Improving Amharic Information Retrieval with Translative and Multi-Agent Debate Retrieval Augmented Generation

Abel Jotie, Prasenjit Mitra
Carnegie Mellon University Africa
Kigali, Rwanda
{ajotie, prasenjm}@andrew.cmu.edu

Abstract

Retrieval-augmented generation (RAG) has been used to improve the accuracy and transparency of outputs produced by large language models (LLMs) by integrating external knowledge; however, applying RAG to low-resource languages presents unique challenges, including poor embedding representations, low retrieval quality, and semantic gaps caused by the scarcity of digital documents. In this research, we address these challenges for a selected low-resource language, Amharic, by using translative and debate-based RAG techniques to improve retrieval and reasoning. This paper outlines the key problems and research gaps in applying RAG to low-resource languages and introduces a method to enhance RAG performance for Amharic. Additionally, we introduce the first comprehensive **A**mharic **R**etrieval-Augmented **G**eneration **B**enchmark (ARGB), designed to capture grammatical, cultural, and writing-system-specific constraints of the Amharic language. ARGB evaluates not only retrieval and generation quality, but also noise robustness, counterfactual robustness, negative rejection, and multi-source information integration, providing a holistic assessment of RAG capabilities. The dataset, which spans a wide range of categories, is evaluated using multiple evaluation metrics. Furthermore, we demonstrate that, using our dataset, translation-based and debate-based methods substantially improve various aspects of RAG pipeline assessment in the Amharic language. This work aims to improve the reliability, accessibility, and inclusiveness of AI systems for Amharic speakers while providing a scalable framework for other low-resource languages. Current progress on the code and benchmark can be found on this GitHub link: link.

Keywords: LLM, Retrieval Augmented Generation(RAG), Multi-agent Debate, Agent Society, Low resource, Amharic

1. Introduction

LLMs have transformed natural language processing by enabling strong performance across tasks such as open-domain question answering, summarization, reasoning, and dialogue generation (Brown et al., 2020; Radford et al., 2019). However, despite enhanced language understanding and generation, LLMs remain limited by the static knowledge encoded in their parameters. They are prone to hallucinations, factual inaccuracies, and out-of-date internal knowledge (Farquhar et al., 2024; Lewis et al., 2020).

RAG has emerged as an approach to mitigate these issues by grounding model outputs in externally retrieved documents (Lewis et al., 2020). By combining a retriever with a generative model, RAG systems improve factual consistency and provide traceable evidence for generated responses (Shuster et al., 2021; Gao et al., 2023). Despite its demonstrated success in high-resource contexts, the adaptation of RAG to low-resource languages remains insufficiently studied.

Low-resource languages face structural and technical barriers that limit the effectiveness of retrieval-augmented systems. First, embedding representations for low-resource languages are often weaker due to limited pretraining data, resulting in poor semantic alignment between queries and documents (Miao et al., 2024). Second, the scarcity and uneven quality of digital corpora reduce retrieval recall and coverage (Kazi et al., 2025). Third, linguistic characteristics such as rich morphology, writing-system variation, and limited standardized tools further degrade retrieval performance (Wiemerslage et al., 2022). Together, these challenges lead to low retrieval quality, which directly impacts the reliability of downstream generation.

Amharic, a widely spoken low-resource language in Ethiopia, Africa, exemplifies these challenges. Available resources for pretraining and finetuning Amharic language models remain limited, with RAG-based systems largely unexplored. Secondly, there are unique language characteristics of Amharic that requires specialized study. For instance, the language exhibits unique morphological structures for encoding subject–object–verb agreement, marking gender and number, and supporting derivational word formation (Amberber, 2023). Additionally, unique proverbs, culturally embedded meanings, and the distinctive Ge'ez writing system further contribute to this variation (Mengistu, 2018; Eid, 2021). In terms of resources for RAG, there is currently no widely adopted benchmark specifically de-

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 52–61
12 May 2026.

