Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

MULTIMODAL LARGE LANGUAGE MODELS FOR LOW-RESOURCE LANGUAGES AND GLOBAL SOUTH DEPLOYMENT: A COMPREHENSIVE SURVEY OF ARCHITECTURES, BENCHMARKS, AND SOCIOTECHNICAL CHALLENGES

*[1]Austin Olom Ogar, [1]Abah Joshua, [1]Aliyu Suleiman Muhammed, [1]Oluwatobi Noah Akande, [1]Faruk Obansa Muhammed, and [2]Ibrahim Anka Salihu

[1]Department of Computer Science, Nile University of Nigeria, Plot 681 Cadastral Zone C-OO, Abuja 900001, Nigeria
[2]Department of Software Engineering, Nile University of Nigeria, Plot 681 Cadastral Zone C-OO, Abuja 900001, Nigeria

*Corresponding Author Email Address: austinolomogar@gmail.com

ABSTRACT
Multimodal Large Language Models (MLLMs) such as LLaVA, InstructBLIP, and Qwen2-VL have unlocked joint reasoning over text and images at an unprecedented scale. The technical literature on MLLM efficiency, domain adaptation, and benchmarking has matured into a substantial corpus. However, the overwhelming majority of surveys are written from the vantage point of high-resource English-language datasets and well-resourced computing infrastructures. This survey takes a different angle. We consolidate the recent literature on multimodal language understanding through the lens of low-resource languages and Global South deployment contexts, where data scarcity, compute constraints, intermittent connectivity, and sociotechnical-trust considerations interact in ways that high-resource surveys rarely address. We propose a four-quadrant taxonomy that organises the field around linguistic coverage, data scarcity, compute constraints, and sociotechnical trust. We trace the evolution of multilingual multimodal architectures across an eight-year arc, map the benchmark landscape against nine languages and five modalities, and describe a three-tier edge-regional-global deployment topology suited to low-resource environments. Five high-impact application domains are surveyed: healthcare, agriculture, education, disaster response, and public service. A dedicated section examines sociotechnical considerations, including linguistic justice, algorithmic fairness, and regulatory readiness. We identify six concrete open research problems and outline a future research agenda. The survey is intended as a single-source reference for researchers, policy makers, and practitioners pursuing equitable multimodal AI deployment in low-resource contexts.

Keyword: Multimodal large language models; low-resource languages; Global South deployment; multilingual benchmarks; edge AI; linguistic justice; sociotechnical trust; healthcare AI; agricultural AI; equitable artificial intelligence.

INTRODUCTION

Multimodal Large Language Models (MLLMs) have reshaped what artificial-intelligence systems can do. By processing text and images jointly within a single generative model, contemporary systems such as LLaVA, InstructBLIP, and Qwen2-VL (Liu *et al.*, 2024)–(Wang *et al.*, 2024a) support tasks ranging from image captioning and visual question answering to clinical decision support, autonomous perception, and document understanding (Zhang *et al.*, 2023a), (European Commission, 2024). The pace of architectural and benchmark progress has been remarkable. A substantial survey literature now maps the field along technical axes such as efficiency (Pope *et al.*, 2023), domain adaptation (Zhai *et al.*, 2023), conditional computation (Fedus *et al.*, 2022), and architectural taxonomy (Liang *et al.*, 2024).

The overwhelming majority of those surveys, however, are written from the vantage point of high-resource English-language datasets, well-resourced computing infrastructures, and production environments in North America, Europe, and East Asia. Less attention has been paid to the constraints that dominate deployment in the Global South: linguistic diversity with limited annotated corpora; intermittent power and bandwidth; mobile-first user populations; data-sovereignty and privacy concerns; and a sociotechnical environment in which fairness, bias, and regulatory readiness are first-class deployment requirements. These are not mere engineering troubles; they determine which architectural options may be feasible, which benchmarks may be viable, and which application areas may realistically take advantage of multimodal AI (Adelani et al., 2021; Aji *et al.*, 2022).

Several recent reviews discuss multilingual multimodal models (Liang *et al.*, 2024), parameter-efficient adaptation (Chen *et al.*, 2024), and edge deployment (Poornashree *et al.*, 2025a). However, they have not yet combined these topics specifically from the perspective of low-resource languages and deployment contexts of the Global South. So, the reader will not find any one resource that covers the intersection of architectural, benchmark, deployment, and sociotechnical considerations that are most relevant when multimodal AI is released from well-resourced environments. This survey is the one to fill that void.

