

MURMUR: Cross-Lingual and Multimodal Retrieval-Augmented Reasoning for Open Question Answering in Tamil and Yoruba

Kiran Raja¹, Mobareji Abejide², Arya Ram³, Utkarsh Sharma⁴, Kevin Zhu⁵

¹Alcuin School, ²Noblesville High School, ³McNair Academic High School

⁴UNSW Canberra, ⁵AlgoVerse AI

{kiran.rajatx1,mobarejiab,aryaram.ams,sharmautkarsh0504}@gmail.com,
zhu502846@berkeley.edu

Abstract

As large language models (LLMs) with retrieval augmented generation (RAG) gain traction in multimodal knowledge base question answering (KBQA), concerns about their transfer to low resource languages (LRLs) remain unaddressed. We introduce **MURMUR**¹, a benchmark evaluating multimodal cross-lingual retrieval and reasoning in Tamil and Yoruba. Using the hardest examples from WebQA and MultimodalQA, we build a high quality LRL benchmark through LLM assisted translation, human validation, and culturally aligned rewriting that reflects native speaker phrasing while preserving answerability. MURMUR exposes major performance gaps and failure modes in current systems. We provide **XM-RAG** as a supporting cross-lingual multimodal RAG baseline, which reaches 38.1 answer accuracy and exceeds the next best baseline by more than 6.3 points. XM-RAG still performs far below top English results on WebQA and MultimodalQA, showing that existing methods struggle with complex tasks in LRL settings. By releasing MURMUR with a reproducible baseline, we offer a resource for measuring these gaps and guiding progress toward equitable multimodal KBQA.

1 Introduction

We introduce **MURMUR** (Multimodal, Under-Resourced Multilingual Understanding and Reasoning), a benchmark for open cross-lingual multimodal KBQA in Tamil and Yoruba. MURMUR measures retrieval and reasoning under realistic language mismatch, where questions are written in an LRL and supporting evidence remains largely in English. Large language models have made rapid progress on open-domain question answering through retrieval-augmented generation (RAG), where external evidence is retrieved and fused during generation (Lewis et al., 2020). This progress has extended into multimodal settings, where systems retrieve and reason over both text and visual evidence for knowledge-intensive tasks (Chen et al., 2022; Ding et al., 2024; Yarrabelly & Mittal, 2025). However, most gains remain concentrated in high-resource languages, while LRL settings are under-evaluated and substantially more error-prone (Lupaşcu et al., 2026; Qi et al., 2025).

A central challenge is that cross-lingual pipelines in LRLs often depend on translate-then-retrieve strategies (Asai et al., 2021; 2022). In multimodal KBQA, this can introduce compounding errors at multiple stages: semantic drift during query translation, retrieval mismatch in high-resource semantic spaces, and answer degradation during back-generation into the original language. These issues become more severe for multi-hop and cross-modal questions, where small translation shifts can break evidence chaining across modalities.

Existing benchmark ecosystems do not fully expose this failure regime. Datasets such as WebQA and MultimodalQA have been foundational for multimodal KBQA progress (Chang

¹You can find the dataset here: [MURMUR on Hugging Face](#).

et al., 2022; Talmor et al., 2021), but they are predominantly HRL-centric. Conversely, many multilingual QA resources in low-resource settings are unimodal and therefore cannot measure cross-modal retrieval-reasoning behavior (Longpre et al., 2020; Sawczyn et al., 2024; Rohera et al., 2024). As a result, current evaluations can overestimate practical robustness for multilingual multimodal deployment.

MURMUR is built from hard examples mined from WebQA and MultimodalQA, followed by translation, native-speaker validation, and culturally aligned rewriting. This design preserves answerability while improving linguistic and cultural naturalness, yielding a benchmark that better reflects real multilingual reasoning conditions.

To support evaluation on the benchmark, we provide **XM-RAG** as a modular baseline designed for direct LRL understanding rather than mandatory translation-first inference. XM-RAG combines multilingual query encoding, cross-modal retrieval, reranking, and multilingual generation. It achieves 38.1 QA-Accuracy on MURMUR and outperforms strong translation-based and RAG-augmented baselines. The remaining gap to English settings is still large, reinforcing that equitable multimodal QA in low-resource languages remains an open research problem.

2 Related works

Multimodal Retrieval-Augmented Generation for Knowledge Base Question Answering

Despite the many advances of LLMs with multimodal RAG, they still struggle to use external knowledge and unseen data, both of which are necessary for KBQA (Zhang et al., 2024). RAG addresses this issue by retrieving external evidence from a corpus of knowledge, in turn increasing accuracy and grounding of model responses to knowledge-base questions (Lewis et al., 2020). These results have seen further improvements due to multimodal retrieval. Methods such as Multimodal Multihop, a methodology used to gather data from multiple sources to formulate an answer, show evidence of these promising results when built upon baseline models (Yarrabelly & Mittal, 2025). Multimodal retrieval allows for the vast multimodal evidence to be leveraged to answer questions that cannot be fully answered with only text. MuRAG (Chen et al., 2022) does this by treating images as visual tokens. RA-BLIP (Ding et al., 2024) projects retrieved text and images into a shared space before fusion. No matter how these methods treat images, they only retrieve content from the language of the query, meaning the quality of the answers is dependent on the quantity and quality of retrievable evidence present, both of which are lacking in low-resource languages.

Cross-Lingual Retrieval Work has shown that retrieval models struggle in LRLs due to lack of high-quality retrieval content (Qi et al., 2025; Lupaşcu et al., 2026). A commonly explored solution to this problem has been a pipeline in which the original LRL query is used to retrieve content from a high-resource knowledge base (KB). That content is used to generate an answer in the original LRL. Quite a lot of work has been done exploring this solution. The most common approach to this is the translate-then-retrieve pipeline, as seen in XOR-RETRIEVE (Asai et al., 2021). This approach translates the query to a high resource language, uses this embedding to retrieve text evidence, and feeds it into a multilingual pre-trained model to generate an answer in the LRL. A substantial improvement on this approach can be seen in CORA, where a multilingual embedding of the query is used, reducing errors that arise with machine translation (Asai et al., 2022). This direction, leveraging multilingual models like BERT for direct cross-lingual information retrieval, has been explored in various works, demonstrating effectiveness in matching queries across languages (Jiang et al., 2020). The embedded query is then fed into a pre-trained retrieval algorithm and the retrieved evidence is fed into a multilingual auto-regressive generation model to produce an answer. However, all these works only take in, reason on, and retrieve text, significantly limiting the type of questions they can answer accurately.