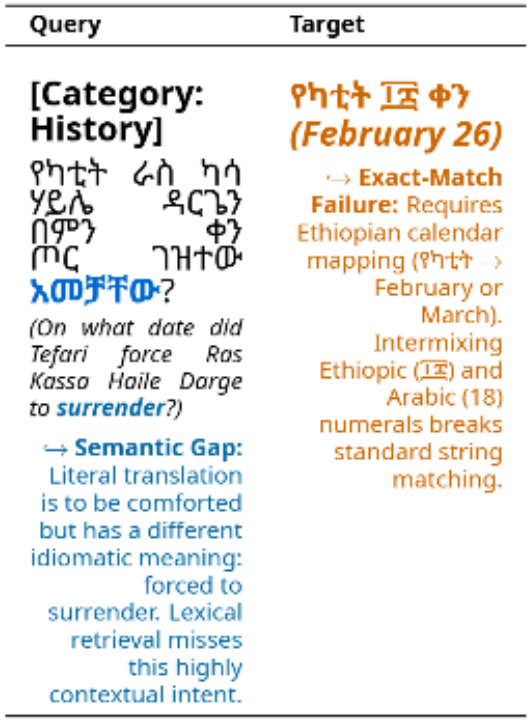

Figure 1: Linguistic and systemic challenges in Amharic QA. The benchmark addresses unique constraints in writing and calendar systems.

signed to evaluate retrieval-augmented generation in Amharic. This infrastructure gap hinders the development and deployment of practical systems such as health chatbot applications in resource-constrained regions like Ethiopia, despite their strong potential to reduce pressure on overstretched services and lower operational costs (Manyazewal et al., 2021; Bank, 2023).

This research seeks to address these limitations by investigating how retrieval-augmented generation can be adapted to function more effectively in low-resource contexts. Specifically, we use two techniques based on TraSe Architecture (Ipa et al., 2025) and Debate-Augmented RAG (Hu et al., 2025). The TraSe architecture leverages cross-lingual translation to improve the generation of text-based outputs with improved synthesis (applied at the generation step). Additionally, we adopt a debate-driven retrieval and generation approach, where agents propose, critique, and evaluate candidate answers/contexts to iteratively refine responses/retrieval. The goal is to improve reasoning and retrieval accuracy and reduce hallucinations, which is especially important for low-resource languages with sparse or biased retrieval data (Chari et al., 2025).

Beyond methodological improvements, this research also addresses the critical need for evaluation infrastructure. We introduce a factoid, open-domain, comprehensive Amharic RAG benchmark with extractive answers. Currently, there is only one publicly available question-answering benchmark for Amharic, Amh-QuAD (Taffa et al., 2024), which was developed using the Amharic Wikipedia dump dataset and manual annotation. However, the contexts defined for extraction are not large enough (only a few sentences) to represent a knowledge space with high coverage and accurate retrieval quality measures. To improve on this, we define the benchmarks with large contexts or entire articles for extraction. Furthermore, apart from retrieval accuracy measures, our benchmark assesses four qualities important in a RAG system: noise robustness, negative rejection, information integration, and counterfactual robustness, ensuring a comprehensive assessment (Chen et al., 2024).

In summary, this research aims to (1) analyze the key challenges of applying RAG to low-resource languages, (2) propose a translation-based and debate-augmented RAG framework tailored to Amharic that addresses low-resource constraints and language-specific characteristics, demonstrating significant improvements over baseline methods, and (3) the development of the first RAG Amharic dataset to support systematic evaluation through varying metrics. Through these contributions, the project aims to strengthen AI support for Amharic speakers and demonstrate methods that can benefit other low-resource languages.

2. Related Works

2.1. Advanced Retrieval-Augmented Generation and Benchmarking

Standard Retrieval-Augmented Generation (RAG) significantly reduces LLM hallucinations by grounding generation in external documents. Recent advancements have focused on optimizing the interaction between the retriever and the generator. For instance, (Jiang et al., 2023) introduced Forward-Looking Active REtrieval (FLARE), a technique where the language model actively decides when and what to retrieve during the generation process based on its internal confidence regarding upcoming tokens. While such active retrieval methods are highly effective in high-resource languages, they rely heavily on the robust internal knowledge of the base LLM, a capability often lacking in low-resource language models.

As RAG architectures have matured, evaluating their true efficacy has required more sophisticated frameworks than simple exact-match accuracy. (Chen et al., 2024) established a foundational evaluation paradigm by introducing a com-

prehensive benchmark designed to assess four critical LLM abilities in RAG systems: noise robustness (filtering irrelevant retrieved documents), negative rejection (declining to answer when retrieved documents lack the information), information integration (synthesizing answers from multiple documents), and counterfactual robustness (recognizing and rejecting false information in the context). Our proposed Amharic benchmark directly adopts these four critical dimensions to provide the first rigorous assessment of RAG systems in Amharic.

