BambaraMLLM: A Unified Multilingual Multimodal Large Language Model for Comprehensive Bambara Language Processing

Seydou Diallo[1] **Allahsera Auguste Tapo**[2] **Kevin Assogba**[2] **Christopher Homan**[2]
[1]Dakar American University of Science & Technology, Senegal
[2]Rochester Institute of Technology, NY, USA
sdiallo@mydaust.org aat3261@rit.edu kta7930@rit.edu cmhvcs@rit.edu

1 Motivation and Problem Statement

Current technologies for under-resourced languages often employ a fragmented paradigm, using separate models for natural language processing (NLP) tasks, including automatic speech recognition (ASR), machine translation (MT), text generation (TG), and text-to-speech (TTS) synthesis (Amalas et al., 2024; de Zuazo et al., 2025; Liang et al., 2025). This approach creates resource inefficiencies, prevents cross-task learning, and challenges adoption in the absence of a platform offering access to all modalities. To bridge the digital divide for the Bambara community, there is a need for a unified system that leverages shared multimodal representations to improve performance across all tasks while reducing computational cost.

2 BambaraMLLM

This paper introduces BambaraMLLM, a multilingual multimodal large language model for the Bambara language, built with two architectural foundations: Mixture of Experts (MoE) (Cai et al., 2025) and MatFormer (Devvrit et al., 2024). These architectures balance parameter efficiency and performance under hardware resource constraints.

2.1 Design Principles

BambaraMLLM incorporates three key principles: **Efficient Multimodal Modeling with Architectural Tradeoff:** To address the challenge of parameter inefficiency in multimodal models due to divergence between modalities, we combine shared representations from MoE and MatFormer to achieve a high-quality model at reduced computation cost. **Multi-Task Learning via Expert Routing and Cross-Task Transfer:** To overcome the imbalanced multi-task learning under limited data, we combine expert routing with cross-task transfer, from models based on high-resource languages, for stable performance across modalities.

Context-Oriented Training and Evaluation: To account for the lack of Bambara instruction datasets for deployment, we propose data creation and evaluation frameworks to support BambaraMLLM.

2.2 System Architecture

BambaraMLLM integrates these principles into a unified architecture for seamless task switching. Given an input prompt, BambaraMLLM leverages an MoE model (Voxtral (Liu et al., 2025)) for ASR and TG, and a MatFormer model (Gemma-3n (Team et al., 2025a)) for audio understanding. This routing supports ASR, TG, and MT without manual reconfiguration, but does not support TTS tasks. We thus introduce an additional module (Ref. Section 2.2.3) to extend with TTS capabilities.

2.2.1 MoE Architecture: Voxtral

The 3.6B parameter Voxtral MoE model combines a 640M parameter Whisper-based audio encoder with a Ministral backbone, allowing ASR and TG with only task-relevant parameters activated.

2.2.2 MatFormer Architecture: Gemma-3n

This variant uses nested matrix factorization to compress 8B parameters into 4B effective parameters. The architecture is optimized for mobile deployment and excels in audio understanding.

2.2.3 Novel TTS Integration Module

To address the absence of TTS capabilities in Voxtral and Gemma-3n, BambaraMLLM extends the base LLM with an output head that predicts audio bicodec tokens, following Spark TTS approach (Wang et al., 2025). Thereby, BambaraMLLM treats TTS as a specialized generation task for bidirectional processing in one framework.

3 Instruction and Data Resources

To address data scarcity, we compiled and cleaned instruction tuning and language understanding datasets for BambaraMLLM fine-tuning.

3.1 Instruction Tuning Data

We implemented a linguistically informed instruction data generation framework (see Figure 1 in the Appendix) that injects structured Bambara linguistic knowledge into LLM-based generation. It integrates lexical resources, grammatical rules, and annotated examples to align with Bambara morphology, syntax, and cultural context, and performs deliberate linguistic transformation rather than direct translation. Using this framework, we generated over 2 million Bambara conversational instruction examples to guide BambaraMLLM.

3.2 Speech and Parallel Corpora

To support the training of BambaraMLLM, we compiled 200+ hours of transcribed Bambara speech, including community contributions of the Bible, Jeli-ASR (Diarra et al., 2022) and Bambara TTS dataset from MALIBA-AI (MALIBA-AI, 2025). Additionally, we also compiled 200K+ parallel Bambara-French/English sentences from various sources including Google-SMoL (Team et al., 2025b), Bayelemabaga (Tapo et al., 2025), MAFAND-MT (Adelani et al., 2022). All sources were cleaned, normalized, and unified by the authors. A sample data is provided in Appendix A.2.

