When Retrieval Hurts: Evidence Utilization, Script Fidelity, and Knowledge Conflicts in Multilingual RAG

Gadha Saji Menon[1*] **Swathi Jayakumar**[1*] **Varalekshmy M Mohan**[1*]
Sachin Kurup[1*] **Veena G**[1] **Vani Kanjirangat**[2]

[1]Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India
[2]SUPSI, IDSIA, Switzerland
*

Abstract

The problem of extractive multilingual QA with LLMs is characterized by complex interactions among retrieval mechanisms, knowledge source configurations, prompting techniques, and scripting biases. Despite high retrieval quality, multilingual RAG often degrades performance, revealing a gap between retrieved evidence and its effective utilization. To address this issue, this paper offers an extensive empirical study that examines these components by comparing retrieval-augmented generation (RAG) with a non-RAG baseline across 21 typologically diverse languages and 5 leading LLMs. Our analysis includes five prompting strategies and multiple retrieval configurations, which enable a unified evaluation across diverse linguistic settings. We have also observed an evidence utilization gap in RAG settings, where RAG underperforms despite high retrieval hit rates due to models' inefficiency in leveraging the retrieved evidence. We also introduce lightweight inference-time metrics to better characterize retrieval usage and conflict patterns.We also highlight script fidelity as a crucial factor in our experiments, as non-Latin-script languages experience significant performance drops and increased hallucinations without proper grounding. Further, we analyzed generator language preferences, systematically examined conflicts, and identified mechanisms for the effective detection and resolution of conflicts. The study further details how prompting strategies affect language families and script types, offering a detailed analysis for optimizing future multilingual RAG settings.

1 Introduction

Extractive multilingual QA involves complex interactions among retrieval methods, prompting strategies, and language-specific factors like script, family, and structures (Hu et al., 2020; Clark et al., 2020). Despite strong performance on English benchmarks, LLMs struggle in multilingual settings due to script diversity, morphological variation, and inconsistent retrieval behavior (Ahuja et al., 2023; Wu et al., 2024). These challenges are further compounded in extractive settings, where models must identify precise answer spans across typologically distant languages. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) was introduced to ground model outputs in external evidence, reducing hallucination and improving factual accuracy. However, in multilingual settings, the benefits of RAG are highly variable. Models frequently fail to use retrieved evidence effectively, often defaulting instead to parametric memory even when relevant evidence is available (Longpre et al., 2021; Shi et al., 2023a). Notably, we find that retrieval itself is not the primary bottleneck—despite high retrieval coverage, performance often degrades under RAG, indicating a gap in the effective utilization of evidence. Prior work on multilingual RAG has primarily focused on aggregate performance across high-resource languages (Ahuja et al., 2023), leaving several gaps. Retrievers exhibit systematic biases toward certain language families and scripts (Ki et al., 2025; Wu et al., 2024). Inconsistencies between generated outputs and retrieved evidence vary significantly across language families, with non-Latin scripts showing greater drift and hallucination (Qi et al., 2025; Chirkova et al., 2024). Standard metrics like Exact Match (EM) and F1 (Rajpurkar et al., 2016) do not capture script-level errors, such as responses generated entirely in the wrong writing system. No prior study provides a unified analysis of retrieval coverage, script fidelity, evidence utilization, and generation inconsistencies across diverse scripts and language families simultaneously.To address these gaps, we conduct a large-scale, controlled study comparing RAG with a non-RAG baseline across 21 typologically diverse languages spanning 9 writing scripts, using the TyDiQA (Clark et al.,

*Equal contribution.

