Comparing Hallucination Detection Metrics for Multilingual Generation

Anonymous ACL submission

Abstract

While many automatic hallucination detection techniques have been proposed for English texts, their effectiveness in multilingual contexts remains unexplored. This paper aims to bridge the gap in understanding how these hallucination detection metrics perform on non-English languages. We evaluate the efficacy of various detection metrics, including lexical metrics like ROUGE and Named Entity Overlap and Natural Language Inference (NLI)-based metrics, at detecting hallucinations in biographical summaries in many languages; we also evaluate how correlated these different metrics are to gauge whether they measure the same phenomena. Our empirical analysis reveals that while lexical metrics show limited effectiveness, NLI-based metrics perform well in high-resource languages at the sentence level. In contrast, NLI-based metrics often fail to detect atomic fact hallucinations. Our findings highlight existing gaps in multilingual hallucination detection and motivate future research to develop more robust detection methods for LLM hallucination in other languages.

1 Introduction

Large Language Models (LLMs) have brought about remarkable advances in text generation. However, they are still prone to factuality hallucination - the generated text conflicts with established world knowledge (Huang et al., 2023; Zhang et al., 2023). Despite considerable research efforts towards understanding and detecting hallucinations, the focus has predominantly been on English texts (Huang et al., 2023; Zhang et al., 2023; Ji et al., 2023). This emphasis has resulted in a significant knowledge gap regarding hallucinations in multilingual contexts, and it is currently unclear whether the methods developed for detecting and addressing hallucinations in English are effective or even applicable in multilingual settings.

In this paper, we address the highlighted issues by evaluating the effectiveness of various metrics (initially proposed for English generation) within a multilingual context. Our approach involves a comparative analysis of traditional lexical metrics, such as ROUGE (Lin, 2004) and Named Entity Overlap, alongside Natural Language Inference (NLI) metrics. Additionally, we compare reference-based metrics with pairwise metrics that are based on the consistency among generated samples; we also present a correlation study between these automated metrics and human factuality verification. These evaluations shed light on the efficacy of these metrics in non-English languages and aim to establish a more robust framework for hallucination detection in multilingual language models.

We empirically evaluate different hallucination detection techniques in the multilingual context and find that:

- Lexical overlap metrics (e.g., ROUGE, named entity overlap) do not correlate with NLI-based metrics in detecting factual hallucinations in both reference-free and reference-based settings.
- Pairwise NLI-based metrics strongly correlate with reference-based metrics in high-resource languages, but this significantly diminishes in low-resource languages.
- Automatic NLI-based metrics effectively identify sentence-level hallucinations in high-resource languages when compared to human evaluations. However, their performance diminishes when assessing simpler atomic facts.

These findings indicate several open problems for multilingual hallucination detection. While traditional lexical overlap methods and pairwise comparisons of multiple generated texts are more accessible approaches for low-resource languages, they are often inadequate for hallucination detection. This highlights that the effectiveness of hallucination detection is closely tied to the availability and quality of language resources, following the

trend observed in English that detection accuracy depends on natural language understanding abilities (Manakul et al., 2023; Min et al., 2023). Our work points to a substantial gap in hallucination detection in multilingual contexts, necessitating focused future research to bridge this divide.

2 Analyzing Methods for Multilingual Hallucination Detection

Hallucination refers to machine-generated content that is not faithful to a reference. This *reference* could be source text, preceding generated context, or world knowledge (Ji et al., 2023; Zhang et al., 2023). Hallucinations can be further classified into two types: *verifiable* hallucinations, which happen when the generated content directly contradicts the reference, and *unverifiable* hallucinations, meanwhile, refer to generated content that cannot be verified with the source material.

In this paper, we measure the efficacy of different metrics on detecting multilingual hallucinations. We focus on biography generation, a domain that is particularly sensitive to factual accuracy and coherence (Min et al., 2023; Dhuliawala et al., 2023). We test a suite of automatic metrics, each of which caters to a different aspect of factual generation: ROUGE (Lin, 2004), named entity overlap, and Natural Language Inference (NLI)-based methods.

2.1 Multilingual Biography Generation

Inspired by prior work measuring factuality in English (Min et al., 2023), we generate parallel biographies in different languages. The generated texts are then compared against a reference text (for *reference-based* metrics) and other generated samples (*pairwise* metrics) to detect hallucinations.

This section characterizes the generation quality of these biographies (Table 1). We consider the average length of each biography (in tokens and sentences), along with estimates of how accurate the generation language is to the prompt language, as in some cases, multilingual LMs will generate continuations in an unexpected language (Kang et al., 2023; Bawden and Yvon, 2023).

The length of the generated texts varies notably across languages. High-resource languages like English and French average 78.3 and 115.8 tokens per response, respectively. However, mid-resource languages such as Thai tend to generate much shorter biographies and incomplete sentences, and low-resource languages fare even worse (for instance, Ukrainian averages just 5.7 tokens and 0.40 sentences), demonstrating the significant gap in generation abilities across languages.

