M4-RAG: A Massive-Scale Multilingual Multi-Cultural Multimodal RAG

Supplementary Material

7. Languages

Table 5 provides a comprehensive breakdown of the languages included in M4-RAG, which ensure rigorous evaluation of cross-lingual generalization. The languages include diverse language families (such as Indo-European, Sino-Tibetan, Afro-Asiatic, Austronesian, Japonic, Koreanic, Niger-Congo, Turkic, and Uralic) and varying resource levels.

We categorize languages based on the taxonomy proposed by Joshi et al. [31], ranging from Class 0 to Class 5. This allows us to analyze how RAG performance correlates with the language vitality. Notably, our benchmark includes significant coverage of low-resource languages (Classes 0–2) such as Oromo, Tigrinya, Sundanese, and Sinhala, which are often underrepresented in standard VQA benchmarks.

Unlike previous benchmarks that treat languages as monoliths, it M4-RAG explicitly annotates regional dialects (e.g., Spanish across Spain, Argentina, Chile, Colombia, Ecuador, Mexico, and Uruguay) and social registers (e.g., formal vs. casual speech in Javanese, Korean, and Indonesian). This granularity is crucial for assessing cultural alignment, as the correct retrieval of cultural context often depends on recognizing dialect-specific nuances in the query.

8. Human Evaluation

8.1. Human Verification on Generated Captions as Oracle Context for CVQA

We use generated captions as oracle context because CVQA does not provide ground-truth evidence passages. The caption serves as a proxy for an upper bound on RAG performance. Note that the CVQA answers themselves are already human-annotated, and we simply generate the caption based on the images, questions, and the human-annotated answers themselves.

To further verify this, we conducted a human verification study by recruiting four annotators who evaluated 200 randomly sampled image-caption pairs on a 1–5 Likert scale for how well the caption supported answering the corresponding question. All samples received a score of 5 with full inter-annotator agreement, which is expected given the setup we have. For example, for a CVQA question asking "What part of the flag reflects the historical period?" given an image of the Romanian flag, the oracle context explicitly describes the central emblem as reflecting the historical period.

8.2. Human Verification on VLM-as-a-judge

We conduct a human validation study to examine whether our VLM-as-a-judge shows consistency with human scoring. Each of the `mmE5` and `B3` embedding models was evaluated on 100 samples, annotated by five human raters, and analyzed using five reliability metrics: Fleiss' κ, Gwet's AC2, Krippendorff's α, Conger's κ, and Brennan-Prediger's coefficient.

Metric	mmE5	B3	All
Fleiss' κ	0.6573	0.4273	0.5488
Gwet's AC2	0.7225	0.5013	0.6179
Krippendorff's α	0.6588	0.4300	0.5498
Conger's κ	0.6591	0.4432	0.5544
Brennan-Prediger	0.7115	0.4881	0.6059

Table 2. Inter-rater agreement between human and model scores using different reliability metrics.

Overall, the results indicate strong agreement for the `mmE5` model and moderate agreement for the `B3` model. This suggests that the lower retrieval performance observed for `B3` may stem from an understanding mismatch, where specific chunks receive higher localized scores despite inconsistent overall perception.

9. Detailed Results

In this section, we provide a granular analysis of the performance metrics reported in Tables 4 and 3. We focus on the interaction between model scale, retrieval modality, and the upper bounds established by oracle contexts.

9.1. Inverse Scaling of Retrieval Benefits

A central finding in our experiments is the inverse correlation between model parameter count and the relative performance gain provided by RAG.

Small Models (<14B). As shown in Table 3, smaller models exhibit substantial gains from multimodal retrieval. For instance, on the CVQA benchmark, `Gemma3 4B` improves from a baseline of 59.22% to 64.96% (+5.74%) when using `mmE5` retrieval. Similarly, `Qwen2.5-VL 3B` sees an improvement of +7.34%. This suggests that smaller models, which lack extensive parametric knowledge, rely heavily on retrieved context to ground their answers.