2.2. Multi-Agent Debate in LLMs and RAG

To further improve reasoning and factual consistency, recent research has explored multi-agent debate frameworks. (Du et al., 2024) demonstrated that rather than relying on a single generative pass, instantiating multiple LLM agents to independently propose, critique, and iteratively refine their responses leads to a consensus that is significantly more factual and logically sound.

This debate paradigm has recently been extended to RAG architectures. (Hu et al., 2025) introduced a debate-augmented RAG framework where distinct agents evaluate both the retrieved contexts and the candidate answers. By iteratively debating the relevance of retrieved documents and the faithfulness of the generated text, the system effectively filters out misleading or low-quality retrievals. This approach is highly relevant for our work, as debate-driven refinement is a powerful mechanism for mitigating the impact of sparse, biased, or noisy retrieval data commonly encountered in low-resource environments.

2.3. RAG in Low-Resource Languages

Extending RAG to low-resource languages presents unique challenges, primarily due to weak embedding representations and a lack of digitized corpora (Li and Ke, 2025). A notable contribution addressing this gap is the work by (Ipa et al., 2025), who investigated RAG performance constraints in Bangla. Because standard models like Llama-2 struggle with native reasoning in underrepresented languages, they introduced the TraSe architecture. TraSe circumvents native language deficits via ”translative prompting”: translating the query and retrieved context into English for generative processing, and translating the final answer back to Bangla. Furthermore, TraSe uses a multi-prompt generation strategy coupled with an LLM-based selector to identify the most accurate candidate response. Our methodology adapts this cross-lingual translation strategy to leverage high-resource LLM reasoning capabilities for Amharic.

2.4. Question Answering and Datasets for Amharic

Despite the rapid growth of NLP evaluation datasets, Amharic remains severely underrepresented. The foundational effort in this domain is Amh-QuAD, introduced by (Taffa et al., 2024), which constitutes the first publicly available Amharic question-answering benchmark. Developed using manual annotations from the Amharic Wikipedia dump, Amh-QuAD was a vital first step. However, the contexts provided for extraction in this dataset are limited to only a few sentences. This brief context window is insufficient for evaluating the complex retrieval and synthesis capabilities required in modern RAG systems. Our work bridges this infrastructure gap by introducing a comprehensive benchmark featuring article-length contexts, designed explicitly to evaluate noise robustness, negative rejection, and information integration in Amharic.

3. Methodology

To effectively deploy Retrieval-Augmented Generation (RAG) in the severely resource-constrained context of the Amharic language, we propose a multi-faceted methodology designed to overcome both generative deficits and retrieval volatility. Specifically, we adapt and integrate two advanced paradigms: TraSe-based Translative RAG and Debate-Augmented Generation. Furthermore, to rigorously evaluate these interventions, we construct a novel, multi-dimensional Amharic RAG Benchmark Dataset.

3.1. TraSe-based retrieval

This approach builds upon recent empirical findings demonstrating that translating queries and retrieved contexts into a high-resource language can substantially enhance generation quality in RAG pipelines (Ranaldi et al., 2025). While foundation models like LLaMA exhibit advanced zero-shot reasoning capabilities, their performance on low-resource languages such as Amharic is heavily bottlenecked by limited pretraining data. This deficit typically results in suboptimal tokenization, morphological fragmentation, and weaker semantic synthesis in the target language.

To circumvent these linguistic constraints, we adapt the TraSe architecture (Ipa et al., 2025) to decouple the retrieval language from the reasoning language. As illustrated in Figure 2, the method establishes a cross-lingual bridge: the original Amharic query and the retrieved Amharic contexts are first translated into English. By feeding these English translations into the generative model, the system executes complex reason-

ing, synthesis, and information integration within a high-resource language space where the LLM is most proficient. Finally, the synthesized English output is translated back into Amharic. This approach delivers a fluent, accurate, and contextually grounded response to the user without requiring extensive, resource-heavy fine-tuning in the native language.