Table 1: Generation quality statistics for BLOOMZ-mt. The languages in **bold** are the most frequent in the ROOTS pretraining corpus (Laurençon et al., 2022), and the underlined languages are covered in the xP3mt fine-tuning dataset (Muennighoff et al., 2023). "FLang." refers to the most frequently generated language for each prompt language.

Lang.	#Token	#Sent.	Valid %	FLang.	Acc.
en	78.3	2.64	99.97	en	96.0
zh	115.8	4.30	100.00	zh	92.43
es	62.8	2.01	100.00	es	92.33
fr	71.3	2.24	100.00	fr	93.23
vi	45.6	1.66	98.92	vi	71.67
id	46.3	1.76	98.30	en	36.45
de	63.3	2.33	99.58	en	2.79
it	58.1	1.94	99.76	en	3.31
ja	50.3	1.97	90.73	zh	21.85
bg	17.4	1.15	86.74	en	13.69
ro	9.6	0.93	80.24	en	2.68
sv	7.6	0.51	40.73	en	1.79
th	14.8	0.81	77.08	th	94.96
ru	10.2	0.68	55.49	ru	50.44
uk	5.7	0.40	35.24	uk	41.87
fa	3.2	0.13	10.80	ur	29.90
fi	1.7	0.11	9.76	fi	34.52
ko	2.0	0.09	8.37	ko	47.30
hu	0.8	0.05	6.28	pt	14.36
Avg.	34.8	1.35	31.65	-	50.01

We assess the accuracy of the generated languages through three metrics: the percentage of valid generations that is detectable for the langdetect package (Valid %)[1], the most frequently generated language for a given target language (Flang), and the accuracy of generated language out of the valid generations (Acc.). For high-resource languages like English, Chinese, Spanish, and French, the models generally generate text in the correct language; however, for the languages highlighted with an underwave the model generates in the wrong language the majority of the time. Often, this is due to the model generating in a closely related high-resource language. For languages such as Italian and Bulgarian, many inaccurate generations are in English. Similarly, Japanese generations often switch to Chinese when mistakes occur. Languages with more distinctive linguistic features —such as Thai's unique script—facilitate more accurate model generations.

[1]Some generations are incomplete, or their language is undetectable.

2.2 Automatic Methods for Multilingual Factuality Detection

Following the language verification and ruling out examples where the generations are in the wrong languages, the next phase involves the detection of hallucinations in a multilingual environment by assessing the consistency between a target generation and either a reference text or its other generations. This section considers various automatic metrics for detecting hallucinations in long-form generations.

ROUGE The ROUGE metric is employed to assess the token-level similarity between texts. We consider the ROUGE 1 (R1), 2 (R2), L (RL), and Lsum (RLsum) scores of the generated text against the reference.

Named Entity Overlap (NEO) We hypothesize that the sets of named entities in the gold and generated text will differ if there is hallucination in the generation (Nan et al., 2021). We calculate the F1, precision, and recall scores of named entities between the generated and reference text as an estimate for factual hallucinations.

NLI-based Detection Following Manakul et al. (2023) and Elaraby et al. (2023), we adopt the NLI-based zero-shot sentence-level SUMMAC ($SummaC_{zs}$) scoring system (Laban et al., 2021) to evaluate hallucinations. The $SummaC_{zs}$ method was originally developed to gauge the consistency between a summary S and a document D, by segmenting them into sentences $S_1, \ldots, S_N$ and $D_1, \ldots, D_M$ respectively. Aligning with the optimal configuration in Laban et al. (2021), we employ the max operator to compute the score for a sentence. Denote $e_{S_n}^{D_m}$ and $c_{S_n}^{D_m}$ as the entailment and contradiction score for the generated sentence S_n given the reference sentence D_m, respectively.

We define three metrics to quantify verifiable hallucination and one metric to quantify unverifiable hallucination, respectively. At sentence-level detection, for a generated sentence S_i and a reference D, to detect verifiable hallucination, we define the following three metrics: $\mathbf{ENT_{S_i}} = \max_m e_{S_i}^{D_m}$, $\mathbf{CON_{S_i}} = \max_m c_{S_i}^{D_m}$, and $\mathbf{DIFF_{S_i}} = \max_m e_{S_i}^{D_m} - \max_m c_{S_i}^{D_m}$. To detect unverifiable hallucination, we define the following metric:

$$\mathbf{UNV_{S_i}} = 1 - \max(\max_m e_{S_i}^{D_m}, \max_m c_{S_i}^{D_m})$$

When evaluating each of the above hallucination metrics on a generated text $\hat{t}$, we consider two settings as the reference text t:

Reference-based This setting compares $\hat{t}$ against the relevant biographical article in Wikipedia.

Pairwise We generate k samples for each biography. In this setting, we compare $\hat{t}$ against the other generated samples for the same person and calculate the average score across all generations.