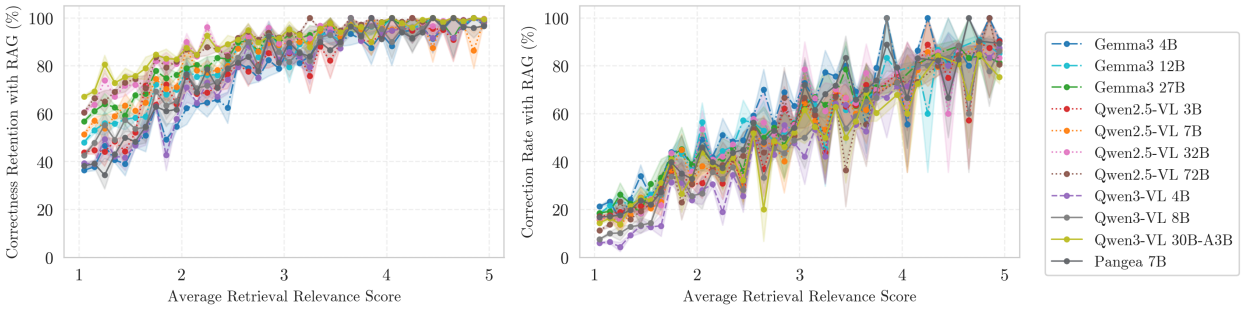


Figure 7. The effect of retrieval quality on RAG performance for various models on the WORLDCUISINES dataset, using mmE5 for multimodal retrieval. **Left:** The "Correctness Retention" rate measures the percentage of responses that were correct without RAG and remained correct with RAG. **Right:** The "Correction Rate" measures the percentage of responses that were incorrect without RAG but were successfully corrected by RAG.

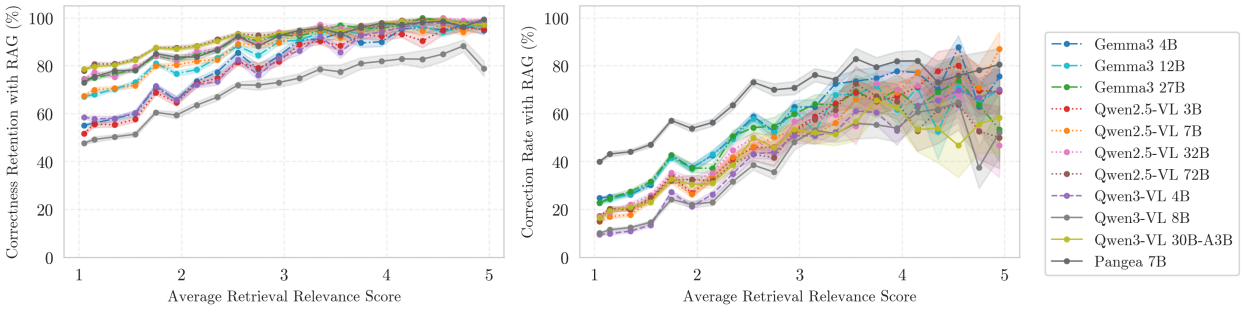


Figure 8. The effect of retrieval quality on RAG performance for various models on the WORLDCUISINES dataset, using B3 for multimodal retrieval. **Left:** The "Correctness Retention" rate measures the percentage of responses that were correct without RAG and remained correct with RAG. **Right:** The "Correction Rate" measures the percentage of responses that were incorrect without RAG but were successfully corrected by RAG.

Large Models (>14B). Conversely, larger models show diminishing returns or performance degradation. `Gemma3 27B` on CVQA regresses from 74.34% (Baseline) to 72.59% with `mmE5` RAG. `Qwen2.5-VL 72B` exhibits a similar pattern. This implies that for large models, imperfect retrieval acts as a distractor rather than an aid; the model's internal parametric knowledge is often more accurate than the noisy context retrieved.