4 Training Methodology

We employ a multi-stage training pipeline to minimize interference between modalities.

4.1 Expert Configuration

BambaraMLLM utilizes MoE experts for specific tasks. A two-expert configuration uses a speech expert for ASR and a unified text expert for language and TTS token prediction. A three-expert alternative adds a dedicated expert solely for TTS.

4.2 Phased Training

The training proceeds in four stages: initial phases focus on weight initialization and modality adaptation, followed by an integration phase in which combined datasets are used with task-specific tokens (e.g., `<task_asr>`, `<task_tts>`) and gradient accumulation to accommodate varying batch sizes, and a final fine-tuning phase that specializes the model for Bambara prosody and translation.

5 Preliminary Results

In this section, we report the validation error observed after training BambaraMLLM and present the results of both automatic and human evaluations across ASR and MT tasks. All evaluations are conducted on a held-out set of 100 samples, comprising audio inputs for ASR and parallel Bambara-English-French sentence pairs for MT, to provide a consistent and controlled basis for assessing model performance and alignment with human preferences. The evaluation results demonstrate that a unified multimodal model for the Bambara language benefits from cross-task learning.

(a) **TG & MT Validation Convergence:** The validation loss with BambaraMMLM for TG and MT tasks reduced by up to 94.9% after 5.5 epochs, as shown in Table 2 (Appendix A.3).

(b) **ASR Evaluation:** Fine-tuning for the ASR task achieves a 96.2% reduction in validation loss over 5.5 epochs, leading to a Word Error Rate of 40. (Appendix A.4).

(c) **Cross-Modal Transfer:** Models pre-trained on Bambara text showed significantly lower starting losses for ASR tasks compared to base models (2.36 vs 7.0), confirming that shared representations enhance learning efficiency.

(d) **MT Machine Evaluation:** BambaraMLLM achieves a BLEU score of 30 and a chrF score of 54 in MT from Bambara to French/English, with > 10 and > 30 BLEU and chrF translating into Bambara (Appendix A.5).

(e) **MT Human Evaluation:** Results indicate BambaraMLLM performs best in short-output tasks, but generates acceptable outputs in a dialogue setting (Appendix A.6).

6 Conclusion

This paper introduces BambaraMLLM, demonstrating the feasibility of multimodal modeling for under-resourced African languages under deployment constraints. The model integrates an ASR-LLM framework, a memory-efficient multimodal backbone, and a Bambara TTS module to enable unified speech, text, and language synthesis. Initial results show consistent improvements over task-specific systems across all tasks, and future work will extend to full-scale training and evaluation, increase language coverage, optimize deployment, and adapt to public-sector applications.

Limitations

Our work has several limitations to acknowledge:

- **Implementation Status**: The complete unified model is still under development, with preliminary results based on sub-components rather than the fully integrated architecture.
- **Data Constraints**: Despite leveraging available open-source Bambara resources, the available Bambara data remains limited compared to high-resource languages, potentially affecting model performance.
- **Evaluation Scope**: Automatic and human evaluation were conducted on a subset of 100 samples. While this provides an initial estimate of performance, a larger and more diverse evaluation set would yield stronger statistical significance and may reveal additional strengths or weaknesses of the system.
- **Computational Requirements**: The full model training requires significant computational resources, which may limit replicability for other under-resourced languages without similar infrastructure access.
- **Dialect Coverage**: Our current implementation focuses primarily on standard Bambara, with limited coverage of regional dialects and variations.
- **Cultural Nuance**: Despite the involvement of native speakers in our evaluation, fully capturing cultural nuances in speech synthesis remains challenging.

Acknowledgments

This work was supported in part by a Google PhD Fellowship awarded to the second author, Allahsera Tapo. We thank Caytu Robotics for providing computational support for the initial experiments. We also thank the reviewers for their constructive comments to enhance the manuscript.

References

David Ifeoluwa Adelani and 1 others. 2022. Mafand-mt: A multi-domain parallel corpus for african languages. *arXiv preprint arXiv:2205.02022*.

Asma Amalas, Mounir Ghogho, Mohamed Chetouani, and Rachid Oulad Haj Thami. 2024. A multilingual training strategy for low resource text to speech. *arXiv preprint arXiv:2409.01217*.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. 2025. Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*.

Fnu Devvrit, Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hanna Hajishirzi, Sham Kakade, Ali Farhadi, and 1 others. 2024. Matformer: Nested transformer for elastic inference. *Advances in Neural Information Processing Systems*, 37:140535–140564.

Sebastien Diarra, Michael Leventhal, and Allahsera Auguste Tapo. 2022. Robotsmali griots speech dataset, and asr. https://github.com/robotsmali-ai/jeli-asr/.

