hebrew and Arabic (e.g., ROUGE-L in high-fusional languages improves from -0.23 to 0.08, lines 5 & 11). Also, applying BLEU to the lemmatized text shows a significant improvement for fusional languages, with the correlation increasing from -0.10 to 0.40 for high-fusional languages (line 6 vs. 16).

Notably, for isolating and agglutinative, correlations decrease, favoring the space-delimited ROUGE variation. We hypothesize that tokenizers struggle with the long morphological sequences in agglutinative languages, making it difficult to split morphemes correctly. As a result, tokenization with space delimitation may be more effective. However, for completeness, the adapted variations have shown better performance. The inverse correlation is also observed, with positive correlations for BERTScore variations and MoverScore in high-fusional languages (lines 18-20). Additionally, using models not trained on non-English languages is suboptimal, as shown in Table 6, where MoverScore—untrained on non-English—performs worst for both coherence and completeness.

6 Conclusion

In this work, we systematically evaluate the reliability of automatic metrics of evaluation for generation in non-English languages through a comprehensive correlation analysis with human annotations. We aim to identify the linguistic factors that influence these correlation and asses new metrics and approaches designed for the multilingual summarization evaluation task.

Our annotation protocol addresses previous weaknesses, including limited typological family and resource type coverage, insufficient evaluation of diverse metrics (particularly neural-network-based models trained for evaluation), and adaptation of general-purpose metrics to non-English languages. We also provide statistical validation, unlike prior non-English evaluations lacking statistical significance reports. We crowd-sourced rank annotations for eight languages representing diverse typological families, each with different word boundaries, a key factor for n-gram-based metrics. To bridge the gap in evaluating neural network metrics, we included both high- and low-resource languages within each typological group.

Based on our findings, which highlight the limited ability of n-gram metrics to handle complex linguistic structures, such as those found in fusional languages, compared to neural network-based metrics—particularly those trained for multilingual evaluation of generative models—we recommend transitioning from n-grams to neural networks specifically trained for multilingual summarization tasks. As an intermediate solution during this transition, when using n-grams for fusional languages, we suggest employing tokenization techniques that can break down complex linguistic structures.

[14]A comprehensive list of the BERT models employed in this study is provided in Appendix B.1.

Limitations

Evaluation Criteria Although we have used coherence and consistency as evaluation criteria (Han et al., 2024; Forde et al., 2024), we acknowledge that the common approach, based on SummEval (Fabbri et al., 2021), typically incorporates fluency, coherence, consistency, and relevance. However, our previous experiments revealed an extremely low inter-annotator agreement rate (~0), suggesting that annotators struggled to distinguish subtle differences among all four metrics. To mitigate this issue, we narrowed our focus to coherence and consistency, as they offer a more straightforward and reliable basis for evaluation.

Number of Samples To cover diverse typological groups and resource levels while relying on available crowd workers, sample sizes vary across languages. For example, Japanese had only one worker, leading to a smaller dataset than other languages.

References

Gulinigeer Abudouwaili, Wayit Ablez, Kahaerjiang Abiderexiti, Aishan Wumaier, and Nian Yi. 2023. Strategies to improve low-resource agglutinative languages morphological inflection. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 508–520.

Anton Alexandrov, Veselin Raychev, Dimitar I Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Bggpt 1.0: Extending english-centric llms to other languages. *arXiv preprint arXiv:2412.10893*.

Abdulaziz Alhamadani, Xuchao Zhang, Jianfeng He, and Chang-Tien Lu. 2022. Lans: Large-scale arabic news summarization corpus. *arXiv preprint arXiv:2210.13600*.

Giorgio Francesco Arcodia et al. 2007. Chinese: A language of compound words. *Selected proceedings of the 5th Décembrettes: Morphology in Toulouse*, pages 79–90.

Satanjeev Banerjee and Alon Lavie. 2004. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Anouschka Bergmann, KC Hall, and SM Ross. 2007. Language files. In *Materials for an Introduction to Language and Linguistics*. The Ohio State University Press.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Georgie Botev, Arya D McCarthy, Winston Wu, and David Yarowsky. 2022. Deciphering and characterizing out-of-vocabulary words for morphologically rich languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5309–5326.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for arabic mt evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213.

