Multi-Hall-SA: A Cross-lingual Benchmark for Multi-Type Hallucination Detection in Low-Resource South African Languages

Anonymous ACL submission

Abstract

Hallucinations generated by Large Language Models (LLMs) pose significant challenges for their application to low-resources languages. We present Multi-Hall-SA, a cross-lingual benchmark for hallucination detection spanning English and four low-resource South African languages: isiZulu, isiXhosa, Sepedi, and Sesotho. Derived from government texts, this benchmark categorizes hallucinations into four types: temporal shifts, entity errors, numerical inaccuracies, and location mistakes. Our cross-lingual alignment methodology enables direct performance comparison between high-resource and low-resource languages, revealing significant gaps in detection capabilities. Evaluation across four state-of-the-art models shows they detect up to 23.6% fewer hallucinations in South African languages compared to English. Knowledge augmentation substantially reduces this disparity, decreasing cross-lingual performance gaps by 59.4% on average. Beyond introducing a new resource for low-resource languages, Multi-Hall-SA provides a systematic framework for evaluating and improving factual reliability across linguistic boundaries, advancing more inclusive and equitable AI development.

1 Introduction

Large Language Models (LLMs) have transformed natural language processing, yet their tendency to generate hallucinations (false or unsupported information) poses significant challenges, particularly for low-resource languages (Maynez et al., 2020; Filippova, 2020; Zhou et al., 2021). This challenge is especially acute for African languages where limited training data and computational resources increase hallucination frequency and complicate detection efforts (Xu et al., 2023; Raunak et al., 2021). In critical domains such as healthcare, education, and public communication, these risks are amplified, as misinformation can have severe societal consequences (Maynez et al., 2020; Falke et al., 2019).

This challenge is particularly pressing for South African languages which, despite serving millions of speakers and holding official status, remain underserved by current NLP technologies. To address this critical gap, we present **Multi-Hall-SA**, a multilingual hallucination detection benchmark derived from government sources across four major South African languages: isiZulu, isiXhosa, Sepedi, and Sesotho.

Multi-Hall-SA advances beyond existing hallucination detection approaches through a novel taxonomy specifically designed for low-resource African languages. Our framework identifies and categorizes four distinct types of hallucinations: entity-based, temporal, numerical, and location-based.

By leveraging these high-quality sources, we ensure the benchmark's reliability while maintaining cultural and linguistic appropriateness. A distinctive feature of Multi-Hall-SA is its **cross-lingual alignment methodology**, which enables direct comparison of model performance between high-resource (English) and low-resource languages. This parallel structure across languages provides insights into how hallucination detection capabilities vary across linguistic boundaries, revealing systematic disparities that remain hidden in monolingual benchmarks.

Our work contributes to both hallucination detection and low-resource language processing by: (1) providing a structured framework for categorizing and detecting multiple hallucination types, (2) creating a parallel dataset for English and four South African languages, (3) establishing a methodology for generating controlled hallucinations suitable for cross-lingual evaluation, and (4) introducing a knowledge-augmented evaluation approach that substantially reduces cross-lingual performance gaps.

Our extensive evaluations reveal significant cross-lingual performance gaps, with models detecting up to 23.6% fewer hallucinations in South African languages compared to English. Knowledge augmentation emerges as a useful mitigation strategy, reducing this gap by 59.4% on average across all languages and models. These findings highlight the importance of developing specialized techniques for low-resource languages to ensure reliable hallucination detection across diverse linguistic contexts.

2 Related Work

Recent advancements in natural language generation have brought hallucination detection to the forefront of NLP research. We examine current approaches to hallucination detection, mitigation strategies, and their limitations in low-resource contexts.

2.1 Hallucination Detection Frameworks

Hallucination detection methods have evolved from simple overlap metrics to sophisticated neural approaches (Pagnoni et al., 2021; Dhingra et al., 2019). Reference-dependent methods utilize ground truth comparisons to identify inconsistencies, exemplified by PARENT and PARENT-T (Dhingra et al., 2019; Wang et al., 2020b), which evaluate faithfulness by measuring alignment with both source documents and references. In summarization, specialized metrics like FEQA (Durmus et al., 2020), QAGS (Wang et al., 2020a), and QuestEval (Scialom et al., 2021) use question generation and answering techniques.

Reference-free methods offer solutions when ground truth is unavailable, using uncertainty quantification (Huang et al., 2025b; Manakul et al., 2023) and internal consistency checks (Elaraby et al., 2023; Raj et al., 2022). Recent advancements include self-consistency approaches (Manakul et al., 2023), fine-grained atomic evaluation (Min et al., 2023), task-specific benchmarks (Li et al., 2023), taxonomic frameworks (Huang et al., 2025a), and multimodal extensions (Gunjal et al., 2024).

These approaches, while effective for high-resource languages, remain largely unevaluated in low-resource contexts. Our work addresses this gap by providing a benchmark specifically designed for cross-lingual evaluation with controlled hallucination types.

2.2 Mitigation Strategies and Applications

The field has developed various hallucination mitigation strategies across NLP applications. For abstractive summarization, researchers have proposed architectural modifications (Aralikatte et al., 2021; Cao et al., 2018; Li et al., 2018) and contrastive learning techniques (Cao and Wang, 2021). Post-processing approaches (Cao et al., 2020; Dong et al., 2020) have shown effectiveness, though their computational requirements limit application in resource-constrained environments.

