INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

MULTIMODAL LARGE LANGUAGE MODELS FOR LOW-RESOURCE LANGUAGES: CROSS-LINGUAL REASONING, RETRIEVAL-AUGMENTED GENERATION, AND HALLUCINATION-AWARE EVALUATION

Moiz Ul Rehman[1], Muhammad Huzaifa Ali[1], Jawad Shafi[*1]

[1]*COMSATS University Islamabad, Lahore Campus, Pakistan.*

[1]moizmalik283@gmail.com, [1]huzaifaali2190@gmail.com, [*1]jawadshafi@cuilahore.edu.pk

DOI: https://doi.org/10.5281/zenodo.23101045

Keywords
Multimodal Large Language Models; Low-Resource Languages; Cross-Lingual Reasoning; Retrieval-Augmented Generation; Hallucination Mitigation; Tokenization Efficiency; Cultural Grounding; Parameter-Efficient Fine-Tuning

Article History
Received: 01 June 2026
Accepted: 23 August 2026
Published: 30 September 2026

Copyright @Author
Corresponding Author: *
Jawad Shafi

Abstract

Multimodal Large Language Models (MLLMs) have changed the face of artificial intelligence by integrating text, vision, audio, and video modalities, but their advantages are mostly limited to high-resource languages like English, Mandarin, and major European languages, leaving the other 7,000+ languages of the world behind. This review provides a systematic overview of the architectural designs, cross-lingual tokenization strategies, culturally-aware instruction tuning, multimodal Retrieval-Augmented Generation (RAG) pipelines, and hallucination evaluation frameworks at the forefront of low-resource MLLM research. We explore three main types of architectural connectors projection-based, query-based, and fusion-based and their respective trade-offs in terms of alignment effectiveness and computational cost. We also examine the tokenization challenges, where subword fragmentation can increase sequence lengths up to 15 times, and consider solutions such as anchor-weighted SentencePiece tokenizers, International Phonetic Alphabet transcriptions, and text-as-image rendering, a vision-only approach that can achieve an 86% fertility reduction and 7.85x sequence compression in low-resource languages. We summarize training methodologies, including cross-modal feature alignment, parameter-efficient fine-tuning with LoRA, and Mixture of Language Experts. We evaluate anchor language priming and culturally relevant dataset generation pipelines like CulturalGround, which generated 22 million multimodal examples in 39 languages, for cross-lingual reasoning. We investigate multimodal RAG elements, such as BGE-M3, LaBSE, and M-CLIP encoders, and the XM-RAG pipeline, which achieved 38. 1% accuracy on LR-MMQA. Lastly, we provide a comparative analysis of hallucination mitigation techniques, including Visual Contrastive Decoding, OPERA, VEGAS, Oh Defense, SENTINEL, and Woodpecker, based on intervention stages and computational costs. We conclude by identifying critical challenges in funding equity, edge-tier deployment, and quantization for Global South applications, charting pathways toward equitable, hallucination-aware multimodal AI for underrepresented linguistic communities.

1. INTRODUCTION

The Emergence of Multimodal Large Language Models (MLLMs) has reshaped computational linguistics and AI by integrating text with vision, audio, and video modalities (Yan et al., 2025). Proprietary and open weights-based architectures achieve strong reasoning capabilities across visual question answering (VQA), optical character recognition (OCR), document understanding, and cross-modal translation. Nevertheless, these milestones in technology are still skewed towards a handful of high-resource languages mainly English, Mandarin, and major European languages leaving behind a significant digital gulf across global linguistic communities (Goworek et al., 2025). Furthermore, there is also a dearth of such resources for under-resourced Sout East Asian languages (Shafi et al., 2025).

The divide in the HR v/s LR language dichotomy arises from structural imbalance in being digitally represented. The high resource languages have large amounts of crawlable corpora, aligned parallel text, and human-annotated multimodal datasets. By contrast, low-resource languages, which encompass the majority of the world's 7,000+ spoken languages, have significant data scarcity (Shafi et al., 2023a) and unstandardized orthographies, and have very little paired multimodal data, such as image-text captions or speech-text alignments (Rugemalila et al., 2026). The comprehensive survey analyses of a large number (117) of studies pertaining to 96 of these LR languages reveal a strong bias in research investment towards geopolitical visibility rather than linguistic need. For instance, three languages Hindi (26 studies), Arabic (18 studies), and Bengali (16 studies) constitute the core of the extant literature, while there are more than 40 under-studied languages represented by a singular paper. In addition, text-image pairs account for 63% of all low-resource multimodal work, thus voice, and more complex video-visual relationships are far less studied (Rao et al., 2025). Building MLLMs for the long tail of languages faces technical challenges beyond simple text translation. Subword tokenizers that are pre-trained on English-centric corpora split morphologically rich or non-Latin low-resource scripts into even smaller units, resulting in high token fertility that exhausts the context window and significantly increases the inference cost (Huang et al., 2026). Moreover, catastrophic forgetting and cultural misalignment, where the translated benchmarks do not fully reflect local scenarios, burden cross-lingual transfer. Low-resource MLLMs are also characteristically more susceptible to object hallucinations producing texts not aligned with visual inputs due to the backbone language model's increased dependency on strong textual priors to make up for weak cross-modal alignment in these models (Gao et al., 2025).