3 Experiment Setup

3.1 Dataset

Our curated dataset encompasses 19 languages: English, Spanish, Russian, Indonesian, Vietnamese, Persian, Ukrainian, Swedish, Thai, Japanese, German, Romanian, Hungarian, Bulgarian, French, Finnish, Korean, Italian, and Chinese. Using WikiData, we extract 500 human names that are covered by all of these languages on Wikipedia, based on diverse page view counts from 2022-01-01 to 2023-01-01. For our reference text, we use the Wikipedia API to obtain the full-page content. We detect instances where the LLMs generate text in an incorrect language with langdetect.[2]

3.2 Models and Prompting Details

In our experiments, we deploy the pretrained multilingual BLOOM model as well as the BLOOMZ-mt model, which is fine-tuned with machine-translated prompts (Workshop, 2023). We use nucleus decoding (Holtzman et al., 2020) with $top\_p = 0.9$ and generate five responses per prompt. For each evaluation language, we generate a prompt template with Google Translate. The template in English is *"Tell me a biography of <Name>."*; all templates are in Appendix (Figure 2).

3.3 Detection Metric Details

ROUGE scores are calculated with TorchMetrics[3], and we remove all stopwords before calculating ROUGE-1. Entities are extracted with Spacy's named entity recognizer[4]; we note that this tagger only covers 13 of the 19 languages considered in our experiments. For the NLI-based metric, we finetune the XLMR-large model (Conneau et al., 2020) on the subset of the XNLI dataset (Conneau

[2]APIs: https://query.wikidata.org/, https://pypi.org/project/wikipedia/, and https://pypi.org/project/langdetect/, respectively

[3]https://github.com/Lightning-AI/torchmetrics

[4]https://spacy.io/api/entityrecognizer

et al., 2018) that intersects with the languages used in our experiments. The finetuned model has an average validation accuracy of 85.4% for the nine intersecting languages.

Language	ENT	DIFF	UNV
# examples in correct language > 1,000			
English	0.55	0.38	0.19
French	0.52	0.40	0.15
Chinese	**0.56**	**0.41**	0.21
Spanish	0.46	0.41	0.17
Thai	0.36	0.39	**0.32**
Vietnamese	0.35	0.31	0.00
Indonesian	0.28	0.31	0.09
# examples in correct language < 1,000			
Russian	0.16	0.21	0.11
Japanese	0.37	0.40	0.07
Ukrainian	0.23	0.19	0.17
Bulgarian	0.42	0.32	0.28
Korean	0.05	0.08	-0.01
# examples in correct language < 100			
Finnish	0.09	0.12	0.01
Italian	0.12	0.13	0.14
Persian	0.13	0.15	0.02
German	0.50	0.45	0.11
Romanian	0.00	0.00	0.14
Hungarian	0.24	0.21	0.11
Swedish	0.30	0.27	-0.29

Table 2: The correlation between the reference-based NLI result and the pairwise NLI result across different languages. The languages with underline are covered in the XNLI finetuing dataset. The numbers in yellow have the *p-values* larger than 0.05.

4 Results

In this section, we compare evaluation metrics to estimate hallucinations in our generated biographical corpus (Section 4.1). We then perform a correlation study to test whether the proposed metrics *agree* when hallucination occurs (Section 4.2). We find that traditional, reference-based lexical metrics (i.e., ROGUE and named entity overlap) are likely insufficient to capture hallucinations, as they rarely correlate with stronger NLI-based approaches. We also observe that pairwise metrics detect hallucinations less often in lower-resource languages when compared to reference-based metrics.

4.1 Comparison of Automatic Hallucination Metrics

We first consider how different automatic methods for detecting hallucination perform across languages on the generated biographical data from the BLOOMZ-mt model (Table 3). High-resource languages, such as English, Chinese, Spanish, French, Vietnamese, and Indonesian, exhibit higher recall scores, which suggests that the text generated in these languages has better coverage of the corresponding Wikipedia reference content. This contrasts sharply with lower-resource languages, which demonstrate significantly diminished recall. Interestingly, languages that frequently produce incorrect language outputs (e.g., German and Italian) or often yield empty or incomplete generations (e.g., Swedish and Hungarian) still have relatively high precision scores. While these generations seem to contain few explicit hallucinations, they also exclude many facts from the reference, as indicated by their correspondingly low recall scores.

When considering the DIFF metrics, all languages under scrutiny resulted in negative scores, which was consistent even among the higher-resource languages like English and Chinese. This indicates a tendency towards contradictions in the generated text with their respective reference texts. For the UNV scores, higher and middle-resource languages (ranging from English to Romanian in the table 3) cluster within a similar scope of 0.15 to 0.25. In contrast, low-resource languages that often produce empty or incomplete generations, such as Ukrainian, Persian, Finnish, and Korean, show much higher UNV scores. This implies that the UNV metric is sensitive to incomplete text generations and missing information and may indicate the model's generation errors beyond hallucination. As for the reference-free pairwise metrics, we observe similar trends across different languages (Table 9).

4.2 Correlation Study Across Metrics

In this section, we conduct a correlation analysis to determine how various metrics align in measuring hallucination in multilingual contexts. This includes (1) the correlation between lexical hallucination metrics and NLI-based metrics, (2) the performance of the four reference-based NLI metrics, and (3) the relationship between pairwise metrics and reference-based metrics.