Dialogue systems have benefited from knowledge grounding (Shuster et al., 2021) and controlled generation (Rashkin et al., 2021), while machine translation has explored corpus filtering (Raunak et al., 2021), factorized divergence (Briakou and Carpuat, 2021), and specialized training objectives (Wang and Sennrich, 2020). These approaches often rely on extensive data and computational resources, limiting their applicability in low-resource settings.

2.3 Challenges in Low-Resource Contexts

The intersection of low-resource languages and hallucination detection presents unique challenges that remain largely unaddressed (Xu et al., 2023; Raunak et al., 2021). Existing benchmarks predominantly focus on high-resource languages, creating a gap in understanding hallucination patterns in low-resource contexts. This disparity is particularly evident for African languages, where limited NLP resources compound detection challenges.

Prior work has primarily focused on data augmentation (Xu et al., 2023) and cross-lingual transfer learning (Raunak et al., 2021) but lacks systematic evaluation frameworks. Recently proposed hallucination detection benchmarks like HaluEval (Li et al., 2023), FactScore (Min et al., 2023), and SelfCheckGPT (Manakul et al., 2023) offer improved evaluation capabilities but overlook cross-lingual assessment, especially for low-resource languages.

Multi-Hall-SA addresses these limitations by introducing specialized techniques for low-resource African languages. Unlike previous approaches requiring extensive training data (Feng et al., 2020; Zhou et al., 2021), our framework operates effectively within low-resource constraints. By focusing on isiZulu, isiXhosa, Sepedi, and Sesotho, we contribute to developing more inclusive NLP technologies while introducing a structured taxonomy that enables precise identification of hallucination types most susceptible to cross-lingual performance gaps.

3 Methodology

3.1 Benchmark Overview

We present Multi-Hall-SA, a novel multilingual benchmark for hallucination detection across English and four South African languages: isiZulu, isiXhosa, Sepedi, and Sesotho. The benchmark enables rigorous evaluation of hallucination detection capabilities in cross-lingual, low-resource settings through two distinctive aspects: (1) cross-lingual alignment, where each hallucination instance exists in parallel across language pairs, enabling direct comparison between high-resource and low-resource languages; and (2) controlled hallucination typology across four distinct categories (temporal, entity, numerical, and location errors), enabling fine-grained analysis of model performance.

3.2 Data Sources and Model Verification

We collect parallel documents from the South African government services portal,[1] which provides information across multiple domains including services for residents, organizations, foreign nationals, and online services. These domains cover topics from education and driving licenses to business procedures and citizenship requirements, providing diverse content for our benchmark.

Before implementing our benchmark creation pipeline, we conducted preliminary evaluations to verify the multilingual capabilities of candidate models. We tested Claude-3.7-Sonnet and GPT-4o on manually translated isiZulu and Sepedi versions of CommonsenseQA and OpenBookQA obtained from Ralethe and Buys (2025). Both models obtained perfect performance (100% accuracy) on both languages, confirming their suitability for benchmark generation. More details are given in Appendix A

3.3 Benchmark Generation Pipeline

The Multi-Hall-SA benchmark generation pipeline consists of two main phases: (1) aligned fact extraction and (2) controlled hallucination generation, as illustrated in Figure 1.

3.3.1 Aligned Fact Extraction

A key technical challenge is ensuring semantic alignment between facts across languages. Our approach uses parallel processing to extract semantically equivalent facts across language pairs by simultaneously considering both languages during extraction. The system processes English and target-language texts with explicit instructions to identify statements present in both texts.

[1] https://www.gov.za/services

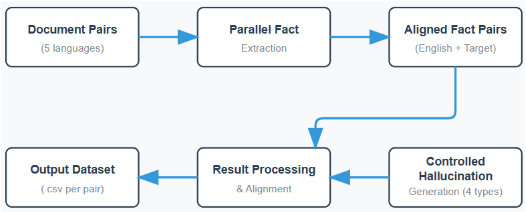

Figure 1: Processing architecture for Multi-Hall-SA benchmark generation

This approach ensures semantic alignment through three mechanisms: (1) explicit cross-lingual verification, requiring that extracted facts must be present in both languages; (2) structural alignment, maintaining identical fact counts across languages; and (3) preservation of original language characteristics without translation artifacts. The system outputs numbered fact pairs with each English statement followed by its semantic equivalent in the target language. Detailed prompt templates are provided in Appendix B.