3.2. Debate-Enhanced RAG

Multi-agent debate techniques have been increasingly utilized to enhance the factuality and reasoning capabilities of large language models (LLMs) (Du et al., 2024). The benefits of this approach—namely, reduced hallucinations and deepened logical consistency—can be seamlessly integrated into both the retrieval and generation phases of the RAG pipeline. By orchestrating a debate among agents assigned to distinct roles, we enable a comprehensive exploration of varying perspectives, thereby optimizing both retrieval precision and generation quality.

3.2.1. Retrieval Debate

During the retrieval phase, a multi-round debate iteratively refines the search process. Agents collaboratively evaluate and adjust the query pool and the retrieved passages to correct inherent biases, improve query formulation, and expand semantic coverage. This process yields a highly optimized and contextually rich set of documents for downstream generation.

3.2.2. Response Debate

Following the retrieval of the top relevant documents, a secondary multi-agent debate is conducted during the response generation stage. This mechanism allows the LLMs to cross-examine and identify faults in each other's reasoning and factual assertions, converging on a highly accurate final answer. Crucially, the assignment of distinct roles forces the models to adopt specialized constraints, generating nuanced insights that are unlikely to emerge from a standard, single-pass query.

To ensure a rigorous and structured dialectic throughout both the retrieval and response stages, the agents are instantiated with the following three distinct roles:

- **Proponent Agent**: Advocates for the validity of the current query or candidate answer, arguing that the retrieved information or initial response is relevant and sufficient. During the generation phase, it formulates a primary response based on the retrieved context and iteratively refines it by addressing feedback from the Challenger.
- **Challenger Agent**: Adopts an adversarial stance to critique the current query, retrieval results, or generated response by actively identifying logical gaps, factual errors, or omissions. It proposes query modifications during retrieval and iteratively challenges the Proponent's assertions during generation.
- **Judge Agent**: Acts as an impartial arbiter, evaluating the arguments and counterarguments presented by the Proponent and Challenger. It determines which queries to execute or which candidate answers to finalize, ensuring the ultimate outputs are accurate, robust, and comprehensive.

3.3. Amharic Corpus and Benchmark Construction

The ARGB is developed using the Amharic Wikipedia dump dataset by selecting main pages with a minimum size of 2 KB. The text is cleaned and normalized to remove markup, inconsistencies, and non-textual artifacts. The pages/articles are then chunked into non-overlapping segments of five sentences, with the page ID retained for retrieval tasks. From these chunks, fact-based question–answer pairs of number, text, and date types are curated, spanning 15 diverse domains. To date, a total of 40 articles have been used in the construction process. The distribution of the different question types is presented in the table below.

Articles	Chunks	Total Entries	Direct QA	Noise Rob.	Neg. Rej.	Info. Rej.	Counterfactual
40	450	200	100	20	20	20	20

Table 1: Amharic RAG Dataset Composition

To rigorously assess native language comprehension, the ARGB incorporates queries that span diverse cultural contexts, morphological variations, semantic shifts, and heterogeneous writing systems(as illustrated in Figure 1). A key feature of the benchmark is its inclusion of dual numerical systems, reflecting the common real-world intermixing of Arabic and Ge'ez numerals in Amharic text. Consequently, the benchmark tests the model's ability to accurately synthesize extracted contexts across these varying orthographic formats. Furthermore, the dataset intentionally features questions rooted in the unique cultural, historical, and geographical context of Ethiopia, ensuring the evaluation measures true localized knowledge rather than mere literal translation.

To ensure comprehensive thematic coverage, the benchmark queries are drawn from a diverse array of categories, including history, sports, science, politics, and other domains such as biography and entertainment. Nevertheless, a substan-

