Mwando: Leveraging AI to Preserve and Teach shiKomori

Naira Abdou Mohamed[1*], **Haidar Nassur Said Ali**[2], **Mohamed Hazra**[3], **Naoufal Mohamed Soibira**[1,4], **Roushnaty Ali Yamani**[3]
[1] Rifai, Moroni, Comoros
[2] CP2BM Lab - Hassan II University, Casablanca, Morocco
[3] Ibn Tofail University, Kenitra, Morocco
[4] Sciences Po Grenoble, Grenoble, France

Abstract

This paper presents *Mwando*, a virtual educational assistant designed to support the teaching and preservation of shiKomori, the language of the Comoros Islands. The system covers the four main dialectal variants (shiNgazidja, shiMwali, shiNdzuani and shiMaore) through a knowledge base constructed from phrases, proverbs, dictionaries and grammar lessons. A multi-agent architecture combining vector search, a knowledge graph and web search fallback enables accurate and context-aware responses. Evaluation on 500 queries demonstrates strong performance on vocabulary lookup and grammar explanations, while qualitative case studies illustrate both capabilities and current limitations. This work represents an initial step toward computational support for shiKomori and provides a blueprint for developing AI-powered educational tools for other low-resource languages.

1 Introduction

Language preservation is a central concern for international organizations such as the United Nations Development Programme (UNDP), which emphasizes African languages as key drivers of the continent development (United Nations Development Programme and Ministry of Enterprises and Made in Italy, 2024). However, this potential remains largely unrealized, as very few technological solutions currently support these languages (Wild, 2025). This situation is problematic, as language and culture are major drivers of development (Rotondo, 2016) and play a central role in everyday life across African societies. However, advances in language processing technologies, particularly generative AI systems, offer a promising opportunity: when appropriately adapted to cultural and linguistic specificities, these models can serve as effective conduits for the preservation of cultural heritage (Colace et al., 2025; Koc, 2025).

In this work, we aim to contribute to the initial development of Natural Language Processing (NLP) technologies for shiKomori, the language spoken in the Comoros Islands. Building on our previous work on foundational NLP resources for Comorian and its dialectal variations (Naira et al., 2024, 2025), we shift the focus to an applied educational setting.

Specifically, we introduce Mwando, a virtual assistant designed for teaching shiKomori, with the goal of supporting the transmission and preservation of Comorian culture. Mwando ("beginning" in shiKomori) honors Sheikh Ahmed Kamar-Eddine's pioneering work on Comorian language documentation and writing standardization (Lafon, 2007; Naira et al., 2025). Our contributions are threefold:

- We compile and organize a multi-source corpus for shiKomori, covering phrases, proverbs, dictionaries and grammar lessons and make these data publicly available to support future research.
- We develop a virtual assistant for educational purposes, capable of interacting with learners in shiKomori.
- We evaluate the assistant's potential for supporting language learning and cultural preservation, providing insights for further development of NLP resources for extremely low-resource languages.

2 Motivations

The arrival of generative AI has been a serious game changer in education (Team et al., 2025) and it was immediately adopted by students and language teachers (Zaim et al., 2025). Among the reasons justifying this rapid adoption is the possibility for students to obtain quick feedback through

*Contact: naira@rifailabs.com

a virtual teacher and for teachers, the ability to quickly and effectively customize lessons according to different cases (Galaczi and Luckin, 2024). It is precisely along these lines that a project was adopted in Senegal in elementary education to support French teachers in better teaching in classrooms composed of students with significant cultural and linguistic diversity (Agence Française de Développement, 2025).

In the Comoros, although the use of AI and information technology in education is very poorly documented (Roukiyat, 2026), there are previous studies that have emphasized the importance of taking shiKomori into account in teaching for greater effectiveness (DANIEL, 2024). And local initiatives such as the Maecha association have set themselves the goal of literacy in Comorian for the adult population, which is composed of approximately 49.7% illiterate individuals (Chauvet, 2015).

Consequently, it is crucial to consider modern solutions for learning shiKomori. Such initiatives could not only benefit other categories of people and sectors, such as improving the experience of tourists in the archipelago, but they would also play an important role in the country's sustainable development (Roukiyat, 2026). Furthermore, they would hold strong potential for the inclusion of descendants of the Comorian diaspora, particularly those in France, whose population is estimated at more than 300,000 people (Le Monde, 2024) and who contribute significantly to the development of the archipelago (Abdillahi, 2012).