3.3.2 Controlled Hallucination Generation

For hallucination generation, we implement a controlled modification strategy that systematically alters specific information types while preserving overall statement structure. For each fact pair, we generate four hallucinated versions corresponding to our taxonomy:

1. **Temporal modifications** alter dates or time periods while preserving event relationships (e.g., changing "established in 2001" to "established in 1989")

2. **Entity alterations** replace organizations or persons with plausible but incorrect alternatives (e.g., substituting "Department of Home Affairs" with "Department of Social Development")

3. **Numerical adjustments** modify quantities or statistics while maintaining plausibility (e.g., changing contribution rates from 2% to 5%)

4. **Location substitutions** replace geographical references with incorrect locations within the

Type	Example (English / Target Language)
Temporal	**Original**: The UIF must be claimed within six months of becoming unemployed. **Hallucinated**: The UIF must be claimed within two years of becoming unemployed. **isiZulu Hallucinated**: I-UIF kumele ifakwe singakapheli iminyaka emibili uthola ukungasebenzi.
Entity	**Original**: The Department of Home Affairs issues identity documents. **Hallucinated**: The Department of Social Development issues identity documents. **Sepedi Hallucinated**: Kgoro ya Tlhabollo ya Leago e ntšha dipampiri tša boitsebišo.
Numerical	**Original**: Employers and employees each contribute 1% of the employee's salary to the UIF. **Hallucinated**: Employers and employees each contribute 3.5% of the employee's salary to the UIF. **isiXhosa Hallucinated**: Abaqashi nabasebenzi banikezela nge-3.5% ngabanye kwimali yomvuzo womsebenzi kwi-UIF.
Location	**Original**: SASSA offices in Pretoria process social grant applications. **Hallucinated**: SASSA offices in Durban process social grant applications. **Sesotho Hallucinated**: Diofisi tsa SASSA tse Durban di sebetsa dikopo tsa dithuso tsa mmuso.

Table 1: Example hallucinations from the Multi-Hall-SA benchmark. Each row shows an original statement in English, its hallucinated version, and the corresponding hallucinated statement in one of the target languages, demonstrating the parallel nature of hallucination generation.

same context (e.g., shifting from "Pretoria" to "Cape Town")

Detailed prompting strategies are provided in Appendix B.

3.4 Dataset Structure

Each entry in the Multi-Hall-SA benchmark contains a source fact index, and hallucination category, followed by the original and hallucinated versions in both English and the target language. This structure enables both monolingual and cross-lingual evaluation across semantically equivalent content. Table 1 provides examples of each hallucination type from our benchmark, illustrating how controlled modifications preserve cross-lingual alignment. This approach ensures both control over hallucination types and cross-lingual alignment, as each hallucination is generated in parallel across languages.

4 Experimental Setup

Our study systematically evaluates large language models' capabilities in detecting hallucinations across multiple languages, specifically comparing performance between English and four South African languages. We aim to establish benchmark metrics, investigate performance variations by hallucination type, and analyze cross-lingual detection discrepancies.

4.1 Evaluation Scenarios

We implement two distinct evaluation scenarios to comprehensively assess cross-lingual hallucination detection capabilities:

4.1.1 Zero-shot Hallucination Detection

The first scenario tests models' inherent ability to detect hallucinations across languages without additional context. This approach uses zero-shot prompting, where models receive only the statement to be evaluated and instructions to determine if it contains factual errors. For non-English statements, minimal language context is provided to inform the model about the language being processed. This baseline evaluation establishes each model's core capability in cross-lingual hallucination detection without external support.

4.1.2 Knowledge-augmented Evaluation

The second scenario enhances models with relevant factual information retrieved from a knowledge base. This approach simulates real-world scenarios where models have access to retrieval systems that provide contextual knowledge. For each statement, we retrieve relevant semantic triples from existing multilingual knowledge bases, which are provided in the same language as the statement being evaluated, enabling assessment of how external knowledge affects hallucination detection across languages.

4.2 Models and Implementation

We evaluate four state-of-the-art language models with varying architectures and sizes: Gemma 3 (12B), Aya-101 (11B), Llama 3.1 (8B), and T0++ (11B). The latter is an instruction-tuned model from the BigScience project based on the T5 architecture.

4.3 Evaluation Metrics and Analysis Methodology

We use a set of metrics to evaluate hallucination detection performance across languages. In addition to per-language classification, we also use a number of cross-lingual discrepancy metrics.

- **Standard classification metrics**: Accuracy, precision, recall, and F1 score provide baseline performance assessment for each model and language.
- **Missed hallucination rate**: The percentage of actual hallucinations that the model correctly identifies in English but fails to detect in the target language.
- **False hallucination rate**: The percentage of factual statements that the model correctly identifies in English but incorrectly flags as hallucinations in the target language.
- **Overall discrepancy rate**: The proportion of statements where a model's prediction differs between English and the target language for the same semantic content.

These metrics enable comprehensive analysis of how model performance varies across languages and hallucination types, with particular focus on identifying systematic disparities in detection capabilities.

For cross-lingual performance analysis, we calculate the average performance gap between English and each target language as the difference in F1 scores. This gap is reported both in absolute percentage points and as a relative percentage of the English performance to quantify the disparity magnitude.

For knowledge augmentation experiments, we measure both absolute performance (F1 scores) and relative improvement ($\Delta\%$), calculated as $(F1_{augmented} - F1_{base})/F1_{base} \times 100\%$. This enables quantification of the differential impact of knowledge augmentation across languages. Similarly, we calculate reduction in missed hallucination rates as $(Rate_{base} - Rate_{augmented})/Rate_{base} \times 100\%$ to measure how effectively knowledge augmentation improves cross-lingual consistency.