Multimodal Reasoning and Retrieval Knowledge Base Question Answering Benchmarks

Many KBQA benchmarks contain or are entirely composed of questions that require retrieval and understanding across data modalities (Marino et al., 2019; Talmor et al., 2021; Chang et al., 2022). However, these datasets solely contain questions and answers in high-resource languages (HRLs). On the other hand, there are KBQA benchmarks that contain questions

Dataset	LRL	MM	Open Ret.	KIQA
XOR-QA (Asai et al., 2021)	Yes	No	Yes	Yes
MKQA (Longpre et al., 2020)	Yes	No	No	Yes
OpenViVQA (Nguyen et al., 2023)	Yes	Yes	No	No
BOK-VQA (Kim et al., 2024)	Yes	Yes	No	No
CVQA (Romero et al., 2024)	Yes	Yes	No	No
Kaleidoscope (Salazar et al., 2025)	Yes	Yes	No	No
WorldMedQA-V (Matos et al., 2025)	No	Yes	No	No
WebQA (Chang et al., 2022)	No	Yes	Yes	Yes
MultimodalQA (Talmor et al., 2021)	No	Yes	Yes	Yes
MURMUR (ours)	Yes	Yes	Yes	Yes

Table 1: Positioning MURMUR against representative multilingual and multimodal QA datasets. Columns: LRL = low-resource language support, MM = multimodal input, Open Ret. = open external retrieval required, KIQA = knowledge-intensive QA.

and answers for LRLs, but only provide the dataset in a single modality, normally text (Longpre et al., 2020; Sawczyn et al., 2024; Rohera et al., 2024). While VQA evaluation in low-resource languages has been studied (Nguyen et al., 2023; Kim et al., 2024; Salazar et al., 2025; Romero et al., 2024; Matos et al., 2025), these datasets are typically image-grounded VQA tasks and do not require open external retrieval over a knowledge source. By contrast, multilingual retrieval QA datasets such as XOR-QA are knowledge-intensive but text-only (Asai et al., 2021). This creates a structural gap: prior work generally covers either multilingual retrieval without multimodality, or multilingual multimodality without retrieval-intensive KBQA.

To make this distinction explicit, Table 1 positions MURMUR against representative multilingual and multimodal QA datasets. The table highlights that MURMUR is the first benchmark in this set that jointly requires (i) low-resource language understanding, (ii) multimodal reasoning, and (iii) knowledge-intensive open retrieval.

3 Dataset

3.1 Task Definition and Scope

MURMUR evaluates open, cross-lingual, multimodal KBQA in low-resource languages. Each query is in Tamil or Yoruba. Ground truth documents remain in English. A system must (i) understand the LRL query, (ii) cross-lingually retrieve relevant text and image evidence, and (iii) generate an answer in the same LRL.

This setup is intentionally strict. It tests retrieval, grounding, and generation under language mismatch. It does not reward systems that rely only on fluent generation without correct evidence support.

The two-language design is deliberate and typologically motivated. Tamil (Dravidian, agglutinative, non-Indo-European) and Yoruba (Niger-Congo, tonal, distinct morphosyntax) stress-test cross-lingual transfer along very different linguistic axes, both relative to English and relative to each other.

Keeping evidence in English is also a realistic constraint rather than an arbitrary simplification: multilingual QA in low-resource settings often retrieves from English-dominant corpora because high-coverage LRL evidence stores remain limited, as reflected in cross-lingual retrieval setups such as XOR-QA and CORA (Asai et al., 2021; 2022).

3.2 Source Data and Hard-Case Mining

We build MURMUR from WebQA and MultimodalQA. We run strong multimodal baselines (SKuRG and RAMQA) on source questions and keep failure cases, i.e., questions that both

strong models get wrong. This produces a deliberately hard subset and reduces ceiling effects.

Origin	Text	Image	Image + Text
WebQA	104	488	0
MultimodalQA	41	61	13
Total	145	549	13

Table 2: Origin and required retrieval modalities for the selected question-answer pairs. (Total unique questions: 707)

The benchmark has 707 unique questions. We create one Tamil and one Yoruba version per question. This gives 1,414 language-specific instances. This scale is intentional: MURMUR is designed as a diagnostic hard-case benchmark where curated difficulty is often more informative than raw size for measuring retrieval-and-reasoning robustness, in line with challenge-focused evaluation paradigms such as Dynabench and BIG-Bench Hard (Kielbaso et al., 2021; Suzgun et al., 2022).

3.3 Translation and Native-Speaker Validation

We first translate each English QA pair into Tamil and Yoruba with few-shot prompts (Appendix A, Appendix B) using Claude 3.5 Sonnet. We then recruited native-speaker volunteers to conduct validation for adequacy, fluency, and answerability.

A translated item is accepted only when it passes all three checks. Items that fail are discarded or rewritten before inclusion. This step prevents noisy translations from inflating downstream error rates.

Translation quality remains strong (Adequacy 8.2/10, Fluency 8.0/10; Table 11 in Appendix F by native-speaker volunteers).

3.4 Culturally Aligned Rewriting

Direct translation preserves literal meaning but often preserves anglicized entities and phrasing that are unnatural for native users. To reduce this translation barrier, we add a cultural alignment stage.

The stage has three components:

- **Step 1 (Fluency):** a lightweight LoRA pass on large monolingual Tamil and Yoruba corpora to adapt Qwen-2.5-7B-Instruct toward native LRL syntax and fluent local phrasing. We selected Qwen-2.5-7B-Instruct for the cultural alignment stage because it offered stable multilingual instruction-following behavior under LoRA adaptation at the time of our experiments in December 2025. While Qwen 3 was available, its specific capabilities for Yoruba were not yet established or benchmarked. Our LoRA adaptation trains directly on native Yoruba and Tamil text, which mitigates limitations in the base model’s language coverage and makes adaptation data quality more important than the specific base model.
- **Step 2 (Task/Culture):** a second LoRA pass with supervised targets from a gold dataset built as follows: native speakers rewrote WebQA and MultimodalQA questions that were initially translated by Claude 3.5 Sonnet, using questions that were not included in MURMUR. These human rewrites were used as gold outputs to teach authentic culturally grounded question style while preserving answerability and generation quality.
- **Step 3 (Entity alignment):** manual replacement of culturally mismatched entities while preserving answerability.

For cultural analogues where the swapped entity is well represented in English-language sources, answerability is preserved naturally through retrieval. When the swapped entity is

not sufficiently covered in the existing retrieval corpus, we add relevant English-language documents covering the new entity to the evidence corpus. This procedure ensures that every culturally rewritten question remains answerable.

The first LoRA pass uses IndicCorp v2 (Doddapaneni et al., 2023), a cleaned Tamil corpus of approximately 3 to 4 billion tokens, for Tamil fluency adaptation. For Yoruba, we use three complementary sources. These are CC-100 Yoruba (Conneau et al., 2020), MasakhaNEWS (Adelani et al., 2023), and Yankari (Akpobi, 2025). Yankari contains 30 million tokens drawn from 13 diverse sources and explicitly excludes religious and machine-translated text.

The second LoRA pass is trained on 500 human-rewritten examples per language and 1,000 examples in total. Native speakers produced these examples by rewriting WebQA and MultimodalQA questions that were not included in MURMUR, which prevents contamination between the adaptation data and benchmark evaluation set.

The goal is controlled rewriting, not free paraphrase. We preserve answer spans and evidence linkage so that benchmark difficulty comes from reasoning and retrieval, not annotation drift. In practice, LoRA adaptation without this supervised stage degrades coherent generation quality; the staged fluency-then-task design avoids that failure mode.

For the cultural-alignment evaluation in Table 3, we used two native speakers as annotators; these annotators were different from the native-speaker volunteers used in Appendix F to reduce confirmation bias.

Metric	Original Translation	Culturally Aligned Rewrite	Δ
Avg. Cultural Relevance (1–5)	2.41	4.12	+1.71
% Rated “Culturally Natural” (≥ 4)	18%	83%	+65%
Cohen’s κ Tamil	0.58	0.78	+0.20
Cohen’s κ Yoruba	0.60	0.74	+0.14
Cohen’s κ Overall	0.59	0.76	+0.17

Table 3: Cultural alignment quality improves strongly after rewriting. Cultural Relevance is scored on a 1–5 scale (1=not authentic / native would never ask, 5=highly authentic). The rewrite step raises relevance and naturalness while also improving annotator agreement.