3 Linguistic Resources

One of the biggest challenges for this research was obtaining the data. Our approach was to consult every possible resource available in order to build our knowledge base. However, to ensure the reliability of the dataset, manual checks were necessary at each stage of processing. In this section, we therefore describe all the corpora used, from the raw data to the transformed data for this work.

3.1 Language Overview

Spoken exclusively in the Comoros archipelago, shiKomori is a Bantu language, closely related to Swahili (Ahmed Chamanga, 2022) and to the Sabaki language group (Serva and Pasquini, 2021). It consists of four dialectal varieties, each primarily associated with one island, yet exhibiting a high degree of mutual intelligibility among speakers (Ahmed Chamanga, 2022; Naira et al., 2024). ShiKomori holds the status of a national language in the archipelago, while French and Arabic are the official languages. In practice, shiKomori is mainly used orally in everyday communication and informal media, French dominates administrative and formal educational domains and Arabic is predominantly used in religious contexts (Chauvet, 2015).

In written form, shiKomori can be transcribed using either the Latin alphabet or the Arabic script (Lafon, 2007; Naira et al., 2025). However, its written use remains limited and is characterized by a lack of orthographic standardization. Despite recent initiatives aimed at improving the representation of African languages in NLP (Adebara et al., 2025), shiKomori remains largely underrepresented in the field, except for a few prior works (Naira et al., 2024, 2025; Abdourahamane et al., 2016).

3.2 Data Sources

This work relies on six primary data sources, illustrated in Figure 1. These sources are grouped into four main categories of data:

- **Useful phrases**: These consist of commonly used phrases in shiKomori, without distinction between dialectal varieties. The phrases are translated into English and cover a wide range of topics, from simple greetings to expressions commonly used in contexts such as markets, travel and everyday interactions. The data were obtained from a Google Drive repository gathering various resources related to the Comoros Islands[1].

- **Proverbs**: The proverbs were collected from the Instagram page *ProverbesComoriens*[2] using the Apify scraping tool. Each post contains proverbs in shiKomori along with their French and English translations. However, the posts are published as images. To extract the textual content, we used DeepSeek-OCR (Wei et al., 2025). This OCR-based extraction occasionally introduced minor errors, particularly with diacritics and special characters, which were manually corrected during the data cleaning phase.

[1] https://drive.google.com/drive/folders/17C_03qCMm2rGDhgGqi_8PKJfR30V_S96

[2] https://www.instagram.com/proverbescomoriens/

- **Dictionaries**: Two dictionaries were used. The first, focusing on shiNgazidja, is a notable work produced by the Bahari Foundation[3], which provides shiNgazidja words and expressions translated into English, with explicit annotation of nominal and grammatical classes. The second dictionary was obtained from the same Google Drive repository mentioned above and contains shiMwali words translated into English, along with grammatical class annotations.

- **Grammar lessons**: The grammar lessons are in shiMaore and shiNdzuani and were obtained from the platforms The Swiss Bay[4] and LiveLingua[5], respectively.

Finally, all data sources used in this work are publicly available online. For the Instagram data, we only collected publicly accessible posts and complied with the platform's terms of service. The dictionaries and grammar manuals are freely distributed by their respective publishers or hosting platforms. We release our processed dataset under an open license to facilitate future research on Comorian language processing (see Table 1).

3.3 Knowledge Base Construction

A processing step was necessary to better structure the data. Indeed, among the collected data, for example, in the case of grammar lessons, some data were in the form of tables, as illustrated in Figure 1. Although Large Language Models (LLM) are effective at understanding textual data, they exhibit limitations when processing text mixed with tabular content (Liu et al., 2024). This issue primarily arises during the data tokenization phase. To mitigate this, we first applied a document structuring step by converting the documents into Markdown format using the Word2md platform[6].

Subsequently, the Markdown content was fed into the ChatGPT API to restructure the data into a more easily exploitable knowledge base. This process involved extracting elements such as chapter sections, covered concepts, tables, definitions and related content and organizing them into JSON files. The specific structuring strategy varied depending on the type of data. In all cases, a manual validation step was conducted to review and correct the outputs generated by ChatGPT.