Lexical hallucination metrics do not correlate with NLI-based metrics. Figure 1 shows that

Table 3: Results of different reference-based metrics for the BLOOMZ-mt model. "-" indicates the language is not covered by the Spacy NER tool. All of the ROUGE and Named Entity Overlap (NEO) results are in percentage (%).

Language	R1-F1	R1-P	R1-R	R2-F1	R2-P	R2-R	NEO-F1	NEO-P	NEO-R	DIFF	UNV
High-Resource Languages											
English	1.83	87.58	0.94	0.87	47.38	0.44	4.27	53.41	2.26	-0.60	0.19
Chinese	6.43	57.34	3.76	2.07	23.22	1.17	4.69	35.27	2.79	-0.62	0.21
Spanish	2.77	85.86	1.47	1.35	49.10	0.71	3.28	48.48	1.76	-0.51	0.18
French	2.18	87.78	1.13	1.06	51.41	0.55	4.35	57.41	2.31	-0.54	0.16
Vietnamese	6.82	92.92	4.22	4.10	73.28	2.43	-	-	-	-0.49	0.15
Indonesian	7.51	68.51	4.87	2.36	26.39	1.53	-	-	-	-0.45	0.22
Middle-Resource Languages											
German	0.38	71.34	0.19	0.11	35.22	0.05	0.83	36.06	0.42	-0.65	0.15
Italian	0.50	69.13	0.25	0.13	29.82	0.07	1.00	30.26	0.52	-0.58	0.17
Japanese	0.73	14.62	0.40	0.05	1.50	0.03	0.47	15.52	0.25	-0.72	0.21
Bulgarian	0.16	4.92	0.09	0.01	0.37	0.01	-	-	-	-0.61	0.19
Romanian	1.02	69.75	0.53	0.42	45.79	0.22	0.39	17.47	0.20	-0.29	0.24
Swedish	0.66	86.37	0.33	0.32	77.23	0.16	1.28	45.24	0.66	-0.40	0.63
Low-Resource Languages											
Thai	0.04	1.14	0.02	0.00	0.00	0.00	-	-	-	-0.56	0.38
Russian	0.09	4.69	0.05	0.01	0.28	0.00	0.48	11.28	0.25	-0.58	0.47
Ukrainian	0.04	1.53	0.02	0.00	0.00	0.00	0.70	20.64	0.36	-0.53	0.66
Persian	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-0.50	0.92
Finnish	0.89	37.70	0.46	0.20	10.03	0.10	0.58	23.71	0.30	-0.59	0.91
Korean	0.18	6.58	0.09	0.01	0.88	0.00	0.24	8.48	0.12	-0.53	0.94
Hungarian	0.74	64.74	0.37	0.16	23.23	0.08	-	-	-	-0.53	0.97

in high-resource languages (i.e., English, Chinese, French, Spanish, Vietnamese, and Indonesian), ROUGE-1 and ROUGE-L metrics demonstrate a high degree of correlation, as do the ROUGE metrics and Named Entity Overlap (NEO). However, we generally find no correlation between lexical- and NLI-based metrics, indicating that while both lexical- and NLI-based approaches are commonly proposed as automatic methods for hallucination detection, they do not measure the same deviations from a reference text.

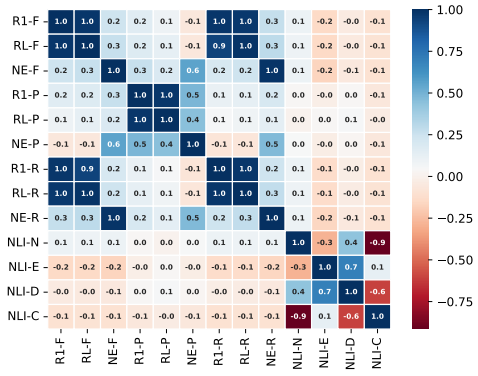

Figure 1: Heat map of the Pearson Correlation between reference-based metrics averaged over high-resource languages. All the P-values are less than 0.05.

Reference-based NLI-based metrics. We also observe interesting trends regarding the relationship between different NLI-based metrics (Figure 1). We find that ENT scores are highly correlated with the DIFF score, indicating that these metrics identify similar artifacts in the text. Surprisingly, no correlation is observed between UNV and CON scores. Moreover, we find a negative correlation between UNV and CON scores. This is because sentences that include verifiable hallucinations likely contradict the reference text. In contrast, sentences with information that is unsubstantiated by the reference (e.g., unverifiable) will be identified as neutral instead.