Recent empirical results show that these obstacles can be overcome with appropriate training regimes and gathering open-source data. A case study on Basque – a language with ca. 1,000 times less text data than English in Common Crawl – shows that adding small amounts of target-language multimodal instruction data to late-fusion MLLMs result in strong performance on standardized VQA benchmarks such as A-OKVQA (Do et al., 2025). Similarly, developments in cross-lingual retrieval-augmented generation (RAG) and hallucination-aware decoding offer scalable routes to fair AI. This survey methodically relates architectural paradigms, cross-lingual tokenization mechanisms, culture-based instruction tuning, multimodal RAG pipelines, and hallucination assessment framework to state-of-the-art in low-resource MLLM, (Kurup et al., 2026).

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 5 • September 2026

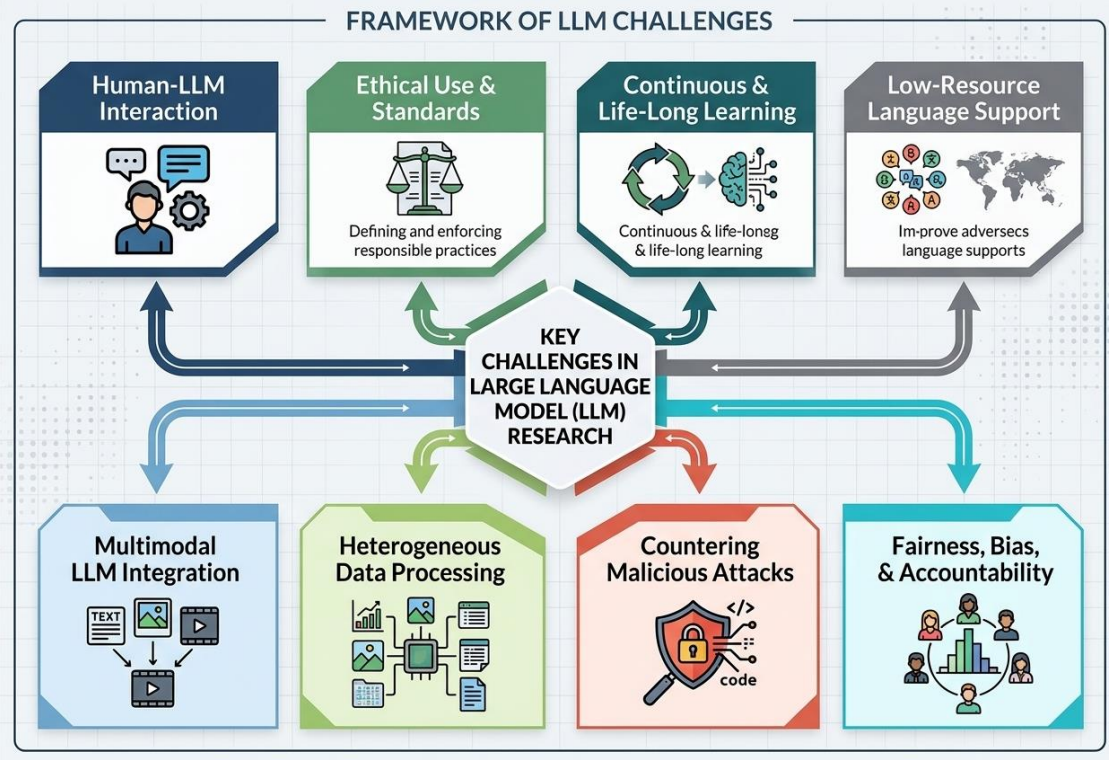


Figure Title: Figure 1: Taxonomy of Core Challenges in Large Language Model (LLM) Research Framework

2. Architectural Frameworks and Cross-Lingual Alignment

2.1 Modality Connectors and Cross-Modal Projection

Most multimodal large language models (MLLMs) use a three-part design: a vision encoder to identify features in images, a language model to process text, and a connector that translates visual features into a format the language model can use. The connector's design affects how well the two modalities work together, as well as the amount of computation required (Liang et al., 2026).

One common option is a projection-based connector, which uses a linear layer or a small neural network—often called an MLP—to map visual tokens into the language model's embedding space. Models such as LLaVA, Shikra, and PandaGPT use this approach because it is relatively lightweight and straightforward to train. It can retain detailed visual information, particularly when resources are limited, but depends on the language model to make sense of the resulting representations (Li & Tang, 2026).