4 Chatbot Workflow

In recent years, the use of multi-agent approaches has significantly improved the reasoning capabilities of virtual assistants, particularly in situations where information must be retrieved from multiple sources (Salve et al., 2024). In our context, where we need to handle multiple dialectal variants and types of datasets, adopting a multi-agent approach was a natural choice. Figure 2 presents the overall pipeline that summarizes the core of our reasoning and information retrieval system.

To address the challenges of handling multiple Comorian dialects and heterogeneous data sources, we designed a multi-agent system that orchestrates specialized agents, each responsible for a specific type of query or knowledge domain. This architecture, illustrated in Figure 2, enables flexible and robust information retrieval by combining vector search, knowledge graphs and external web resources.

4.1 Overall Pipeline

The system operates as follows: when a user submits a query, it is first processed by a central Retrieval Agent. This agent acts as an orchestrator: it analyzes the query and determines which specialized agent should be invoked to provide the most accurate and comprehensive answer. The response is then generated by an LLM based on the retrieved information and returned to the user.

4.2 Core Components and Specialized Agents

The system comprises five main specialized agents, each leveraging different tools and knowledge bases:

- **Retrieval Agent**: As the entry point for all queries, this agent is responsible for query understanding and dynamic routing. It decides whether to forward the request to a single specialist or to combine results from multiple agents. It relies on an LLM for intent classification and on a Vector Search engine to retrieve relevant passages from a precomputed embedding database covering all dialectal variants.

[3] https://fr.scribd.com/document/619345660/ShiNgazidja-English-Dictionary

[4] https://theswissbay.ch/pdf/Books/Linguistics/Mega%20linguistics%20pack/African/Niger-Congo/Bantu/Shimaore%2C%20Manuel%20Grammatical%20de.pdf

[5] https://www.livelingua.com/peace-corps/Comorian/Shinzwani%20Grammar%20Book-%201st%20ed%202015.pdf

[6] https://word2md.com/

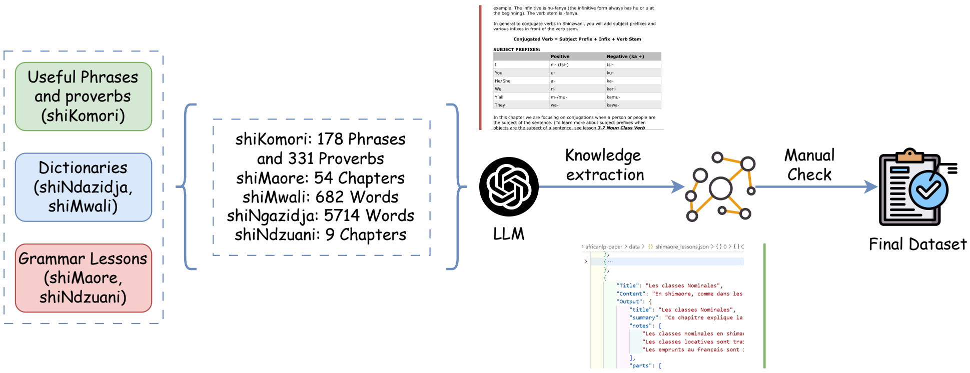

Figure 1: Data Preparation Pipeline.

Table 1: Overview of the datasets used in this work.

Category	Dialect(s)	# Rows	Link
Useful Phrases	All varieties	178	https://huggingface.co/datasets/rifailabs/shikomori_phrases
Proverbs	All varieties	329	https://huggingface.co/datasets/rifailabs/shikomori_proverbs
Dictionary	shiNgazidja	5,710	https://huggingface.co/datasets/rifailabs/shingazidja_dictionary
Dictionary	shiMwali	682	https://huggingface.co/datasets/rifailabs/shimwali_dictionary
Grammar Lessons	shiMaore	54	https://huggingface.co/datasets/rifailabs/shimaore_lessons
Grammar Lessons	shiNdzuani	67	https://huggingface.co/datasets/rifailabs/shindzuani_lessons

- **Translator/Proverb Agent**: This agent handles two closely related tasks. For translation queries, it accesses bilingual dictionaries (shiNgazidja-English, shiMwali-English) and phrase collections. For the queries related to proverb, we use a dedicated graph database that stores proverbs along with their meanings, cultural contexts and translations. The graph structure allows for semantic navigation (e.g., finding proverbs related to a specific theme).