The contributions of this article are:

i. Four-quadrant categorization of problems that create the challenge for multimodal AI in low-resource settings, linguistic coverage, data scarcity, compute constraints, and sociotechnical trust, highlighted in Figure 1.
ii. The timeline of multilingual and multimodal architectures, Figure 2, shows the progression of the field over the last eight years up to the recent emergence of region-specific MLLMs, the current focus of the era.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

iii. The map showing regional benchmark coverage (Figure 3) and the matrix comparing benchmark vs. language coverage (Figure 5) reveal particular shortcomings in the evaluation infrastructure.
iv. A network of three tiers: edge-regional-global (Figure 6), with a design for low-resource settings and real-world latency targets for each raises example concrete cases for low-resource situations.
v. Five main areas where multimodal AI can bring a substantial change in low-resource settings through disproportionate advantages are highlighted in the paper: healthcare, agriculture, education, disaster response, and public service.
vi. A sociotechnical trust model (Figure 8) that links linguistic justice, algorithmic fairness, and regulatory readiness as inseparable sides.
vii. A research agenda for the future consisting of six pillars (Figure 9), which, in our opinion, are the areas that must be focused on during the next two years.

This article is organised as follows. Section II discusses the survey methodology. Section III presents the four-quadrant taxonomy. Section IV highlights the architectural history of multilingual MLLMs. Section V presents a survey of the benchmark landscape. Section VI discusses deployment topologies in low-resource settings. Section VII presents a survey of major application domains. Section VIII discusses sociotechnical considerations. Section IX lists six open research problems. Section X summarizes.

SURVEY METHODOLOGY

A. Search Strategy and Databases

A PRISMA-style literature review was used to perform a literature search using IEEE Xplore, ACM Digital Library, arXiv, The ACL Anthology, and the Web of Science, as well as the main conferences of NeurIPS, ICML, ICLR, CVPR, ECCV, ACL, EMNLP, NAACL, AAAI, and INTERSPEECH in the period 2018-2026. The combined search expression was "(multimodal OR vision-language OR multilingual OR low-resource) AND (large language model OR LLM OR multimodal language understanding) AND (deployment OR benchmark OR application OR Global South OR African OR Asian OR Indigenous)". 412 candidate records were preserved after removing duplicates.

B. Inclusion and Exclusion Criteria

We included papers if: (i) the paper focused on a multimodal language model or a structure very closely related to that, plus at least one of the model evaluations had to be low-resource or from a non-Western point of view; (ii) a paper had to be in a peer-reviewed journal or conference, or a well-argued preprint in arXiv accompanied by public code or data; and (iii) papers had to be written in English or have an English translation of the highest standards. Reasons for exclusion were: single-modality NLP or single-modality vision papers without multimodal context; opinion papers without empirical grounding; and two-page workshop abstracts with no follow-up papers. Reasons for exclusion were: single-modality NLP or single-modality vision papers without multimodal context; opinion papers without empirical grounding; and two-page workshop abstracts with no follow-up papers.

C. Analytical Framework

We arranged the included literature according to four main aspects: language coverage, lack of data, computational restrictions, and sociotechnical trust, and at the same time, the evaluation was based on four main aspects: precision, response time, power consumption, and availability. The four-part classification chart and the architectural chronology served as the main ideas of the review. The quality of the selected studies was assessed using three criteria: whether the study was peer-reviewed by the venue, whether the code or data were made publicly available, and whether an evaluation was conducted on low-resource or non-Western data.

A FOUR-QUADRANT TAXONOMY OF LOW-RESOURCE MULTIMODAL AI

Figure 1 introduces the four-quadrant taxonomy that frames the rest of the survey.

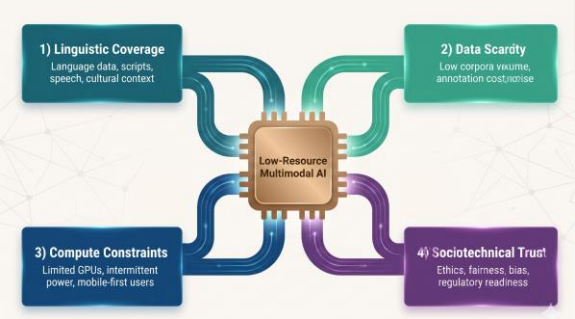

Figure 1. *A four-quadrant taxonomy of challenges for multimodal AI in low-resource settings.*

A. Linguistic Coverage

Language coverage is a measure of the amount, quality, and type of language data available for a specific language. English accounts for more than 60% of publicly available multimodal training databases (Joshi et al., 2020), whereas the other 7,000 languages of the world together account for the rest. The lack of language coverage shows itself in several areas: the availability of image-text pairs is limited in most languages; speech datasets are even more limited; and signed languages and low-resource scripts are almost completely missing from today's MLLM benchmarks (Liang et al., 2024).