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 92–107

2020) and XQuAD (Artetxe et al., 2020) benchmarks. We evaluate both English-centric general-purpose models (Gemma-2-9B-IT (Google DeepMind, 2024), Qwen2.5-7B-Instruct (Qwen Team, 2025), LLaMA-2-7B-Chat (Touvron et al., 2023) and multilingual-specialized models (Aya-23-8B (Üstün et al., 2024), EuroLLM-9B-Instruct (Utter Project, 2024)) under five prompting strategies (Liu et al., 2023; Lin et al., 2022), ranging from minimal instruction to few-shot demonstration, to examine how strategy interacts with language family and script type. To capture failure modes invisible to standard metrics, we introduce three novel script fidelity diagnostics: Language Match Rate (LMR), Script Consistency (SC), and Transfer Rate (TR), which quantify the degree to which model outputs conform to the expected writing system. Together, these components enable a systematic analysis of where and why multilingual RAG fails.

Contributions. Our contributions are summarized as follows:

- **Evidence-Utilization Gap:** Our analysis shows that retrieval may not always be the primary bottleneck in multilingual QA. Despite high retrieval quality (RHR = 0.927), performance often degrades with RAG, highlighting failures in effectively utilizing evidence.
- **Model and Retrieval Interaction:** We observe that stronger LLMs, like Gemma and Qwen, are more affected by retrieval interference. Also, high overlap with retrieved evidence does not ensure correctness, especially in non-Latin languages.
- **Novel Script Fidelity Metrics:** Due to the limitations of existing metrics, we introduce three complementary diagnostics—Language Match Rate (LMR), Script Consistency (SC), and Transfer Rate (TR)—to quantify the degree to which model outputs adhere to the expected writing system. These metrics enable fine-grained analysis of multilingual generation failures that remain undetected by standard evaluation measures.
- **Script Fidelity and Diagnostics:** Our analysis identifies script fidelity as an important failure area in multilingual QA. Simple diagnostics like LMR effectively reveal these inconsistencies, leading to significant improvements in script alignment, such as the Russian LMR improving from 0.565 to 0.984.

2 Related Work

Early multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) established strong cross-lingual transfer baselines, later extended by instruction-tuned Large Language Models (LLMs), including Aya (Üstün et al., 2024), Gemma (Google DeepMind, 2024), Qwen (Qwen Team, 2025), EuroLLM (Utter Project, 2024), and LLaMA (Touvron et al., 2023). Despite these advances, a substantial body of work highlights persistent disparities in multilingual performance (Hu et al., 2020; Wu et al., 2024). MEGA (Ahuja et al., 2023) shows that high- vs. low-resource performance gaps remain even in large models, while performance also varies across language families and scripts. In particular, non-Latin scripts exhibit greater inconsistency and degradation (Wu et al., 2024; Qi et al., 2025; Chirkova et al., 2024), retrievers demonstrate language preference biases (Ki et al., 2025), and tokenizer fertility across scripts has been shown to predict downstream degradation (Rust et al., 2021).To systematically evaluate cross-lingual performance, several benchmarks have been introduced. TyDiQA (Clark et al., 2020), XQuAD (Artetxe et al., 2020), XTREME (Hu et al., 2020), and XCOPA (Ponti et al., 2020) provides a standardized evaluation across diverse languages. However, these benchmarks primarily report aggregate performance and do not explicitly capture script-level inconsistencies. Prompting strategies have emerged as a de facto approach for adapting LLMs to new tasks. Surveys (Liu et al., 2023; Reynolds and McDonell, 2021) and few-shot studies (Lin et al., 2022; Shi et al., 2023b) demonstrate strong cross-lingual generalization. However, prompting behavior varies across languages: Ahuja et al. (2023) shows that native-language prompts do not consistently outperform English prompts, indicating a complex interaction between prompting strategy and language characteristics that our five-strategy design directly investigates. Evaluation in multilingual QA requires metrics beyond exact string matching. Standard metrics such as EM and F1 (Rajpurkar et al., 2016) fail to capture deeper generation errors. Alternative metrics such as BERTScore (Zhang et al., 2020), chrF (Popović, 2015), RAGAS (Es et al., 2023), and FActScore (Min et al., 2023) ad-