Svitlana Budzhak-Jones. 1998. Against word-internal codeswitching: Evidence from ukrainian-english bilingualism. *International Journal of Bilingualism*, 2(2):161–182.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Etienne Denoual and Yves Lepage. 2005. Bleu in characters: towards automatic mt evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 9, pages 48–59.

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

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)*.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-examining summarization evaluation across multiple quality criteria. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13829–13838, Singapore. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jessica Forde, Ruochen Zhang, Lintang Sutawika, Alham Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19476–19493.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Rilyn Han, Jiawen Chen, Yixin Liu, and Arman Cohan. 2024. Rethinking efficient multilingual text summarization meta-evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15739–15746.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ozlem Istek and Ilyas Cicekli. 2007. A link grammar for an agglutinative language. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 285–290.

Farayi Kambarami, Scott McLachlan, Bojan Bozic, Kudakwashe Dube, and Herbert Chimhundu. 2021. Computational modeling of agglutinative languages: the challenge for southern bantu languages. *Arusha Work. Pap. Afr. Linguist*, 3(1):52–81.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-based evaluation metrics by disentangling along linguistic factors. *arXiv preprint arXiv:2110.04399*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. *arXiv preprint arXiv:2106.01478*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sandeep Kumar and Arun Solanki. 2023. Rouge-ss: A new rouge variant for evaluation of text summarization. *Authorea Preprints*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, et al. 2024. On the challenges and opportunities in generative ai. *arXiv preprint arXiv:2403.00025*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Jolaade Okanlawon. 2016. An analysis of the yoruba language with english. *Phonetics, Phonology, Morphology and Syntax. NorthEastern University*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

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.

Tzuf Paz-Argaman, Itai Mondshine, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. Hesum: a novel dataset for abstractive text summarization in hebrew. *arXiv preprint arXiv:2406.03897*.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Masayoshi Shibatani and Taro Kageyama. 2015. Introduction to the handbooks of japanese language and linguistics. *Kubozono, Haruo (Hg.): Handbook of Japanese Phonetics and Phonology. Berlin Ua: De Gruyter, S. Vii–Xxxiii*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *arXiv preprint arXiv:2308.16687*.

Otakar Smrž. 2007. Functional arabic morphology. *The Prague Bulletin of Mathematical Linguistics*, (88):5–30.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2019. What's wrong with hebrew nlp? and how to make it right. *arXiv preprint arXiv:1908.05453*.

AA Vetrov and EA Gorn. 2022. A new approach to calculating bertscore for automatic assessment of translation quality. *arXiv preprint arXiv:2203.05598*.

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, and Lei Li. 2021. Cnewsum: a large-scale chinese news summarization dataset with human-annotated adequacy and deducibility level. *arXiv preprint arXiv:2110.10874*.

Oreen Yousuf, Gongbo Tang, and Zeying Jin. 2024. Improving bertscore for machine translation evaluation through contrastive learning. *IEEE Access*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Data Collection

A.1 Language Selection

Table 8 displays the full resource-type categorization per language we have defined using GPT-3 pre-trained data.

A.2 Power Analysis for Sample Size

To ensure the reliability of our statistical tests, we conducted a power analysis to determine the minimum required sample size for detecting a statistical correlation ($p - value \leq 0.05$). we applied a t-test power analysis and computed the required sample size per group to achieve these conditions. The analysis revealed that a minimum of ~400 samples per language is necessary for a well-powered correlation.

A.3 Participant Interface

The tasks are performed using a custom-built application displayed via mTurk, as shown in Figures 2-5. The task is in Arabic, for example; see Figure 6 for a Spanish example.

A.4 Data Collection Details

We utilized Amazon Mechanical Turk (MTurk) to distribute the task to various workers. For the student participants, all were undergraduate students from the linguistics field. To provide a custom user interface (UI) for our evaluation, we developed a JavaScript application and deployed it as a service using Google Cloud Run.[15]. Subsequently, we connected the MTurk participants to this service.

All participants were compensated in full, regardless of whether they correctly completed the task. The payment was set at $2.5 for rating 5 pairs of summaries, which we estimated would take approximately 10–15 minutes to complete.

Drawing lessons from previous studies, we invested significant effort into enhancing the user experience (UX) and the visual design of the application. This focus ensured that the interface was both intuitive and visually appealing, thereby improving participant engagement and task performance.