Another approach uses learned queries to select information from the image. The Q-Former in BLIP-2 and InstructBLIP, for example, uses cross-attention to gather a fixed set of visual features before sending them to the language model. This keeps the input sequence shorter, but may omit specific visual details needed to identify uncommon cultural objects, especially those with limited representation in training data (Yi et al., 2025).

Fusion-based connectors integrate cross-attention layers directly inside the Transformer blocks of the LLM backbone, as seen in Flamingo. While expressive, fusion-based architectures require training specialized attention layers across billions of parameters, rendering them computationally prohibitive for resource-constrained adaptation projects (Aldabbas et al., 2025). Pre-trained vision encoders such as CLIP ViT-L/14, OpenCLIP ViT-

H, and EVA ViT-g are generally kept frozen during early adaptation stages to preserve zero-shot visual representation capabilities. Alignment is achieved by training the connector module to map visual features into the embedding space of a pre-trained multilingual text backbone (Zeng et al., 2023).

2.2 Tokenization Bottlenecks and Subword Efficiency

Standard subword tokenization algorithms (e.g., Byte-Pair Encoding, Unigram Language Models) are trained in corpus statistics heavily biased towards high-resource languages. As a result, low-resource languages, particularly those with non-Latin scripts (e.g., Assamese, Meitei Mayek) or agglutinative morphologies (e.g., Basque, Khasi, Garo), have high token fertility, which is defined as the average number of subword tokens needed to represent a single word (Tay et al., 2021).

Extreme vocabulary fragmentation leads to character- or sub-character-level splits. This over-segmentation inflates input sequence lengths by as much as 15x over English for the same semantic content. This sequence length inflation is a major source of computational burden, as standard self-attention mechanisms have a quadratic complexity with respect to sequence length, and it also weakens the LLM's ability to learn long-range contextual dependencies and cross-lingual semantics (Yang, 2024).

Researchers have identified three main approaches to alleviating tokenization bottlenecks: Custom sentence-piece tokenizers trained on carefully curated multilingual corpora employ a strategy of sample weighting. NE-BERT models downweight anchor languages (e.g., Hindi to 0.05×, English to 0.2×) when building their vocabulary. This approach focuses subword allocation to low-resource languages (Shafi et al., 2023), decreasing average fertility from 2.48 tokens/word in mBERT to 1. 72 tokens/word for indigenous Northeast Indian scripts (Teklehaymanot & Nejdl, 2025).

Substituting orthographic text tokenization with International Phonetic Alphabet (IPA) transcriptions or raw byte streams also eliminates unmapped vocabulary entries. IPA tokenizers are shown to consistently enhance the equitability of vocabulary allocation across various language families without compromising downstream transfer (Schmidt et al., 2024).

A radical break from textual subword tokenization is vision-centric tokenization (SeeTok), which converts raw text into images and feeds them to the frozen visual encoder of an MLLM. Completely avoiding the use of text tokenizers, vision-centric tokenization removes subword fragmentation, with an average fertility reduction of 86% across 13 languages. In high-resource scenarios, this leads to a sequence compression ratio of 5.71×, and for low-resource languages, it is 7.85×, significantly reducing costs of inference while enhancing cross-lingual translation quality (Gigant et al., 2026).

2.3 Training Paradigms: Alignment and Parameter-Efficient Adaptation

The adaptation of pre-trained MLLMs to low-resource languages follows a well-defined multi-stage training pipeline. In the initial phase of cross-modal feature alignment, the connector module is trained on large-scale image-text pairs (e.g., WebLI, MMC4, OBELICS) with an auto-regressive, next-token prediction cross-entropy loss while freezing the vision encoder and the multilingual LLM backbone. It matches raw visual features straight into the embedding space of the language backbone (Wang et al., 2025).

In the second phase of multimodal instruction tuning, the model is trained to follow user prompts on task-specific instructions to theends (e.g., VQA, image captioning, complex reasoning). In resource-constrained scenarios, Parameter Efficient Fine-Tuning (PEFT) methods specifically Low-Rank Adaptation (LoRA) are utilized to fine-tune the connector and a small number of LLM attention layers. Experimental results show that adapter-only fine-tuning achieves 70–90% of full fine-tuning performance by learning less than 5% of the total parameters (Sun et al., 2026).

MoLE and Conditional Mixture of LoRA (mixedLoRA) are employed for routing of highly multilingual or codemixed input to specialized low-rank parameter blocks in such cases. This

alleviates the "curse of multilinguality" where increasing language coverage leads to worse performance in individual languages by separating target-language representations from core backbone weights (Han et al., 2025).