dress semantic similarity, grounding, and hallucination (Ji et al., 2023). However, none explicitly evaluate the script fidelity, leaving a gap in assessing whether outputs conform to the expected writing system. RAG (Lewis et al., 2020) grounds outputs in retrieved evidence via dense retrieval (Karpukhin et al., 2020), and has been extended to multilingual settings through BGE-M3 (Chen et al., 2024), FAISS (Johnson et al., 2021), REPLUG (Shi et al., 2024), and Self-RAG (Asai et al., 2024). Despite these advances, multilingual RAG is not uniformly effective: evidence utilization remains inconsistent (Qi et al., 2025; Chirkova et al., 2024), models favor parametric memory over retrieved context (Longpre et al., 2021), and are sensitive to irrelevant passages (Shi et al., 2023a). Recent benchmarks XRAG (Liu et al., 2025) and MEMERAG (Talur et al., 2025) target cross-lingual RAG evaluation but focus primarily on high-resource settings, while knowledge conflict studies (Xu et al., 2024; Du et al., 2024) remain largely English-centric and entity-focused. In contrast, our work identifies script- and syntax-specific conflict types and shows that retrieval interference increases with stronger parametric knowledge.Representation biases in low-resource settings (Kanjirangat et al., 2025) are addressed through fine-tuning (Kohli et al., 2024; Nair and Gupta, 2024) and preference optimization (Rajagopalan Nair et al., 2026). Multilingual LLM applications span translation (Padmavilochanan et al., 2025), speech (R. et al., 2024), event understanding (Hrudya et al., 2025), and QA generation (Pradeesh et al., 2025), reflecting their use across a wide range of tasks.

Unlike prior work studying these components in isolation, we jointly evaluate retrieval, prompting, script fidelity, and hallucination across 21 languages.

3 Methodology

3.1 Overview

This study presents a systematic empirical comparison of five instruction-tuned LLMs with varying multilingual capabilities on extractive QA under two paradigms: *zero-shot context-based generation* (No-RAG) and *retrieval-augmented generation* (RAG) (Lewis et al., 2020). Experiments span two benchmarks, TyDiQA (Clark et al., 2020) and XQuAD (Artetxe et al., 2020), covering 21 typologically diverse languages across 9 writing scripts.

3.2 Datasets

TyDiQA provides naturally authored questions across 9 typologically diverse languages (Arabic, Bengali, English, Finnish, Indonesian, Korean, Russian, Swahili, Telugu). We use two configurations: the *secondary task* supplies the gold passage directly (No-RAG), while the *primary task* requires full corpus retrieval (RAG). We evaluate 1,730 samples in total, with 200 per language except Bengali (130).

XQuAD consists of SQuAD passages professionally translated into 11 languages (Arabic, German, Greek, English, Spanish, Hindi, Russian, Thai, Turkish, Vietnamese, Chinese), evaluated zero-shot (No-RAG). We use 2,200 samples (200 per language). Overall, the evaluation comprises 3,930 samples across both datasets.

3.3 Models

We evaluate five instruction-tuned LLMs spanning diverse architectural families and multilingual training regimes. All models are evaluated across both benchmarks and all experimental conditions. Model selection covers a range from broadly multilingual (Aya-23-8B (Üstün et al., 2024), EuroLLM-9B-Instruct (Utter Project, 2024),) to general-purpose instruction-tuned (Gemma-2-9B-IT(Google DeepMind, 2024), Qwen (Qwen Team, 2025), LLaMA-2-7B-Chat(Touvron et al., 2023)), enabling analysis of how multilingual specialization affects performance across scripts and retrieval settings. Implementation details are provided in Appendix A.

3.4 Retrieval Pipeline

We implement a multilingual RAG pipeline with sentence-aware chunking and dense retrieval, following the architecture illustrated in Figure 1 for the TyDiQA primary task.

Chunking. Documents are segmented into overlapping chunks using Unicode-aware sentence boundary detection. Chunk sizes are language-specific, calibrated via empirical character-to-token ratios (e.g., ∼450 characters for Arabic, ∼810 for Finnish). Adjacent chunks overlap by 1–2 sentences to preserve cross-sentence answer spans, with longer overlaps for morphologically rich languages (Finnish, Telugu, Bengali, Russian).