B. Data Scarcity

These days, data scarcity is more than just a lack of data; it is also an uneven distribution of data quality. In many cases, data sets are available, but the quality of their annotations varies; excellent test sets that can serve as a standard are rare, and culturally aligned evaluation rubrics may be difficult to obtain (Ahmad *et al.*, 2024). Scarcity is not limited to finance: the costs of assembling a top-notch multimodal corpus in a low-resource language can exceed the annual AI budget of a Global South research entity.

C. Compute Constraints

In terms of computing constraints, low-resource settings are characterised by limited GPU availability, intermittent power supply, expensive bandwidth, and a mobile-first end-user

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

community (Ali *et al.*, 2024). Edge devices prevalent in such environments include basic smartphones, affordable tablets, and Raspberry Pi-class single-board computers. All of these devices alone are incapable of running modern 7-billion-parameter MLLMs without undergoing intensive quantisation, pruning, or task-specific distillations (Abdin *et al.*, 2024).

D. Sociotechnical Trust

Trust in sociotechnical systems covers linguistic justice, algorithmic fairness, regulatory readiness, data sovereignty, and the appropriate cultural use (Gebru *et al.*, 2021; Norgeot *et al.*, 2020). Where national AI governance frameworks are still evolving, the lack of clearly defined trust norms can be a double-edged sword: an opportunity (to establish inclusive standards from the outset) and a threat (deployment may proceed without sufficient protections). Hence, sociotechnical trust should not be viewed as a mere consequence of the deployment; rather, it is the condition that enables adoption.

MULTILINGUAL AND MULTIMODAL ARCHITECTURES

Figure 2 traces the eight-year arc from cross-lingual encoders (mBERT, XLM-R) through the first wave of multimodal foundation models (Flamingo, BLIP, CLIP) and instruction-tuned MLLMs (LLaVA, InstructBLIP) to the current era of region-specific and language-specialised multimodal LLMs.

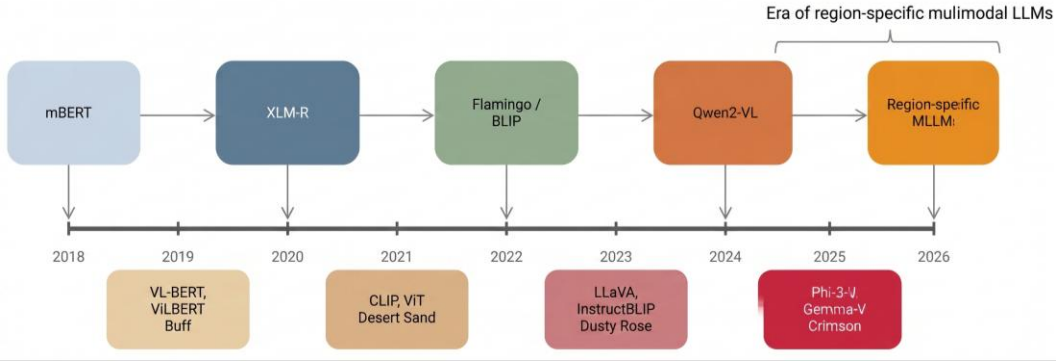

Figure 2. *Eight-year timeline of multilingual and multimodal language model development (2018-2026).*

A. Architectural Taxonomy

Figure 4 organises contemporary multilingual MLLMs along five architectural patterns: modular dual-tower designs (separate encoders for text and image with a thin alignment head), joint unified backbones (shared transformer body), cross-lingual transfer using an anchor language pivot, adapter-based designs that add small language-family-specific modules to a shared backbone, and the emerging mixture-of-language-experts (MoLE) pattern that routes inputs to language-specific expert subnetworks.