9.2. Oracle-RAG Performance Gap

Figures 9a and 9b illustrate the substantial performance gap between providing oracle context and retrieval-augmented generation across both benchmarks. On CVQA (Figure 9a), oracle context consistently achieves 94–99% accuracy across all models, establishing a clear upper bound. In contrast, even the best RAG configurations using multimodal retrieval (`mmE5` or Oracle-Query RAG) achieve only 64–74% accuracy for the largest models, revealing a gap of 20–30%. This disparity is even more pronounced on WORLDCUISINES (Figure 9b), where oracle performance reaches 74–80%, while RAG variants plateau at 62–68%. The caption-based RAG approach consistently underperforms, often falling below the baseline. Notably, the gap between oracle and RAG widens as model size increases, indicating that while larger models can effectively leverage perfect context, they struggle to extract useful information from imperfect retrieval. This underscores that current retrieval systems are far from providing the quality of evidence that VLMs can utilize, hence pointing to retrieval quality as the primary bottleneck in multilingual multimodal RAG pipelines.

9.3. Language-Wise Performance Analysis

To further investigate the multilingual capabilities of current VLMs, Figures 11 (WORLDCUISINES) and 10 (CVQA) break down the performance impact of language choice on instructions. These figures represent the performance change when switching from English instructions to the target language. We observe similar patterns across both benchmarks. While high-resource languages like Chinese, Spanish, and French maintain relatively same performance, low-resource languages such as Amharic, Telugu, and Oromo suffer significant degradation, often dropping by over 5–10%. This confirms an inherent bias in cur-

Model	No RAG		Oracle Context		RAG			
	Baseline	+ Multilingual Prompt	Eng.	Multilingual	Eng. Cap.	Oracle Eng.	mmE5	B3
Gemma3 4B	59.22	59.32	95.01	94.50	53.16	82.02	64.96	56.71
Gemma3 12B	69.43	69.43	98.09	97.31	61.50	85.33	69.99	63.05
Gemma3 27B	74.34	73.89	98.61	92.13	66.04	86.86	72.59	68.03
Qwen2.5-VL 3B	56.29	55.09	93.97	91.59	52.63	79.68	63.63	52.85
Qwen2.5-VL 7B	62.26	61.47	95.32	93.46	59.26	82.17	67.05	59.04
Qwen2.5-VL 32B	68.75	65.37	97.14	92.12	65.44	85.88	71.72	65.49
Qwen2.5-VL 72B	73.51	71.19	97.48	94.52	68.38	86.23	72.03	68.73
Qwen3-VL 4B Think	58.48	57.88	94.65	93.94	50.95	78.97	62.00	53.28
Qwen3-VL 8B Think	64.10	63.54	96.25	95.36	55.95	82.10	66.21	58.33
Qwen3-VL 30B A3B Think	72.34	72.35	97.51	96.72	68.82	87.14	74.38	69.80
Pangea 7B	48.99	45.45	94.33	87.94	46.86	78.63	61.93	50.11

Table 3. Detailed results for CVQA across different multilingual settings and RAG settings.

Model	No RAG		Oracle Context		RAG			
	Baseline	+ Multilingual Prompt	Eng.	Multilingual	Eng. Cap.	Oracle Eng.	mmE5	B3
Gemma3 4B	48.26	47.22	57.19	54.39	39.60	47.91	52.73	47.20
Gemma3 12B	62.46	62.71	74.24	70.97	49.08	56.76	59.45	57.25
Gemma3 27B	66.20	66.24	78.43	76.50	55.56	62.70	63.83	62.66
Qwen2.5-VL 3B	46.22	44.95	57.27	52.07	39.43	46.38	51.08	41.49
Qwen2.5-VL 7B	53.87	52.32	64.22	58.28	47.96	55.08	56.02	50.08
Qwen2.5-VL 32B	60.00	55.75	74.31	66.35	53.39	61.94	62.89	57.48
Qwen2.5-VL 72B	65.14	62.67	79.68	74.64	58.03	65.95	63.68	61.76
Qwen3-VL 4B Think	47.22	46.34	59.39	52.93	34.86	44.37	45.93	39.29
Qwen3-VL 8B Think	53.79	52.70	68.05	61.93	40.84	49.69	51.09	42.42
Qwen3-VL 30B A3B Think	65.54	64.77	77.61	74.35	59.69	66.00	65.68	62.26
Pangea 7B	47.05	36.53	61.80	47.32	35.88	44.68	50.99	40.54

Table 4. Detailed results for WORLDCUISINES across different multilingual settings and RAG settings.

rent instruction-tuning approaches: despite being capable of generating multilingual text, these models follow reasoning instructions significantly better when presented in English.

