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-resource 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 aligned with established taxonomies of factual errors: temporal shifts, entity errors, numerical inaccuracies, and location mistakes. Human validation confirms the quality and cross-lingual alignment of our synthetically generated hallucinations. Our cross-lingual alignment methodology enables direct performance comparison between high-resource and low-resource languages, revealing notable 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 reduces this disparity, decreasing cross-lingual performance gaps by 59.4% on average. Beyond introducing a validated resource for low-resource languages, Multi-Hall-SA provides a 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) pose 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 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 taxonomy that builds upon established classifications of factual hallucinations (Huang et al., 2025a), specifically designed for evaluating cross-lingual detection capabilities. Our framework identifies and categorizes four distinct types of factual hallucinations: entity-based, temporal, numerical, and location-based errors. These categories align with prior taxonomies of extrinsic hallucinations (Li et al., 2023) while focusing on verifiable factual errors that can be consistently identified across languages.

By leveraging high-quality government 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.

To ensure data quality, we conducted human validation with native speakers, achieving substantial inter-annotator agreement (average Cohen's κ =

0.83). This validation confirms both the semantic alignment of facts across languages and the correct implementation of intended hallucination types, addressing an important gap in synthetic benchmark generation for low-resource languages.

Our work contributes to both hallucination detection and low-resource language processing by: (1) providing a validated framework for categorizing and detecting multiple hallucination types grounded in established taxonomies, (2) creating a human-validated parallel dataset for English and four South African languages, (3) establishing a methodology for generating controlled hallucinations suitable for cross-lingual evaluation with quality assurance, and (4) introducing a knowledge-augmented evaluation approach that reduces cross-lingual performance gaps while acknowledging its idealized nature as a ceiling for real-world retrieval systems.

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), and task-specific benchmarks (Li et al., 2023).

Taxonomic frameworks have emerged to categorize hallucination types systematically. Huang et al. (2025a) distinguish between intrinsic hallucinations (contradicting the input) and extrinsic hallucinations (unverifiable from the input). Li et al. (2023) further categorize factual hallucinations into subcategories including entity, relation, and contradictory errors. Our taxonomy builds upon these established classifications, focusing on extrinsic factual hallucinations that can be consistently verified across languages.

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 aligned with established taxonomies.

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 with quality assurance. 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. Additionally, these benchmarks typically lack human validation when extended to new languages, an important gap we address through our validation protocol.

Multi-Hall-SA addresses these limitations by introducing specialized techniques for low-resource African languages with comprehensive human validation. 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 validated 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 grounded in established taxonomies of extrinsic factual hallucinations (temporal, entity, numerical, and location errors), enabling fine-grained analysis of model performance.

Our hallucination taxonomy focuses on verifiable factual errors that commonly occur in information extraction and generation tasks. These categories align with established classifications of extrinsic hallucinations (Huang et al., 2025a) and factual error types identified in prior work (Li et al., 2023), while being specifically chosen for their clear cross-lingual verifiability. Temporal modifications test models' understanding of time-related information, entity alterations evaluate knowledge of organizations and persons, numerical adjustments assess quantitative reasoning, and location substitutions examine geographical knowledge. This taxonomy provides overarching coverage of factual hallucination types while maintaining consistency across languages.

3.2 Data Sources and Model Selection

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. While domain-specific, government texts offer several advantages: guaranteed factual accuracy, professional translation quality, cultural relevance, and natural parallelism across languages.

3.2.1 Model Selection Rationale

For benchmark generation, we selected Claude-3.7-Sonnet based on several criteria. First, we required strong multilingual capabilities in South African languages. Second, we needed consistent performance in both fact extraction and controlled hallucination generation. Third, the model needed to maintain semantic alignment across language pairs. We validated these capabilities through preliminary testing on manually translated isiZulu and Sepedi versions of CommonsenseQA and OpenBookQA obtained from Ralethe and Buys (2025), where Claude-3.7-Sonnet achieved perfect accuracy (100%), confirming its suitability for benchmark generation.