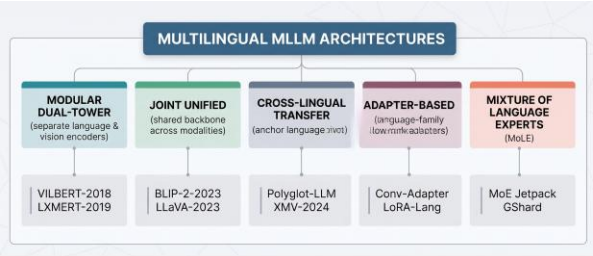

Figure 4. *Architectural taxonomy of multilingual multimodal language models with representative examples.*

B. Adapter-Based and Parameter-Efficient Approaches

Adapter-based methods such as LoRA (Hu *et al.*, 2022), Conv-Adapter, and language-family adapters are particularly attractive for low-resource settings because they preserve the shared backbone (which carries the bulk of pre-trained knowledge) and inject small task- or language-specific modules. Empirically, adapter-only fine-tuning has been reported to close 70-90% of the gap to full fine-tuning while training fewer than 5% of parameters (Hu *et al.*, 2022; Houlsby *et al.*, 2019). The relative advantage grows as the target language drifts further from the backbone's pre-training distribution.

C. Mixture-of-Language-Experts (MoLE)

The recent MoLE pattern extends the mixture-of-experts paradigm (Shazeer *et al.*, 2017) to multilingual settings by dedicating expert subnetworks to language families or scripts. While theoretically attractive — only the relevant experts are activated per input — MoLE introduces challenges around gating overhead, load imbalance, and the difficulty of defining experts when language boundaries are fluid (as in code-mixing) (Soliman *et al.*, 2025).

D. Cross-Lingual Transfer through Anchor Languages

A widely used pragmatic strategy is to fine-tune the model in a well-resourced anchor language (typically English) and rely on cross-lingual transfer to target languages. This works adequately for typologically similar languages but degrades sharply for low-resource languages whose grammar, script, or sociolinguistic context differs from the anchor (Conneau *et al.*, 2020). The architectural choice between MoLE and anchor-language transfer is therefore not purely technical; it carries cultural and equity implications.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Table I. Architectural approaches to multilingual multimodal llms

Pattern	Strengths	Limitations	Suitability for low-resource
Modular dual-tower	Simple; transferable encoders	Shallow cross-modal fusion	Moderate
Joint unified	Deep multimodal interaction	Costly to retrain	Low
Cross-lingual transfer	Reuses anchor model	Quality drops for distant languages	Variable
Adapter-based PEFT	Cheap; preserves backbone	Routing decisions per language	High
Mixture-of-language-experts	Sparse compute per input	Gating overhead; load imbalance	High (with care)

BENCHMARK LANDSCAPE

A. Regional Coverage

Figure 3 visualises the stylised regional coverage of major multilingual MLLM benchmarks. North America, Europe, and East Asia enjoy Tier-1 coverage (more than 100 million training pairs); Latin America and the Middle East / North Africa region are Tier-2 (10-100 million pairs); Sub-Saharan Africa and South Asia are Tier-3 (1-10 million pairs); and Southeast Asia, the Pacific, and indigenous micro-languages are Tier-4 (fewer than 1 million pairs).

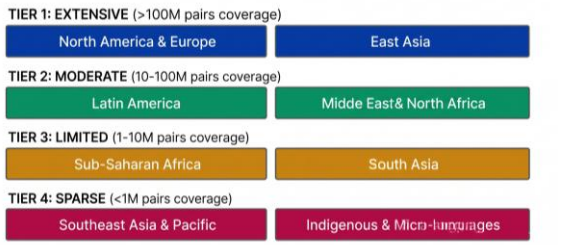

***Figure 3**. Stylised regional coverage of major multilingual MLLM benchmarks.*

B. Coverage Matrix

Figure 5 presents a coverage matrix for nine illustrative languages across five modalities (image-text, speech-text, video, document, sign language). The matrix exposes specific gaps that point to where benchmark construction would have the most impact: Hausa is essentially uncovered across all five modalities; Yoruba has speech-text but little image-text data; Bengali, Hindi, and Swahili sit at the boundary between partial and sparse coverage; and sign-language coverage is sparse for every non-English language surveyed.