- **Teacher Agent**: Focused on pedagogical needs, this agent provides grammar explanations, conjugation tables and language learning resources. It draws upon the grammar manuals (shiMaore and shiNdzuani) and can generate structured lessons or exercises. It also maintains a database of common learner errors and misconceptions.

- **Culture Agent**: For queries that fall outside the scope of other agents or require general information not covered by specialized knowledge bases, this agent performs live web searches. It serves as a fallback mechanism, ensuring that the system can still provide relevant answers even when internal resources are insufficient.

4.3 Technical Implementation

The system is built around several key technologies. For embeddings and vector search, all textual resources including dictionaries, proverbs, phrase collections and grammar lessons, were embedded using Qwen3 Embedding (Zhang et al., 2025), a multilingual sentence transformer. This model was selected for two main reasons. First, it is among the top-performing embedding models on recent benchmarks, offering high-quality semantic representations across multiple languages. Second, with only 2 billion parameters, it is lightweight enough to run in resource-constrained environments, making it suitable for deployment in regions with limited computational infrastructure. The model was deployed locally using Ollama, a lightweight framework for running large language models efficiently on consumer hardware. The resulting embeddings were indexed in FAISS, a vector database (Douze et al., 2025) for efficient similarity search, enabling rapid retrieval of relevant passages based on query semantics rather than keyword matching alone.

Regarding the knowledge graph, proverbs and

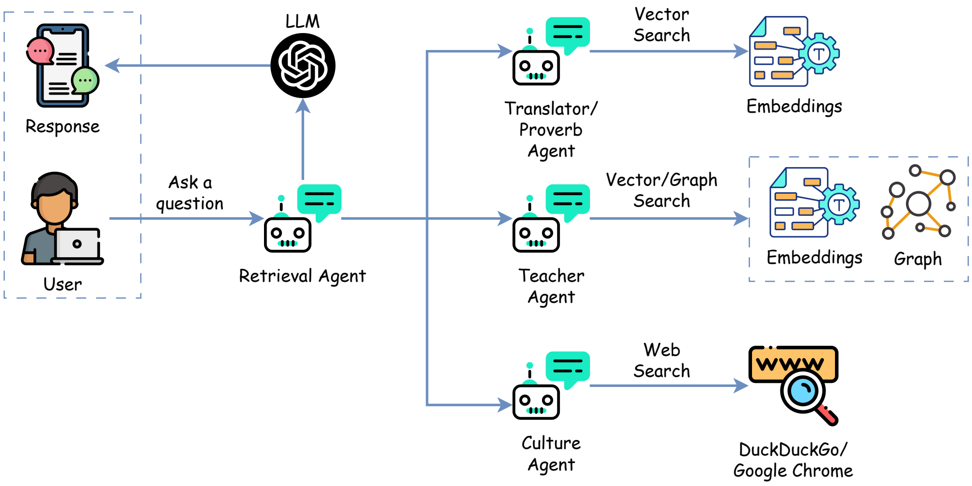

Figure 2: RAG Pipeline.

cultural knowledge are stored in a graph database using Neo4j. In this structure, nodes represent proverbs, definitions and everyday life entities, while edges capture semantic and thematic relationships. This design allows for nuanced navigation and reasoning across the knowledge base, which is particularly valuable for answering queries.

For orchestration and reasoning, agent coordination is implemented using LangChain. This handles prompt engineering, tool invocation, dynamic query routing and response synthesis. The core reasoning and language generation tasks are performed using Groq's ultra-fast inference platform, which hosts gpt-oss, an open-source LLM optimized for conversational AI. Additionally, the Culture Agent can perform live searches using DuckDuckGo when information is unavailable internally. Retrieved web content is summarized by gpt-osss before integration into the final response.