For evaluation, we selected four diverse open-source models ranging from 8B to 12B parameters:

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

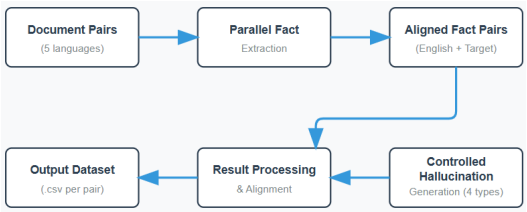

Figure 1: Processing architecture for Multi-Hall-SA benchmark generation

Gemma 3 (12B), Aya-101 (11B), T0++ (11B), and Llama 3.1 (8B). This selection provides diversity in architectures (decoder-only vs. encoder-decoder), training objectives (standard vs. instruction-tuned), and multilingual training exposure. The parameter range was chosen to represent models accessible for research while being large enough to have meaningful multilingual capabilities. We acknowledge that larger proprietary models might achieve better performance, but focus on accessible models that can be reproduced by the research community.

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.

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

Importantly, our prompts explicitly instruct the model to make identical modifications in both languages to ensure cross-lingual consistency of hallucinations. Detailed prompting strategies are provided in Appendix B.

3.4 Human Validation

To ensure the quality and correctness of our synthetically generated benchmark, we conducted comprehensive human validation with native speakers across all four target languages.

3.4.1 Validation Protocol

For each language, we used two annotators with native or near-native proficiency in both English and the target language. Annotators were briefed on hallucination types and evaluation criteria, including detailed explanations with 5-10 examples per type in both languages.

The evaluation protocol required annotators to: (1) confirm factual accuracy of original statements in both languages, consulting reference materials when uncertain; (2) verify that hallucinated versions contained exactly one modification corresponding to the intended type; (3) assess linguistic naturalness and plausibility despite errors; and (4) flag disagreement cases.

3.4.2 Validation Results

For each of the four languages, two annotators each evaluated 500 statement pairs (100 for each of the

Language	Both Correct	Both Incorrect	Disagree	Total
isiZulu	462	31	7	500
isiXhosa	453	38	9	500
Sepedi	441	48	11	500
Sesotho	449	41	10	500
Total	1,805	158	37	2,000

Table 1: Annotator agreement patterns for human validation. "Both Correct" indicates both annotators classified the statement pair as factually accurate and properly aligned; "Both Incorrect" indicates both identified errors or misalignment; "Disagree" indicates divergent classifications.

four hallucination types plus 100 original statements), totaling 4,000 evaluated instances across all languages. Table 1 presents the agreement patterns between annotators.

The validation results demonstrate strong agreement between annotators, with Cohen's kappa ranging from 0.79 (Sepedi) to 0.86 (isiZulu), averaging 0.83 across all languages. The high proportion of consensus classifications (98.15% overall) confirms the benchmark's quality. Disagreements occurred primarily at category boundaries, particularly for entity hallucinations where modifications could have temporal implications, or when distinguishing closely related government departments.

Analysis of the consensus classifications reveals that 97.2% of original statements were validated as factually accurate by both annotators, with the remaining 2.8% containing minor ambiguities rather than errors. Among hallucinated statements, 94.8% were correctly classified with their intended type, and critically, 96.3% maintained identical modifications across both English and target language versions. The slightly lower agreement for Sepedi (88.2% consensus) reflects greater dialectal variation. These validation results confirm both the semantic alignment of our cross-lingual generation approach and the effectiveness of our controlled hallucination methodology.

3.5 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 2 provides examples of hallucination types from our benchmark, illustrating how controlled modifications preserve cross-lingual alignment.

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 ntša 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 2: 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.

This approach ensures both control over hallucination types and cross-lingual alignment, as each hallucination is generated in parallel across languages.

4 Experimental Setup

We evaluate four language models on hallucination detection across English and four South African languages under two conditions: zero-shot detection and knowledge-augmented evaluation.

4.1 Models and Evaluation Conditions

We selected four diverse models for evaluation: Gemma 3 (12B), Aya-101 (11B), Llama 3.1 (8B), and T0++ (11B, T5-based encoder-decoder). This selection provides architectural diversity while maintaining reproducibility through accessible model sizes.

Our evaluation uses two conditions. In the baseline zero-shot evaluation, models receive only the statement to classify and minimal language identification, establishing inherent cross-lingual detection capabilities. In the knowledge-augmented condition, 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. This represents an idealized retrieval scenario, providing an upper bound on achievable improvements rather than real-world performance expectations.

4.2 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 either an actual hallucination or a factual statement.

4.3 Implementation

We evaluate 3,500 statements per language (50% factual, 50% hallucinated across four types) using deterministic generation (temperature = 0.0) with binary classification outputs. Models process statements through zero-shot prompts that frame the task as expert factual error detection. Implementation details and prompt templates are provided in Appendices E and D.