For hallucination type analysis, we separate the evaluation data into four subsets corresponding to our taxonomy (temporal, entity, numerical, and location). We calculate F1 scores for each model on each subset, both for English and target languages (reported as the average across all four South African languages). This enables identification of which hallucination types are most challenging across languages and which benefit most from knowledge augmentation.

Model	Acc.	P	R	F1
Gemma 3 (12B)	78	81	73	**76**
Aya-101	74	76	68	71
T0++	69	72	61	65
Llama 3.1 (8B)	64	67	54	59

Table 2: Overall hallucination detection performance across models (averaged across all languages) reporting accuracy, precision, recall, and F1 as percentages.

Model	EN	ZU	XH	NSO	ST
Gemma 3 (12B)	**86.4**	75.1	78.1	71.3	73.2
Aya-101	78.1	70.3	73.2	67.2	69.1
T0++	76.2	64.4	68.2	59.1	62.3
Llama 3.1 (8B)	72.3	55.4	59.2	51.2	53.4
Avg. Gap	—	-11.9	-8.5	-16.1	-13.7

Table 3: Hallucination detection F1 (%) scores performance per language. The average gap in performance between English and each of the other langauges are also given.

4.4 Experimental Conditions

We implement two experimental conditions:

Baseline Evaluation (Zero-shot): Models are provided only with the statement to evaluate and minimal language context for non-English statements. This establishes each model's inherent cross-lingual hallucination detection capabilities without external support.

Knowledge-augmented Evaluation: Models are provided with relevant factual information retrieved from a knowledge base before evaluating each statement. We utilize the cross-lingual knowledge bases developed by Ralethe and Buys (2025), which provide parallel semantic triples across English and South African languages projected using their LeNS-Align methodology. These knowledge bases, derived from ConceptNet and DBpedia, were specifically designed for low-resource South African languages. For each statement, we retrieve up to 5 relevant triples using a two-hop retrieval process detailed in Appendix C.

4.5 Prompting and Evaluation Protocol

We implement zero-shot prompting approaches to evaluate models' ability to detect hallucinations without specific examples. The prompt template

Model	ZU	XH	NSO	ST	Avg
Gemma 3 (12b)	4.3	3.9	5.1	4.7	4.5
Aya-101	**3.8**	**3.5**	**4.6**	**4.2**	**4.0**
T0++	4.6	4.2	5.4	5.0	4.8
Llama 3.1 (8B)	5.3	4.8	6.1	5.7	5.5
Avg	4.5	4.1	5.3	4.9	4.7

Table 4: False hallucination rates by model and language.

Model	ZU	XH	NSO	ST	Avg
Gemma 3 (12b)	16.7	14.5	21.6	18.4	17.8
Aya-101	**12.8**	**11.3**	**19.7**	**16.9**	**15.2**
T0++	19.3	17.8	25.2	23.6	21.5
Llama 3.1 (8B)	22.5	21.3	27.8	25.2	24.2
Avg	17.8	16.2	23.6	21.0	19.7

Table 5: Overall cross-lingual discrepancy rates by model and language.

Model	ZU	XH	NSO	ST	Avg
Gemma 3 (12B)	17.2	15.9	21.3	18.7	18.3
Aya-101	**13.8**	**12.5**	**17.9**	**15.3**	**14.9**
T0++	21.4	19.7	29.8	25.1	24.0
Llama 3.1 (8B)	26.8	24.3	35.7	31.2	29.5
Avg	19.8	18.1	26.2	22.6	21.7

Table 6: Missed hallucination rates by model and language

includes a system message defining the assistant's role as an expert at identifying factual errors, followed by instructions to determine if the statement contains hallucinations.

For knowledge-augmented evaluations, we modify this template to include retrieved knowledge triples in the same language as the statement being evaluated. Full prompt templates are detailed in Appendix D.

For each model, language, and condition, we evaluate the complete benchmark dataset of 3,500 statements, comprising both factual statements (to test for false positives) and statements with introduced errors across all four hallucination types (to test for true positives). All evaluations use deterministic generation settings (temperature = 0.0) for reproducibility. Model responses are constrained to binary classifications ("FACTUAL" or "HALLUCINATION"), enabling automated evaluation and analysis of cross-lingual discrepancies. The complete evaluation implementation details, including API configurations and processing architecture, are documented in Appendix E.

5 Results

We present an analysis of hallucination detection performance across models, languages, and experimental conditions, examining four key aspects: overall model performance, cross-lingual detection disparities, knowledge augmentation impact, and performance variations by hallucination type.

5.1 Overall Performance Across Models

In our baseline evaluation (Table 2), we observe significant variation in hallucination detection performance across models. Gemma 3 demonstrates the strongest overall performance with an average F1 score of 76.0% across all languages, followed by Aya-101 (71.0%), T0++ (65.0%), and Llama 3.1 (59.0%). Precision scores consistently exceed recall across all models, indicating models are more likely to miss hallucinations (false negatives) than to incorrectly flag factual statements (false positives).

5.2 Cross-Lingual Performance Analysis

The cross-lingual analysis (Table 3) reveals a consistent performance gap between English and target languages across all models. English detection performance significantly exceeds that of all target languages, with isiXhosa showing the smallest gap (average of 8 percentage points) and Sepedi exhibiting the largest (average of 15 percentage points). This suggests that linguistic proximity to high-resource languages may influence hallucination detection capabilities.