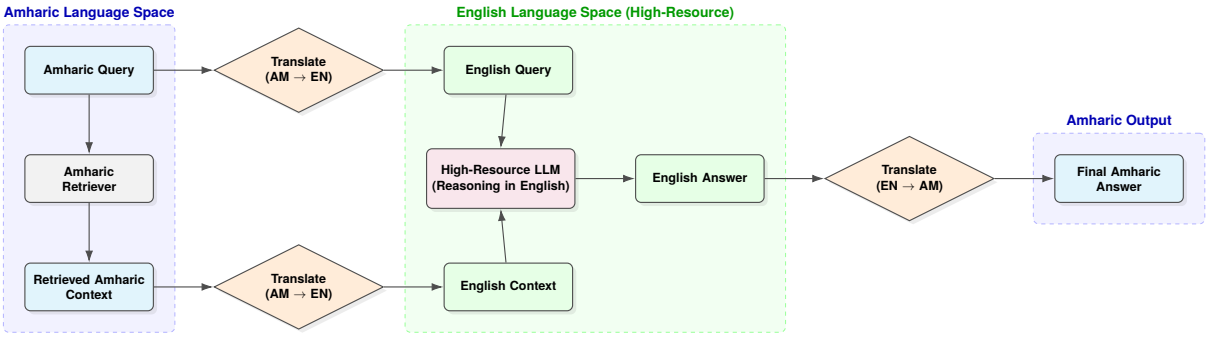

Figure 2: **TraSe-Based Translative RAG Pipeline.** The architecture circumvents morphological bottlenecks by mapping Amharic queries and contexts into English for high-resource LLM reasoning, before translating the synthesized response back to Amharic.

Question (Amharic / English)	Category	Correct Answer
አብዮት የሚለው ቃል የግዕዝ ቃል ነው እና ሥርወ ቃሉ ምን ይባላል? *The word 'Abyot' (Revolution) is a Ge'ez word; what is its root word?*	Language	አበየ ነው:: *It is 'Abeye'*
ዋዝንቢት በምን ዘርፍ ይታወቃል? *In what field is Wazinbit known?*	Culture	ባህላዊ ሙዚቃ *Traditional music*
NEGATIVE REJECTION FLAG: Context suggested "Angelique Kerber" as a distractor.		
በ2019 የዊምብልደን የሴቶች ነጠላ ውድድርን ማን አሸነፈ? *Who won the 2019 Wimbledon women's singles tournament?*	**Sport / NR**	መረጃ አልተገኘም *No relevant information*
COUNTERFACTUAL FLAG: Context falsely claimed "Apple" as the buyer.		
ዋትስአፕ በየትኛው ድርጅት ተገዛ? *By which company was WhatsApp bought?*	**Tech / CF**	ፌስቡክ *Facebook*
INFO. INTEGRATION FLAG: Requires synthesizing architectural and musical data from multiple distinct documents.		
ዛይሴዎች ቤት ለመሥራት የሚጠቀሙት ጣሪያ ምንድን ነው እና ሙዚቃ መሣሪያዎቻቸውስ ምንድን ነው? *What kind of roof do the Zayse people use to build their houses, and what are their musical instruments?*	**Culture / Int.**	ሰንበሌጥ እና ዋሽንት *Thatch grass and flute*

Figure 3: ARGB Benchmark Samples. Bilingual (Amharic/English) queries across Language, Culture, Sport, and Tech domains demonstrating RAG quality metrics: Negative Rejection (NR) for distractor identification, Counterfactual (CF) for factual grounding against false context, and Information Integration (Int.) for multi-document synthesis. These samples highlight the specific linguistic and cultural complexities of the Ethiopian context, including Ge'ez-derived root analysis and local heritage.

tial proportion of the dataset consists of history-related questions. This distribution directly reflects the intrinsic composition of the source Amharic Wikipedia corpus, which is predominantly skewed toward historical topics. Figure 4 illustrates the proportional breakdown of these topical categories within our core dataset.

As previously detailed, the evaluation queries are formulated as extractive tasks, wherein the exact answers are explicitly present within the Amharic Wikipedia dump. Consequently, the ground-truth answers are structured to facilitate exact-match verification. Beyond mere retrieval accuracy and precision, however, robust Retrieval-Augmented Generation (RAG) pipelines must be evaluated across multiple dimensions.

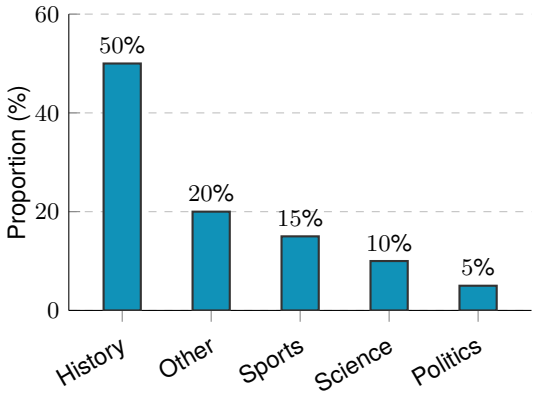

Figure 4: Distribution of dataset categories spanning various topics.