Table 1: Architectural Paradigms and Tokenization Strategies for Low-Resource MLLMs

Strategy / Model	Connector Type	Tokenization Approach	Primary Language Focus	Key Architectural / Efficiency Impact
Basque MLLM	Late-Fusion MLP Connector	Standard Subword BPE (Latxa)	Basque	Low target multimodal data ratios ($< 10\%$) suffice for strong close-ended benchmark performance.
NE-BERT	Encoder-level Alignment	SentencePiece Unigram (Anchor Weighting)	9 Northeast Indian Languages	Reduces average fertility to 1.72 tokens/word; improves efficiency by up to $3.08\times$ on non-Latin scripts.
SeeTok	Vision Encoder Projection	Vision-Centric Text-as-Image Rendering	Multilingual (13+ Languages)	Achieves $7.85\times$ sequence compression in LRLs; cuts fertility by 86% across scripts.
CulturalPangea-7B	Linear Connector (LLaVA-NeXT)	Multilingual Subword Tokenizer	Global (39+ Languages)	Fine-tuned on 18M VQA instances; state-of-the-art open-model visual cultural reasoning.
Aya Vision (8B/32B)	High-Capacity Projection	Multilingual Subword Tokenizer	23 Languages	Balanced multilingual VQA performance across high- and low-resource language pairs.

3. Cross-Lingual Reasoning and Cultural Grounding

3.1 Cross-Lingual Modality Transfer and Anchor Language Priming

Cross-lingual transfer of MLLMs is enabled by the shared semantic representations within pre-trained multimodal spaces. Visual features serve as a pivot bridge across languages: since an image of an object encapsulates the same semantic context whether the prompt language is English or Italian, visual inputs ensure semantic alignment during cross-lingual generation (Zheng et al., 2026).

Anchor language priming is useful when instruction data in the target language is limited. Through preserving downweighted representations of high-resource anchor languages (e.g. English or regional vehicular languages such as Hindi) in the process of training, the model transfers complicated reasoning chain patterns learned in the anchor language to the target low-resource language (Nyandwi et al., 2025).

In order to take full advantage of instruction tuning for mixed modalities, Multimodal Instruction Grouping (MINT) provides a taxonomy of an instruction's tasks based on modality interaction: redundant information in image and text tasks; exclusive information source of key information in a modality in unique tasks; and in synergistic tasks, a correct answer is a result of integrating non-overlapping visual and textual evidence in the reasoning process (Shen et al., 2026). The categorization of instruction tuning data according to these interaction patterns mitigates negative task interference and enables low-resource models to acquire a more balanced

reason- ing ability without overfitting on textual language priors (Liu et al., 2021).

3.2 Culturally Grounded Dataset Curation and De-Anglicization

A major flaw in early low-resource multimodal benchmarks was an over-reliance on direct machine translation of Western datasets, such as translating English VQA pairs. Direct translation introduces severe cultural artifacts, name leakage, and Western-centric conceptual biases (Yue et al., 2025).

To solve this, modern frameworks like CulturalGround utilize structured knowledge bases such as Wikidata to construct culturally authentic datasets. The CulturalGround curation pipeline extracts entities connected to specific regions via culturally meaningful properties, including local architecture, culinary traditions, traditional dress, and historical rituals (Alam et al., 2025).

The structured dataset curation process progresses through five defined operational phases:

1. **Entity Selection**: Systematic identification of regional entities across target countries using structured Wikidata metadata.
2. **Image Harvesting**: Collection of licensed images from Wikimedia Commons directly mapped to extracted Wikidata IDs.
3. **Template-Based Question-Answering**: Creation of native-language question-answer pairs corresponding to entity properties.
4. **De-Anglicization and LLM Refinement**: Automated and manual rewriting of translated or template questions to replace Western entities with local equivalents, such as replacing "Arthur Ashe" with Nigerian tennis player "Nduka Odizor", or the "Point Skyhawks" logo with the "Shooting Stars" football team logo (Maji et al., 2025).
5. **Relevance Filtering and Native Validation**: Application of vision-language classifiers to filter out misaligned image-text pairs, followed by validation from native speakers.

This process produced the CulturalGround dataset, which contains 22 million multimodal instances (13 million open-ended VQA pairs and 5 million multiple-choice items in its primary training split) spanning 42 countries and 39 languages. Fine-tuning the Pangea-7B baseline on this data yielded CulturalPangea-7B, which achieves state-of-the-art performance on cross-cultural benchmarks like ALMBench and MERLIN while preserving broad vision-language reasoning capabilities (Goworek et al., 2025).

4. Multimodal Retrieval-Augmented Generation (RAG) in Low-Resource Domains

4.1 Dense and Hybrid Cross-Lingual Vector Retrieval

Multimodal Retrieval-Augmented Generation (RAG) alleviates parametric storage limitations of LLMs by fetching external knowledge during inference. In such low-resource, cross-lingual scenarios, the retrieval pipeline is required to operate on multi-source knowledge bases that include text, images, as well as structured tables, and that span across languages (Yan et al., 2025).