Xiao Liang, Yen-Min Jasmina Khaw, Soung-Yue Liew, Tien-Ping Tan, and Donghong Qin. 2025. Towards low-resource languages machine translation: A language-specific fine-tuning with lora for specialized large language models. *IEEE Access*.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

MALIBA-AI. 2025. Bambara text-to-speech: Open-source high-quality tts for bambara language. https://huggingface.co/MALIBA-AI/bambara-tts. Hugging Face Model Repository.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri. 2025. Bayelemabaga: Creating resources for Bambara NLP. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

NLLB Team and 1 others. 2025b. Smol: Scaling multilingual machine translation for low-resource languages. *arXiv preprint arXiv:2502.12301*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Table 1: Sample parallel data from compiled corpora. Token and phrase-level translations demonstrate basic vocabulary and common expressions. Sentence-level translations from oral tradition transcriptions illustrate natural Bambara discourse patterns.

Bambara	Translation	Type
Token/Phrase-level (Bambara ↔ English)		
aw ni baara	Hello	Token
i ka kɛnɛ wa	How are you?	Phrase
n b'i fɛ	I love you	Phrase
a' ni sɔgɔma	Good morning	Phrase
N m'a faamu	I don't understand	Phrase
i ye san joli ye	How old are you?	Phrase
i dɔnnin diyaran ye	Nice to meet you	Phrase
Sentence-level (Bambara ↔ English)		
BM: N'i ma fɛn min tɔ o ye, i t'o tɔ fɛn ye tugun de. **EN:** What you do not leave them, you will not give to someone else.		Sentence
BM: Alihamdulayi Arabil Alamina, an bɛ Ala tandu. **EN:** Praise be to God, Lord of the Worlds, we thank God.		Sentence
Sentence-level (Bambara ↔ French)		
BM: O y'a sɔrɔ Tanbage banna, Masa Makan banna, an fa banna. **FR:** À ce moment, Tanbage était décédé, Massa Makan était décédé, notre père était décédé.		Sentence
BM: Jawoyi fa ko Bolijigi, Karala, o sigira o nɔ rɔ. **FR:** Alors le père de Jawoyi nommé Bolijigi, de Karala, il a été installé.		Sentence
BM: Ala ka Mali suma, ka hɛrɛ don a kɔnɔ. **FR:** Que Dieu bénisse le Mali et y apporte la paix.		Sentence

A Appendix

A.1 Instruction Generation Framework

We designed and implemented an instruction generation procedure that combines data acquisition, transformation, and validation to ensure alignment with human preference. The framework is illustrated in Figure 1.

A.2 Sample Parallel Corpora

Table 1 presents representative examples from our compiled parallel corpora, illustrating the diversity of translation pairs at different granularities.

A.3 Text Generation and MT Validation Convergence

Table 2 presents the training progression for text generation and machine translation tasks, demonstrating rapid convergence with a 94.9% reduction in validation loss.

Table 2: Training convergence for text generation and MT tasks. Validation loss decreased from 7.47 to 0.38 (94.9% reduction) over 5.5 epochs.

Epoch	Step	Train Loss	Val Loss
0.0	0	–	7.4726
0.5	3,521	0.8556	0.7959
1.0	7,042	0.7163	0.6873
1.5	10,563	0.6039	0.6272
2.0	14,084	0.6421	0.5720
2.5	17,605	0.6149	0.5307
3.0	21,126	0.5277	0.4907
3.5	24,647	0.4892	0.4540
4.0	28,168	0.4957	0.4261
4.5	31,689	0.4304	0.4030
5.0	35,210	0.4572	0.3880
5.5	38,731	0.4399	0.3810
Total Reduction		48.6%	**94.9%**

A.4 ASR Evaluation

Table 3 shows the fine-tuning progression for the Automatic Speech Recognition (ASR) task. Validation loss decreases rapidly from 2.3603 to 0.08887 (96.2% reduction) over 5.5 epochs. BambaraM-LLM achieves WER and CER of 39.78% and 21.40%, respectively, on 100 audio samples.