[15] https://cloud.google.com/run

A.5 Data Corruption

We have experimented with the following corruptions of the generated summaries concerning each quality criteria: **Coherence**: All verbs were replaced with their lemma forms, resulting in ungrammatical sentences. We removed random words from each sentence and replaced conjunctions with alternatives for languages without a lemmatizer (e.g., Chinese, Japanese, and Yoruba). In addition, reorder two sentences that are not adjacent. This corruption is inspired by the Shuffle Test Barzilay and Lapata (2008) used to evaluate whether models can detect incoherent text. **Completness**: Named entities with the same labels (e.g., PERSON and LOCATION) were shuffled within the summary. This is a common factual mistake of models (Pagnoni et al., 2021). Additionally, a random sentence from another article was inserted into the summary. Table 9 provides an example for a clean sentence and it's corrupted version.

A.6 Qualification Task

To filter out unqualified annotators, each was required to answer a generated question about the article in their native language. The model was prompted as follows: `Given the text: <TEXT> in <LANGUAGE>, generate a single-sentence question whose answer is found in the text.`

B Correlation Analysis

B.1 Implementation Details

Language-Specific BERT Models See Table 7 for the list of Bert models we used for each language.

Python Libraries To use BERTScore (mBERT), we employed the official implementation. For ROUGE (mBERT) and BPE tokenization, we used *Multilingual-Rouge-Scorer*.[20] For ROUGE (Language Tokenizer), we used the standard ROUGE package commonly applied in non-English papers.[21] For other metrics, we used the implementation from SummEval (Fabbri et al., 2021).[22]. We have used ChatGPT for assistance in coding the evaluation framework.

[20] https://github.com/faisaltareque/Multilingual-Rouge-Scorer/tree/main

[21] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

[22] https://github.com/Yale-LILY/SummEval

B.2 Results

See Table 11 for the full correlation for each language and metric. Also, Table 10 shows the correlation measured by Spearman's rank correlation coefficient.

Language	BERT Model	NER Model	Lemmatizer
Turkish	bert-base-turkish-cased	bert-base-turkish-cased-ner [16]	zeyrek [17]
Hebrew	DiktaBERT	DiktaBERT Shmidman et al. (2023)	DiktaBERT
Arabic	bert-base-arabic	CAMeL-Lab/bert-base-arabic-camelbert-msa-ner [18]	qalsadi [19]
Chinese	bert-base-chinese	zh_core_web_sm (spacy)	N.A
Japanese	bert-base-japanese-v3	ja_core_news_sm (spacy)	N.A
Spanish	bert-base-spanish-wwm-cased	es_core_news_sm (spacy)	es_core_news_md
Ukrainian	bert-base-multilingual-cased	uk_core_news_sm (spacy)	uk_core_news_sm
Yoruba	bert-base-multilingual-cased	N.A	N.A

Table 7: Language-specific BERT models, NER models, and lemmatizers.

Language	Lang Code	Number of Tokens	Percentage of Tokens ($p\%$)	Class
English	en	181,015	92.64%	A+
Spanish	es	1,510	0.77289%	A
Japanese	ja	217	0.11109%	A
Chinese	zh	194	0.09905%	A
Turkish	tr	116	0.05944%	B
Arabic	ar	61	0.03114%	A
Hebrew	he	15	0.00769%	B
Ukrainian	ukr	14	0.00763%	B
Yoruba	yor	0	0.00000%	B

Table 8: List of languages, language codes, number of tokens in pre-trained GPT-3 data, data ratios. The languages are grouped into two classes based on their data ratios in the GPT-3 pre-trained data: High Resource ($p > 0.1\%$), Low Resource ($p < 0.1\%$)

Criterion	Rule	Example
Coherence	Replace with lemmas	*Clean:* The athletes are preparing for the championship.*Corrupt:* The athlete be prepare for the championship.
	Replace conjunctions	*Clean:* Policies address rising inflation.*Corrupt:* Policies however address rising inflation.
	Reorder non-adjacent sentences	*Clean:* The center is hosting a charity event. Volunteers are needed.*Corrupt:* Volunteers are needed. The center is hosting a charity event.
Completness	Replace named entities	*Clean:* Joe Biden met Britney Spears at a charity event.*Corrupt:* Britney Spears, former president, met Joe Biden.
	Insert irrelevant sentence	*Clean:* Scientists found a new fish species in the Amazon.*Corrupt:* Scientists found a new fish species. A bakery is giving free cake samples.