Table 3 confirms the cultural alignment design. Cultural Relevance improves from 2.41 to 4.12 and “Culturally Natural” ratings rise from 18% to 83%. Agreement also increases in both languages, with Tamil at $\kappa = 0.78$ and Yoruba at $\kappa = 0.74$, and the pooled overall score at $\kappa = 0.76$. Rewritten examples are in Appendix L.

The downstream effect of cultural rewriting is reported in Table 26 in Appendix U. The comparison uses identical gold answers and evidence corpora for the pre-rewrite and post-rewrite question sets.

3.5 Quality Control and Traceability

Each final item must pass three checks: semantic fidelity, fluent local phrasing, and preserved answerability with ground-truth evidence. We also store traceability metadata that links each final sample to its source QA pair and post-translation stage (pre-rewrite) answer. This metadata enables controlled comparisons between pre-rewrite and post-rewrite question sets because each final item is linked to its post-translation pre-rewrite version with identical gold answers and evidence corpora.

This audit trail is important for reviewer-facing reliability. It allows us to inspect whether model errors come from retrieval and reasoning, rather than hidden dataset corruption.

4 XM-RAG

4.1 Design Goal

XM-RAG is a strong baseline for MURMUR. The model is built to answer low-resource multilingual multimodal KBQA questions without forcing an LRL→HRL→LRL translation loop. The design follows a practical constraint: strong accuracy under limited compute and moderate memory budgets, while preserving direct LRL semantics.

4.2 Pipeline

Given an input query q in Tamil or Yoruba and optional input image x , XM-RAG runs five stages.

(1) Input encoding. We encode the query text and query image in a shared multilingual space. If no query image exists, the pipeline uses only text features.

(2) Candidate retrieval. We retrieve top- K text candidates and top- K image candidates from FAISS indices over high-resource corpora. We denote retrieved text as t_i and retrieved images as v_i .

(3) Similarity scoring and reranking. We compute query-conditioned similarity scores for text and image candidates. A lightweight reranker then combines text similarity, image similarity, and simple consistency signals (e.g., lexical overlap and entity consistency). Candidates are re-ordered and truncated to a compact evidence set.

(4) Evidence fusion. Retrieved image evidence is summarized into short textual captions with BLIP-2. Retrieved text evidence is deduplicated and trimmed. Both streams are merged into a single generation context.

(5) Answer generation. The final answer is generated in the original LRL with language tags, so output stays in Tamil or Yoruba.

4.3 mT5 and Caption-Based Fusion Design

We use mT5 because the project target is not only raw peak accuracy, it is strong multilingual QA under realistic resource limits. mT5 is robust across many languages and gives stable behavior for Tamil and Yoruba with moderate compute cost. As shown in Table 4, mT5 does not need to outperform every frontier model in pure generation to remain competitive end to end, since retrieval quality and cross modal evidence ordering drive large gains in final QA accuracy. This makes the system easier to reproduce and deploy than heavier frontier models that often require larger memory budgets and higher inference latency.

We also use caption-based visual fusion instead of sending every retrieved raw image to a large MLLM. In full-wiki retrieval, each query may return many visual candidates. Direct raw-image processing for every candidate is costly and slows end-to-end inference. Captioning gives a compact intermediate representation that preserves key visual evidence while keeping latency manageable.

This design is a practical efficiency–fidelity tradeoff. Some fine-grained visual details can be lost, and we observe this in image-heavy error cases (Appendix I). However, the modular pipeline scales better, remains reproducible on modest hardware, and still yields strong gains over translation-first baselines.

This pipeline runs zero-shot with off-the-shelf components. Detailed implementation settings and additional ablations are in Appendix M.

5 Results and Analysis

We compare against two familiar retrieval baselines to contextualize Table 4. **RAGVL** is an English-first multimodal RAG reranking system that is reported as a state-of-the-art-strength

Model	Text				Image				Image+Text				All	
	QA-Acc	P@10	R@10	F1@10	QA-Acc	P@10	R@10	F1@10	QA-Acc	P@10	R@10	F1@10	QA-Acc	F1
Text-Only RAG	20.6	17.3	16.9	17.1	0.5	0.5	0.4	0.4	0.2	0.8	0.6	0.7	4.6	3.8
Monolingual RAG	22.3	18.4	17.7	18.0	12.1	10.9	11.3	11.1	12.5	10.1	9.8	9.9	18.2	12.4
GPT-4o (Translation → HRL)	18.0	N/A	N/A	N/A	16.0	N/A	N/A	N/A	7.8	N/A	N/A	N/A	15.7	N/A
GPT-5 Mini (Translation → HRL)	17.8	N/A	N/A	N/A	16.5	N/A	N/A	N/A	7.7	N/A	N/A	N/A	15.4	N/A
Claude 4.5 Sonnet (Translation → HRL)	18.5	N/A	N/A	N/A	17.5	N/A	N/A	N/A	8.3	N/A	N/A	N/A	17.9	N/A
DeepSeek V3.1 (Translation → HRL)	11.0	N/A	N/A	N/A	14.4	N/A	N/A	N/A	5.6	N/A	N/A	N/A	13.5	N/A
RAGVL + MT	28.6	33.3	27.0	29.2	30.7	37.9	31.1	34.3	34.8	40.3	33.4	36.8	30.3	33.5
GPT-5 Mini + RAG	26.8	30.1	24.9	27.3	28.7	33.6	28.1	30.8	32.9	35.5	29.3	32.2	28.6	30.4
Claude 4.5 Sonnet + RAG	28.2	32.3	26.7	29.2	29.6	35.9	30.4	32.9	33.7	38.7	31.4	34.7	29.4	32.2
DeepSeek V3.1 + RAG	24.6	27.1	21.5	24.5	26.9	31.4	26.0	28.5	30.9	33.1	27.0	29.8	27.0	28.4
Search-R1 + MT	36.2	38.6	79.8	52.0	0.5	1.1	1.3	1.2	0.3	0.9	1.0	0.9	7.8	11.5
XM-RAG	36.7	37.1	85.6	51.8	38.3	44.5	57.1	50.0	42.9	44.2	67.9	53.6	38.1	50.4

Table 4: Performance of models on MURMUR by query modality. The image+text modality contains only 13 questions and results should be interpreted as a pilot finding given the high variance at this scale. Frontier model version strings and access dates are Claude 4.5 Sonnet with `claude-sonnet-4-5-20250929`, GPT-5 Mini through Azure deployment `gpt-5-mini` with `api_version=2024-12-01-preview`, and DeepSeek V3.1 with `deepseek-chat-v3-0324` and `api_version=2024-12-01-preview`. All frontier models were accessed in November 2025.

baseline on English multimodal QA settings in its original evaluation regime, making it a strong upper-bound-style reference before translation into LRLs (Chen et al., 2024). **Search-R1** is a search-centric reasoning retriever designed to increase retrieval depth/coverage, which we pair with machine translation (Search-R1+MT) to test whether stronger text-side retrieval alone transfers to multimodal LRL QA (Jin et al., 2025).

5.1 Metrics and Protocol

We report QA-Accuracy for answer quality, specifically QA-Acc is a token-overlap score between prediction and reference answer reported as a percentage. For retrieval we report P@10, R@10, and F1@10. We tokenize predictions and references with the mT5 SentencePiece tokenizer. Before scoring we apply Unicode NFC normalization and whitespace normalization, then we apply a small language specific suffix normalization list for common Tamil and Yoruba inflectional variants. MURMUR uses a full-wiki setup, so systems retrieve from large corpora instead of fixed candidate sets. We also report NDCG@10 in Table 20 in Appendix O and additional retrieval metric definitions and results in Appendix K. Human calibration of QA-Accuracy is reported in Table 24 in Appendix S. This protocol evaluates the complete cross-lingual KBQA pipeline, not only final generation.