To understand the nature of these performance gaps, we examine cross-lingual discrepancies (cases where models make different predictions between English and the target language for the same semantic content). Table 5 shows that overall discrepancy rates range from 11.3% (Aya-101 on isiXhosa) to 27.8% (Llama 3.1 on Sepedi), with an average of 19.7% across all models and languages. Aya-101 demonstrates the most cross-lingual consistency with the lowest average discrepancy rate (15.2%), while Llama 3.1 shows the highest inconsistency (24.2%).

Further analysis reveals a striking asymmetry in the direction of these discrepancies. As shown in Table 4, the false hallucination rate (cases where models classify factual statements as hallucinations in the target language but correctly as factual in English) is relatively rare, averaging just 4.7% across all models and languages. In contrast, Table 6 demonstrates that the missed hallucination rate (cases where models correctly identify hallucinations in English but miss them in the target

Model	Setup	EN (%)	ZU (%)	XH (%)	NSO (%)	ST (%)
Gemma 3	Base	86	75	78	71	73
	+Know	**91**	**87**	**89**	**85**	**86**
Aya-101	Base	78	70	73	67	69
	+Know	83	79	81	77	78
T0++	Base	76	64	68	59	62
	+Know	82	77	80	74	76
Llama 3.1	Base	72	55	59	51	53
	+Know	76	64	68	62	63
Avg	Base	78	66	70	62	64
	+Know	83	77	80	75	76

Table 7: Impact of knowledge augmentation on hallucination detection (F1 scores)

Model/Setup	English				Target Language Avg			
	Temp	Entity	Num	Loc	Temp	Entity	Num	Loc
Gemma 3	0.85	0.84	**0.89**	0.83	0.72	0.68	**0.83**	0.71
Gemma 3+Know	0.89	**0.93**	**0.93**	0.91	0.85	0.87	**0.89**	0.86
Aya-101	0.77	0.76	**0.83**	0.75	0.69	0.64	**0.76**	0.67
Aya-101+Know	0.81	**0.86**	**0.87**	0.84	0.77	0.78	**0.82**	0.78
T0++	0.75	0.74	**0.82**	0.73	0.63	0.57	**0.72**	0.60
T0++ +Know	0.80	**0.85**	**0.86**	0.83	0.76	0.79	**0.80**	0.77
Llama 3.1	0.71	0.70	**0.77**	0.69	0.55	0.49	**0.64**	0.52
Llama 3.1+Know	0.74	**0.78**	**0.79**	0.76	0.64	0.64	**0.71**	0.65

Table 8: Hallucination detection F1 scores by hallucination type for English and target language average, with and without knowledge augmentation.

language) is substantially higher, averaging 21.7%.

This 4.6:1 ratio between missed hallucinations and false hallucinations indicates a systematic bias in cross-lingual reliability. Gemma 3 misses 18.3% of hallucinations across South African languages that it correctly identifies in English, with this pattern more pronounced for T0++ (24.0%) and Llama 3.1 (29.5%). Aya-101 shows the greatest cross-lingual consistency with the lowest missed hallucination rate (14.9%), though the disparity remains substantial.

These findings highlight a concerning reliability gap in multilingual contexts, where models that appear capable in English may fail to maintain that capability in other languages. The asymmetric pattern suggests models exhibit greater skepticism in English, potentially reflecting the English-centric nature of their training data. Appendix F provides additional analysis of these cross-lingual discrepancies, including language-specific patterns and more detailed error distributions.

5.3 Impact of Knowledge Augmentation

Knowledge augmentation substantially improves hallucination detection performance across all models and languages (Table 7), with significantly larger gains for South African languages (ranging from +11.0% to +25.4%) compared to English (+5.6% to +7.9%). This disparity suggests knowledge augmentation particularly benefits low-resource languages, potentially compensating for the inherent English-centric biases in models' pre-trained parameters.

Sepedi consistently shows the greatest improvement with knowledge augmentation across all models (average F1 score increase of 21.0%). This is particularly significant as Sepedi has the lowest baseline performance, suggesting knowledge augmentation is most beneficial for the most challenging languages. T0++ demonstrates the most improvement with knowledge augmentation (average increase of 18.2% across all languages), suggesting it may have reasoning capabilities that effectively leverage external knowledge despite weaker baseline multilingual performance.

The impact on missed hallucination rates (Table 9) is even more notable. T0++ shows the most transformation, with its average missed hallucination rate dropping from 24.0% to just 5.0% (a 79.2% relative reduction). For Sepedi, T0++'s missed hallucination rate falls from 29.8% to 5.7% (an 80.9% reduction). All models show substantial improvements across all languages, with Sepedi experiencing the largest absolute reductions.