Pairwise and reference-based metrics correlate in higher-resource languages. For high-resource languages in the XNLI finetuning dataset (English, French, Chinese, Spanish, Bulgarian, and German), we observe a higher correlation in the range of 0.35 to 0.56 for pairwise NLI metrics when it comes to detecting verifiable hallucinations by ENT score (Table 2). This suggests pairwise metrics can identify generated content that deviates from the reference. However, the Pearson Correlation Coefficient shows lower correlation values (in the range of 0.15 to 0.21) compared to pairwise UNV. For lower-resource languages, such as Finnish, Italian, Persian, and others, the correla-

Entity	Generation
	Example of Annotation
Alessandro Del Piero	Gen: Alessandro Del Piero, born on September 28, 1976 in Brescia, Italy, is a former Italian professional football player who served as a forward. Wiki: Alessandro Del Piero, Italian male football player...The old Maldini, who was the head coach of the national team at the time, appointed newcomers Del Piero and Vieri as the main forwards... Comment: There are 4 facts in this sentence, with 1 contradictory hallucination and 1 unverifiable hallucination. The birth date is wrong; Wikipedia doesn't mention the birth place; for the last entity, the evidence indirectly support it.
	Example of Instruction-Conflict Hallucination
Carl Edward Sagan	Gen: Stephen Hawking (January 8, 1942 -) is a famous British physicist and cosmologist.
	Example of Preceding-Context-Conflict Hallucination
Blanca of Castile	Gen: Blanca of Castile (Blanca of Castile), born in 1188 and died in 1252, was the princess of Castile, queen and regent of France. Blanca is a city in the Autonomous Community of Castilla-La Mancha, Spain.

Table 4: Examples of annotation process and types of hallucinations. Sentences highlighted in red represent verifiable hallucinations that contradict evidence found in Wikipedia; those marked in yellow denote unverifiable hallucinations which lack clear evidence for verification and content highlighted in green is considered factual.

tion coefficients with the entailment score are often in the range of 0.00 to 0.30 and not statistically significant, highlighting the challenge of effective hallucination detection in contexts with limited language resources.

5 Human Evaluation of Multilingual Hallucination Metrics

Currently, there are no standard multilingual hallucination detection datasets available on which to compare the considered methods. We, therefore, annotate model generations in English and Chinese; annotations are performed by native speakers of each language. They manually find all verifiable and unverifiable hallucinations by checking if the generation is supported by the Wikipedia reference at both the sentence- and atomic-fact-level. Examples of resulting annotations are shown in Table 4.

5.1 Experimental Setup

Annotators manually label all the seven elements mentioned above in each sentence of the generation. For each sentence, we annotate 1) all relevant evidence sentences from the entire Wikipedia page; 2) the number of total facts in a sentence (N_t); 3) the number of verifiable supported facts (N_{vs}); 4) the number of verifiable contradictory facts (N_{vc}); 5) the number of non-verifiable facts (N_{nv}); 6) yes/no if the generated sentence has a conflict with the instruction (e.g., generating in wrong human entity); 7) yes/no if the generated sentence has a conflict with its preceding generated context. When annotating atomic facts, the annotators also rephrase the atomic facts from a given sentence into simpler single-fact sentences. Annotation statistics can be found in Table 8.

Metrics We compare automatic metric performance with human annotations using correlation and classification. For correlation, we investigate the relationship of the metrics with the support rate (N_{vs}/N_t) for verifiable hallucination detection and with the unverified rate (N_{nv}/N_t) for unverifiable hallucination detection using their Pearson correlations. When considering the classification performance of these metrics, we calculate the Precision-Recall area under the curve (AUC-PR) between the human annotations and the automatic continuous metrics. We convert the human annotations into classification labels by labeling an example as factual only if all its facts are supported by evidence (for verifiable hallucinations); we consider three label thresholds for estimating unverifiable hallucination: $N_{nv} \geq 1$, $N_{nv}/N_t \geq 50\%$, or $N_{nv} = N_t$ (Table 6). We also discretize classification accuracy by setting thresholds for NLI-based metrics with 0.5 for the entailment and contradictory scores and 0 for the difference between these two scores. The thresholds were selected based on the different degrees of tolerance for the proportions of unverifiable hallucinations in a sentence.

We observe distinct patterns for sentence- and atomic-level annotations: sentence-level examples typically encompass multiple facts and require several sentences of evidence, whereas a single sen-

Metric	Sentence-Level Pearson	Sentence-Level AUC_F	Sentence-Level AUC_{NF}	Sentence-Level Accuracy	Atomic-Fact-Level Pearson	Atomic-Fact-Level AUC_F	Atomic-Fact-Level AUC_{NF}	Atomic-Fact-Level Accuracy
Random	-	10.84	82.86	50.00	-	52.11	43.56	50.00
Pairwise								
R2-P.	0.08	19.78	81.48	-	0.10	51.23	44.23	-
RL-P.	0.11	20.04	83.10	-	0.12	52.18	42.83	-
NEO-P.	0.14	17.49	80.84	-	0.09	53.09	45.32	-
DIFF	0.21	38.46	89.49	68.21	0.19	57.46	54.41	56.26
ENT	0.31	40.32	90.86	70.97	0.23	60.71	57.48	53.47
CON	0.11	16.47	80.41	49.13	-0.01	51.49	52.16	50.48
Reference								
R2-P.	0.21	30.05	89.08	-	0.19	53.28	46.25	-
RL-P.	0.17	28.54	85.35	-	0.13	50.31	49.93	-
NEO-P.	0.17	16.15	83.75	-	0.12	57.54	47.51	-
DIFF	0.34	56.11	94.14	75.00	0.31	65.85	60.90	63.19
ENT	**0.49**	**65.32**	**94.96**	**78.97**	0.35	68.00	63.69	64.18
CON	0.08	31.56	87.49	69.77	-0.19	53.18	57.43	59.70

Table 5: Sentence- and atomic-fact-level verifiable hallucination detection in human evaluation. *F* denotes factual examples and *NF* denotes non-factual examples. The numbers in yellow have *p-values* of the correlation is larger than 0.05. Note that the correlation coefficient of CON should be expected as negative.

tence of evidence generally suffices for atomic-fact verification (Table 8). We also note that unverifiable hallucinations are much more common than supported or contradictory instances of factuality.