Furthermore, figures reveal a critical limitation regarding contextual grounding. Intuitively, one might expect that answering a culture-specific question would be easier when the supporting evidence (oracle context) is provided in that culture's native language. However, our results indicate the opposite. Across the majority of languages, particularly in the WORLDCUISINES for languages like Yoruba and Marathi, providing oracle context in the target language causes a sharper performance decline than simply changing the prompt language. This inverse effect indicates that VLMs treat English as a reasoning pivot: they struggle to integrate non-English evidence, preferring English context even for culture-specific queries where native-language grounding should be advantageous.

10. Prompts

10.1. Translation Prompts

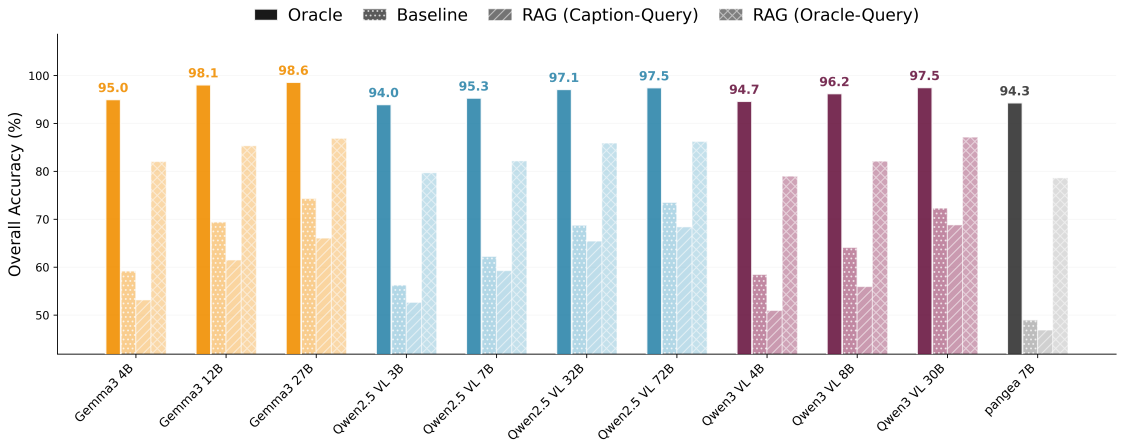
To generate the multilingual instructional prompt, we utilized an LLM with the prompt structure shown in Figure 12. The prompt is designed to ensure the translation maintains the specific formatting required for template substitution (e.g., preserving double curly braces).

10.2. Evaluation Prompts

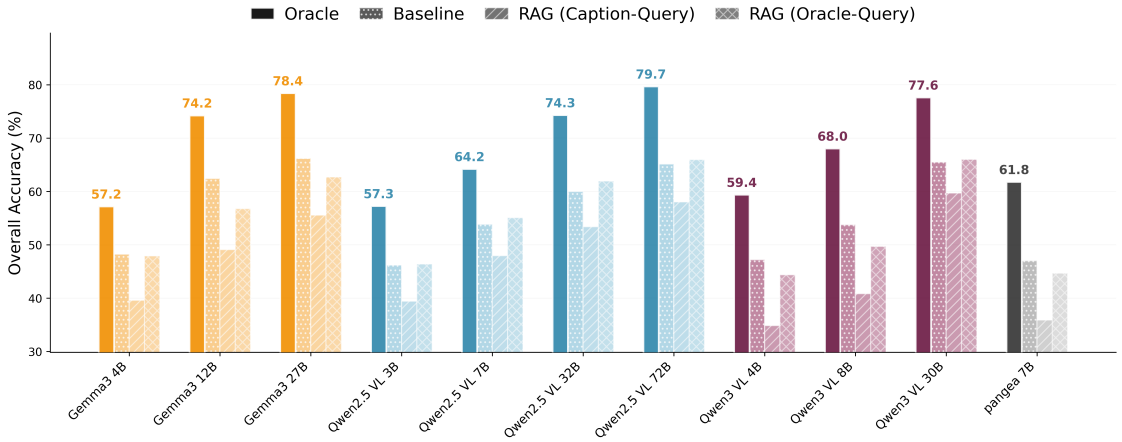
To assess performance and quality, we utilized two distinct prompts. The first is a "VLM-as-a-judge" prompt used to evaluate the relevance of retrieved context (Figure 13). The second is the inference prompt used to generate the final multiple-choice answer given the context (Figure 14).