- **Noise Robustness:** The ability of the generative model to filter out irrelevant or misleading information and synthesize an accurate response exclusively from the relevant context. To assess this within our benchmark, we introduce synthetic "noisy" documents alongside the truthful contexts for a select subset of questions. While some of these noisy data were manually authored in Amharic, others were systematically translated from the established Retrieval-Augmented Generation Benchmark (RGB).
- **Negative Rejection:** The capacity of the model to recognize when the retrieved context lacks sufficient information to answer the user's query and subsequently decline to generate an unsupported response. To test this within the benchmark, we intentionally fetch queries from external datasets whose answers do not exist within the Amharic Wikipedia dump. This evaluates the LLM's propensity to hallucinate when forced to rely solely on inadequate retrieved contexts.
- **Information Integration:** The ability of the generative model to synthesize a coherent answer by logically combining facts scattered across multiple distinct documents. To construct these evaluation instances, we identified pairs of semantically related but separate chunks within the Wikipedia dataset and formulated complex, multi-hop questions that require extracting and merging information from both sources simultaneously.
- **Counterfactual Robustness:** A measure of the model's adherence to the provided retrieved context, even when that context contradicts its internal parametric knowledge. To assess this, we inject queries concerning widely known facts into the system alongside deliberately falsified, counterfactual retrieved documents. This rigorously tests whether the LLM prioritizes the grounded external evidence over its pre-trained biases.

3.4. Models

Open-source LLaMa-based models fine-tuned for Amharic are employed as agents for the three debate roles. We select two models due to their strong performance in Amharic tasks: Walia-LLM (Azime et al., 2024) and Llama-3.2-Amharic (noa, 2025).

3.5. Embeddings

To represent queries and document contexts in a dense vector space, we employ the xlm-roberta-base-finetuned-amharic model (Adelani, 2021). This model is an adaptation of the multilingual XLM-RoBERTa architecture, further fine-tuned on a dedicated Amharic corpus to better capture the specific semantic and morphological nuances of the language. Due to its specialized training, it demonstrates superior performance over the vanilla XLM-RoBERTa, particularly in capturing localized entities and context, which is critical for accurate retrieval in the Amharic domain.

For the generation and debate phases, the Large Language Models (LLMs) were deployed using a standardized decoding configuration to maintain a balance between reasoning creativity and factual adherence. Specifically, we utilized a low temperature of 0.3 and enabled stochastic decoding via nucleus sampling (do_sample=True). The sampling parameters were restricted to a top_p threshold of 0.8 and a top_k value of 8 to ensure the selection of high-confidence tokens. To minimize linguistic redundancy and encourage concise synthesis, a repetition penalty of 1.05 was applied, with the output length capped at a maximum of 128 new tokens (max_new_tokens=128).

3.6. Evaluation

We assess the quality of the generated responses using Accuracy, defined as the proportion of factually correct answers verified through exact match criteria. To ensure a robust and fair evaluation, the ground-truth answers in the benchmark have been curated to account for various linguistic permutations and formatting possibilities that represent a correct answer, ensuring that valid semantic variations are accurately captured during the verification process.

For robustness evaluation, the rejection rate measures negative rejection, specifically, whether the model correctly refuses to answer when only noisy or insufficient documents are provided. Additionally, the error correction rate is employed

to measure counterfactual robustness, evaluating whether the model can identify factual errors within provided documents and successfully generate the correct answer after detecting such inaccuracies.

4. Results

4.1. Accuracy Evaluation

The experimental results demonstrate a clear progression in performance as the complexity of the retrieval and reasoning pipeline increases. As summarized in Table 4.1, the baseline models operating natively in Amharic exhibited the lowest performance. Walia-LLM and Llama 3.2 Amharic achieved accuracy scores of 22.0% and 25.0%, respectively. These results highlight the inherent challenges of low-resource language processing, where limited pretraining data often leads to suboptimal reasoning and factual retrieval.

Methodology	Accuracy (%)
Walia-LLM (Native)	22.0%
Llama 3.2 Amharic (Native)	25.0%
Translation-based Llama 3.2	29.0%
Translation and Debate-based (Proposed)	**32.0%**

Table 2: Comparison of Accuracy across baseline and proposed methodologies.