5.2 Analysis Findings

NLI entailment outperforms lexical metrics on sentence-level verification. We observe low correlation coefficients between lexical metrics like ROUGE-2 (R2) and Named Entity Overlap (NEO) with the human-annotated support rate (SR; Table 5). Moreover, AUC-PR for these metrics exhibits minor improvements over the random classification baseline, particularly for non-factual examples (*NF*), underscoring the limit of lexical metrics for accurately detecting factual hallucinations.

In contrast, we observe that NLI-based metrics, particularly ENT, outperform other metrics in detecting verifiable hallucinations at the sentence level due to its high correlation with different measures of human verification (Table 5). These results corroborate the NLI metric results in English in Manakul et al. (2023). Furthermore, the DIFF score shows comparable performance, whereas CON alone does not demonstrate the same level of effectiveness.

Pairwise metrics underperform reference-based metrics. When compared against human annotations, pairwise metrics underperform the reference-based ones across the board: their Pearson correlations, AUC_F, AUC_{NF}, and accuracies are all worse than the metrics using references, and this holds at both the sentence- and atomic-fact-levels (Table 5). Notably, pairwise lexical-based metrics like R2-P, RL-P, and NEO-P show relatively weaker performance than reference-based metrics, indicating a limited capability in accurately evaluating hallucinations. However, we note that pairwise NLI-based metrics, particularly the entailment and difference scores, demonstrate significantly better performance than other pairwise settings (though they still underperform comparable reference metrics). This suggests pairwise NLI-based approaches may remain useful in hallucination detection settings where references are unavailable.

Metric	Pearson	AUC_U	AUC_{NU}	Accuracy
At least one unverifiable fact				
$Random_{Sent}$	-	25.45	68.72	50.00
UNV_{Sent}	0.02	22.25	77.50	45.07
$Random_{Atom}$	-	71.93	17.88	50.00
UNV_{Atom}	0.12	79.26	27.36	57.74
At least 50% unverifiable facts				
$Random_{Sent}$	-	42.02	44.01	50.00
UNV_{Sent}	-	47.25	48.88	57.36
100% unverifiable facts				
$Random_{Sent}$	-	25.01	67.27	50.00
UNV_{Sent}	-	26.79	69.75	59.35

Table 6: Results of unverifiable hallucination detection in human evaluation. All the *p-values* of the correlation is less than 0.05. *Sent* denotes sentence-level detection, and *Atom* denotes atomic-fact-level detection. *U* denotes unverifiable hallucination, and *NU* denotes non-unverifiable (i.e., factual or verifiable).

NLI-based metrics struggle to detect unverifiable hallucination. Comparison of NLI-based metrics against human annotators on unverifiable hallucination detection reveals significant challenges. Automatic metrics show marginal improvements over random classification and low correlation with the human *unverifiable* annotation. This indicates difficulties in discerning the nuanced differences between evidence and generated text (e.g., the first example in Table 4).

NLI-based metrics fails to evaluate factuality on simple atomic facts. Unlike sentence-level verification, NLI metrics encounter significant challenges in accurately verifying the factuality of simple atomic facts (Table 5). Metrics such as AUC-PR and accuracy demonstrate a marked decrease in the effectiveness of NLI-based metrics on atomic facts. This aligns with Luo et al. (2022), which also highlighted similar limitations in English NLI metrics for verification, and these findings underscore the need for further refinement in NLI approaches to handle atomic factuality. This finding underscores the potential of alternative metrics in addressing the limitations of current NLI approaches, especially in the critical area of atomic factuality evaluation.

6 Related Work

6.1 Factuality Hallucination Detection

Detecting hallucinations in LLMs is crucial for ensuring the reliability of machine-generated content. One line of work utilizes the uncertainty of LLM generation. Some studies analyze the probability of tokens (Mielke et al., 2022; Kadavath et al., 2022; Varshney et al., 2023), whereas other methods evaluate the inconsistency between multiple generated samples, including NLI-based approaches (Elaraby et al., 2023; Manakul et al., 2023). Other work utilizes external retrieved evidence as reference text and verifies if the generated text is supported by the reference text (Chern et al., 2023; Min et al., 2023). However, such factuality hallucination evaluations are limited in multilingual settings. Existing works in this area include prompting methods (Ahuja et al., 2023) and machine translation (Lai et al., 2023). Our work aims to bridge the gap between English and multilingual generation evaluation settings.