10.3. Inference Prompts

To generate the final answer for the visual question answering task, we employ the structured inference prompt displayed in Figure 14. This prompt aggregates the input question, the retrieved context passages (if available), and the multiple-choice options. The model is instructed to reason based on the provided context and output the answer in a strict JSON format to facilitate automated parsing.



(a) CVQA.



(b) WORLDCUISINES.

Figure 9. Comparison of oracle context versus RAG performance across model families on **(a)** CVQA and **(b)** WORLDCUISINES. The performance gap widens with model scale, indicating that while larger VLMs can effectively leverage perfect context, current retrieval systems fail to provide evidence of sufficient quality to match oracle performance.

11. Hyper-parameters

For all inference runs, we use 4 NVIDIA H100 80GB GPUs with vLLM and set the maximum output length to 16,384 tokens. For `Qwen3-VL`, we use the recommended generation settings: temperature = 1.0, presence penalty = 0.0, repetition penalty = 1.0, top-k = 20, and top-p = 0.95. For `Gemma3`, we use top-k = 64 and top-p = 0.95. For `Qwen2.5-VL` and `Pangea`, we follow the recommended settings of repetition penalty = 1.05 and temperature = 0.

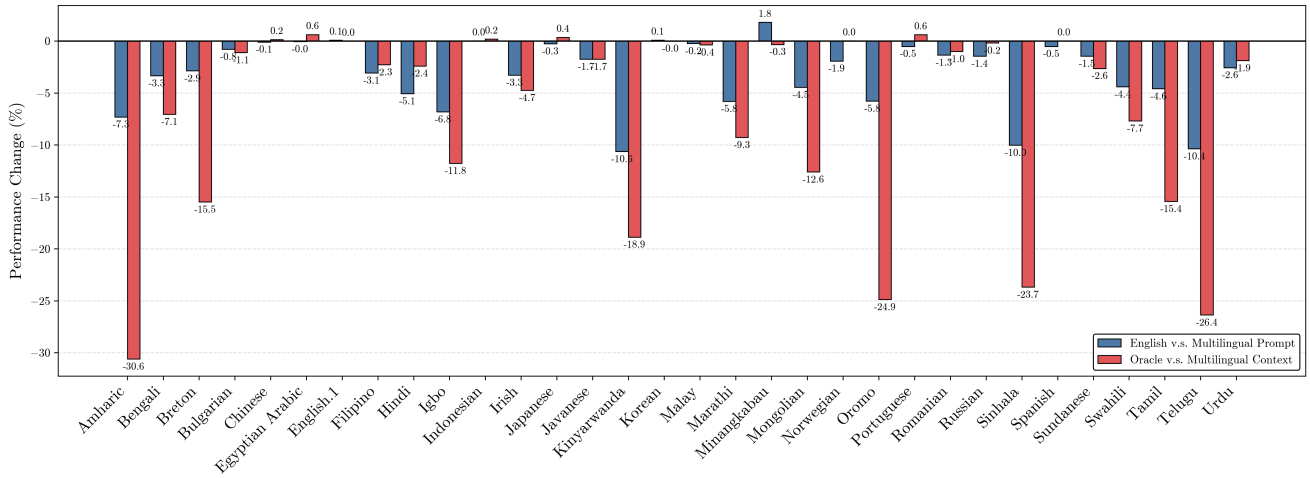


Figure 10. Language-wise performance change on CVQA when switching from English to multilingual prompts. Similar to WORLD-CUISINES, low-resource languages exhibit substantial performance degradation.