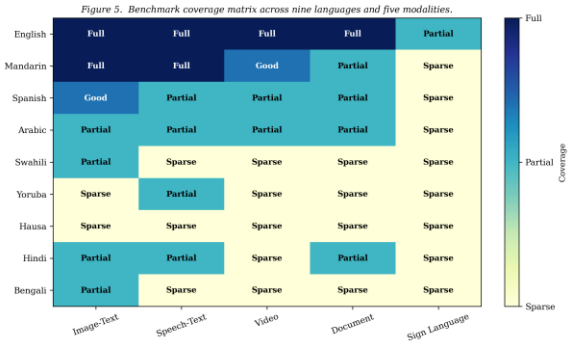

***Figure 5**. Benchmark coverage matrix across nine languages and five modalities.*

C. Benchmark Construction Recommendations

Three priorities for benchmark construction emerge. First, paired audio-image-text corpora in Yoruba, Hausa, Swahili, and Igbo would unlock Sub-Saharan-Africa multimodal benchmarks at a stroke. Second, document-image benchmarks for Hindi, Bengali, and Mandarin script variants would close the document-AI gap. Third, sign-language video corpora for non-English languages remain almost entirely absent and should be a long-term priority.

Table II. Notable multilingual multimodal benchmarks

Benchmark	Languages	Modalities	Size	Focus
XVNLI / xGQA	8-32	Image+Text	100k-2M	Cross-lingual VQA
MaRVL	5	Image+Text	2k	Culturally-grounded reasoning
FLEURS	102	Speech+Text	6k/lang	Cross-lingual ASR
MMBench-multi	6	Image+Text	2.4k	Multilingual MLLM bench
AfriBench	20+	Image+Text	Pilot	African-language MLLM
IndicGen-VL	11	Image+Text+Doc	30k	Indic-language MLLM

DEPLOYMENT TOPOLOGY FOR LOW-RESOURCE SETTINGS

Figure 6 illustrates a three-tier deployment topology adapted to low-resource and Global South contexts. The edge tier hosts compact models (under 3 billion parameters, INT8 quantised) on mobile devices, IoT sensors, and community kiosks. The regional hub tier hosts 7-30 billion-parameter models on shared infrastructure operated by universities, national research and education networks (NRENs), or regional cloud providers. The global cloud tier hosts the largest dense or MoE models on hyperscale infrastructure (Agarwal *et al.*, 2024).

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

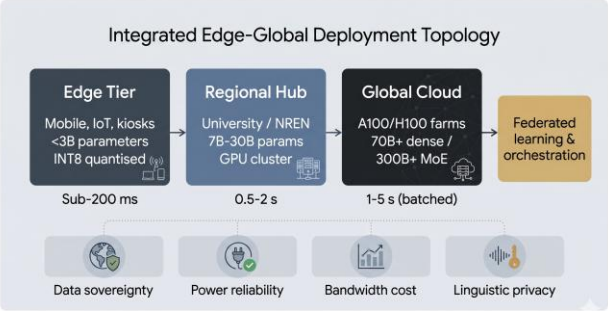

Figure 6. *Three-tier edge-regional-global deployment topology for multilingual MLLMs in Global South contexts.*

A. Latency and Connectivity Budgets

Latency budgets differ markedly between tiers. Edge-tier inference must complete within 200 milliseconds for conversational interaction; regional hubs operate within 0.5-2 seconds; global cloud responses may take 1-5 seconds even with batching (Xu *et al.*, 2025). In low-resource settings, however, the dominant variable is often connectivity rather than compute: a request that would take 500 milliseconds round-trip on a fibre link may take 5-30 seconds on a 3G mobile network common in rural areas.

B. Energy and Power Reliability

Edge devices in low-resource settings frequently operate under intermittent grid power, sometimes with solar augmentation. Energy efficiency, therefore, matters not only for sustainability but for basic operability. INT8 and INT4 quantisation, structured pruning, and on-device caching of frequent queries are particularly valuable in this context (Krishnamoorthi, 2018).

C. Data Sovereignty and Federated Learning

Data-sovereignty concerns can preclude sending raw user data to global cloud providers. Federated learning, where model updates rather than raw data are exchanged, is therefore particularly appealing for low-resource multilingual deployment (Bonawitz *et al.*, 2019). Federated multilingual fine-tuning has shown promising results in clinical and educational pilots and is gaining policy traction.

HIGH-IMPACT APPLICATION DOMAINS

Figure 7 illustrates five application domains in which multimodal AI offers disproportionate benefits in low-resource settings.