The central building block of cross-lingual RAG is the dense sentence encoder. Leading multilingual embedding families are, LaBSE (Feng et al., 2022), which is focused on bitext mining, as well as on fine-grained cross-lingual distance measurement, mE5 (Pan et al., 2023), trained on massive-scale contrastive multilingual pairs and high-performing on dense passage retrieval, BGE-M3 as a one-for-all model that can be used to complete the workflow of dense vector retrieval, sparse lexical matching and multi-vector late-interaction re-ranking (Rugemalila et al., 2026).

The cited study conducted an ablation analysis in which we replace Tutor with those fewer variants. In cross-lingual cross-modal retrieval, replacing a modality-aligned M-CLIP encoder with a text-only BGE encoder resulted in a Recall@10 decline from 67.9% to 54.3% in Text+Image mode, which corroborates that dense encoders need to explicitly uphold joint vision- language-cross-lingual spaces (Rao et al., 2025).

Table 2: Performance and Features of Leading Cross-Lingual Multimodal RAG Components and Encoders

Component / Model	Modality Support	Cross-Lingual Retrieval Mechanism	Target Languages / Benchmark	Key Empirical Performance Metric
BGE-M3	Text (Multi-vector)	Hybrid (Dense + Sparse + ColBERT)	100+ Languages	Achieves 82.49% baseline retrieval accuracy paired with light MLP classifiers.
LaBSE	Text	Dual-Encoder Cosine Alignment	109 Languages	Strongest baseline for bitext mapping and fine-grained semantic distance.
M-CLIP Encoder	Image + Text	Joint Vision-Language Embedding	Multilingual	Key driver in XM-RAG; replacing with text-only BGE drops Recall@10 from 67.9% to 54.3%.
XM-RAG Pipeline	Image + Text + Table	Multi-Evidence Cross-Lingual RAG	Tamil, Yoruba (LR-MMQA)	**38.1% Accuracy** on LR-MMQA vs. 29.4% (Claude 4.5 Sonnet) and 28.6% (GPT-5 Mini).

5. Hallucination-Aware Evaluation and Mitigation

5.1 Taxonomy and Vulnerabilities of Multimodal Hallucinations

Hallucination in MLLMs is the production of text responses that are fluent and semantically coherent but are not grounded in the visual context they receive. In low-resource language scenarios, the hallucination rates are further exacerbated by poor cross-modal alignment as well as skewed pre-training data distributions (Yang, 2024).

Multimodal hallucination is divided, empirically, into three main structural types: object category hallucination, where the model asserts that objects or classes of objects are in the image when they are not; object attribute hallucination, where the model identifies an object‟s class correctly but hallucinates its colour, shape, material, number or quantity and object relation hallucination, in which entities and attributes are correctly recognized but spatial positions or human-object interactions are inaccurately described (Rugemalila et al., 2026) or more Recently.

An important failure mode is the escalation of early hallucination into snowball hallucination. Once an MLLM produces a first hallucinatory token declaring a non-existent object, the token is used as part of the prompt history in the auto-regressive decoding loop. To stay logically consistent with the context that it has created itself, the model conjures up more detailed errors of explanation. For low-resource languages, the strong textual language priors easily overpower the weak visual attention signals, leading to compounding error bias for the models to seize on (Gao et al., 2025).

5.2 Diagnostic Probing and Evaluation Benchmarks

Assessing hallucination vulnerabilities calls for rigorous diagnostic methodologies:

Polling-based Object Probe Evaluation (POPE) assesses the existence of an object by querying models with a set of structured binary queries ("Is there a [category] in the image?") on random, popular and adversarial co-occurrence distributions (Li & Tang, 2026).

Generative Hallucination Optimization Through Subtle Transformations (GHOST) is an adversarial probing framework that leverages diffusion models to search for subtle latents optimizable in the input images. GHOST adds subtle semantics that confuse MLLM decision boundaries into perceiving non-existent objects but does not impact human vision. On SOTA vision models (e.g., Qwen2.5-VL), GHOST is able to cause object hallucination with a 29% success rate (0.1% for normal baseline perturbations). Furthermore, GHOST-transformed images have

transferable weaknesses, causing object hallucinations on proprietary models like GPT-4o at a 66.5% success rate (Gigant et al., 2026).

5.3 Training-Free Decoding and Activation Steering Strategies

To reduce hallucinations without high computational overhead, researchers have proposed training-free inference-time intervention methods:

Visual Contrastive Decoding (VCD) compares output logit distributions from the original visual input with output logit distributions from a visually corrupted or noise-added version. The noise-conditioned distribution subtraction extracts visually grounded tokens and negates visual-agnostic language priors (Han et al., 2025).

OPERA (Over-trust Penalty in Self-Attention) detects attention collapse, in which self-attention distributions in long-sequence generation become excessively concentrated on summary tokens. It adds an over-trust penalty during beam search decoding to reduce repetitive attention loops and recover trust on visual tokens (Sun et al., 2026).