Retrieval. We use BGE-M3 (Chen et al., 2024) as the multilingual dense retriever. For each query,

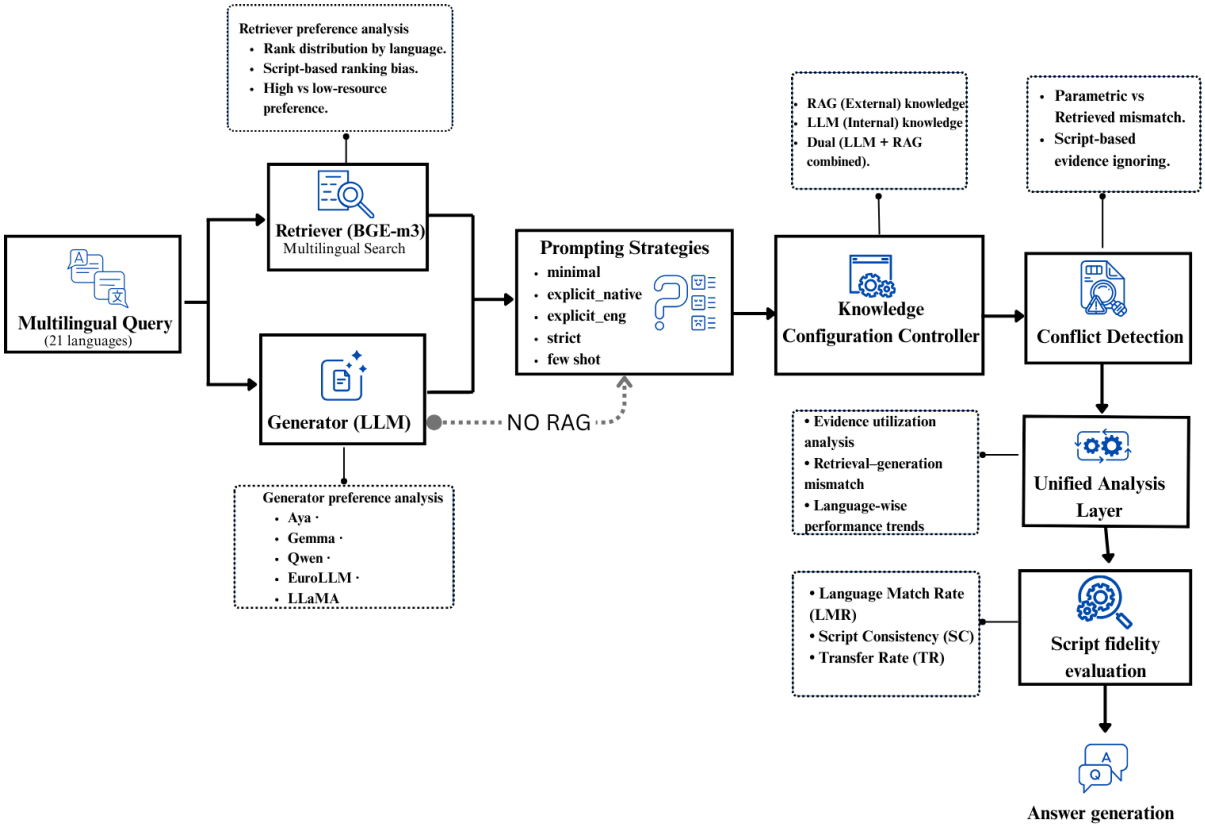


Figure 1: Overview of the multilingual RAG framework with a non-RAG baseline, illustrating retrieval (BGE-M3), LLM generation, prompting, knowledge integration, conflict detection, unified analysis (RHR, RUR, EV, HALL, CONF), and script fidelity evaluation before final answer generation.

top-k chunks are retrieved ($k \in [6, 12]$ depending on language), re-ranked by cosine similarity, and the top-k_{keep} chunks (3–5) are concatenated as the generator context. FAISS (Johnson et al., 2021) is used for efficient similarity search.