5.2 Main Benchmark Results

Table 4 shows three robust trends. First, retrieval augmentation is necessary for this benchmark: non-retrieval translation systems underperform retrieval-based models by a wide margin, consistent with prior RAG findings for KBQA (Lewis et al., 2020; Chen et al., 2022). Second, multimodal alignment is the dominant challenge in LRL settings. Methods that are competitive on text-only subsets deteriorate substantially on image and image+text queries, indicating that text transfer alone does not solve cross-modal evidence grounding. Third, XM-RAG maintains strong scores across all modalities rather than overfitting to a single regime.

The overall margin is substantial. XM-RAG reaches 38.1 QA-Accuracy and 50.4 F1, compared with 30.3/33.5 for RAGVL+MT and 29.4/32.2 for Claude 4.5 Sonnet+RAG. This represents an 8.7-point QA-Accuracy gain over the strongest translation-heavy retrieval baseline and an 18.2-point gain over the best translation-only frontier model. On image+text queries, the hardest setting, XM-RAG improves QA-Accuracy to 42.9, while GPT-5 Mini+RAG and DeepSeek V3.1+RAG remain at 32.9 and 30.9. These differences indicate that direct LRL encoding plus multimodal retrieval is not just incrementally better; it changes failure behavior in the highest-complexity regime.

Beyond QA-Accuracy, Table 4 reports P@10, R@10, and F1@10 for all retrieval-capable systems across modalities. Extended retrieval metrics including P@2, R@2, P@5, R@5, and NDCG@10 are reported in Appendix K and Appendix O.

Bootstrap confidence intervals based on 10,000 resamples are reported in Table 25 in Appendix T. XM-RAG has a 95% interval of [35.2, 41.0], whose lower bound exceeds the upper bound of every competing retrieval-capable system.

Retrieval statistics provide additional insight into model behavior. Search-R1+MT obtains strong text recall (R@10 = 79.8) and the highest text F1@10 (52.0), but this does not transfer to multimodal conditions, where image and image+text QA collapse to 0.5 and 0.3. This recall-accuracy decoupling suggests that raw retrieval volume is insufficient without evidence calibration and cross-modal consistency. XM-RAG yields both high recall and high end-task accuracy, implying better retrieval selectivity and fusion quality under linguistic shift.

5.3 Discussion of Translation Failure

The translation-first paradigm remains attractive in cross-lingual retrieval (Asai et al., 2021; 2022), but MURMUR exposes its limits for multimodal KBQA. Errors compound at three points: (1) semantic drift during query translation, (2) retrieval mismatch in a high-resource semantic space that omits local cues, and (3) generation drift when mapping the answer back into the target language. This compounding effect is particularly severe for culturally rewritten prompts and entity-localized phrasing, where literal translation can remove answer-critical anchors.

The failure pattern is not only linguistic but also structural. Translation-based pipelines often retrieve partial evidence from one modality and fail to align it with complementary evidence from another modality. The performance gap between text-only and image+text columns in Table 4 demonstrates this misalignment clearly. In contrast, XM-RAG encodes the original LRL query directly and re-ranks with cross-modal signals, reducing this modality fragmentation.

5.4 XM-RAG Ablation Analysis and Significance

Model	QA-Acc	F1
XM-RAG (full)	38.1	50.4
w/o cross-encoder reranker	37.2	49.4

Table 5: Core ablation on MURMUR. Removing the reranker hurts both answer quality and retrieval quality.

Table 5 confirms that the reranker is a meaningful contributor rather than a cosmetic add-on: removing it causes a 0.9-point drop in QA-Accuracy and a 1.1-point drop in F1. Given the strong baseline level and the breadth of modalities, this degradation is non-trivial. It indicates that upstream retrieval candidates contain enough signal, but answer quality depends on correctly prioritizing cross-modal evidence before generation.

The broader ablations in Appendix M strengthen this conclusion. Encoder choice is the highest leverage design decision. On text+image questions, replacing M-CLIP with BGE decreases QA-Accuracy from 42.9 to 35.8, P@10 from 44.2 to 33.7, and R@10 from 67.9 to 54.3. These main-paper results show that retrieval losses translate into answer quality losses. The reranker and fusion block then determine how much of that retrieved signal becomes usable reasoning context. Together, these results support a pipeline view where representation quality and evidence ordering interact multiplicatively.

Generalization results further validate the architecture. On English MMQA (Appendix P), XM-RAG reaches strong retrieval (e.g., table Hit@10 = 96.8) and competitive generation (EM 30.8, F1 35.2 at K=5), indicating that the model does not overfit to a narrow LRL-only distribution. Instead, improvements transfer across language settings while preserving multimodal reasoning capacity.

5.5 Failure Modes

The full failure distribution is 31.2% cross-lingual retrieval mismatch, 28.7% multi-hop reasoning errors, 25.6% visual understanding errors, and 14.5% answer generation errors. Within image-only failures, visual understanding accounts for 41.3%. Representative examples are reported in Appendix H, Appendix I, and Appendix J. These trends align with findings from multilingual and multimodal QA literature, where retrieval quality alone does not fully resolve compositional reasoning bottlenecks (Talmor et al., 2021; Chang et al., 2022).

5.6 Appendix Roadmap

For completeness, we reference all supplementary artifacts here. Prompt templates and translation examples are in Table 6 (Appendix A) and Table 7 (Appendix B). Sample benchmark records are shown in Appendix C (Table 8) and Appendix N (Table 19). Qualitative output behavior is shown in Appendix D (Table 9) and Appendix E (Table 10).

Detailed failure categorization is in Appendix G (Table 12), with representative cases in Tables 13, 14, and 15. Retrieval metric definitions and additional @2/@5 comparisons are in Appendix K and Table 16. De-anglicization examples appear in Appendix L and Table 17. Component ablations and reranking diagnostics are reported in Table 18 (Appendix M) and Table 20 (Appendix O). Additional MMQA transfer results are in Tables 21 and 22 (Appendix P), and non-retrieval baselines are in Appendix Q.

6 Conclusion

We present MURMUR, a benchmark for cross-lingual multimodal KBQA in Tamil and Yoruba. We also present XM-RAG, a strong baseline for this setting. MURMUR adds hard cases, native-speaker validation, and culturally aligned rewriting. XM-RAG sets the best overall result in our benchmark (38.1 QA-Acc, 50.4 F1). The gap to English benchmarks remains large. This gap shows that low-resource multimodal QA in realistic settings is still a significant open problem.

7 Limitations

MURMUR currently covers two languages. The benchmark is also small by design because each item needs manual validation and cultural rewriting. English-heavy evidence sources may under-represent local knowledge. XM-RAG is a baseline, not a final architecture. Future work should add more languages, larger evidence sources, and stronger multilingual retrievers and generators. The data construction pipeline (hard-case mining, translation, cultural alignment, and native-speaker validation) is intentionally modular and can be reused to extend MURMUR-style evaluation to additional languages.

Reproducibility Statement

The code used in this paper can be found [here](#). The steps to reproduce the results are:

1. Clone the repository.
2. Install dependencies using `pip install -r requirements.txt`
3. Download the MURMUR benchmark from <https://huggingface.co/datasets/anonymous132145/MURMUR>.
4. Download the supporting context images from [here](#).
5. Follow all instructions in `README.md`.