VEGAS (Vision Attention Block Entropy) tracks the Vision Attention Block Entropy (VABE) along the transformer layers in the process of generation. A sudden rise in VABE indicates a hallucination is about to occur. VEGAS replaces attention maps at middle transformer layers (Layers 14 and 15) with early vision-encoder attention distributions in an adaptive fashion when such switching is triggered by VEGAS during inference, imposing the visual grounding in those layers which are the highest at which early vision-encoder attention maps can be swapped out for full attention maps without needless reduction of model's inference parallelizability. FarSight mitigates attention collapse by introducing special attention registers to hold the overflowing attention weights, as well as absolute positional focus encodings that counteract long-term Rotary Position Embedding (RoPE) decay (Liang et al., 2026). Oh Defense reveals that most of the safety and visual truthfulness capabilities in MLLMs are stored in a few, important attention heads. By observing and enhancing the discriminative activation signal of a single key attention head during decoding, Oh Defense can nearly perfectly suppress hallucinations (>98% defense success rate) with minimal false-positive rates (<5%) on standard benchmarks (Wang et al., 2025).

5.4 Preference Optimization and Post-Hoc Verification

In addition to decoding interventions, fine-tuning through Direct Preference Optimization (DPO) can also align model generation with factual outputs. Frameworks like SENTINEL form a richly contextualized preference dataset using an automated object detector to verify generated entities. SENTINEL intervenes at the first instance of a hallucinated entity, creating negative preference pairs and positive pairs that are causally consistent with the surrounding context to train policy models using DPO loss.

Another approach is post-hoc pipelines like Woodpecker, which apply multi-stage visual verification. Woodpecker identifies major claims in the generated text, applies open-vocabulary object detectors and visual grounding models (e.g., Grounding DINO, SAM) to obtain regional visual evidence, and then rewrites hallucinatory claims. Although this significantly improves POPE accuracy (by as much as 30.66% over MiniGPT-4 baselines), post-hoc verification also multiplies inference latency by 5 to 10 times due to external detector calls (Do et al., 2025).

Table 3: Comparative Analysis of Hallucination Mitigation Approaches in MLLMs

Method	Intervention Stage	Compute / Retraining Cost	Architectural Mechanism	Primary Strength / Limitation

Visual Contrastive Decoding (VCD)	Decoding Time	Low (Inference Only)	Logit contrast between clean and noisy visual inputs.	Effective penalization of language priors; slightly increases inference latency.
OPERA	Decoding Time	Low (Beam Search)	Rollback penalty on self-attention summary token concentration.	Prevents snowballing in long generations; requires beam search capability.
VEGAS	Decoding Time	Low (Adaptive Layer-Shift)	Vision Attention Block Entropy (VABE) adaptive layer substitution.	Targets middle transformer layers (Layers 14–15) to restore visual focus dynamically.
Oh Defense (One-Head)	Inference Steering	Zero Retraining	Amplification of a single critical safety attention head.	**>98% defense success rate** with zero additional parameter updates.
SENTINEL	Training Phase	Medium (DPO Fine-Tuning)	Sentence-level early preference optimization.	Prevents early hallucination propagation without degrading general capabilities.
Woodpecker	Post-Processing	High (5 × ~10 × Latency)	External object detection, visual grounding, and text rewriting.	High precision gains (+30.66% on POPE); high latency from external model pipeline.

6. Benchmarks, Cultural Grounding, and Empirical Evaluation

Standardized benchmarks are essential for tracking the progress of multimodal models across diverse languages. Early evaluations relied on translation-based adaptations of standard English datasets such as A-OKVQA, xGQA, and XM3600. However, recent efforts focus on native, culturally grounded, and multi-discipline suites that evaluate real-world low-resource performance (Maji et al., 2025).

Table 4: Representative Multilingual Multimodal Evaluation Benchmarks

Benchmark	Primary Modalities	Language Coverage	Primary Evaluation Focus	Key Distinctive Characteristics
XM3600	Image + Text	36 Languages	Cross-Lingual Image Captioning	Human-annotated diverse images worldwide with localized captions.
xGQA	Image + Text	7+ Languages	Structured Visual Compositional QA	Evaluates compositional reasoning across translated/localized visual graphs.
M3Exam	Image + Text	Multilingual	Multi-Level Academic Exam Questions	Real human exam questions require integrated visual-text reasoning.
MUNIChus	Image + Text	9 Languages (inc. 3 LRLs)	News Image Captioning	Largest news dataset with 700k+ images across diverse language families.

LR-MMQA	Image + Text + Table	Tamil, Yoruba	Open-Domain Multimodal KBQA	718 de-anglicized multi-hop question pairs per language.
CulturalGround / PangeaBench	Image + Text	39+ Languages (42 Countries)	Localized Cultural Knowledge VQA	22M instances grounded in Wikidata cultural properties.