Knowledge Configuration. As shown in Figure 1, the Knowledge Configuration Controller switches between three modes: RAG (external knowledge only), parametric (LLM internal knowledge only), and dual (LLM + RAG combined). This enables controlled analysis of retrieval contribution per model and language. Due to non-uniform sample distribution in TyDiQA streaming,[1]

3.5 Prompting Strategies

Five prompt templates are evaluated per language, all instructing the model to produce a short extractive answer using exact words from context (Table 1).

Language-specific adaptations follow pilot findings: Russian prompts include a Cyrillic enforcement clause to correct script-switching. Indonesian uses minimal prompting only, as explicit instructions increase hallucination. Few-shot demonstrations are manually constructed with no overlap with evaluation data (Lin et al., 2022). Prompt template details are provided in Appendix C.

Table 1: Prompting strategies.

Strategy	Description
`minimal`	Minimal instruction, answer label only
`explicit_en`	Explicit instruction in English
`explicit_native`	Explicit instruction in native script
`strict`	Script + word-count constraint
`few_shot`	One in-language demonstration

3.6 Preference Analysis

Following the architecture in Figure 1, two analysis blocks quantify systematic model preferences:

Retriever Preference Analysis measures rank distribution by language, script-based ranking bias, and high- vs. low-resource retrieval preference, capturing whether BGE-M3 systematically favors certain language families during chunk ranking.

[1] Bengali data is loaded in non-streaming mode to ensure sufficient sample coverage.

Generator Preference Analysis tracks evidence usage (EV), script preference in generation, and language drift and omission, quantifying the degree to which each model exploits retrieved context versus relying on parametric memory. Evidence usage (EV) is defined in Section 4.1.3.

3.7 Conflict Detection

The conflict detection module, shown in Figure 1, identifies discrepancies between generated answers and retrieved evidence (Shi et al., 2023a). We focus on two primary conflict types:

- **Parametric–retrieved mismatch:** The model generates answers based on internal parametric knowledge rather than retrieved context, detected when Evidence Usage (EV) falls below a threshold.
- **Script-based evidence ignoring:** The model generates answers in an incorrect script despite correctly scripted retrieved evidence. We identify this behaviour using our proposed Language Match Rate (LMR) metric (Section 4.1.4), which measures alignment between the script of the generated output and the retrieved evidence.

4 Results

We evaluate five instruction-tuned LLMs under three conditions: RAG on TyDiQA (9 languages), No-RAG on TyDiQA (gold passage supplied), and No-RAG on XQuAD (12 languages, zero-shot). All metrics are macro-averaged over the best-strategy Token F1.

4.1 Evaluation Metrics

A unified set of metrics is computed for each language, prompting strategy, and experimental condition, covering four evaluation dimensions.

4.1.1 Exact Match and Token-Level Metrics

Exact Match (EM) (Rajpurkar et al., 2016): Binary exact string match after normalization.

Token F1 (Rajpurkar et al., 2016): Harmonic mean of token-level precision and recall, maximized over all gold answers. Tokenization is character-level for Korean, Telugu, Bengali, Thai, and Chinese, where whitespace segmentation is unreliable or absent; whitespace-based tokenization is used otherwise. This ensures fair token-level comparison across scripts with different word boundary conventions.

Semantic Overlap (SEM): Recall-oriented soft token overlap:

$$SEM = \frac{|\text{pred}_{\text{tok}} \cap \text{gold}_{\text{tok}}|}{|\text{gold}_{\text{tok}}|} \quad (1)$$

chrF (Popović, 2015): Character n-gram F-score, particularly informative for morphologically rich and agglutinative languages (Finnish, Turkish, Korean).