Table 9: Examples of clean and corrupt sentences based on coherence and completeness criteria.

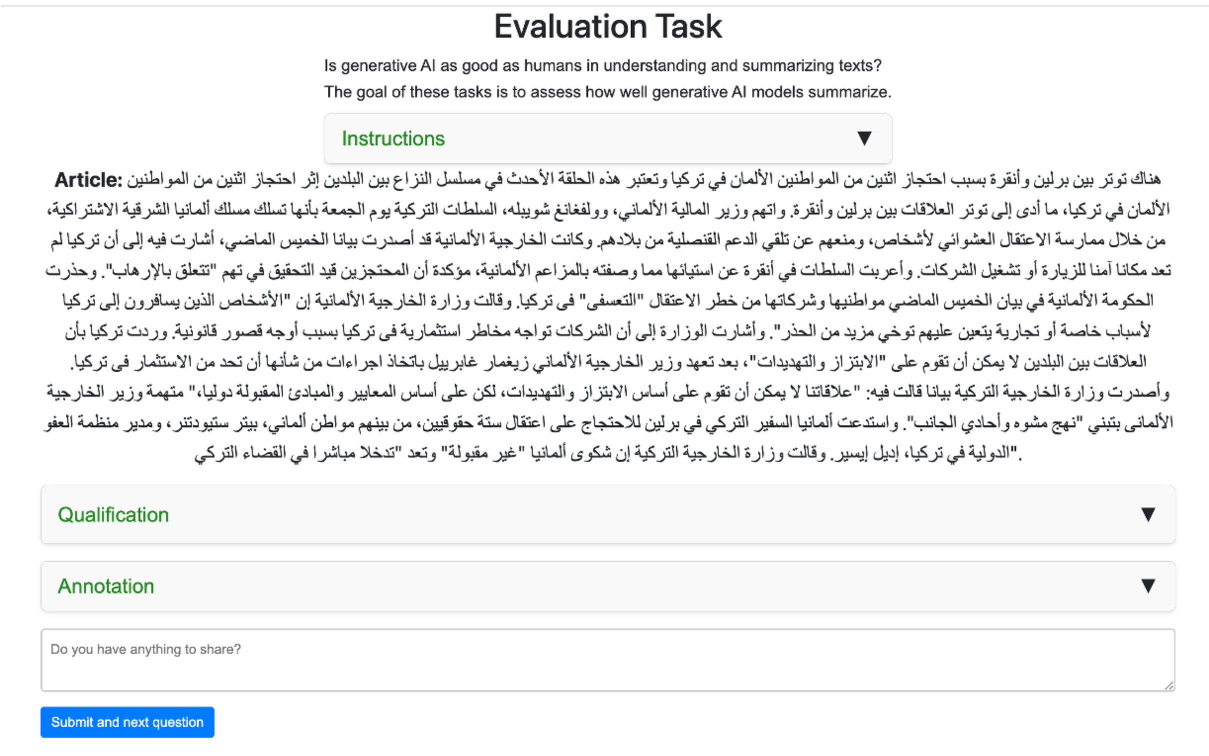

Figure 2: *Participant Interface in a closed mode:* The interface includes three drop-down sections: Instructions, Qualification and the Annotation task.