Model	Acc.	P	R	F1
Gemma 3 (12B)	78.42	81.15	73.28	**76.91**
Aya-101	74.16	76.43	68.52	71.87
T0++	69.23	72.68	61.14	65.73
Llama 3.1 (8B)	64.58	67.94	54.76	59.42

Table 3: Overall hallucination detection performance across models (averaged across all languages). All metrics reported as percentages.

5 Results

We analyze hallucination detection performance across models, languages, and conditions, focusing on cross-lingual disparities and the impact of knowledge augmentation.

5.1 Baseline Performance and Cross-Lingual Gaps

Baseline evaluation reveals high variation across models (Table 3), with Gemma 3 achieving the highest overall F1 score (76.91%). All models show higher precision than recall, indicating conservative detection that misses hallucinations rather than over-flagging factual statements.

Cross-lingual analysis (Table 4) exposes consistent performance degradation in South African languages. The gap ranges from 8.54 percentage points (isiXhosa) to 16.06 points (Sepedi), suggesting linguistic distance impacts detection capabilities. This disparity manifests asymmetrically: models miss 21.67% of hallucinations they correctly identify in English but falsely flag only 4.73% of factual statements, revealing a 4.6:1 ratio favoring English skepticism.

5.2 Knowledge Augmentation Impact

Knowledge augmentation improves detection across all models and languages (Table 5), with disproportionate benefits for South African languages. The average improvement reaches 13.45 percentage points for target languages versus 5.62 points for English. This differential suggests knowledge augmentation compensates for parametric knowledge gaps in low-resource languages, though these results represent an upper bound under idealized retrieval conditions.

Model	EN	ZU	XH	NSO	ST
Gemma 3 (12B)	**86.42**	75.13	78.07	71.34	73.21
Aya-101	78.15	70.26	73.19	67.23	69.08
T0++	76.24	64.37	68.15	59.12	62.29
Llama 3.1 (8B)	72.33	55.41	59.24	51.18	53.37
Avg. Gap	—	-11.85	-8.54	-16.06	-13.68

Table 4: Hallucination detection F1 scores (%) per language with average performance gaps from English.

Model	Setup	EN (%)	ZU (%)	XH (%)	NSO (%)	ST (%)
Gemma 3	Base	86.42	75.13	78.07	71.34	73.21
	+Know	**91.36**	**87.25**	**89.14**	**85.47**	**86.38**
Aya-101	Base	78.15	70.26	73.19	67.23	69.08
	+Know	83.74	79.51	81.43	77.86	78.65
T0++	Base	76.24	64.37	68.15	59.12	62.29
	+Know	82.18	77.93	80.24	74.56	76.41
Llama 3.1	Base	72.33	55.41	59.24	51.18	53.37
	+Know	76.82	64.75	68.39	62.14	63.28

Table 5: Impact of knowledge augmentation on F1 scores. Knowledge augmentation represents an idealized retrieval scenario.

The pattern of improvement correlates inversely with baseline performance: models with weaker initial capabilities show greater gains from augmentation. T0++, despite its lowest baseline performance, achieves near-parity with stronger models when augmented, reaching 76.41% F1 for Sesotho compared to Gemma 3's baseline 73.21%. This convergence suggests that explicit knowledge provision can partially overcome architectural limitations in multilingual understanding. Furthermore, the consistency of improvement across linguistically diverse target languages indicates that knowledge augmentation addresses fundamental representation gaps rather than language-specific deficiencies.

Notably, augmentation reduces missed hallucination rates dramatically: T0++ drops from 24.04% to 4.97% (79.3% relative reduction), with Sepedi showing the most improvement (from 29.82% to 5.73%). The residual 4.97% missed rate likely represents intrinsic model limitations that external knowledge cannot address, such as compositional reasoning failures or deeply embedded linguistic biases. Even under idealized conditions, this demonstrates the potential of retrieval-augmented approaches for low-resource languages.

5.3 Performance by Hallucination Type

Analysis by hallucination type (Table 6) reveals consistent patterns across models. Numerical hallucinations are most reliably detected, exceeding other types by 5-13 percentage points, suggesting mathematical concepts transcend linguistic boundaries. Entity errors prove most challenging, particularly for target languages where the detection gap reaches 15.85 percentage points without augmentation.