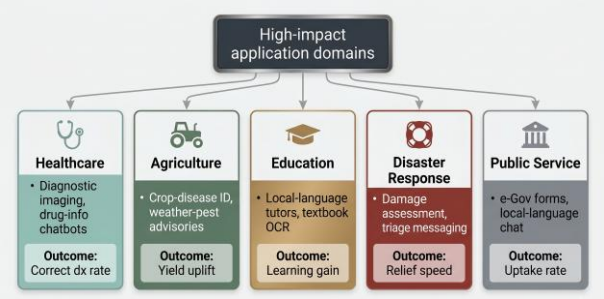

Figure 7*. Five high-impact application domains for multimodal AI in low-resource contexts.*

A. Healthcare

Multimodal AI in low-resource healthcare contexts spans diagnostic imaging review (chest X-ray triage, dermoscopic lesion screening, ophthalmologic fundus analysis), local-language drug-information chatbots, and multimodal symptom-checker interfaces. In settings where a single radiologist may cover an entire province, AI-assisted triage can materially shorten time to diagnosis (Pfohl *et al.*, 2024). Outcome metrics include correct-diagnosis rate, time-to-diagnosis, and clinician workload reduction.

B. Agriculture

Agricultural applications include crop disease identification from smartphone photographs, weather-conditioned pest advisories, and multimodal yield prediction dashboards. Plant-disease classifiers trained on globally diverse leaf imagery can be deployed via low-bandwidth mobile applications that work offline once downloaded. Outcome metrics include yield uplift, input cost reduction, and farmer adoption rate.

C. Education

Educational applications include local-language tutoring assistants, textbook OCR systems that ingest scanned curricula into searchable knowledge bases, and multimodal homework helpers that ground language explanations in diagrams. Effective deployment requires not only translation but cultural and curricular alignment. Outcome metrics include learning gain on standardised assessments and engagement duration.

D. Disaster Response

Disaster-response applications include rapid damage assessment from satellite and drone imagery, multilingual triage messaging that bridges first-responder languages, and resource-allocation dashboards that integrate aerial imagery with population data. Outcome metrics include relief-delivery speed, prioritisation accuracy, and mis-routing-loss reduction (Patterson *et al.*, 2021).

E. Public Service

Public-service applications include local-language e-government chatbots, multimodal form-filling assistants for citizens with limited literacy, and audio-grounded service-discovery interfaces. Outcome metrics include service uptake rate, error reduction in submitted forms, and citizen satisfaction scores.

SOCIOTECHNICAL CONSIDERATIONS

Figure 8 organises the sociotechnical considerations into a three-circle framework: linguistic justice, algorithmic fairness, and regulatory readiness. Trustworthy deployment requires attention to all three and explicitly to their intersections.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.42
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

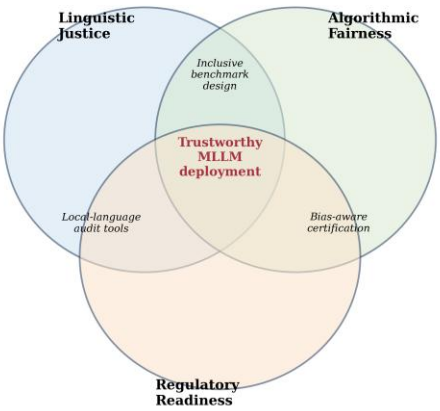

***Figure 8**. Sociotechnical trust framework for multimodal AI in low-resource and global-equity contexts.*

A. Linguistic Justice
Linguistic justice means that speakers of low-resource languages have equal access to AI capabilities that work in their language and that respect their cultural context. Benchmark construction that excludes a language is a form of structural injustice; deployment that requires English-only interaction marginalises non-English speakers. The Masakhane initiative for African NLP, the AI4Bharat initiative for Indic languages, and the SEACrowd initiative for Southeast Asian languages illustrate community-led responses to this challenge (Adelani *et al*., 2021; Aji *et al*., 2022).

B. Algorithmic Fairness
Algorithmic fairness considerations are sharper in low-resource contexts because the costs of error are often borne disproportionately by already-marginalised populations. Bias evaluation rubrics from high-resource benchmarks may not transfer; culturally calibrated rubrics must be co-developed with local stakeholders (Gebru *et al*., 2021). Counterfactual fairness, calibration under distribution shift, and subgroup-level error analysis are first-class concerns.

C. Regulatory Readiness
Regulatory frameworks for AI are emerging at different places across regions. The EU AI Act (European Commission, 2024), the African Union AI Continental Strategy, and national frameworks under development across the Global South establish baseline requirements for high-risk AI applications. Compliance preparation should be designed in, not retrofitted: data provenance, model documentation, and post-deployment monitoring are now mandatory elements of credible deployment (Norgeot *et al*., 2020).