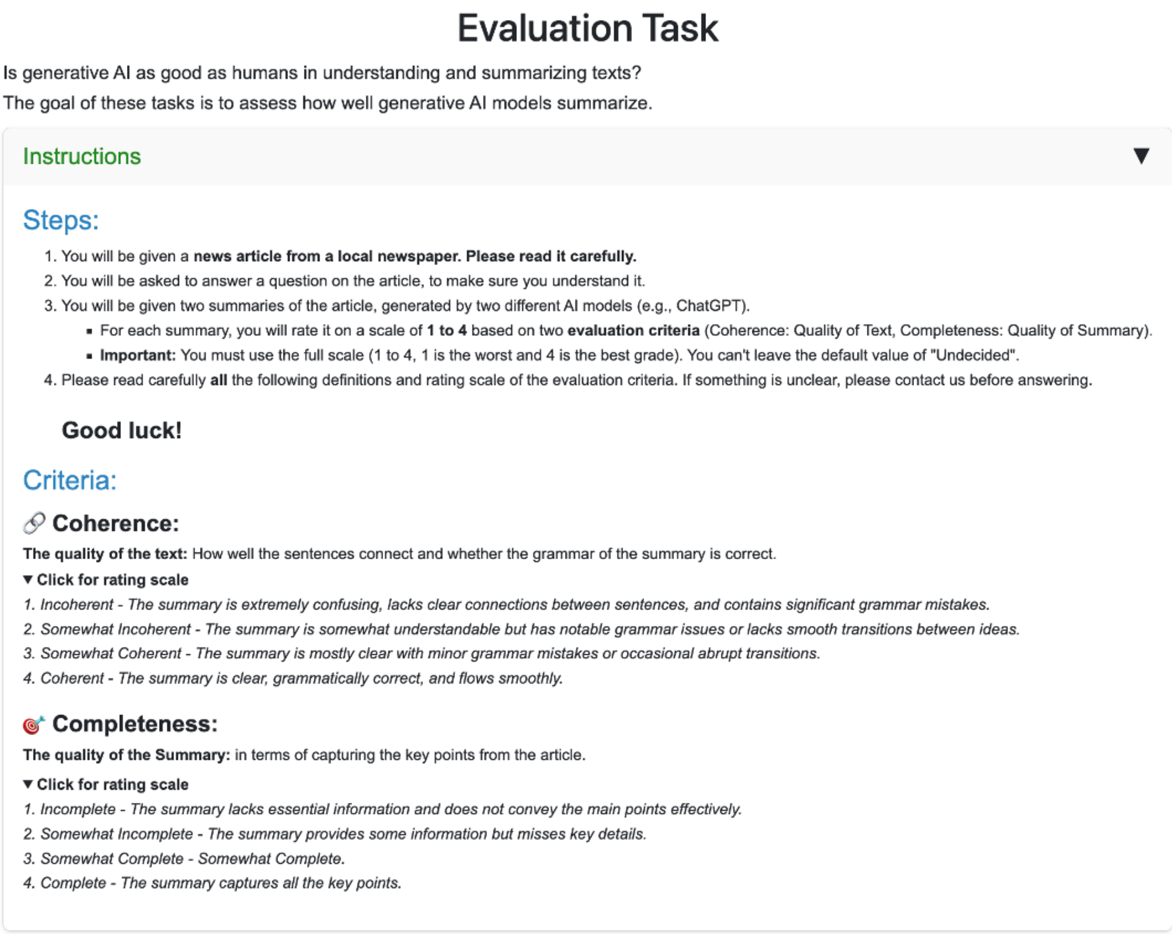

Figure 3: *The Participant Instructions Interface:* The participant has general steps and a detailed explanation and examples of each tested criteria.

Figure 4: *The Participant Qualification Interface:* The human summary is displayed at the top (the example is in Arabic), while the Qualification section below requires the participant to fill in their home state and answer a question generated by Gemini based on the human summary, designed to assess basic comprehension of the provided summary.