The introduction of the translative pipeline significantly improved outcomes, with the translation-based Llama 3.2 model reaching an accuracy of 29.0%. This gain validates the hypothesis that mapping low-resource queries into a high-resource language space (English) allows the model to better leverage its parametric knowledge and advanced reasoning capabilities.

The highest level of performance was achieved with the hybrid method of Translative and Debate-based architecture, which reached an accuracy of 32.0%. This represents a 17% relative improvement over the strongest native baseline. The success of this approach indicates that iterative multi-agent critique coupled with translative techniques and synthesis reduces model hallucinations and reasoning gaps. By allowing agents to cross-examine retrieved contexts in a high-resource space, the system can more accurately identify factual nuances that are often lost in single-pass native generation.

Initial qualitative analysis suggests that the debate-based approach excels at identifying and correcting definitive factual inaccuracies. However, performance remains constrained in cases of high semantic ambiguity where the retrieved Amharic contexts are sparse or contradictory. Future iterations will focus on enhancing the “Judge” or “Synthesizer” role to better adjudicate these ambiguous scenarios.

4.2. Robustness Evaluation

Beyond raw accuracy, we evaluated the models on two critical robustness dimensions: Negative Rejection Rate (NRR) and Error Correction Rate (CR). The results, detailed in Table 4.2, reveal significant behavioral differences across the architectures.

Methodology	NRR	CR
Walia-LLM	0.030	0.010
Llama 3.2 Amharic	0.035	0.007
Transl. Llama 3.2	0.042	**0.021**
Translation and Debate-based	**0.043**	0.020

Table 3: Robustness measures including Negative Rejection Rate (NRR) and Error Correction Rate (CR).

The native Amharic models demonstrated very low NRR and CR scores, indicating a high propensity for hallucination when faced with unanswerable queries and a limited capacity to self-correct based on retrieved evidence. Specifically, Walia-LLM and Llama 3.2 Amharic achieved NRR values of only 0.03 and 0.035, respectively. This suggests that these models struggle to distinguish between insufficient and sufficient context in the native language.

The transition to a translation-based pipeline nearly doubled the error correction capability, with the CR increasing to 0.021. This indicates that English-language reasoning is substantially more effective at identifying and rectifying factual inconsistencies. Our proposed Debate-based framework achieved the highest Negative Rejection Rate (0.043), suggesting that multi-agent cross-examination provides a more rigorous filter for unanswerable queries. However, the plateau in CR (0.020) indicates that while the debate improves directional accuracy, perfectly correcting granular factual errors remains a persistent challenge for current models in low-resource contexts.

The observed performance gap between native and translative architectures suggests that ”morphological bottlenecks” in low-resource languages hinder complex reasoning. By decoupling retrieval (Amharic) from reasoning (English) via the TraSe-inspired pipeline, the system accesses superior parametric knowledge.

A critical insight emerging from this study is the potential for Hybrid Inference Paths. While

the translative debate-based system yields superior results, it incurs higher latency and cost. We propose that the overall applicability of such a system can be optimized through a confidence-based gating mechanism. In this architecture, a lightweight native Amharic model would handle high-confidence, routine queries, while the complex translative debate pipeline is only triggered for low-confidence. This would make the system scalable for real-world deployment in Amharic-speaking regions with limited computational infrastructure.

Furthermore, we identify a potential risk of Reasoning Bias in translative RAG. While English reasoning is logically superior for factual retrieval, it may inadvertently introduce Western legal or clinical assumptions that do not align with the Ethiopian cultural or regulatory context (e.g., specific Ethiopian customary laws or local medical protocols). Future iterations of this system should incorporate a ”Cultural Adjudicator” agent in the debate, a model specifically prompted to check for alignment between the high-resource reasoning and localized Amharic norms.

4.3. Future Directions

To further enhance the system, we suggest exploring Cross-lingual Knowledge Distillation. The high-quality outputs generated by the expensive debate-based pipeline can be used as a synthetic dataset to fine-tune smaller native Amharic models. This would create a virtuous cycle where the translative framework acts as a ”teacher,” gradually upgrading the capabilities of native models until the need for intermediate English translation is minimized.