Table III. Sociotechnical concerns, manifestations, and mitigations

Concern	Manifestation	Mitigation
Language exclusion	No model coverage for L1	Community-led benchmark co-design
Cultural mis-grounding	Image labels miss cultural cues	Local rubric co-development
Subgroup bias	Accuracy drops for demographic group	Disaggregated evaluation
Data-sovereignty risk	Raw data leaves jurisdiction	Federated learning, edge inference
Inadequate documentation	Model cards missing	Standardised model cards
Audit gap	No post-deployment monitoring	Continuous performance & fairness audit

OPEN RESEARCH PROBLEMS AND FUTURE DIRECTIONS

Figure 9 visualises six pillars of a future research agenda for equitable multimodal AI.

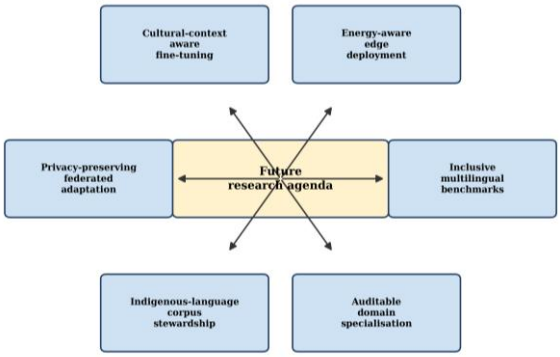

***Figure 9**. Six-pillar future research agenda for equitable multimodal AI.*

A. Inclusive Multilingual Benchmarks
Constructing benchmarks for the languages and modalities highlighted in Figure 5 is the most immediate engineering priority. AfriBench, IndicGen-VL, and SEACrowd point the way; the next two years should produce comparable resources for at least twenty underserved languages.

B. Energy-Aware Edge Deployment
Edge inference for multimodal models remains energy-intensive. New quantisation schemes, structured pruning, and on-device caching are needed to support sustained operation under intermittent power.

C. Cultural-Context-Aware Fine-Tuning
Beyond translation, models must adapt to cultural context: regional measurement units, naming conventions, holidays, food, dress, and indigenous knowledge systems. Cultural-context-aware fine-tuning is an active research frontier.

D. Privacy-Preserving Federated Adaptation
Federated learning protocols that preserve linguistic privacy and respect data sovereignty constraints are essential for Global South deployment. Standardised federated multilingual training recipes do not yet exist.

E. Indigenous-Language Corpus Stewardship
Indigenous-language data raises distinct questions of community consent, cultural appropriateness, and intellectual property. Stewardship frameworks co-designed with indigenous communities are urgently needed.

F. Auditable Domain Specialisation
As MLLMs are specialised for healthcare, education, agriculture, or public service applications, auditability becomes essential. Domain specialisation that can be verified, reversed, and re-audited is the foundation of trustworthy deployment.

Conclusion
Multimodal Large Language Models are a major technological achievement, but the engineering literature has so far been written predominantly through a high-resource, English-centric lens. This survey has taken a deliberately different angle, consolidating the recent literature on multimodal language understanding through the lens of low-resource languages and Global South deployment contexts. We organised the field around a four-quadrant taxonomy of linguistic coverage, data scarcity, compute constraints, and sociotechnical trust; traced an eight-year architectural arc; mapped benchmark coverage across nine languages and five modalities; described a three-tier deployment topology adapted to low-resource environments; surveyed five high-impact application domains; and outlined a sociotechnical-trust framework and a six-pillar future-research agenda.

Three messages stand out. First, the most pressing engineering problem in multilingual multimodal AI is not raw model capability but benchmark and deployment infrastructure for the languages and regions that current corpora underserve. Second, the sociotechnical and engineering dimensions of trust are inseparable; technical choices have justice implications and vice versa. Third, the research community has the methods to address these problems; what is missing is the alignment of agendas, funding, and stakeholder co-design. We hope this survey contributes a shared vocabulary and reference frame for the practitioners and researchers who will tackle these directions over the coming years.

Funding
This research received no specific external funding.

Conflicts of Interest
The authors declare no conflict of interest.

Data Availability
This is a survey article that does not produce original empirical data. All benchmarks, datasets, and initiatives discussed are cited in the reference list and are openly available from their respective project pages.

Ethics Statement
This study did not involve human subjects, animals, or sensitive personal data and did not require institutional review board approval.

Acknowledgments
The authors thank the Faculty of Computing at Nile University of Nigeria for institutional support and library access.