4.1.2 Semantic Similarity Metrics

BERTScore F1 (BS-F1) (Zhang et al., 2020): Semantic similarity via contextual embeddings from `bert-base-multilingual-cased`. Token-level greedy alignment is used for TyDiQA; mean-pooled cosine similarity for XQuAD.

4.1.3 Grounding and Hallucination

Evidence Usage (EV): Inspired by RAGAS (Es et al., 2023), EV measures the fraction of predicted tokens found in the context, using deterministic token overlap rather than LLM-based scoring:

$$EV = \frac{|\text{pred}_{\text{tok}} \cap \text{ctx}_{\text{tok}}|}{|\text{pred}_{\text{tok}}|} \quad (2)$$

Hallucination Rate (HALL): A conservative binary indicator inspired by prior work on hallucination in text generation (Ji et al., 2023; Es et al., 2023) that fires only when the prediction is simultaneously inaccurate *and* ungrounded:

$$HALL = \mathbf{1}\left[F_1 < 0.3 \ \wedge \ EV < 0.5\right] \quad (3)$$

The thresholds ($F_1 < 0.3$, $EV < 0.5$) are chosen empirically based on pilot experiments, where this region consistently corresponds to clearly incorrect and unsupported predictions.

Conflict Rate (CONF): A window-restricted negation-mismatch heuristic (Longpre et al., 2021). A conflict is flagged when a language-specific negation word (e.g., *not*, *ne*, *la*) appears within a ± 8-token window around the answer span in the retrieved context but is absent from the prediction. The local window reduces false positives compared to full-passage scanning, which was observed to inflate conflict rates in pilot experiments (Shi et al., 2023a).

4.1.4 Script Fidelity

We introduce a set of metrics to quantify *script fidelity*, a dimension not captured by standard QA evaluation metrics.

Language Match Rate (LMR):

$$LMR = \frac{\sum_{c \in \mathcal{S}(\hat{y})} \mathbf{1}[c \in \mathcal{U}_l]}{|\mathcal{S}(\hat{y})|} \quad (4)$$

LMR measures the fraction of characters in the prediction that belong to the expected script. Here, $\mathcal{S}(\hat{y})$ excludes whitespace, digits, and neutral punctuation, and $\mathcal{U}_l$ denotes the Unicode range for the target language l.

For example, a Spanish answer written fully in Latin script (e.g., *"Madrid es la capital de España"*) will have $LMR \approx 1.0$, while a mixed output such as *"Madrid es the capital"* will result in a lower LMR. Purely numeric answers are assigned $LMR = 1.0$.

Script Consistency (SC):

$$SC = \mathbf{1}[LMR \geq 0.6] \quad (5)$$

SC indicates whether the prediction is predominantly written in the target script. The threshold of 0.6 is chosen empirically to reflect majority script usage, tolerating minor mixing while filtering out predictions dominated by a foreign script. For example, a mostly Indonesian answer such as *"Jakarta adalah the capital"* would still be considered consistent under SC.

Transfer Rate (TR):

$$TR = \mathbf{1}[LMR < 0.7] \quad (6)$$

TR captures cross-language or script leakage using a stricter threshold. Values below 0.7 indicate noticeable mixing, making TR more sensitive than SC. In the above example, *"Jakarta adalah the capital"* would result in $TR = 1$, reflecting partial mixing. SC ensures the prediction is mostly in the target script, while TR detects even minor mixing. Together, they capture both coarse and fine-grained script fidelity in multilingual generation.

4.1.5 Retrieval Metrics

Retrieval Hit Rate (RHR) (Karpukhin et al., 2020; Es et al., 2023): Fraction of queries where at least one retrieved chunk contains the gold answer, providing an upper bound on achievable QA performance:

$$RHR = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[\exists\, c \in C_i : y_i \subseteq c\big] \quad (7)$$

Retrieval Utilization Rate (RUR): We define RUR as a proxy for measuring how effectively the generator uses retrieved evidence, computed via bigram overlap between the generated answer and retrieved context:

$$RUR = \frac{|\text{bigrams}(\hat{y}) \cap \text{bigrams}(C)|}{|\text{bigrams}(\hat{y})|} \quad (8)$$

The gap between RHR and F1, interpreted alongside RUR, captures the *evidence utilization gap* between retrieval quality and generation performance.