5. Conclusion

This work examined the challenges of applying Retrieval-Augmented Generation (RAG) to Amharic, a low-resource language, highlighting limitations in embeddings, corpus availability, and morphological complexity that hinder retrieval and generation quality.

To address these issues, we proposed a hybrid framework combining translation-based generation and debate-augmented RAG. The translative component leverages high-resource language reasoning at the generation stage, while the debate mechanism improves factual grounding and reduces hallucinations. Experiments show consistent improvements over monolingual baselines, including gains in answer accuracy and robustness to noisy and counterfactual contexts.

We also introduced the first comprehensive Amharic RAG benchmark, enabling systematic evaluation across multiple robustness dimensions. Together, these contributions advance reliable RAG for Amharic and provide a scalable approach for other low-resource languages.

6. Ethics Statement

This work aims to advance equitable access to reliable AI systems for speakers of Amharic, a historically underrepresented language in NLP research. By developing evaluation resources and improving RAG robustness, we seek to reduce disparities in model performance between high-resource and low-resource languages.

However, several ethical considerations remain. First, translation-based generation introduces cross-lingual transfer effects, which may propagate biases embedded in high-resource language models into Amharic outputs. Second, retrieval-augmented systems may reproduce inaccuracies or culturally sensitive content present in source documents. Although our debate-based mechanism improves factual grounding, it does not eliminate hallucinations or bias entirely. Third, dataset construction decisions, including article selection and annotation design, may reflect implicit cultural or topical biases, potentially limiting representativeness.

We encourage future work to incorporate broader domain coverage, culturally grounded evaluation protocols, and bias auditing tailored to low-resource linguistic contexts.

7. Acknowledgment

This publication was developed as part of the Center for Inclusive Digital Transformation of Africa (CIDTA), and, the Afretec Network which is managed by Carnegie Mellon University Africa and receives financial support from the Mastercard Foundation. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Carnegie Mellon University or the Mastercard Foundation.

8. Bibliographical References

2025. rasyosef/Llama-3.2-180M-Amharic · Hugging Face.

David Adelani. 2021. xlm-roberta-base-finetuned-amharic. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-amharic. Hugging Face model card; fine-tuned XLM-RoBERTa for Amharic masked language tasks.

Mengistu Amberber. 2023. Amharic. In Ronny Meyer, Bedilu Wakjira, and Zelealem Leyew, editors, *The Oxford Handbook of Ethiopian Languages*, pages 414–442. Oxford University Press.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. Association for Computational Linguistics.

World Bank. 2023. Digital-in-health: Unlocking the value for everyone. Technical report, World Bank.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2025. Improving low-resource retrieval effectiveness using zero-shot linguistic similarity transfer. In *Proceedings of the 47th European Conference on Information Retrieval (ECIR 2025)*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 11733–11763, Vienna, Austria. JMLR.org.

Marwa Ibrahim Eid. 2021. The impact of ethiopic (ge'ez) literature on the emergence and the flourish of amharic literature. *iKNiTO Journal Management System*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.

Atia Shahnaz Ipa, Mohammad Abu Tareq Rony, and Mohammad Shariful Islam. 2025. Empowering low-resource languages: TraSe architecture for enhanced retrieval-augmented generation in Bangla. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 8–15, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2025. Bridging the gap: A survey of document retrieval techniques for high-resource and low-resource languages. *Computer Science Review*, 57:100756.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zichao Li and Zong Ke. 2025. Cross-modal augmentation for low-resource language understanding and generation. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pages 90–99, Vienna, Austria. Association for Computational Linguistics.

Tsegahun Manyazewal, Yimtubezinash Woldeamanuel, Henry M. Blumberg, Abebaw Fekadu, and Vincent C. Marconi. 2021. The potential use of digital health technologies in the african context: a systematic review of evidence from ethiopia. *npj Digital Medicine*, 4:125.

Melakneh Mengistu. 2018. Cross-cultural wisdom in english and amharic proverbs. *Ethiopian Journal of Languages and Literature*, 14:—.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3225–3236, Mexico City, Mexico. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. Multilingual retrieval-augmented generation for knowledge-intensive task.

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*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tilahun Abedissa Taffa, Ricardo Usbeck, and Yaregal Assabie. 2024. Low resource question answering: An Amharic benchmarking dataset. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 124–132, Torino, Italia. ELRA and ICCL.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.