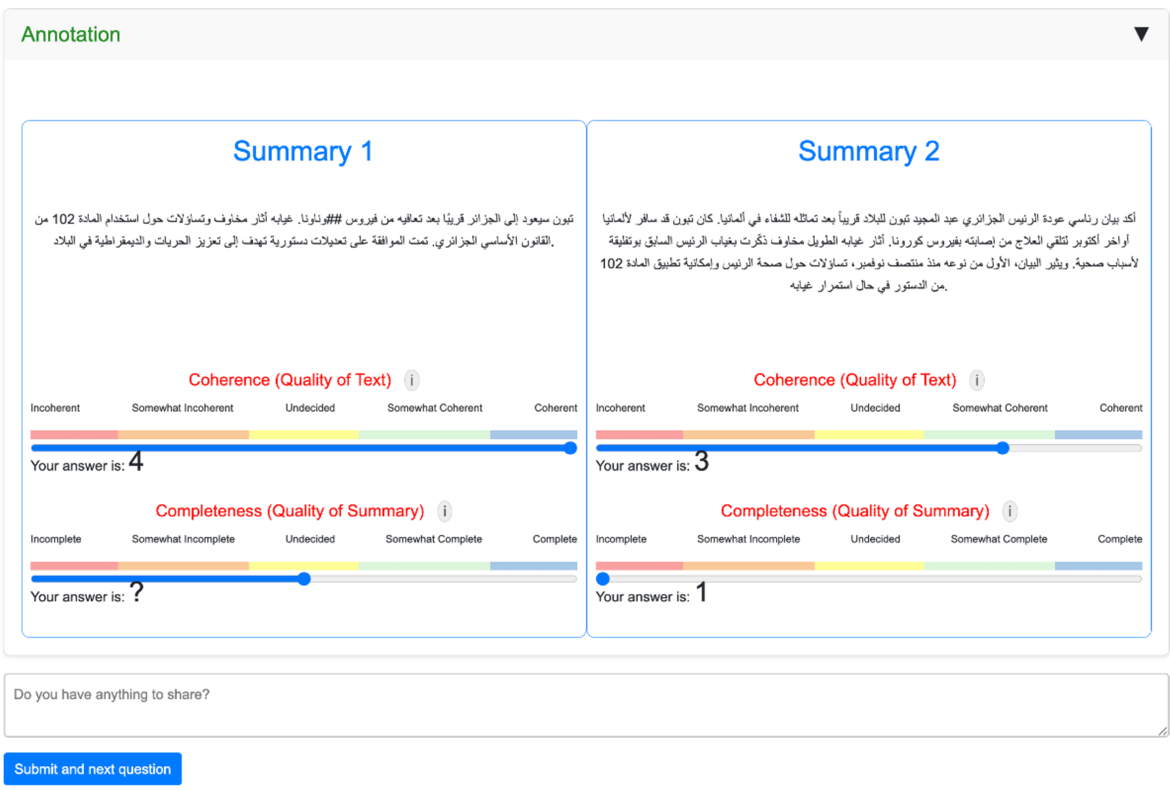

Figure 5: *The Participant Annotation Interface:* Two summaries are displayed side by side. Each criterion includes a slider ranging from 1 to 4, along with an info hover feature providing a reminder of the criterion's definition.