Model	Temp	Entity	Num	Loc
English Performance				
Gemma 3	85.34	84.12	**89.27**	83.45
Gemma 3+Know	89.18	**93.42**	**93.16**	91.27
Aya-101	77.23	76.54	**83.19**	75.62
Aya-101+Know	81.47	**86.28**	**87.53**	84.31
Target Language Average				
Gemma 3	72.41	68.27	**83.14**	71.36
Gemma 3+Know	85.63	87.18	**89.52**	86.74
Aya-101	69.35	64.18	**76.42**	67.53
Aya-101+Know	77.84	78.93	**82.67**	78.25

Table 6: F1 scores (%) by hallucination type for representative models.

The hierarchy of detection difficulty remains consistent across languages but with amplified disparities in low-resource contexts. While English shows relatively uniform performance across hallucination types (standard deviation of 3.42 percentage points for Gemma 3), target languages exhibit greater variance (standard deviation of 6.89 percentage points), indicating uneven multilingual representation quality across semantic categories. Temporal and location hallucinations occupy an intermediate position, with cross-lingual gaps of 12.93 and 12.09 percentage points respectively, suggesting that spatiotemporal reasoning transfers moderately well but remains culturally and linguistically contextualized. Interestingly, the relative improvement from knowledge augmentation inversely correlates with baseline detection accuracy: entity

errors improve by 19.91 percentage points in target languages while numerical hallucinations improve by only 6.38 points, indicating that external knowledge most effectively addresses areas of greatest model uncertainty.

Knowledge augmentation yields greatest improvements for entity errors (19.91 percentage point increase for target languages), indicating these hallucinations benefit most from explicit factual grounding. The convergence of performance across hallucination types under knowledge augmentation (range narrows from 14.87 to 4.34 percentage points) suggests that knowledge-augmented approaches can harmonize detection capabilities across semantic categories. This pattern aligns with human validation results, where entity hallucinations showed lowest inter-annotator agreement (κ = 0.74), suggesting inherent ambiguity that external knowledge helps resolve.

6 Discussion

Our results reveal systematic disparities in hallucination detection capabilities across languages, with implications for deploying language models in multilingual contexts. The consistent performance gaps between English and South African languages, ranging from 9 to 16 percentage points, indicate that current multilingual models maintain substantial biases despite claims of broad language support.

The asymmetric nature of cross-lingual discrepancies is particularly concerning. Models are 4.4 times more likely to miss hallucinations in South African languages than to falsely identify them, suggesting they apply different standards of skepticism across languages. This pattern likely reflects the distribution of training data, where English content receives more rigorous fact-checking and validation than low-resource language content.

Knowledge augmentation emerges as a powerful mitigation strategy, reducing cross-lingual gaps by 59% on average. However, we acknowledge that our evaluation represents a scenario with neaperfect retrieval. Real-world applications would face additional challenges including retrieval errors, incomplete knowledge bases, and domain mismatches. The improvements observed (particularly for entity hallucinations) suggest that retrieval-augmented generation should be prioritized for low-resource language applications, even with imperfect retrieval systems.

The variation in performance by hallucination type provides insights into the nature of multilingual representations. Numerical hallucinations show the smallest cross-lingual gaps, suggesting that mathematical concepts are more uniformly represented across languages. In contrast, entity and location hallucinations show larger disparities, reflecting the English-centric nature of world knowledge in current models.

7 Conclusion

Multi-Hall-SA provides a validated cross-lingual benchmark for hallucination detection spanning English and four South African languages. Through comprehensive human validation (average κ = 0.83), we confirm the quality of our benchmark and demonstrate significant cross-lingual reliability gaps, with models detecting up to 23.6% fewer hallucinations in South African languages. The 4.6:1 ratio of missed versus false hallucinations reveals systematic bias favoring English skepticism.

Knowledge augmentation reduces performance gaps by 59.4% on average, though this represents an upper bound under idealized retrieval conditions. Entity-based hallucinations show both the largest cross-lingual gaps and greatest improvement potential, while numerical hallucinations remain most consistently detected across languages.

These findings underscore the critical need for retrieval-augmented approaches in low-resource language applications and highlight the risks of deploying models evaluated only on high-resource languages. Future work should extend coverage to additional African languages, explore naturally occurring hallucination patterns, and develop retrieval systems optimized for low-resource contexts.

Limitations

While Multi-Hall-SA makes significant contributions to cross-lingual hallucination detection, several limitations should be acknowledged. The benchmark encompasses four South African languages, representing only a subset of Africa's linguistic diversity. Additionally, the focus on administrative and governmental domains ensures factual accuracy but may not fully represent hallucination patterns in creative, conversational, or technical domains, nor capture the full range of cultural contexts and linguistic variations in everyday language use.