5 Evaluation

5.1 Retrieval Performance

For this evaluation, we set a target of testing 500 queries, divided among five evaluators with 100 queries each. Each evaluator categorized their queries according to the following four areas: Proverb Interpretation, Grammar Explanations, Vocabulary Lookup and Cultural Questions. We measured the following metrics, often used to evaluate the quality of a recommendation system (Jadon and Patil, 2024) and more recently for evaluating a RAG system (Gan et al., 2025):

- **Precision@k**: This metric measures the proportion of relevant documents among the top k results.
- **Recall@k**: This metric measures the proportion of relevant items retrieved out of all existing relevant items. In other words, if there are p relevant items in total and among the top k results we find q relevant items, then Recall@k is q/p.
- **Mean Reciprocal Rank (MRR)**: The average of the reciprocal ranks of the first relevant result. It is computed as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$$

  where N is the total number of queries and rank_i is the rank position of the first relevant result for query i.

Table 2 presents retrieval performance across different query categories. The system performs best on vocabulary lookup, which benefits from direct matches in dictionary entries. Grammar explanations and proverb interpretation yield lower scores due to the complexity of these queries and the variability in how concepts are expressed across grammar manuals. Cultural questions are the most challenging, as they often require synthesis from multiple sources or reliance on web search.

5.2 Response Quality

The goal here is to evaluate the quality of the responses provided by the LLM. To do this, we revisited the 500 queries used during the retrieval evaluation to examine the responses generated by

Table 2: Retrieval performance by query category

Query Category	Precision@5	Recall@5	MRR
Vocabulary Lookup	0.82	0.76	0.81
Grammar Explanations	0.71	0.60	0.73
Proverb Interpretation	0.69	0.64	0.78
Cultural Questions	0.65	0.63	0.69
Average	**0.71**	**0.65**	**0.75**

the LLM. These responses were evaluated using a 5-point Likert scale along the following three axes:

- **Accuracy**: Is the information factually correct?
- **Completeness**: Does the response fully address the query?
- **Clarity**: Is the response easy to understand for a learner?

Table 3 summarizes the average scores. The evaluators noted that vocabulary and grammar queries were generally well-handled, while proverb interpretations sometimes lacked the depth of cultural context that a human expert would provide. Cultural questions occasionally suffered from incomplete or overly general answers when relying on web search.

5.3 Case Studies

While the quantitative metrics presented in the previous section provide a high-level overview of system performance, they do not capture the nuances of individual interactions. To complement these findings, we present four representative case studies (See Figures 3, 4, 5 and 6) that illustrate the system's behavior in practice. These examples span the four different query categories analyzed previously. Each case concludes with a brief discussion of strengths and limitations observed during the interaction.

5.4 Discussion of Limitations

Our evaluation reveals several limitations:

- **Coverage gaps**: Despite our efforts, some dialectal variants and specialized vocabulary remain underrepresented, particularly for shiNdzuani and shiMwali.
- **Cultural depth**: Proverb interpretations and cultural explanations sometimes lack the nuanced understanding that a human expert would provide.
- **Web search quality**: When relying on web search, the system occasionally retrieves low-quality or irrelevant information, affecting response accuracy.
- **Evaluation scale**: Our human evaluation involved only five annotators; a larger study with more diverse participants (learners, teachers, elders) would provide more robust insights.

Despite these limitations, the results demonstrate that *Mwando* provides a solid foundation for computer-assisted language learning for shiKomori, with particular strength in vocabulary and grammar support.

6 Conclusion and Future Work

In this paper, we introduced *Mwando*, a virtual educational assistant for shiKomori, supporting its four dialectal variants. We assembled a multi-source corpus comprising phrases, proverbs, dictionaries and grammar lessons and developed a multi-agent architecture combining vector search, a knowledge graph and web search fallback. The system leverages lightweight embeddings (Qwen3 via Ollama) and fast reasoning (Groq with gpt-oss, an open-source LLM).

Evaluation on 500 queries showed strong performance on vocabulary lookup and grammar explanations, with qualitative case studies illustrating both capabilities and current limitations, particularly in cultural depth and proverb interpretation.

Coverage remains uneven across dialects and web search quality is variable. Future work will focus on expanding resources for shiNdzuani and

Table 3: Human evaluation of response quality (average scores, 1-5 scale)

Query Category	Accuracy	Completeness	Clarity
Vocabulary Lookup	4.35	4.26	4.80
Grammar Explanations	4.24	4.76	4.39
Proverb Interpretation	4.08	3.82	4.15
Cultural Questions	3.65	3.24	3.86
Average	**4.08**	**4.02**	**4.30**

Figure 3: Vocabulary Query (shiNgazidja)

User: Translate the word "hospital" in shiNgazidja.