6.2 Task-specific Multilingual Hallucination

There are also some existing multilingual hallucination mitigation methods for specific tasks, such as abstractive summarization and machine translation. Qiu et al. (2023) introduces the mFACT metric to evaluate the faithfulness of non-English summaries by transferring English metrics to target languages via machine translation. To enhance the faithfulness of multilingual summarization, Aharoni et al. (2022) leverages factual consistency models, highlighting the importance of addressing hallucinations in languages beyond English. In the realm of neural machine translation (NMT), Dale et al. (2022) emphasizes the role of the source's contribution to detect and alleviate hallucinations, while another line of works (Lee et al., 2019; Raunak et al., 2021) analyze the NMT's susceptibility to hallucinations and impact of data-generation processes like back-translation on hallucination. Xu et al. (2023) takes a deep dive into the internal workings of neural sequence generation models. Together, these works underline the pervasive challenge of hallucinations in multilingual tasks.

7 Conclusion

This study investigates the effectiveness of automatic metrics for detecting factual hallucinations in multilingual contexts. We document the limitations of traditional lexical metrics, such as ROUGE and Named Entity Overlap, in detecting factuality hallucinations in these multilingual settings. Specifically, our work reveals that while NLI-based metrics show promise in high-resource languages at the sentence level, their effectiveness diminishes when applied to simpler atomic facts. Moreover, the reliability of NLI-based metrics is tied to the performance of NLI models, which poses a significant challenge in languages with lower resources. Therefore, our work highlights that the ability to detect hallucinations effectively in a multilingual setting is intrinsically linked to the availability and quality of linguistic resources in a given langauge. Therefore, automatically detecting factuality hallucinations in lower-resource languages remains a significant gap in current NLP methods.

Limitations

This study focuses on text generation and hallucination in a specific setting (namely, generating biographical summaries with the BLOOM-mt model)

in order to perform a controlled study on how different automatic metrics detect factual hallucinations. It remains an open question whether these findings hold in other generation settings, particularly when there is less reliance on factual knowledge (such as story generation).

Additionally, portions of our experimental setup rely on automatic methods. Specifically, we use machine translation to construct the prompt templates, which may introduce noise into the prompts. Furthermore, due to the unavailability of native speakers for other languages, our human evaluation and comparison against automated metrics is limited to Chinese and English.

Finally, we note that we do not include more complex, LLM-based methods of hallucination detection. The efficacy of these methods is directly linked to the performance of these models in the target language, which remains an open question for popular LLMs on many of the languages we consider. We therefore limit our analysis to simpler lexical- and NLI-based approaches.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. arXiv preprint arXiv:2212.10622.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. arXiv preprint arXiv:2303.12528.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. arXiv preprint arXiv:2212.08597.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2023. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. arXiv preprint arXiv:2304.13803.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 13171–13189, Singapore. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. Advances in Neural Information Processing Systems, 35:31809–31826.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation.

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

Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but challenging: Natural language inference models fail on simple sentences. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3449–3462, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. Transactions of the Association for Computational Linguistics, 10:857–872.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2727–2733, Online. Association for Computational Linguistics.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. arXiv preprint arXiv:2305.13632.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1172–1183, Online. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. arXiv preprint arXiv:2307.03987.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. Transactions of the Association for Computational Linguistics, 11:546–564.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

A Qualitative Analysis of Challenging Cases in Annotation

We identify four challenging categories of hallucination detection for annotators and NLI metrics (Table 7).

- *Inferred*: Implicit fact connections between generation and evidence.
- *Subjective*: Generation contains subjective content, which is challenging for both human annotators and fact-based NLI models.
- *Nuanced Difference*: There are subtle distinctions between evidence and generated text, which is often missed by surface-level text classification in NLI models.
- *Temporal Information*: Generation contains time-sensitive information, which requires models to have an understanding of temporal context.

Each category presents unique difficulties in determining the factuality of generated content with its evidence source.

Type	Example
Inferred	Gen: Alessandro Del Piero is a former Italian professional football player and a forward. Wiki: Maldini, who was the head coach of the national team, appointed Del Piero as the main forward.
Subjective	Gen: Frida Kahlo is widely regarded as the most influential painter of the 20th century.
Nuanced Difference	Gen: Louis Pasteur is known as the "Father of Modern Microbiology". Wiki: ... has been honored as the "father of microbiology"...
Temporal Information	Gen: Michelle Bachelet is currently the President of Chile. Wiki: She served as President of Chile from 2006 to 2010 and from 2014 to 2018...

Table 7: Categories of special case in annotation.

- *Inferred* connection between generated content and evidence is one of the biggest challenges for both annotators and NLI models, since they need to infer the relationship or the factual basis that links them. This requires a deep understanding of context and the ability to draw inferences from potentially sparse or indirect evidence.

- *Subjective* content in generations poses a significant challenge because it introduces personal opinions, emotions, or interpretations that are inherently difficult to verify against factual evidence. For human annotators, this can lead to variability in judgments based on personal biases or interpretations. For NLI models, which are primarily designed for fact-based analysis, handling subjective content requires advanced understanding of sentiment, opinion, and cultural context, areas where current models may fall short.