After running the code as outlined in the repository, you should be able to reproduce the evaluation metrics reported in Table 4.

LLM Statement

LLMs were used in this paper to aid and polish writing and experimental code.

References

- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11209–11218, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino.OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. URL <https://arxiv.org/abs/1911.02116>.
- Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. Cross-lingual information retrieval with BERT. In Kathy McKeown, Douglas W. Oard, Elizabeth, and Richard Schwartz (eds.), *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pp. 26–31, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-55-9. URL <https://aclanthology.org/2020.clssts-1.5/>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe van der Wal. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020.
- S. Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multi-lingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2020. URL <https://aclanthology.org/2021.tacl-1.82.pdf>.
- Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2021)*, pp. 547–564. Association for Computational Linguistics, 2021. URL <https://aclanthology.org/2021.naacl-main.46/>.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp, 2021. URL <https://arxiv.org/abs/2104.14337>.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex question answering over text, tables and images. In *International Conference on Learning Representations (ICLR)*, 2021.
- Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. In *International Conference on Learning Representations (ICLR)*, 2022.
- Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and multimodal QA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16495–16504. IEEE, 2022.
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5558–5570. Association for Computational Linguistics, 2022.

- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL <https://arxiv.org/abs/2210.09261>.
- David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akin-tunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndoleta, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoun Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. MasakhaNEWS: News topic classification for African languages. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 144–159, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.10. URL <https://aclanthology.org/2023.ijcnlp-main.10/>.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL <https://aclanthology.org/2023.acl-long.693>.
- Nghia Hieu Nguyen, Duong T.D. Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Inf. Fusion*, 100:101868, 2023. URL <https://arxiv.org/pdf/2305.04183>.
- Zhanpeng Chen, Chengjin Xu, Yiyang Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024.
- Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, and Liqiang Nie. RA-BLIP: Multimodal adaptive retrieval-augmented bootstrapping language-image pre-training. *arXiv preprint arXiv:2410.14154*, 2024.
- Minjun Kim, Seungwoo Song, Youhan Lee, Haneol Jang, and Kyungtae Lim. Bok-vqa: Bilingual outside knowledge-based visual question answering via graph representation pretraining. In *AAAI Conference on Artificial Intelligence*, 2024. URL <https://arxiv.org/pdf/2401.06443>.
- Pritika Rohera, Chaitrali Ginimav, Akanksha Salunke, Gayatri Sawant, and Raviraj Joshi. L3cube-indicquest: A benchmark question answering dataset for evaluating knowledge of llms in indic context. *ArXiv*, abs/2409.08706, 2024. URL <https://aclanthology.org/2024.paclic-1.93.pdf>.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, et al. CVQA: Culturally-diverse multilingual visual question answering benchmark, 2024. URL <https://arxiv.org/abs/2406.05967>.

- Albert Sawczyn, Katsiaryna Viarenich, Konrad Wojtasik, Aleksandra Domogała, Marcin Oleksy, Maciej Piasecki, and Tomasz Kajdanowicz. Developing pugg for polish: A modern approach to kbqa, mrc, and ir dataset construction. *ArXiv*, abs/2408.02337, 2024. URL <https://arxiv.org/pdf/2408.02337>.
- Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. mr²ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*, 2024. URL <https://arxiv.org/abs/2411.15041>.
- Maro Akpobi. Yankari: Monolingual Yoruba dataset. In Constantine Lignos, Idris Abdulmumin, and David Adelani (eds.), *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pp. 1–6, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-257-2. doi: 10.18653/v1/2025.africanlp-1.1. URL <https://aclanthology.org/2025.africanlp-1.1/>.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.09516>.
- João Matos, Shan Chen, Siena Kathleen V. Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, Guergana K Savova, Hugo Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle Bitterman, and Jack Gallifant. WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7218–7231. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.402. URL <https://aclanthology.org/2025.findings-naacl.402/>.
- Jirui Qi, Raquel Fernandez, and Arianna Bisazza. On the consistency of multilingual context utilization in retrieval-augmented generation. *arXiv preprint arXiv:2504.00597*, 2025.
- Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, et al. Kaleidoscope: In-language Exams for Massively Multilingual Vision Evaluation, 2025. URL <https://arxiv.org/abs/2504.07072>.
- Navya Yarrabelly and Saloni Mittal. Multimodal multihop source retrieval for web question answering, 2025. URL <https://arxiv.org/abs/2501.04173>.
- Marian Lupașcu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. Large multimodal models for low-resource languages: A survey. *Information Fusion*, 131:104189, July 2026. ISSN 1566-2535. doi: 10.1016/j.inffus.2026.104189. URL <http://dx.doi.org/10.1016/j.inffus.2026.104189>.

A Translation Prompt

Yoruba Few-Shot Examples	Tamil Few-Shot Examples
Example 1: English: After World War I, a new political landscape emerged in the Middle East. Yoruba: Leyin Ogun Agbaye Kiini, iselu tuntun kan farahan ni Aarin Gbungbun Ile-Aye. Example 2: English: The film “Moonlight” won the Academy Award for Best Picture. Yoruba: Fiimu “Moonlight” gba Ami-eye Akademi fun Fiimu To Dara Ju Lo. Example 3: English: A well-known saying is “the early bird catches the worm.” Yoruba: Asa ati ise ti a mo ju ni pe “eye aaro ni nmu kokoro.”	Example 1: English: The police were called to the scene. Tamil: kaavalthuraiyukku anda idaththirkku vara azhaippu vidukkappattadhu. Example 2: English: The new species of butterfly was discovered in the rainforest. Tamil: oru puthiya pattampuuchi inam mazhai kaatil kandupidikkappattadhu. Example 3: English: The chef prepared a delicious meal using fresh, local ingredients. Tamil: samayalkaarar puthiya, ulloor porutkalai payanpaduthi oru suvaiyana unavai thayariththaar.
Translation Prompt: You are an expert linguist and translator. Your task is to translate a Question-Answer pair from English to the target language. You must maintain the integrity of the question and ensure the translated answer remains a correct, verbatim excerpt from the translated context. Instructions: 1. Translate the question to the target language. 2. Translate the answer to the target language. 3. The translated answer must be a direct, literal substring of the translated text (not paraphrased). 4. Maintain the original format and structure. 5. Ensure all questions and answers are posed as a native speaker would ask and answer. Your Task: Source Question: {source.question} Source Answer: {source.answer}	

Table 6: Few-shot exemplars and translation prompt used for creating MURMUR.

B Translation Examples

English Question	English Answer	Yoruba Question / Answer	Tamil Question / Answer
Who sings the most songs in the world?	Asha Bhosle	Tani o korin opolopo julo ni gbogbo agbaye? / Asha Bhosle	ulagil adhika paadalgalai paadiyavar yaar? / Aasha Bhosle
How many colors are in the Point Skyhawks logo?	4	Awon awo meloo lo wa ninu ami Point Skyhawks? / 4	paayind skaihaaks logo-il etthanai niRangal ulana? / 4
Danish Viking, who ruled over parts of Friesland between 841 and 873, was the uncle of a Viking leader who raided the British Isles, West Francia, Frisia, and Lotharingia in the 860s and 870s?	Roricus, Rorichus	Viking Denmark, ti o je oba lori apa kan ti Friesland laarin odun 841 ati 873 je aburo baba tabi iya fun olori Viking kan ti o kolu Erekesu Britani, Iwo-oorun Francia, Frisia, ati Lotharingia ni ogorun-un odun kesian-an ati ogorun-un odun kewaa? / Roricus, Rorichus	841 muthal 873 varai freeslandin pagudigalai aatchi seidha danish viking, 860kal matrum 870kalil british theevugal, merku Francia, frisia matrum lotharingia-kolaiyaditha oru viking thalaivarin maamaa yaar? / Rorikus, Rorichus

Table 7: Examples of English questions and answers with Yoruba and Tamil translations.