Assistant: Word: lopitali / lapitali (malapitali)
Meaning: "hospital"
Grammatical class: noun (undefined)
Noun class: 9-10, 5-6
Origin: French (fr)

Analysis: *The system correctly retrieves the dictionary entry with morphological variations (singular/plural) and etymology.*

shiMwali, enriching cultural knowledge with expert input, improving web search filtering and deploying *Mwando* in real-world educational settings. We hope this work contributes to the preservation of Comorian heritage and serves as a blueprint for AI support in other low-resource languages.

References

Youssouf Abdillahi. 2012. *La diaspora de la Grande Comore à Marseille et son apport sur le développement de l'île*. Theses, Université de la Réunion.

Moneim Abdourahamane, Christian Boitet, Valérie Bellynck, Lingxiao Wang, and Hervé Blanchon. 2016. Construction d'un corpus parallèle français-comorien en utilisant de la TA français-swahili. In *TALAf (Traitement Automatique des Langues africaines)*, Paris, France.

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating LLM performance on African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.

Agence Française de Développement. 2025. Using AI to improve language learning in Senegal. `https://www.afd.fr/en/AI-for-language-learning-in-senegal`. Accessed: 2026-02-14.

Mohamed Ahmed Chamanga. 2022. *ShiKomori, the Bantu Language of the Comoros: Status and Perspectives*, page 79–98. BRILL.

Aurélie Chauvet. 2015. Statuts des langues et éducation de base aux comores. *Rev. int. d éduc. Sèvres*, 70(70):77–84.

Francesco Colace, Rosario Gaeta, Angelo Lorusso, Michele Pellegrino, and Domenico Santaniello. 2025. New ai challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, 75:168–193.

Riziki Souf DANIEL. 2024. Le shikomor pour enseignement / apprentissage du français langue étrangère : issue interculturelle de l'insularité de ndzuani. *Revue Internationale du Chercheur*, 5(2).

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Evelina Galaczi and Rose Luckin. 2024. Generative AI and language education: Opportunities, challenges and the need for critical perspectives. Cambridge papers in english language education, Cambridge University Press & Assessment. Accessed: 2026-02-14.

Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025.

Figure 4: Proverb Interpretation

User: I want a proverb about loss and its meaning.

Assistant: Shikomori proverb: *Namdodjipviwa ne shahanyu owakati mlonisho.*
Meaning: We only realize the value of something after we have lost it; therefore, we must appreciate it while it is still there.
Topic: Appreciate the present
Category: Time & Present

Analysis: *The system successfully retrieves a relevant proverb based on semantic matching of "loss". The response includes both the original proverb and its French translation, along with thematic tags. The informal query syntax ("i want a proverb about loss") is correctly interpreted.*

Figure 5: Cultural Questions

User: What is the best way to greet an elderly person in ShiNdzuani?

Assistant: The most appropriate greeting for an older person in ShiNdzuani is:
Mungu nahunusuru.
This is the customary salutation used when addressing someone older than yourself.

Analysis: *The system correctly identifies the culturally appropriate greeting for addressing elders in ShiNdzuani. This demonstrates its ability to capture social norms and respectful language use, which is essential for learners navigating Comorian cultural contexts. However, the response could be enriched with additional context, such as the literal meaning of the phrase ("May God protect you") or the expected response from the elder.*

Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *Preprint*, arXiv:2504.14891.

Aryan Jadon and Avinash Patil. 2024. *A Comprehensive Survey of Evaluation Techniques for Recommendation Systems*, page 281–304. Springer Nature Switzerland.

Vincent Koc. 2025. Generative ai and large language models in language preservation: Opportunities and challenges. *Preprint*, arXiv:2501.11496.

Michel Lafon. 2007. Le système Kamar-Eddine : une tentative originale d'écriture du comorien en graphie arabe. *Ya Mkobe*, 14-15:29–48.