7. Challenges and Future Directions

Low-resource multimodal language modeling still has critical bottlenecks in its structure and technology, although some recent advances have been made. Most of the research activity remains strongly decoupled from linguistic need. As regional vehicular languages grow in number and strength of funding, some tens of indigenous, endangered and unwritten languages are left out. Closing that gap requires funding streams that foreground marginalized rather than geopolitically dominant language families (Gigant et al., 2026).

The deployment of multi-billion parameter MLLMs in rural or offline areas, for example, at healthcare kiosks in the Global South or on mobile phones, is constrained by substantial memory and power requirements. In the future, research needs to be done on edge-tier deployment topologies, on INT8/INT4 quantization, structured pruning, and on ultra-compact sub-3B architectures capable of running on low-power devices (Teklehaymanot & Nejdl, 2025).



Figure 2: Overview of Key Technical and Operational Challenges in Multimodal Large Language Model (MLLM) Development

Conclusion

Based on the mapping of the current research landscape in Multimodal Large Language Models for low-resource languages, this survey exposed both a range of exciting new advances and difficult-to-overcome structural challenges. Our findings highlight that the high-resource/low-resource divide is not about data scarcity alone: it is about inefficiencies in tokenization, misalignments in modalities, cultural misalignment, and lacunae in evaluation. The good news from the empirical literature reviewed here is that these obstacles can be overcome, at least in part, by grass roots measures. Vision-

centric tokenization reaches a fertility reduction of 86% and sequence length compression of 7.85×, and weighted anchor-based SentencePiece tokenizers achieve the highest average compression of 1.72 tokens per word for native scripts. LoRA-based parameter-efficient fine-tuning achieves 70–90% of the full fine-tuning performance with less than 5% of parameters updated, and Mixture of Language Experts (MoLE) architectures address the curse of multilinguality. Systems such as CulturalGround have yielded state-of-the-art cross-cultural reasoning for CulturalPangea-7B through a culturally informed curation pipeline that has generated 22 million instances in 39 languages. Multimodal pipelines such as XM-RAG demonstrate, however, that retrieval-based approaches can at least in part compensate for parametric deficiencies in low-resource contexts. Hallucination mitigation techniques encompassing training-free decoding modifications (VCD, OPERA, VEGAS, Oh Defense) and preference optimization (SENTINEL) and subsequent verification (Woodpecker) facilitate multiple avenues of enhancing truthfulness, however Oh Defense has a defense success rate of >98% with zero retraining cost and training-free decoding interventions have higher costs. However, critical challenges remain. Research investment continues to be driven by geopolitical visibility rather than linguistic need, leaving dozens of indigenous and endangered languages entirely unsupported. Deployment in rural and disconnected settings is hampered by memory and power demands, necessitating INT8/INT4 quantization, structured pruning, and sub-3B parameter architectures. Future work must prioritize equitable funding pipelines, edge-tier deployment topologies, and native-speaker validation protocols. Only through coordinated architectural innovation, culturally authentic data curation, and inclusive evaluation frameworks can MLLMs fulfill their promise as truly global technologies serving all linguistic communities.

REFERENCES

Alam, F., Shahroor, A. E., Hasan, M. A., Ali, Z. S., Bhatti, H. H., Kmainasi, M. B., ... & Milic-Frayling, N. (2025). EverydayMMQA: A Multimodal and Multilingual Framework for Culturally Grounded VQA.

Aldabbas, F., Elsafty, H., & Sifa, R. (2025). It's All Connected: A Survey for Multimodal Arabic AI.

Do, X. L., Dinh, D., Nguyen, N. H., Kawaguchi, K., Chen, N., Joty, S., & Kan, M. Y. (2025, July). What makes a good natural language prompt?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5835-5873).

Gao, C., Fan, G., Chong, C. Y., Chen, S., Liu, C., Lo, D., ... & Liao, Q. (2025). A systematic literature review of code hallucinations in llms: Characterization, mitigation methods, challenges, and future directions for reliable ai. *arXiv preprint arXiv:2511.00776*.

Gigant, T., Peng, B., & Quesnelle, J. (2026). Decoupling the Benefits of Subword Tokenization for Language Model Training via Byte-level Simulation. *arXiv preprint arXiv:2604.27263*.

Goworek, R., Macmillan-Scott, O., & Özyiğit, E. B. (2025). Bridging language gaps: Advances in cross-lingual information retrieval with multilingual llms. *arXiv preprint arXiv:2510.00908*.

Han, X., Yang, J., Wang, T., Bi, Z., Song, X., Hao, J., & Song, J. (2025). Towards alignment-centric paradigm: A survey of instruction tuning in large language models. *arXiv preprint arXiv:2508.17184*.