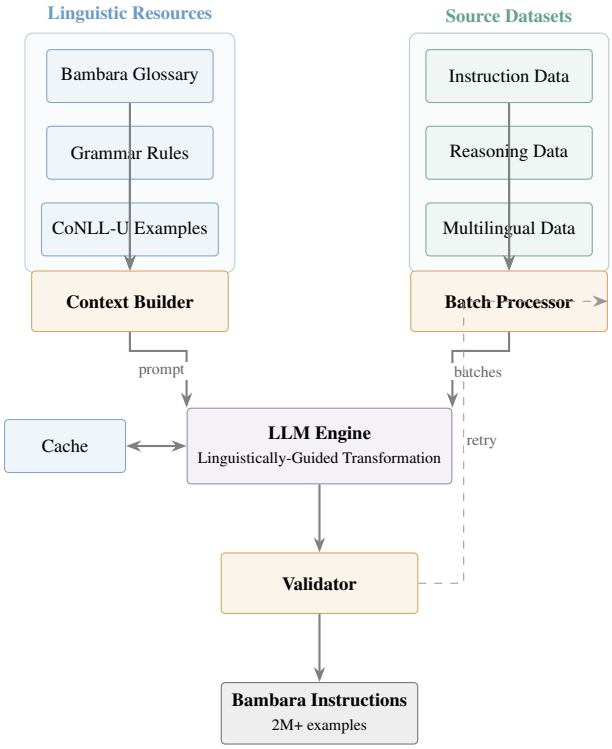

Figure 1: Linguistically informed instruction generation framework. Bambara linguistic resources (glossary, grammar rules, CoNLL-U annotations) are embedded in the system prompt to guide the structured transformation of source instructions instead of direct translation.

Epoch	Step	Train Loss	Val Loss	
0.0	0	No log	2.3603	
0.50	293	0.5423	0.5847	
1.00	586	0.3592	0.4356	
1.50	879	0.3555	0.3764	
2.00	1172	0.2928	0.3247	
2.50	1465	0.2411	0.2867	
3.00	1758	0.2800	0.2314	
3.50	2051	0.1091	0.2033	
4.00	2344	0.1150	0.1539	
4.50	2637	0.0835	0.1185	
5.00	2930	0.0722	0.1021	
5.50	3223	0.0945	0.0889	
Total Reduction		–	**96.2%**	–

Table 3: ASR fine-tuning convergence. Validation loss drops from 2.36 to 0.09 (96.2% reduction) over 5.5 epochs, with stable memory usage throughout training.

A.5 Automatic Evaluation Results

To evaluate BambaraMLLM, we computed BLEU and chrF on 100 bidirectional translation samples (Bambara–French/English and French/English–Bambara). Figure 2 reports BLEU scores of up to 30 and chrF scores of up to 54.

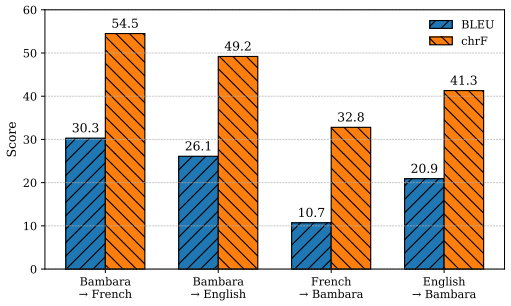

Figure 2: Automatic evaluation scores (BLEU and chrF) for machine translation with BambaraMLLM.

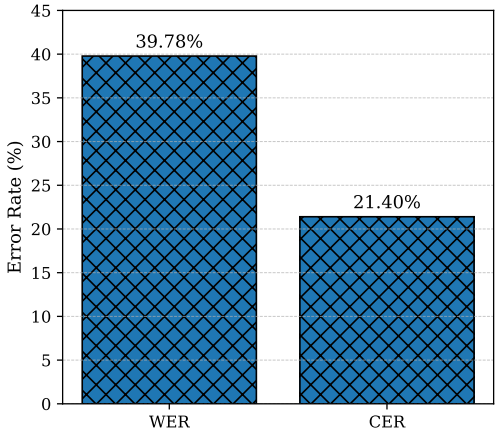

Figure 3: Analysis of word error rate (WER) and character error rate (CER) for ASR with BambaraMLLM.

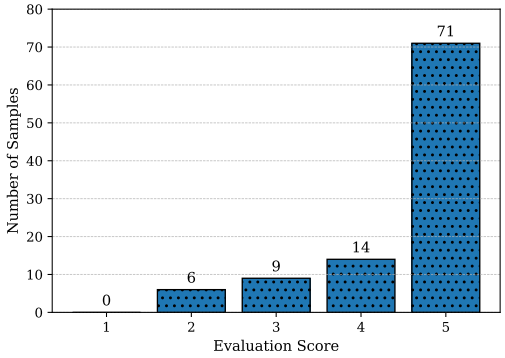

Figure 4: Distribution of scores assigned by human evaluators to BambaraMLLM-generated outputs

A.6 Human Evaluation Results

We collected feedback from Bambara speakers to assess alignment between BambaraMLLM's generated outputs and human preferences using a 1–5 scale that combines grammatical and cultural criteria. Figure 4 shows that 71% of the outputs received the highest score.