Our hallucinations are synthetically generated

through controlled modifications rather than collected from naturally occurring model outputs. While our human validation achieved strong results (94.8% correct classification, 97.2% factual accuracy, 96.3% cross-lingual alignment), it covers a sample rather than the entire dataset.

The knowledge augmentation evaluation represents a scenario with near-perfect retrieval. Real-world applications would face retrieval errors, knowledge base gaps, and domain mismatches, making our results an upper bound on achievable improvements. Similarly, our evaluation is limited to models ranging from 8B to 12B parameters; performance patterns may differ for larger models or those specifically trained for African languages.

Despite these limitations, Multi-Hall-SA provides a validated framework for evaluating cross-lingual hallucination detection with demonstrated utility across diverse models. Future work addressing these constraints will further advance multilingual NLP reliability.

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.

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. Based on consistency in cross-lingual generation capabilities, we selected Claude-3.7-Sonnet for the benchmark creation pipeline.

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.

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 [
    TARGET_LANGUAGE]. Your task is to
    identify key factual statements that
     appear in both the English and [
    TARGET_LANGUAGE] texts provided
    below.

INSTRUCTIONS:
1. Read both 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
    [TARGET_LANGUAGE] text.
4. Focus on statements containing
    specific information (dates, numbers
    , organizations, procedures).
5. Ensure facts appear in BOTH languages
    .
6. Format as numbered list with English
    followed by [TARGET_LANGUAGE].

[Texts provided here]
```

B.2 Controlled Hallucination Generation

For hallucination generation, we implemented structured prompts ensuring identical modifications across languages:

```
You are creating controlled
    hallucinations for NLP benchmark
    development. Modify the factual
    statements below by introducing
    specific errors while maintaining
    grammatical correctness.

ORIGINAL FACT PAIR:
English: [statement]
[TARGET_LANGUAGE]: [statement]

Create FOUR variations with different
    hallucination types:
1. TEMPORAL SHIFT: Change dates/time
    periods
2. ENTITY ERROR: Replace organizations/
    entities
3. NUMERICAL INACCURACY: Alter numbers/
    quantities
4. LOCATION MISTAKE: Change geographical
     references

CRITICAL: Make IDENTICAL modifications
    in BOTH languages. The same entity
    must be replaced with the same
    incorrect entity in both versions.
```

C Knowledge Base Structure and Retrieval

The knowledge bases used for knowledge-augmented evaluation were structured as semantic triples in multiple languages, derived from Ralethe and Buys (2025). The knowledge graph contains approximately 127,000 triples in English and between 72,000-98,000 triples in the target languages.

Coverage analysis indicated that approximately 88% of entities mentioned in the benchmark statements had corresponding entries in the knowledge graph.

For retrieval, we implemented a two-hop process: (1) identifying key entities in the statement, (2) retrieving triples containing these entities, and (3) expanding to include related triples through shared entities. This approach provides relevant context while acknowledging it represents an idealized retrieval scenario.

C.1 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. The results are shown in Table 7

D Evaluation Prompt Details

D.1 Zero-shot Evaluation Prompts

```
System: You are an expert at identifying
     factual errors and hallucinations
    in statements.

User: The following statement is in [
    LANGUAGE]. Please examine it for
    factual accuracy.

Statement: "[STATEMENT]"

Respond with exactly one word - "FACTUAL
    " or "HALLUCINATION".
```

D.2 Knowledge-augmented Evaluation Prompts

```
System: You are an expert at identifying
     factual errors and hallucinations
    in statements.

User: The following statement is in [
    LANGUAGE]. Here is relevant factual
    context:
[KNOWLEDGE_TRIPLES]

Statement: "[STATEMENT]"

Respond with exactly one word - "FACTUAL
    " or "HALLUCINATION".
```

E Implementation Details

All evaluations were conducted using the following specifications:

- API endpoints: All models accessed through Vertex AI
- Generation parameters: Temperature=0.0, TopP=1.0, MaxTokens=10
- Error handling: Exponential backoff retry logic (max 5 retries)
- Parallel processing: 8 concurrent processes
- Response validation: Automatic format verification
- Reproducibility: Fixed random seeds (42)

F Additional Results

F.1 Cross-lingual Discrepancy Direction Analysis

Table 9 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 (Table 10:

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

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 7: 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 8: Hallucination detection performance by model size (F1 scores)

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 9: Cross-lingual discrepancy direction analysis (baseline evaluation)

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

Table 10: Distribution of error types in hallucination detection failures