Model	Setup	ZU (%)	XH (%)	NSO (%)	ST (%)
Gemma 3	Base	17.2	15.9	21.3	18.7
	+Know	5.1	4.3	7.2	6.4
	Δ%	*-70.3*	*-73.0*	*-66.2*	*-65.8*
Aya-101	Base	13.8	12.5	17.9	15.3
	+Know	**4.7**	**3.9**	6.7	5.8
	Δ%	*-65.9*	*-68.8*	*-62.6*	*-62.1*
T0++	Base	21.4	19.7	29.8	25.1
	+Know	4.9	4.1	**5.7**	**5.2**
	Δ%	*-77.1*	*-79.2*	*-80.9*	*-79.3*
Llama 3.1	Base	26.8	24.3	35.7	31.2
	+Know	11.3	9.8	13.2	12.5
	Δ%	*-57.8*	*-59.7*	*-63.0*	*-59.9*

Table 9: Impact of knowledge augmentation on missed hallucination rates

5.4 Performance by Hallucination Type

Table 8 reveals patterns in how models handle various hallucination forms across languages. Numerical inaccuracies are the most successfully detected category, with F1 scores approximately 5.0-13.0 percentage points higher than other categories. This suggests stronger representations of numerical relationships that generalize well across languages, possibly because numbers follow more consistent patterns transcending linguistic boundaries.

Entity errors present the greatest challenge, particularly in non-English languages. The cross-lingual detection gap for entity errors (up to 23% for some models) likely reflects models' stronger grounding in English-language entities compared to entities in South African contexts.

Knowledge augmentation has particularly strong effects on the most challenging hallucination types, with entity errors seeing the most substantial improvements (F1 score increases ranging from 21.9% to 38.6% in target languages). This disproportionate improvement suggests entity-based hallucinations are especially amenable to correction through explicit factual contextualization.

These results demonstrate that knowledge augmentation serves as an effective intervention for improving cross-lingual reliability in hallucination detection. By providing explicit factual information in both languages, knowledge augmentation creates a more level playing field that substantially mitigates cross-lingual biases in models' parametric knowledge. A more detailed analysis of discrepancy patterns and error types is available in Appendix F.3.

6 Conclusion

Multi-Hall-SA is a cross-lingual benchmark for hallucination detection spanning English and four low-resource South African languages. Our evaluation reveals significant cross-lingual reliability gaps, with models detecting up to 23.6% fewer hallucinations in South African languages compared to English. This disparity varies by hallucination type: entity-based errors present the greatest cross-lingual challenge, while numerical hallucinations remain more consistently detected. Knowledge augmentation emerges as a powerful mitigation strategy, reducing performance gaps by 59.4% on average and demonstrating that explicit factual contextualization effectively compensates for inherent model biases.

These findings have significant implications for deploying language models in multilingual contexts. Models evaluated only in high-resource languages may fail to maintain reliability when serving diverse linguistic communities, creating potential harms through uncaught hallucinations. The improvement from knowledge augmentation suggests retrieval-augmented generation approaches should be prioritized for low-resource languages, where parametric knowledge appears substantially less robust than for English.

Limitations

While Multi-Hall-SA makes significant contributions to cross-lingual hallucination detection, several limitations should be acknowledged. The benchmark currently encompasses four South African languages, which represents only a subset of Africa's linguistic diversity. Though these languages were carefully selected to include representatives from major language families, findings may not generalize to all low-resource languages.

The benchmark's current scope focuses primarily on administrative and governmental domains. While this ensures factual accuracy through authoritative sources, it means the benchmark may not fully represent hallucination patterns in other domains.

Our knowledge bases, though carefully constructed, show coverage variations across languages (ranging from 86.2% to 92.9% as detailed in Appendix C.2). These differences in coverage may influence the comparative effectiveness of knowledge augmentation across languages.

The controlled hallucination generation approach focuses on four specific hallucination types. Although this taxonomy enables structured analysis, it may not capture the full spectrum of hallucination patterns that occur in natural language generation contexts.

Finally, our evaluation is limited to four commercial language models selected for their multilingual capabilities. The performance patterns observed may not be representative of all language models, particularly those specifically designed or fine-tuned for individual African languages.

References

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan T. McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6078–6095. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7236–7249. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6251–6258. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6633–6649. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4884–4895. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9320–9331. Association for Computational Linguistics.

Esin Durmus, He He, and Mona T. Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5055–5070. Association for Computational Linguistics.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *CoRR*, abs/2308.11764.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 59–66. AAAI Press.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 864–870. Association for Computational Linguistics.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025b. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Trans. Software Eng.*, 51(2):413–429.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1430–1441. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. Association for Computational Linguistics.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. *CoRR*, abs/2211.05853.

Sello Ralethe and Jan Buys. 2025. Cross-lingual knowledge projection and knowledge enhancement for zero-shot question answering in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10111–10124, Abu Dhabi, UAE. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 704–718. Association for Computational Linguistics.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1172–1183. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6594–6604. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, pages 3784–3803. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3544–3552. Association for Computational Linguistics.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020b. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1072–1086. Association for Computational Linguistics.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Trans. Assoc. Comput. Linguistics*, 11:546–564.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona T. Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1393–1404. Association for Computational Linguistics.

A Model Selection and Verification

We conducted preliminary testing to ensure that foundation models possessed sufficient capabilities in the target South African languages. We tested Claude-3.7-Sonnet and GPT-4o on manually translated isiZulu and Sepedi versions of CommonsenseQA and OpenBookQA obtained from Ralethe and Buys (2025). Both models obtained perfect performance (100% accuracy) on both languages, confirming their suitability for benchmark generation.