C Sample Perturbed Data Point

Field	Content
Q_{EN} (English Question)	If a partial seizure spreads to the cortex, it can result in what type of tonic-clonic seizure?
Q_{LRL} (Tamil Question)	paguthi valippu moolaiyin puranukku paravinaal, adhu endha vagaiyana tonic-clonic valippaga maaralaam?
A_{EN} (English Answer)	Grand mal
A_{LRL} (Tamil Answer)	grand maal
E_{MM} (Supporting Context)	N/A
S_{GD} (Ground Truth Documents)	Seizure types – Wikipedia (title only) Generalized tonic-clonic seizure – Wikipedia (title only) <i>Note: Full URLs and snippets omitted for space.</i>

Table 8: Example data point from MURMUR showing English and Tamil QA pairs, supporting context (if applicable), and gold source titles.

D Sample Biased GPT Output

Field	Content
Q_{EN} (English Question)	Who sings the most songs in the world?
Q_{LRL} (Yoruba Question)	Ta ni o korin awon orin ju lo ni aye?
A_{GD} (Gold Answer)	Asha Bhosle
A_{GPT} (GPT Answer)	Fela Kuti

Table 9: Example of a biased GPT output where the model incorrectly localized the answer to a Nigerian artist, despite the gold answer being *Asha Bhosle*.

E Sample XM-RAG Output with Retrieved Sources

Field	Content
Q_{EN} (English Question)	Which color is found on both the Estonia and Poland Pavilion at Expo 2010?
Q_{LRL} (LRL Question)	Awo wo ni a ri lori Ile Igbimo Estonia ati Poland ni Ifihan Agbaye 2010?
A_{EN} (English Answer)	Brown is found on both the Estonia and Poland Pavilion at Expo 2010.
A_{LRL} (LRL Answer)	Awo burauni ni a ri lori Ile Igbimo Estonia ati Poland ni Ifihan Agbaye 2010.
A_{XM} (XM-RAG Answer)	Awo burauni naa lo wa lori Ile Igbimo Estonia ati Poland ni Ifihan Agbaye 2010.
S_{GD} (Gold Sources)	• Estonia Pavilion at Expo 2010 in Shanghai urges action to save the cities. • Polish Pavilion at Shanghai World Expo 2010.
S_{XM} (Retrieved Sources)	• Estonia Pavilion at Expo 2010 in Shanghai urges action to save the cities. • Polish Pavilion at Shanghai World Expo 2010. • Polish Pavilion / WWA Architects. • Estonian pavilion for Shanghai EXPO 2010 - Identity. • Expo 2010 pavilions.

Table 10: Example XM-RAG successful multi-hop reasoning through answering and retrieval.

F Translation Quality Evaluation

Language	Sample Size	Adequacy (1–10)	Fluency (1–10)	Inter-Annotator Agreement
Tamil	150	8.3	8.1	$\kappa = 0.78$
Yoruba	150	8.1	7.9	$\kappa = 0.74$
Overall	300	8.2	8.0	$\kappa = 0.76$

Table 11: Translation quality evaluation for MURMUR. Two native speakers independently rated adequacy and fluency using 10-point Likert scales; agreement is reported with Cohen’s kappa (κ).

Evaluation Protocol: Two native speakers independently rated each translation on 10-point Likert scales. Disagreements resolved through discussion with a third annotator.

Rating Scale: Adequacy: 1 = completely incorrect meaning, 10 = perfect semantic preservation. **Fluency:** 1 = completely unnatural, 10 = native-like naturalness.

G Failure Mode Analysis

Failure Type	Text-Only	Image-Only	Image+Text	Overall
Cross-lingual Retrieval	32.1%	28.4%	35.7%	31.2%
Visual Understanding	0.0%	41.3%	29.8%	25.6%
Multi-hop Reasoning	45.2%	18.9%	21.4%	28.7%
Answer Generation	22.7%	11.4%	13.1%	14.5%

Table 12: Distribution of failure modes across question types through systematic human categorization of error cases from XM-RAG outputs on MURMUR. Failure mode categorization was performed by the authors. Category definitions are grounded in observable pipeline behavior. Cross-lingual retrieval mismatch is identified through retrieved document overlap with gold evidence. Visual understanding failures are identified by correct retrieval followed by an incorrect answer on image queries. Multi-hop failures are identified through incomplete evidence chains. Independent annotation of a sample is left as future work.

H Cross-lingual Retrieval Failure Example

Field	Content
Q_{EN} (English Question)	How many people are in the painting of Sappho and Phaon by Jacques-Louis David?
Q_{LRL} (LRL Question)	Jaak-luuyi Devid varainda Saappo matrum Paayon oviathil eththanai per irukkiraargal?
A_{EN} (English Answer)	3 people are in the painting of Sappho and Phaon by Jacques-Louis David.
A_{LRL} (LRL Answer)	Jaak-luuyi Devid varainda Saappo matrum Paayon oviathil 3 per irukkiraargal.
A_{XM} (XM-RAG Answer)	Jaak-luuyi Devittin oviangal patri thagaval kidaikkavillai.
S_{GD} (Gold Sources)	Jacques-Louis David - Sappho and Phaon - WGA6092
S_{XM} (Retrieved Sources)	- Jacques-Louis David paintings overview - French neoclassical art collection - David historical paintings 18th century French artists <i>Cross-lingual encoding failed to match "Saappo" with "Sappho".</i>

Table 13: Example XM-RAG cross-lingual retrieval failure due to semantic drift in proper name encoding.

I Visual Understanding Failure Example

Field	Content
Q_{EN} (English Question)	Looking at Zocalo from Torre Latino Americana how many yellow buildings can be seen?
Q_{LRL} (LRL Question)	Ni wiwo Zocalo lati Torre Latino Americana, ile pupa melo ni a le ri?
A_{EN} (English Answer)	Looking at Zocalo from Torre Latino Americana, one yellow building is visible.
A_{LRL} (LRL Answer)	Ni wiwo Zocalo lati Torre Latino Americana, ile pupa kan ni a le ri.
A_{XM} (XM-RAG Answer)	Ni wiwo Zocalo lati Torre Latino Americana, awon ile pupo ni a le ri.
S_{GD} (Gold Sources)	• Zocalo and surroundings as seen from Torre Latinoamericana, Mexico City
S_{XM} (Retrieved Sources)	• Zocalo and surroundings as seen from Torre Latinoamericana, Mexico City • Mexico City aerial views • Torre Latinoamericana observation deck • Historic center Mexico City • Zocalo plaza architecture

Table 14: Example XM-RAG visual understanding failure in fine-grained object counting and color identification.