Author Contributions (CRediT)
Conceptualization, A.O.O. and O.N.A.; methodology, A.O.O.; investigation, A.O.O., A.J., and A.S.M.; writing — original draft, A.O.O.; writing — review and editing, O.N.A., A.J., A.S.M., F.O.M., and I.A.S.; visualisation, A.O.O.; supervision, O.N.A.; project administration, A.O.O. All authors have read and approved the final manuscript.

REFERENCES

Abdin, M., Aneja, J. et al. (2024). Phi-3 technical report: a highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Adelani, D., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J. et al. (2021). MasakhaNER: named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*.

Agarwal, S., Phanishayee, A. & Venkataraman, S. (2024). Blox: a modular toolkit for deep learning schedulers. In *Proceedings of EuroSys*.

Ahmad, F., Frej, B., Adams, M. & Jameel, S. (2024). Beyond English: a global survey of multilingual benchmark coverage. *ACM Computing Surveys*, 56(4):1–39.

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R. et al. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ali, M., Khan, H., Rana, M. T. A., Ali, A., Baig, M. Z., Rehman, S. U., & Alsaawy, Y. (2024). A machine learning approach to reduce latency in edge computing for IoT devices. *Engineering, Technology & Applied Science Research*, 14(5):16751–16756.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V. et al. (2019). Towards federated learning at scale: system design. In *Proceedings of MLSys*.

Boutouta, H., Lakhfif, A., Senator, F. & Mediani, C. (2025). A transformer-based hybrid model for implicit emotion recognition in Arabic text. *Engineering, Technology & Applied Science Research*, 15(3):23834–23839.

Chen, H., Tao, R., Zhang, H., Wang, Y., Li, X., Ye, W. et al. (2024). Conv-adapter: exploring parameter-efficient transfer learning for ConvNets. *International Journal of Computer Vision*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F. et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W. et al. (2023). InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Advances in*

Neural Information Processing Systems (NeurIPS).

Endah, S. N., Suprapto & Suyanto, Y. (2025). Enhancing low-resource dialectal ASR in Indonesian using speech-transformer models and data augmentation. *Engineering, Technology & Applied Science Research*, 15(5):28095-28101.

European Commission (2024). Regulation (EU) 2024/1689 on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Fedus, W., Zoph, B. & Shazeer, N. (2022). Switch Transformer: scaling to trillion-parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A. et al. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. et al. (2022). LoRA: low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Krishnamoorthi, R. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang, C. X., Tian, P., Yin, C. H., Yu, Y., An-Hou, W., Ming, L. et al. (2024). A comprehensive survey and guide to multimodal large language models. *arXiv preprint*.

Liu, H., Li, C., Li, Y. & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R. et al. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26(9):1320–1324.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D. et al. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A. et al. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*.

Poornashree, S. J., Joshi, P. V., Sudharshan, K. M. & George, T. M. (2025). A hybrid AI pipeline for real-time aerial video analytics on resource-limited edge devices with performance profiling. *Engineering, Technology & Applied Science Research*, 15(6):29574–29579.

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J. et al. (2023). Efficiently scaling transformer inference. In *Proceedings of MLSys*.

Salih, M. I., Mohammed, S. M., Ibrahim, A. K., Ahmed, O. M. & Haji, L. M. (2025). Fine-tuning BERT for automated news classification. *Engineering, Technology & Applied Science Research*, 15(3):22953–22959.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G. & Dean, J. (2017). Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Soliman, M., Abdelaziz, O., Radwan, A. & Shehata, M. (2025). GNN-MoE: context-aware patch routing using GNNs for parameter-efficient domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Utami, E., Wahyuni, S. N., Sulistiyono, M., Raharjo, S., Hartanto, A. D., Rohman, A. N. et al. (2026). A multimodal transformer-LSTM framework for cross-lingual lexical alignment of Indonesian regional languages. *Engineering, Technology & Applied Science Research*, 16(2):34283–34292.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J. et al. (2024a). Qwen2-VL: enhancing vision-language models' perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xu, Z., Xu, Y., Li, H. & Sun, T. (2025). Learning to infer adaptively for multimodal large language models (AdaLLaVA). In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J. & Ma, Y. (2023). Investigating the catastrophic forgetting in multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, R., Zhou, Y., Chen, J., Gu, J., Chen, C., & Sun, T. (2024). LLaVA-Read: enhancing the reading ability of multimodal language models. *arXiv preprint*.

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023). PMC-VQA: visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.