To ensure models were genuinely processing content in these languages rather than relying on English instruction understanding, all instructions were given exclusively in the target language:

```
[In isiZulu/Sepedi]
Phendula umbuzo olandelayo: {Question in
    isiZulu/Sepedi}
{Answer choices in isiZulu/Sepedi}
```

B Benchmark Generation Prompts

B.1 Aligned Fact Extraction

The parallel fact extraction process used carefully designed prompts that ensured semantic alignment across languages:

```
You are an expert in both English and
    isiZulu. Your task is to identify
    key factual statements that appear
    in both the English and isiZulu
    texts provided below.

INSTRUCTIONS:
1. Read both the English and isiZulu
    texts carefully.
2. Identify 5-7 clear factual statements
    that appear in BOTH texts.
3. For each fact, provide the exact
    sentence from the English text and
    its corresponding sentence from the
    isiZulu text.
4. Focus on statements that contain
    specific information (dates, numbers
    , organizations, procedures,
    requirements).
5. Ensure the facts you select appear in
    BOTH languages.
6. Format your response as a numbered
    list with the English statement
    followed by its isiZulu equivalent.

ENGLISH TEXT:
{english_text}

ISIZULU TEXT:
{isizulu_text}

Please provide the 5-7 aligned factual
    statements in this format:
1. English: [English factual statement]
   IsiZulu: [Corresponding isiZulu
       statement]
```

The key design elements enabling successful cross-lingual alignment include:

- Explicit instruction to process both languages simultaneously
- Parallel context windows providing both texts
- Structured output format ensuring clear correspondence
- Information-type guidance focusing on verifiable content
- Exact sentence requirement maintaining linguistic authenticity

B.2 Controlled Hallucination Generation

For hallucination generation, we implemented structured prompts for creating specific types of hallucinations while maintaining semantic alignment:

```
You are an expert in creating controlled
    hallucinations for NLP benchmark
    development. Your task is to modify
    the factual statements below by
    introducing specific types of errors
    while maintaining grammatical
    correctness and plausibility.

ORIGINAL FACT PAIR:
English: {english_factual_statement}
{target_language}: {
    target_language_factual_statement}

INSTRUCTIONS:
Create FOUR variations of this fact pair
    , each containing a different type
    of hallucination:

1. TEMPORAL SHIFT: Change dates, time
    periods, or chronological
    information
2. ENTITY ERROR: Replace organizations,
    departments, or named entities with
    incorrect ones
3. NUMERICAL INACCURACY: Alter numbers,
    percentages, or quantities
4. LOCATION MISTAKE: Change geographical
    references or spatial information

IMPORTANT:
- Ensure the same type of error is made
    in BOTH language versions
- Keep the statements grammatically
    correct and plausible
- Preserve the overall structure and
    context of the original statements
- Make only ONE type of change per
    variation
```

C Knowledge Base Structure and Retrieval

C.1 Triple Structure

The knowledge bases used for knowledge-augmented evaluation were structured as semantic triples in multiple languages, following the format (subject, predicate, object). These knowledge bases were derived from the work of Ralethe and Buys (2025), who projected English knowledge from ConceptNet and DBpedia to South African languages.

The knowledge graph contains approximately 127,000 triples in English and 98,000, 94,000, 76,000, and 72,000 triples in isiZulu, isiXhosa, Sepedi, and Sesotho, respectively. Coverage analysis indicated that approximately 88% of entities mentioned in the benchmark statements had corresponding entries in the knowledge graph.

```
# English triples
(Department of Home Affairs, issues,
    identity documents)
(identity documents, required for,
    passport applications)
(identity documents, contain, biometric
    information)

# isiZulu triples
(UMnyango Wezasekhaya, ukhipha,
    amadokhumenti esintu)
(amadokhumenti esintu, adingeka ukuze,
    ufake isicelo sephasipoti)
(amadokhumenti esintu, aqukethe, ulwazi
    lwe-biometric)
```

C.2 Knowledge Coverage Analysis

We conducted a detailed coverage analysis, comparing the entities and relations in the benchmark statements with those available in the knowledge graph:

D Evaluation Prompt Details

D.1 Zero-shot Evaluation Prompts

For the baseline evaluation, we used this prompt template:

```
System: You are an expert at identifying
     factual errors and hallucinations
    in statements. Your job is to
    determine if a given statement
    contains false information.

User: [For non-English statements only]
    The following statement is in {
    LANGUAGE}, which is spoken in South
    Africa. Please examine it for
    factual accuracy.

Statement: "{STATEMENT}"

Is this statement factually accurate or
    does it contain hallucinations or
    errors? Respond with exactly one
    word - either "FACTUAL" if the
    statement is accurate, or "
    HALLUCINATION" if it contains false
    information.
```

D.2 Knowledge-augmented Evaluation Prompts

For knowledge-augmented evaluations, we modified the template to include retrieved knowledge triples:

```
System: You are an expert at identifying
     factual errors and hallucinations
    in statements. Your job is to
    determine if a given statement
    contains false information.
```