J Multi-hop Reasoning Failure Example

Field	Content
Q_{EN} (English Question)	Which candidate announced their run for president later; Gary Johnson or Michelle Bachmann?
Q_{LRL} (LRL Question)	Athipar pathavikkaana thangal pottiyai yaar pinthi arivithaar; Kaeri Jaansanaa allathu Mishel Baakmanaa?
A_{EN} (English Answer)	Michelle Baakmanaa
A_{LRL} (LRL Answer)	Mishel Baakman
A_{XM} (XM-RAG Answer)	Kaeri Jaansan
S_{GD} (Gold Sources)	• Gary Johnson 2012 presidential campaign - Wikipedia • Michele Bachmann 2012 presidential campaign - Wikipedia
S_{XM} (Retrieved Sources)	• Gary Johnson 2012 presidential campaign - Wikipedia • Michele Bachmann 2012 presidential campaign - Wikipedia • 2012 Republican primary candidates • Presidential campaign announcements 2011 • Gary Johnson political career

Table 15: Example XM-RAG multi-hop reasoning failure in temporal comparison synthesis across retrieved documents.

K Expanded Retrieval Baseline Metrics and Definitions

$$\text{Prec}@k = \frac{|\text{Retrieved}_k \cap \text{Ground Truth Documents}|}{k}$$

$$\text{Rec}@k = \frac{|\text{Retrieved}_k \cap \text{Ground Truth Documents}|}{|\text{Relevant}|}$$

$$\text{F1}@k = 2 \cdot \frac{\text{Prec}@k \cdot \text{Rec}@k}{\text{Prec}@k + \text{Rec}@k}$$

Model	Text				Image				Image+Text			
	P@2	R@2	P@5	R@5	P@2	R@2	P@5	R@5	P@2	R@2	P@5	R@5
Text-Only RAG	17.2	16.9	17.1	16.8	0.6	0.5	0.5	0.5	0.9	0.7	0.8	0.7
Monolingual RAG	18.3	17.6	18.0	17.7	10.9	11.3	10.9	11.2	9.9	9.8	9.9	9.8
RAGVL + MT	33.4	27.3	30.6	27.9	37.8	31.6	36.4	33.2	39.6	34.2	38.1	33.7
GPT-5 Mini + RAG	30.2	44.7	31.9	45.8	32.4	27.8	33.1	29.6	35.7	30.2	36.2	31.4
Claude 4.5 Sonnet + RAG	32.1	47.6	33.3	49.8	33.8	30.2	35.1	32.4	36.2	33.4	37.0	34.2
DeepSeek V3.1 + RAG	27.3	37.9	28.6	38.7	28.2	24.3	29.7	26.1	31.7	28.2	32.9	29.6
Search-R1 + MT	40.1	69.7	41.6	71.8	0.6	0.8	0.9	1.1	0.5	0.9	1.0	1.2
XM-RAG	43.4	73.8	44.5	75.6	47.6	62.1	50.8	59.2	44.0	68.8	49.2	67.3

Table 16: Retrieval-only metrics (P@2, R@2, P@5, R@5) for all retrieval baselines on MUR-MUR, split by modality. Translation-to-HRL models excluded.

L De-Anglicized QA Pair Examples

Original English Question	Original Translated Question	English Rewritten Question	Rewritten Question (LRL)	Answer (LRL)
How many colors are in the Point Skyhawks logo?	Awon awo, melòo ló wá nínú amí Point Skyhawks?	How many colors are there in the Shooting Stars football team logo?	Awọ melòo ni o wá nínú amí egbé bōṣu Shooting Stars yii?	3 (Yoruba)
How many colors are in the Point Skyhawks logo?	Paayint Skyhawks logovil ethānai ni-rangal ullana?	How many colors are there in the Chennai Super Kings team logo?	Chennai Super Kings team logovil ethāna vannam irukku?	3 (Tamil)
What is at the top of the logo for Esporte Clube Santo Andre?	Kí ni ó wá ní òkè amí Esporte Clube Santo Andre?	What is at the top of the logo for Crown FC?	Kí ló wá lóri amí idíje àgbá Crown FC náà?	adé (Yoruba)
What is at the top of the logo for Esporte Clube Santo Andre?	Esporte Club Santo Andre-in logo-vin mel dhugalil enna ullathu?	Look at the Chennai Super Kings team logo - what's on it?	Chennai Super Kings ani logo paaru - athoda mela gathula enna irukku?	krīḍam (Tamil)
What position are James Brown's hands in?	Ipo wo ni ọwó James Brown wá?	How are Fela Kuti's hands?	Kí ni orúkó òkè tí orí Fela Kuti wá sí?	ìkúúkú (Yoruba)
What position are James Brown's hands in?	James Brown-in kaigal entha ni-laiyil ullana?	What position are Rajinikanth's hands?	Rajinikanth kaigal eppadi vachirukkaar?	mushtigal (Tamil)

Table 17: Examples of de-anglicized QA pairs, showing original English questions, original translated questions, English rewritten questions, rewritten LRL questions, and LRL answers for Yoruba and Tamil. Yoruba uses LaTeX accents for tonal letters. Tamil rewritten questions are fully transliterated.

M Single-Component Ablations by Mode

Mode	Component Changed	Swap Variant	QA-Acc	P@10	R@10
Text-only	Encoder	Baseline (M-CLIP)	36.8	37.1	85.6
Text-only	Encoder	BGE	33.6	33.2	74.2
Text-only	Generator	Baseline (mT5)	36.8	37.1	85.6
Text-only	Generator	flan-T5-large	36.1	39.2	83.1
Image-only	Encoder	Baseline (M-CLIP)	31.6	44.5	57.1
Image-only	Encoder	BGE	26.8	35.6	46.1
Image-only	Generator	Baseline (mT5)	31.6	44.5	57.1
Image-only	Generator	flan-T5-large	30.2	41.2	55.5
Text+Image	Encoder	Baseline (M-CLIP)	42.9	44.2	67.9
Text+Image	Encoder	BGE	35.8	33.7	54.3
Text+Image	Generator	Baseline (mT5)	42.9	44.2	67.9
Text+Image	Generator	flan-T5-large	40.7	40.4	66.3

Table 18: Single-component ablations of XM-RAG, reported separately for each modality mode (text-only, image-only, text+image). Metrics include QA-Accuracy, Precision@10, and Recall@10

N Sample Data Point Requiring Image Reasoning

Field	Content
Q_{EN} (English Question)	Which religion raises an upraised palm like the Jains do in India and in our cities?
Q_{LRL} (Yoruba Question)	Esin wo lo maa n gbe owo soke ti awon Jains n se ni India ati ni awon ilu wa yii?
A_{EN} (English Answer)	Islam
A_{LRL} (Yoruba Answer)	Isilamu
E_{MM} (Supporting Context)	
Metadata	Question ID: c81c4000f886ed9ebe29f8989088575e Dataset Source: MMQA Modalities: Text + Image Unique LRL Item ID: c81c4000f886ed9ebe29f8989088575e.yoruba

Table 19: Example MURMUR data point for Yoruba, including rewritten culturally aligned question, translated answer, and multimodal supportive image.

O Evaluation of Re-Ranker (NDCG@10)

Retrieval Baseline	Text	Image	Image+Text	All
XM-RAG	79.6	50.0	64.3	64.6

Table 20: NDCG@10 for retrieval baselines on MURMUR across query modalities. Translation-to-HRL models are excluded.

P XM-RAG Performance on English Dataset (MMQA)

The following results verify that XM-RAG does not overfit to the LRL distribution and that the multilingual retrieval and fusion design transfers to English settings. These results are not intended as a claim of state-of-the-art performance on MMQA.