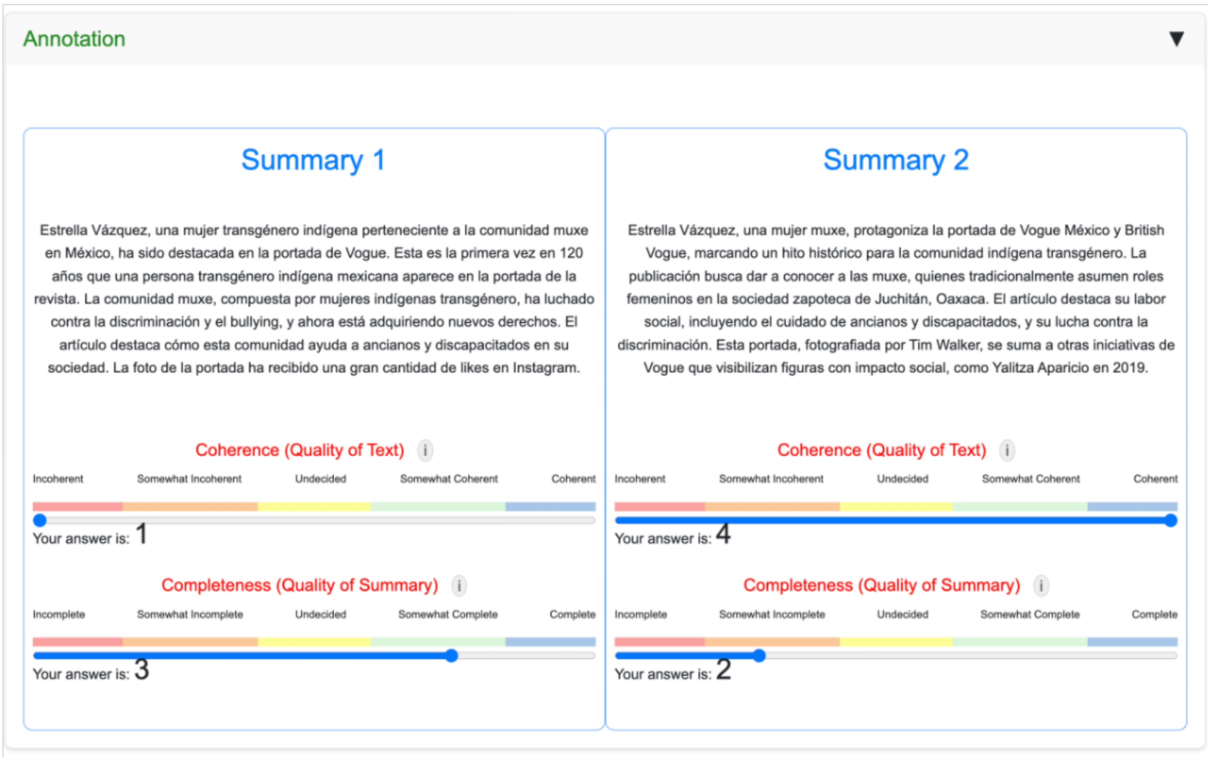

Figure 6: *The Participant Annotation Interface:* displayed in Spanish

		Coherence								Completeness							
	Typological Family	**Isolating**		**Agglutinative**		**High Fusional**		**Low Fusional**		**Isolating**		**Agglutinative**		**High Fusional**		**Low Fusional**	
	Language Code	**ZH**	**YOR**	**JA**	**TR**	**AR**	**HE**	**ES**	**UKR**	**ZH**	**YOR**	**JA**	**TR**	**AR**	**HE**	**ES**	**UKR**
	N-Gram Metrics																
1	**ROUGE1**	0.06	0.06	0.25**	0.28*	0.14*	-0.31**	0.16**	0.13*	0.11**	0.06	0.20*	0.08	0.19**	-0.26**	0.11*	0.16*
2	**ROUGE2**	0.07	0.08*	0.23*	0.31*	0.13*	-0.14*	0.15*	0.08	0.12**	0.10	0.21*	0.13*	0.17*	0.06	0.10	0.08
3	**ROUGE3**	0.08	0.06*	0.23*	0.28*	0.14*	-0.07	0.08*	0.10*	0.12**	0.06*	0.19*	0.06	0.13*	0.18**	0.07	-0.02
4	**ROUGEL**	0.06	0.08	0.28*	0.28*	0.10*	-0.26**	0.17**	0.09	0.10	0.10*	0.25*	0.09*	0.16*	-0.26**	0.12*	0.13*
5	**CHRF**	0.08	0.02	0.27*	0.25*	0.12*	-0.21**	0.15**	0.17**	0.10*	0.02	0.23**	0.19*	0.18*	-0.41**	0.13*	0.17**
6	**BLEU**	0.08	0.10*	N.A	0.24*	0.14*	-0.16*	0.11**	0.15*	-0.05	0.11*	0.24**	0.05	0.12*	-0.38**	0.06	0.04
7	**ROUGEL (mBERT Tokenizer)**	0.10**	0.07*	0.13*	0.21*	0.03	0.36**	0.13**	0.04	0.08	0.09*	0.09*	0.03	0.12*	0.40**	0.09*	0.12*
8	**ROUGEL (Language Tokenizer)**	0.07	-0.02	0.10*	0.20*	0.04	0.30*	0.13*	0.11*	0.04	-0.02	0.12*	0.06	0.17*	0.40**	0.11*	0.12*
	Neural-Based Metrics																
11	**BERTScore Monolingual**	0.10*	-0.02	0.30	0.33*	0.10*	0.0	0.24**	0.12*	0.16	0.01	0.26*	0.11**	0.14*	0.13	0.15*	0.21**
12	**BERTScore (mBERT)**	0.12*	0.02	0.27*	0.25*	0.08*	-0.15*	0.22**	0.11*	0.21	0.03	0.24*	0.10*	0.12*	-0.06	0.15*	0.15*
13	**COMET**	0.13*	0.00	0.27*	0.23*	0.00	0.38	0.27**	0.16	0.23**	0.02	0.24*	0.11*	0.25**	0.49**	0.09	0.25*
14	**Gemini Model**	0.07*	0.11*	0.27	0.08*	0.03	-0.10	0.16**	0.16**	0.05	0.16*	0.23*	0.19**	0.19**	0.12	0.06	0.06

Table 10: Spearman correlation between language and evaluation metrics. Significance levels are denoted by: * $p < 0.05$, ** $p < 0.01$. The dashed line separates the English-based metrics from the multilingual metrics.