Huang, Y., Gao, C., Wu, S., Wang, H., Wang, X., Ye, J., ... & Zhang, X. (2026, April). TrustGen: A Platform of Dynamic Benchmarking on the Trustworthiness of Generative Foundation Models. In *International Conference on Learning Representations* (Vol. 2026, pp. 60745-60905).

Kurup, S., Menon, G. S., Jayakumar, S., Mohan, V. M., Veena, G., & Kanjirangat, V. (2026, September). Scope Matters: A Hallucination-Aware Evaluation of Retrieval Strategies for Multilingual RAG. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 103-117). Cham: Springer Nature Switzerland.

Li, S., & Tang, H. (2026). Multimodal alignment and fusion: A survey: S. li, h. tang. *International Journal of Computer Vision*, *134*(3), 103.

Liang, C. X., Tian, P., Yin, C. H., Yua, Y., Wei, A. H., Li, M., ... & Feng, P. (2026). A comprehensive survey and guide to multimodal large language models in vision–language tasks. *Computation*, *14*(6), 125.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., & Elliott, D. (2021, November). Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 10467-10485).

Maji, A., Kumar, R., Ghosh, A., Shah, N., Borah, A., Shah, V., ... & Saha, S. (2025, November). Drishtikon: A multimodal multilingual benchmark for testing language models' understanding on indian culture. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 1289-1313).

Nyandwi, J. D. D., Song, Y., Khanuja, S., & Neubig, G. (2025, November). Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 24198-24242).

Rao, J., Liu, X., Yan, H., Shen, J., Mo, H., Dong, Y., ... & Zhang, M. (2025). A data-centric perspective on the lifecycle of large language models.

Rugemalila, I. E., Cai, W., Zhang, X., Liu, C., & Wang, B. (2026). Curriculum-Aware Retrieval-Augmented Generation for Bilingual Tutoring in Low-Resource Swahili–English Secondary Schools. *Technologies*, *14*(2), 129.

Schmidt, C. W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., & Tanner, C. (2024, November). Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 678-702).

Shafi, J., Iqbal, H.R., Nawab, R.M.A. and Rayson, P., 2023. UNLT: Urdu natural language toolkit. *Natural language engineering*, *29*(4), (pp. 942-977).

Shafi, J., Hussain, A., Nawab, R.M.A. and Rasool, M., 2025. Multi-label document classification for Urdu: corpus and methods. *Language Resources and Evaluation*, *59*(4), (pp.4253-4283).

Shafi, J., Adeel Nawab, R.M. and Rayson, P., 2023a. Semantic tagging for the Urdu language: Annotated corpus and multi-target classification methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(6), (pp.1-32).

Shen, F., Zeng, Y., Lyu, D., & Wen, W. (2026, April). From Words to Worlds: Measuring Cultural Narrative Bias in LLMs via a Structural–Value Pipeline. In *Proceedings of the ACM Web Conference 2026* (pp. 8961-8972).

Sun, Y., Zhu, J., Cheng, H., Luv, C., Yang, Z., Chen, L., & Wang, Y. (2026). Align then adapt: Rethinking parameter-efficient transfer learning in 4d perception. *IEEE Transactions on Multimedia*.

Tay, Y., Tran, V. Q., Ruder, S., Gupta, J., Chung, H. W., Bahri, D., ... & Metzler, D. (2021). Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Teklehaymanot, H. K., & Nejdl, W. (2025, November). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)* (pp. 822-828). IEEE.

Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2025). Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, *58*(8), 227.

Yan, L., Li, R., Geng, J., Li, Q., Wu, M., Wang, Z., ... & Lyu, C. (2025, October). On the Taxonomy, Tasks, and Open-Challenges for Multimodal Large Language Models. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2518-2525). IEEE.

Yang, J. (2024). Rethinking tokenization: Crafting better tokenizers for large language models. *International Journal of Chinese Linguistics*, *11*(1), 94-109.

Yi, W., Cai, X., Ma, H., Fu, Z., & Zhan, Y. (2025). A Library-Oriented Large Language Model Approach to Cross-Lingual and Cross-Modal Document Retrieval. *Electronics*, *14*(15), 3145.

Yue, X., Song, Y., Asai, A., Kim, S., NYANDWI, J., Khanuja, S., ... & Neubig, G. (2025, May). Pangea: A fully open multilingual multimodal llm for 39 languages. In *International Conference on Learning Representations* (Vol. 2025, pp. 47758-47811).

Zeng, Y., Zhou, W., Luo, A., Cheng, Z., & Zhang, X. (2023, July). Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5731-5746).

Zheng, W., Liu, Z., Chakraborty, T., Xu, W., Gao, X., Tan, B. C. Z., ... & Chen, N. (2026, July). MMAC: A Multilingual, Multimodal Alignment Framework for Cultural Grounding Evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 21667-21709).

International Journal of Advanced Research