Entity Type	EN (%)	ZU (%)	XH (%)	NSO (%)	ST (%)
Organizations	94.3	91.7	90.5	87.2	88.4
Locations	96.8	94.2	93.7	90.1	91.3
Temporal Terms	89.6	85.3	86.9	82.4	83.7
Numerical Concepts	98.2	97.5	96.8	94.3	94.8
Procedures	85.7	80.4	81.2	76.9	77.5
Overall	92.9	89.8	89.8	86.2	87.1

Table 10: Knowledge graph coverage by language and entity type

Model	Size (B)	English (%)	Target Avg. (%)	Gap (%)	Gap %
Gemma 3	12	86.0	74.0	12.0	14.0
Aya-101	11	78.0	70.0	8.0	10.3
T0++	11	76.0	63.0	13.0	17.1
Llama 3.1	8	72.0	55.0	17.0	23.6

Table 11: Hallucination detection performance by model size (F1 scores)

```
User: [For non-English statements only]
    The following statement is in {
    LANGUAGE}, which is spoken in South
    Africa. Please examine it for
    factual accuracy.

Here is some factual context that may be
    relevant:
{RETRIEVED_KNOWLEDGE_TRIPLES}

Statement: "{STATEMENT}"

Is this statement factually accurate or
    does it contain hallucinations or
    errors? Respond with exactly one
    word - either "FACTUAL" if the
    statement is accurate, or "
    HALLUCINATION" if it contains false
    information.
```

E Implementation Details

All evaluations were conducted using the following implementation specifications:

- **API endpoints:** All models were accessed through Vertex AI endpoints, specifically version 2023-06-01
- **Generation parameters:** Temperature=0.0, TopP=1.0, MaxTokens=10
- **Error handling:** Exponential backoff retry logic for API failures (max 5 retries)
- **Parallel processing:** Evaluations distributed across 8 concurrent processes
- **Response validation:** Automatic verification of correct response format
- **Reproducibility:** Fixed random seeds (42) for all randomized processes

F Additional Results

F.1 Cross-lingual Discrepancy Direction Analysis

Table 13 provides a detailed breakdown of cross-lingual discrepancies by direction, showing the proportion of statements where models made different predictions between English and target languages.

The data shows a strong asymmetry in the direction of discrepancies. Cases where models classified statements as hallucinations in the target language but as factual in English (E=F, T=H) were relatively rare (4.7% on average), while the reverse scenario (E=H, T=F) was much more common (14.6% on average). This asymmetry suggests that models have stronger skepticism in English, possibly reflecting their training data distribution.

F.2 Performance by Model Size

We analyzed the relationship between model size and cross-lingual hallucination detection performance:

The results suggest model architecture and training objective influence cross-lingual consistency beyond raw parameter count.

F.3 Error Analysis

We conducted detailed error analysis on randomly sampled detection failures:

In target languages, cultural context misalignment and entity confusion represent a larger proportion of errors, while temporal ambiguity is more prevalent in English errors.

Error Type	English (%)	Target Lang. (%)
Entity confusion	29	36
Numeric reasoning errors	8	11
Location inconsistency	18	23
Temporal ambiguity	31	19

Table 12: Distribution of error types in hallucination detection failures

Language	Overall Discrep.	E=F, T=H	E=H, T=F	Missed Hall. Rate
isiZulu	17.8%	4.5%	13.3%	19.1%
isiXhosa	16.2%	4.1%	12.1%	17.3%
Sepedi	23.6%	5.3%	17.4%	24.9%
Sesotho	21.0%	4.9%	15.6%	22.4%
Average	19.3%	4.7%	14.6%	21.4%

E=F, T=H: English=FACTUAL, Target=HALLUCINATION
E=H, T=F: English=HALLUCINATION, Target=FACTUAL
Missed Hall. Rate: Rate of hallucinations detected in English but missed in target language

Table 13: Cross-lingual discrepancy direction analysis (baseline evaluation)

G Sample Hallucinations

Below are representative examples of each hallucination type from the benchmark across different languages:

G.1 Temporal Hallucination Example

English Original: The UIF must be claimed within six months of becoming unemployed.
English Hallucinated: The UIF must be claimed within two years of becoming unemployed.
isiZulu Hallucinated: I-UIF kumele ifakwe singakapheli iminyaka emibili uthola ukungasebenzi.

G.2 Entity Hallucination Example

English Original: The Department of Home Affairs issues identity documents.
English Hallucinated: The Department of Social Development issues identity documents.
Sepedi Hallucinated: Kgoro ya Tlhabollo ya Leago e ntšha dipampiri tša boitsebišo.

G.3 Numerical Hallucination Example

English Original: Employers and employees each contribute 1% of the employee's salary to the UIF.
English Hallucinated: Employers and employees each contribute 3.5% of the employee's salary to the UIF.
isiXhosa Hallucinated: Abaqashi nabasebenzi banikezela nge-3.5% ngabanye kwimali yomvuzo womsebenzi kwi-UIF.

G.4 Location Hallucination Example

English Original: SASSA offices in Pretoria process social grant applications.
English Hallucinated: SASSA offices in Durban process social grant applications.
Sesotho Hallucinated: Diofisi tsa SASSA tse Durban di sebetsa dikopo tsa dithuso tsa mmuso.