		Coherence								Completeness							
	Typological Family	**Isolating**		**Agglutinative**		**High Fusional**		**Low Fusional**		**Isolating**		**Agglutinative**		**High Fusional**		**Low Fusional**	
	Language Code	**ZH**	**YOR**	**JA**	**TR**	**AR**	**HE**	**ES**	**UKR**	**ZH**	**YOR**	**JA**	**TR**	**AR**	**HE**	**ES**	**UKR**
	N-Gram Metrics																
1	**ROUGE1**	0.09*	0.12*	0.25**	0.33*	0.10*	-0.31**	0.18**	0.09*	0.10**	0.11**	0.15*	0.08**	0.14**	-0.23**	0.13*	0.15*
2	**ROUGE2**	0.11*	0.14*	0.20*	0.45*	0.10*	-0.14*	0.14*	0.06*	0.10**	0.11**	0.16*	0.12*	0.13*	-0.06	0.11*	0.10*
3	**ROUGE3**	0.11*	0.08	0.20*	0.36*	0.12*	-0.07	0.08*	0.08*	0.10**	0.08**	0.16*	0.07	0.11*	-0.01	0.07	0.00
4	**ROUGEL**	0.10*	0.14**	0.23*	0.34*	0.06*	-0.26**	0.19**	0.06*	0.10	0.12**	0.19*	0.09*	0.10*	-0.21**	0.13*	0.11**
5	**CHRF**	0.10*	0.11	0.22*	0.31*	0.09*	-0.21**	0.18**	0.12*	0.10*	-0.16**	0.18*	0.11*	0.14*	-0.12*	0.14*	0.15**
6	**BLEU**	-0.03	0.13*	N.A	0.36*	0.06*	-0.16*	0.10**	0.10*	-0.03	-0.38**	0.10	0.05	0.09(	-0.08	0.10*	0.00*
7	**ROUGEL (mBERT Tokenizer)**	-0.05*	0.10*	0.30*	0.28*	0.09*	0.46**	0.18**	0.12*	0.11*	0.10*	0.21*	0.05	0.12*	0.49**	0.09	0.10
8	**ROUGEL (Language Tokenizer)**	-0.06*	0.14**	0.25*	0.32*	0.11*	0.3*	0.17*	0.10*	0.08*	0.00	0.21*	0.08	0.17*	0.47**	0.09	0.09
9	**ROUGEL (Llema Form)**	N.A	N.A	N.A	0.29*	0.09	0.42*	0.15*	0.14**	N.A	N.A	N.A	0.10*	0.12*	0.46*	0.10	0.09*
	Neural-Based Metrics																
10	**BERTScore**	0.02	0.16*	0.16*	0.07	0.10*	-0.01	0.19**	0.09*	0.20**	0.25**	0.17*	0.12*	0.14*	0.17*	0.17*	0.13*
11	**BERTScore Monolingual**	0.12*	0.04	0.09	0.31*	0.09*	0.0	0.27**	0.07*	0.09	0.07	0.19*	0.11**	0.15*	0.11	0.17*	0.15**
12	**BERTScore (mBERT)**	0.11*	0.09*	0.22*	0.23*	0.05*	-0.15*	0.23**	0.08*	0.10	0.15*	0.17*	0.09*	0.09*	-0.07	0.16*	0.21**
13	**COMET**	0.08*	0.01	0.21*	0.21*	0.10	0.38	0.32**	0.11*	0.24**	0.06**	0.18*	0.09*	0.23**	0.17**	0.14*	0.25*
14	**Gemini Model**	0.16**	0.01	0.23*	0.08*	0.03	-0.10	0.19**	0.16**	0.11*	0.16**	0.20*	0.16**	0.16**	0.00	0.09	0.10*

Table 11: Pearson correlation between language and evaluation metrics. Significance levels are denoted by: * $p < 0.05$, ** $p < 0.01$. The dashed line separates the English-based metrics from the multilingual metrics.