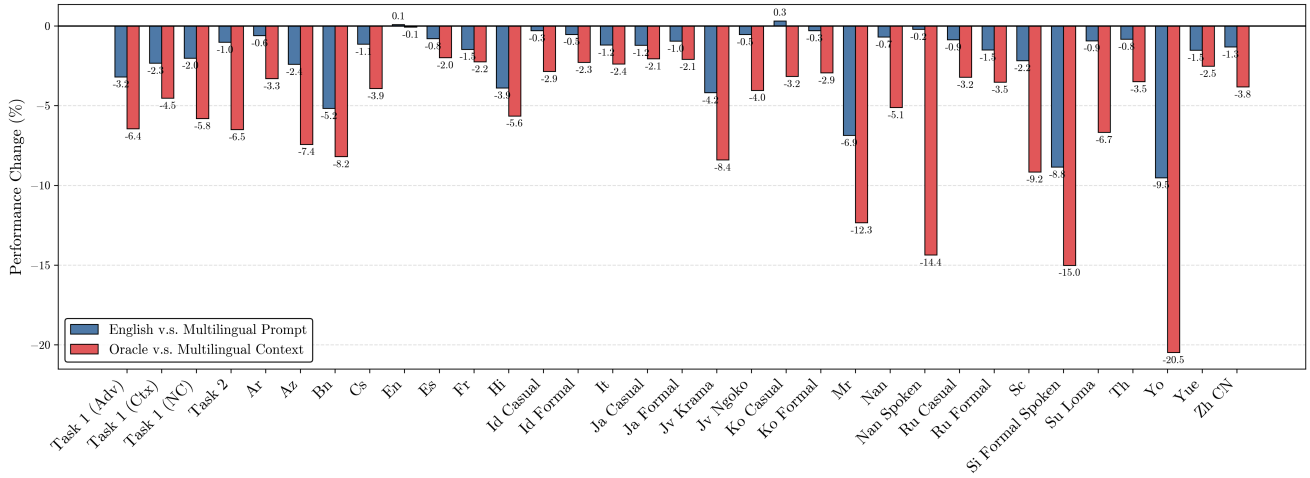


Figure 11. Language-wise performance change on WORLDCUISINES when switching from English to multilingual prompts. Negative values indicate performance degradation, with low-resource languages showing the most significant drops.

```
You are an expert translator with specialization in prompt engineering. Your task is to translate the
    string values of the following JSON object into {target_language}.

### Guidelines:
1. Tone & Style: The text is used to prompt an AI model. Ensure the translation is clear, concise, and
    instructional. It should sound natural and culturally appropriate for a native speaker of {
    target_language}, but maintain the directive nature of the original text.
2. Placeholders: Do NOT translate or alter any text inside curly braces (e.g., keep '{{input}}' or '{{
    name}}' exactly as they are).
3. Structure: Keep the JSON keys exactly the same. Only translate the values.

### Output Format:
Return ONLY the raw JSON string.
   - Do NOT use Markdown code blocks (no ```json).
   - Do NOT add explanations or conversational text.
   - Ensure the output is valid, parseable JSON.

### Input JSON:
{input_json_string}
```

Figure 12. Prompt for translating system instructions to target language.

```
You are an expert evaluator for a Vision-Language RAG system. Given an image and a question, assess how
     well the provided textual context supports answering the image-based question, considering both
    its relevance to the question and its helpfulness in reaching or verifying the ground truth answer.
     You must evaluate the context according to the given rubric by providing a short explanation for
    your reasoning and then assign a single holistic score (1-5).

### Question
{{ question }}

### Ground Truth Answer
{{ ground_truth_answer }}

### Context
{{ context }}

### Evaluation Rubric
1: The context is completely irrelevant or misleading as the context provides no useful information for
     answering the question.
2: The context is slightly related but mostly unhelpful as the context contains minimal connection or
    value toward the answer.
3: The context is somewhat relevant and partially useful as the context offers limited insight or
    indirect clues toward the answer.
4: The context is mostly relevant and helpful as the context supports reasoning toward the correct
    answer though not perfectly comprehensive.
5: The context is highly relevant and directly helpful as the context clearly supports or confirms the
    correct ground truth answer.

### Response Format
Provide your response in the following JSON format:

{{ format | schema }}

### Response
```