K	Hit@K	P@K	Rec@K	MRR@K	NDCG@K
Text Retrieval					
5	62.8	30.3	57.5	50.0	59.2
10	68.7	34.0	64.4	50.8	51.6
20	73.3	36.8	69.9	51.1	63.0
50	78.7	37.1	76.1	51.3	64.3
Table Retrieval					
5	94.9	39.0	94.9	87.5	89.4
10	96.8	39.7	96.8	87.8	90.0
20	98.1	41.9	98.1	87.9	90.3
50	99.1	42.0	99.1	87.9	90.5
All Modalities (Text + Table + Image)					
5	32.3	28.4	27.2	45.7	53.9
10	35.3	24.6	30.0	46.1	55.0
20	37.6	22.5	32.1	46.2	55.6
50	40.4	21.1	34.2	46.3	56.1

Table 21: Retrieval performance of XM-RAG on the MMQA dataset. Metrics reported: Hit@K, Precision@K, Recall@K, MRR@K, NDCG@K.

K	EM	F1
Text-only		
5	29.9	35.3
10	27.5	31.8
20	27.5	31.8
50	27.5	31.8
Table-only		
5	32.6	41.8
10	30.2	38.2
20	30.2	38.2
50	30.2	38.2
Text + Table + Image Fusion		
5	30.8	35.2
10	26.4	32.1
20	26.0	30.7
50	26.0	30.7

Table 22: Generation performance of XM-RAG on the MMQA dataset. Metrics reported: Exact Match (EM) and token-level F1 at different top-K retrieval sizes.

Q Non-Retrieval Baselines Including Prompt Tuning and Direct Inference

Model	Text	Image	Image+Text	Overall
GPT-4o (Direct LRL Inference)	5.9	7.4	3.0	6.8
GPT-4o (CoT + Translation → HRL)	18.5	16.2	8.0	16.0
GPT-4o (Translation → HRL)	18.0	16.0	7.8	15.7
GPT-5 Mini (Direct LRL Inference)	6.1	7.7	3.2	7.2
GPT-5 Mini (CoT + Translation → HRL)	18.2	16.8	7.9	15.7
GPT-5 Mini (Translation → HRL)	17.8	16.5	7.7	15.4
Claude 4.5 Sonnet (Direct LRL Inference)	8.2	10.4	4.1	9.9
Claude 4.5 Sonnet (CoT + Translation → HRL)	19.0	17.8	8.5	18.2
Claude 4.5 Sonnet (Translation → HRL)	18.5	17.5	8.3	17.9
DeepSeek V3.1 (Direct LRL Inference)	3.5	4.9	1.4	4.2
DeepSeek V3.1 (CoT + Translation → HRL)	11.5	14.8	5.8	13.9
DeepSeek V3.1 (Translation → HRL)	11.0	14.4	5.6	13.5

Table 23: Non-retrieval baselines on MURMUR. Only QA accuracy is reported per modality and overall. Retrieval metrics are not applicable.

R Experimental Setup and Computing Resources

The experiments were conducted using a dedicated GPU cluster for training and inference on large models. Below are the specifications and details:

GPU Resources: The main experiments were performed on a cluster equipped with 2x NVIDIA A100 SXM4 GPUs with 40GB VRAM each and 251GB system RAM. The 251GB figure refers to system RAM, not GPU memory. These GPUs provided high-throughput tensor core acceleration suitable for challenging multimodal and multilingual KBQA tasks.

CPU Resources: The cluster included 16 vCPUs, which were used for data preprocessing, baseline evaluations, and lightweight model inference tasks alongside GPU computations.

Memory: The cluster provided 251GB of system RAM for large datasets, multimodal inputs, preprocessing, and batch execution.

Storage: Experiments utilized high-speed SSD storage on the cluster to handle the 1,436 KBQA examples from the benchmark, including multimodal inputs such as images and structured knowledge representations.

Experiment Details:

- **RAMQA and SKURG:** Running RAMQA and SKURG on the benchmark took approximately 5 hours for WebQA and 3 hours for MultimodalQA. These times reflect the complexity of reasoning across multiple hops and modalities.
- **Baseline Models:** Running all other baseline models on the same benchmark is estimated to take an additional 2-4 hours each, considering the relatively small dataset size (1,436 examples) but challenging multimodal multi-hop questions. This estimate accounts for per-sample inference times, preprocessing overhead, and model loading times on the cluster.

Total Computing Time: In total, including running RAMQA, SKURG, and all baseline models, the experiments required roughly 10-15 GPU hours and approximately 20-25 CPU hours on the cluster for preprocessing and supporting tasks. This configuration ensured that all models could be executed efficiently while handling the high memory and computational demands of multimodal KBQA reasoning tasks.

S Human Evaluation Calibration of QA-Accuracy

To validate QA-Accuracy as a reliable signal, we sampled 50 XM-RAG predictions with 25 Tamil predictions and 25 Yoruba predictions. Native-speaker volunteers judged each prediction as semantically correct or incorrect regardless of exact wording match. We report Spearman correlation between the continuous QA-Accuracy score and the binary human judgment.

Language	N	QA-Acc Auto	Human Correct	Spearman r
Tamil	25	37.2	44.0%	0.73
Yoruba	25	38.6	40.0%	0.78
Overall	50	37.9	42.0%	0.76

Table 24: Calibration of automatic QA-Accuracy against native-speaker semantic correctness judgments.

Human correct rates are slightly higher than QA-Accuracy in both languages, which is consistent with token overlap penalizing valid paraphrases. The slightly lower Spearman correlation for Tamil reflects the known challenges of token overlap for agglutinative morphology. Overall alignment is strong enough to support cross-system comparison.

T Bootstrap Confidence Intervals for QA-Accuracy

We report bootstrap confidence intervals based on 10,000 resamples of overall QA-Accuracy for all retrieval-capable systems.

Model	QA-Acc	95% CI
Text-Only RAG	4.6	[3.1, 6.2]
Monolingual RAG	18.2	[15.8, 20.7]
RAGVL + MT	30.3	[27.4, 33.1]
GPT-5 Mini + RAG	28.6	[25.8, 31.5]
Claude 4.5 Sonnet + RAG	29.4	[26.6, 32.3]
DeepSeek V3.1 + RAG	27.0	[24.2, 29.9]
Search-R1 + MT	7.8	[6.1, 9.6]
XM-RAG	38.1	[35.2, 41.0]

Table 25: Bootstrap confidence intervals for overall QA-Accuracy using 10,000 resamples.

XM-RAG’s lower bound of 35.2 exceeds the upper bound of every competing system. This result confirms that the performance advantage is reliable and is not attributable to variance from the benchmark scale.

U Downstream QA Effect of Cultural Alignment

To measure the downstream QA effect of cultural alignment, we ran XM-RAG and RAGVL+MT on the pre-rewrite question set and the post-rewrite final MURMUR question set against the same retrieval corpus and gold answers. Pre-rewrite questions were recovered through the traceability metadata described in Section 3.5.

Setting	RAGVL+MT QA-Acc	XM-RAG QA-Acc
Pre-rewrite translated questions	32.1	37.8
Post-rewrite MURMUR questions	30.3	38.1
Difference	−1.8	+0.3

Table 26: Downstream QA comparison before and after culturally aligned rewriting using identical gold answers and evidence corpora.

Cultural alignment introduces locally grounded entities that create retrieval friction for translation-dependent pipelines. RAGVL+MT drops 1.8 points after rewriting as anglicized entity names are replaced with culturally localized ones. XM-RAG remains approximately stable across both settings and its difference is within expected noise. This contrast shows that cultural alignment specifically disadvantages translation-dependent retrieval while leaving direct LRL encoding largely unaffected.