Le Monde. 2024. Comores : la diaspora installée en France dénonce son exclusion du scrutin présidentiel. *Le Monde*. Accessed: 2026-02-14.

Tianyang Liu, Fei Wang, and Muhao Chen. 2024. Rethinking tabular data understanding with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.

Abdou Mohamed Naira, Abdessalam Bahafid, Zakarya Erraji, Anass Allak, Mohamed Soibira Naoufal, and Imade Benelallam. 2025. Preserving comorian linguistic heritage: Bidirectional transliteration between the Latin alphabet and the Kamar-eddine system. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 11–18, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdou Mohamed Naira, Abdessalam Bahafid, Zakarya Erraji, and Imade Benelallam. 2024. Datasets creation and empirical evaluations of cross-lingual learning on extremely low-resource languages: A focus on comorian dialects. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 140–149, St. Julians, Malta. Association for Computational Linguistics.

Francesco Rotondo. 2016. *Cultural Heritage as a Key for the Development of Cultural and Territorial Integrated Plans*, page 21–27. Springer International Publishing.

Issa Goba Roukiyat. 2026. The incorporation of shikomori to improve ict comprehension, access, and uptake by comorian communities. *Digital Policy Studies*, 4(1):57–83.

Aniruddha Salve, Saba Attar, Mahesh Deshmukh, Sayali Shivpuje, and Arnab Mitra Utsab. 2024. A collaborative multi-agent approach to retrieval-augmented generation across diverse data. *Preprint*, arXiv:2412.05838.

Maurizio Serva and Michele Pasquini. 2021. The sabaki languages of comoros. *INDIAN OCEAN REVIEW OF SCIENCE AND TECHNOLOGY*.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Ankit Anand, Avishkar Bhoopchand, Brett Wiltshire, Daniel Gillick, Daniel Kasenberg, Eleni Sgouritsa, Gal Elidan, Hengrui Liu, Holger Winnemoeller, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, and 18 others. 2025. Evaluating gemini in an arena for learning. *Preprint*, arXiv:2505.24477.

United Nations Development Programme and Ministry of Enterprises and Made in Italy. 2024. Scaling language data ecosystems to drive industrial development growth.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Sarah Wild. 2025. Ai models are neglecting african languages — scientists want to change that. *Nature*.

Muhammad Zaim, Safnil Arsyad, Budi Waluyo, Havid Ardi, Muhd. Al Hafizh, Muflihatuz Zakiyah, Widya Syafitri, Ahmad Nusi, and Mei Hardiah. 2025. Generative ai as a cognitive co-pilot in english language learning in higher education. *Education Sciences*, 15(6):686.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Figure 6: Grammar Explanation (shiNdzuani)

User: How do noun classes work in shiNdzuani?

Assistant: In shiNdzuani, every noun belongs to a **noun class**. The class is signalled by a **prefix** on the noun (and on the plural form of the noun). That same prefix is also used by adjectives, numbers, demonstratives and verbs so that they **agree** with the noun.

How the class is recognised:

- Look at the singular prefix of the noun.
- The corresponding plural prefix tells you which class it is (the "singular-plural prefix correspondence").
- Example: the singular prefix *m-* (or *mw-*) pairs with the plural prefix *wa-* (this is class 1-2).

Examples:

Singular	Plural	Class	Meaning
gari	magari	class ? (m-/ma-)	"voiture / voitures"
dzitso	matso	class ? (d-/ma-)	"œil / yeux"

Agreement:

- An adjective placed after the noun must take the same class prefix: *gari kubwa* (big car). If the noun were plural, the adjective would also change: *magari makubwa*.
- Numbers and ordinal numbers behave the same way; most change form to match the class, except a few invariable ones (e.g., *muhimu*, *sita*, *saba*, *shendra*, *kumi*).
- Verbs also carry the noun-class prefix so that the subject and verb agree in class.

So, to use shiNdzuani correctly you:

1. Identify the noun's class by its singular-plural prefix pair.
2. Apply the same prefix to any modifying adjective, number, demonstrative, or verb.
3. If a word is listed without a prefix change (e.g., the invariable adjectives above), it does not change with the noun class.

Analysis: *This response demonstrates the system's ability to provide structured grammatical explanations. It correctly identifies noun class mechanisms and provides clear examples with agreement rules.*