Figure 13. Prompt for evaluating the relevance of retrieved context (VLM-as-a-judge).

```
Given the multiple-choice question below, choose the single best answer based on the question and any
    relevant context provided. Respond only with the number of the correct option (i.e., 1, 2, 3, or 4)
    . Use the context if helpful, but ignore unrelated information.

### Question
{{ question }}
{% if context_list %}

### Context
{% for context in context_list %}
- {{ context }}
{% endfor %}

{% endif %}

### Options
{% for option in options %}
{{ loop.index }}. {{ option }}
{% endfor %}

### Answer Format
Provide your response in the following JSON format:

{{ format | schema }}

### Response
```

Figure 14. Prompt template for the multiple-choice VQA task with retrieval augmentation.

Language	Family	Resource Class†	Register	Regional Dialects	In CVQA	In WORLDCUISINES
Amharic	Afro-Asiatic	2		Ethiopia	✓	
Arabic	Afro-Asiatic	5		Arab		✓
Azerbaijani	Turkic	1				✓
Bengali	Indo-European	3		India	✓	✓
Breton	Indo-European	1		France	✓	
Bulgarian	Indo-European	3		Bulgaria	✓	
Cantonese	Sino-Tibetan	1				✓
Chinese	Sino-Tibetan	5		China	✓	✓
Chinese	Sino-Tibetan	5		Singapore	✓	✓
Czech	Indo-European	4				✓
Egyptian Arabic	Afro-Asiatic	3		Egypt	✓	
English	Indo-European	5		United States	✓	✓
French	Indo-European	5		France		✓
Hokkien	Sino-Tibetan	0	Written	Medan		✓
Hokkien	Sino-Tibetan	0	Spoken	Medan		✓
Hindi	Indo-European	4		India	✓	✓
Igbo	Niger-Congo	1		Nigeria	✓	
Indonesian	Austronesian	3	Formal	Indonesia	✓	✓
Indonesian	Austronesian	3	Casual	Indonesia		✓
Irish	Indo-European	2		Ireland	✓	
Italian	Indo-European	4				✓
Japanese	Japonic	5	Formal	Japan	✓	✓
Japanese	Japonic	5	Casual	Japan		✓
Javanese	Austronesian	1	Krama	Java	✓	✓
Javanese	Austronesian	1	Ngoko	Java		✓
Kinyarwanda	Niger-Congo	1		Rwanda	✓	
Korean	Koreanic	4	Formal	South Korea	✓	✓
Korean	Koreanic	4	Casual	South Korea		✓
Marathi	Indo-European	2		India	✓	
Malay	Austronesian	3		Malaysia	✓	
Minangkabau	Austronesian	1		Indonesia	✓	
Mongolian	Mongolic	1		Mongolia	✓	
Norwegian	Indo-European	1		Norway	✓	
Oromo	Afro-Asiatic	1		Ethopia	✓	
Portuguese	Indo-European	4		Brazil	✓	
Romanian	Indo-European	3		Romania	✓	
Russian	Indo-European	5	Formal	Russia	✓	✓
Russian	Indo-European	5	Casual	Russia		✓
Sardinian	Indo-European	1		Italy		✓
Sinhala	Indo-European	0	Formal	Sri-Lanka	✓	✓
Spanish	Indo-European	5		Spain	✓	✓
Spanish	Indo-European	5		Argentina	✓	
Spanish	Indo-European	5		Chile	✓	
Spanish	Indo-European	5		Colombia	✓	
Spanish	Indo-European	5		Ecuador	✓	
Spanish	Indo-European	5		Mexico	✓	
Spanish	Indo-European	5		Uruguay	✓	
Sundanese	Austronesian	1	Loma	Indonesia	✓	✓
Swahili	Niger-Congo	2		Kenya	✓	
Tagalog	Austronesian	3		Phillipines	✓	✓
Tamil	Indo-European	3		India	✓	
Telugu	Indo-European	1		India	✓	
Thai	Kra-Dai	3				✓
Urdu	Indo-European	3		India	✓	
Urdu	Indo-European	3		Pakistan	✓	
Yoruba	Niger-Congo	2				✓

Table 5. Languages used in M4-RAG. Resource classes follow the 0–5 scale of Joshi et al. [31].