- *Nuanced difference* between evidence and generated text highlight the limitations of surface-level text classification approaches in NLI models. Detecting nuanced differences demands a granular analysis of semantics, requiring models to understand context, synonyms, and slight variations in meaning. This challenge underscores the need for more sophisticated NLI models capable of deep semantic analysis and the importance of training annotators to pay attention to detail and understand the significance of minor discrepancies.

- *Time-sensitive information* introduces complexity because it requires both annotators and models to have an understanding of temporal context and the ability to evaluate statements within the correct time frame. This can be particularly challenging when information changes over time, requiring up-to-date knowledge and the ability to discern the relevance of temporal qualifiers in text. For NLI models, this underscores the need for dynamic knowledge bases and the ability to reason about time, which are areas where current models may lack proficiency.

Overall, these challenges highlight the complexities involved in hallucination detection and the need for advanced capabilities in human annotators and NLI models.

B Additional Results

We show the pairwise metric results in Table 9. We observe similar trends as the reference-based metrics. Also, the average statistics of annotation result are shown in Table 8.

Metric	Sent-Level	Atomic-Level
# Examples	111	102
# Words	46.21	10.21
# Evidence	2.17	1.00
# Total Facts	4.76	1.00
Support Rate	0.35	0.29
Contradictory Rate	0.15	0.24
Unverified Rate	0.50	0.47
Instruction-conflict Rate	0.03	0.07
Context-conflict Rate	0.13	0.06

Table 8: Average statistics of Chinese and English annotation data.

C Generation Prompt Templates

We present the full set of prompt templates for all languages from Section 3 in Figure 2.

Table 9: Results of different pairwise consistency metrics for the BLOOMZ-mt model. "-" indicates they have no coverage for the NER tool we use. All of the ROUGE and Named Entity Overlap results are in percentage (%).

Language	R1_F1	R1_P	R1_R	R2_F1	R2_P	R2_R	NEO_F1	NEO_P	NEO_R	DIFF	UNV
English	12.03	14.21	14.71	5.91	7.10	7.37	4.27	53.41	2.26	-0.25	0.57
Chinese	7.57	8.56	8.55	4.00	4.55	4.55	4.69	35.28	2.79	-0.29	0.57
Spanish	12.49	14.45	15.05	6.21	7.31	7.68	3.28	48.48	1.76	-0.22	0.52
German	4.42	10.47	4.48	1.98	5.51	2.04	0.83	36.06	0.42	-0.33	0.56
Russian	0.15	0.19	0.13	0.00	0.00	0.00	0.48	11.28	0.25	-0.09	0.55
Indonesian	4.74	6.23	6.52	1.59	2.13	2.27	-	-	-	-0.16	0.68
Vietnamese	11.60	14.36	15.91	6.26	7.91	8.91	-	-	-	-0.32	0.56
Persian	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-0.31	0.52
Ukrainian	0.00	0.00	0.00	0.00	0.00	0.00	0.70	20.64	0.36	0.14	0.42
Swedish	10.04	14.69	10.82	8.53	13.23	9.74	1.28	45.24	0.66	-0.20	0.54
Thai	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.19	0.68
Japanese	0.47	0.75	0.58	0.08	0.13	0.09	0.47	15.52	0.25	-0.07	0.56
Romanian	12.21	13.95	12.68	8.57	9.40	8.70	0.39	17.47	0.20	-0.30	0.44
Hungarian	0.18	0.25	0.46	0.00	0.00	0.00	-	-	-	-0.18	0.58
Bulgarian	0.30	0.44	0.28	0.05	0.08	0.06	-	-	-	-0.02	0.96
French	12.02	14.03	14.61	6.26	7.43	7.76	4.35	57.41	2.31	-0.05	0.92
Finnish	0.46	0.47	0.61	0.17	0.19	0.19	0.58	23.71	0.30	-0.04	0.94
Korean	0.19	0.26	0.17	0.00	0.00	0.00	0.24	8.48	0.12	0.25	0.63
Italian	4.79	7.85	5.08	2.71	4.15	2.77	1.00	30.26	0.52	-0.01	0.97

Language	Prompt Template
EN	Tell me a biography of {}.
ZH	给我写一篇关于{}的传记。
ES	Dime una biografía de {}.
DE	Erzähl mir eine Biografie von {}.
RU	Расскажите мне биографию {}.
ID	Ceritakan tentang biografi {}.
VI	Hãy cho tôi biết tiểu sử của {}.
FA	بیوگرافی {} را به من بگویید.
UK	Розкажіть мені біографію {}.
SV	Berätta en biografi om {}.
TH	บอกเล่าประวัติของ {}.
JA	{} の略歴を教えてください。
RO	Spune-mi o biografie a lui {}.
HU	Mondja el {} életrajzát.
BG	Разкажи ми биография на {}.
FR	Dites-moi une biographie de {}.
FI	Kerro minulle henkilön {} elämäkerta.
KO	{}의 약력을 알려주세요.
IT	Raccontami una biografia di {}.

Figure 2: Prompt templates of all languages used in generating biography. {} represents human names.