4.2 TyDiQA with RAG

Among (Table 2) the evaluated models, Gemma-2-9B-IT achieves the highest performance (F1 = 0.669, EM = 0.460), followed by Aya-23-8B (0.606), Qwen2.5-7B (0.576), EuroLLM-9B (0.480), and LLaMA-2-7B (0.261), which serves as a lower-bound baseline . SEM (Eq. 1) closely tracks F1 across models (within 0.02–0.05), confirming stable rankings and consistent near-miss behavior across conditions. A clear **Latin-script bias** is observed: all models score higher on Indonesian, Finnish, English, and Swahili than on Arabic, Bengali, Korean, and Telugu under identical retrieval conditions, pointing to a systematic generator preference that persists regardless of retrieval quality (Appendix B). SC (Eq. 5) highlights severe failures in weaker models, with LLaMA-2-7B showing substantially lower compliance in Korean and Arabic. Finer-grained analysis further reveals systematic script inconsistencies in models such as EuroLLM-9B for Telugu and Korean, which are not evident from aggregate metrics (Appendix B).

Table 2: TyDiQA RAG, macro-average over 9 languages.

Model	EM	F1	BS	Hall	LMR
Gemma-2-9B-IT	**0.460**	**0.669**	**0.831**	0.058	0.960
Aya-23-8B	0.402	0.606	0.829	**0.036**	**0.986**
Qwen2.5-7B	0.354	0.576	0.807	0.076	0.980
EuroLLM-9B	0.211	0.480	0.767	0.060	0.950
LLaMA-2-7B	0.077	0.261	0.671	0.264	0.686

4.3 Retrieval Coverage and Utilisation

The retriever is not the bottleneck. BGE-M3 achieves macro-averaged RHR = 0.927 (Eq. 7), placing the gold answer in retrieved chunks in over 92% of queries. This already challenges the common framing that RAG failures stem from poor retrieval (Mallen et al., 2023); our findings show that

failure is in generator utilization. As shown in Figure 2b (Appendix E), RUR (Eq. 8) reveals a consistent script-level pattern: models exhibit higher overlap with retrieved content in non-Latin languages (e.g., Bengali and Korean, RUR $>$ 0.87) than in Latin languages (e.g., English and Finnish, RUR $<$ 0.70). This suggests that in non-Latin settings, generators rely more directly on retrieved spans, likely because there are fewer parametric paraphrasing alternatives. EV (Eq. 2) scores (0.90–0.95 overall; near 1.0 for Korean and Bengali) further support this behavior at the span level. However, increased reliance on retrieved text does not translate into improved answer quality. Non-Latin languages still exhibit lower F1 and LMR (Eq. 4), indicating that high lexical overlap reflects surface-level copying rather than effective semantic utilization. In contrast, LLaMA-2-7B shows low RUR (0.463), suggesting under-utilization of retrieved context, which correlates with a higher hallucination rate (HALL = 0.264; Eq. 3).

Table 3: RHR (model-independent) and RUR / RHR−F1 gap (per model).

Model	RHR	RUR	RHR−F1
Gemma-2-9B-IT	0.927	0.795	0.258
Aya-23-8B	0.927	0.711	0.321
Qwen2.5-7B	0.927	**0.835**	0.351
EuroLLM-9B	0.927	0.772	0.447
LLaMA-2-7B	0.927	0.463	**0.666**
Avg.	0.927	0.708	0.389

4.4 Parametric vs. Retrieved Knowledge

When the gold passage is supplied directly (Table 4), all models improve, most strikingly Gemma (F1 rises from 0.669 to 0.765) and EuroLLM (0.480 to 0.600). As illustrated in Figure 2a (Appendix E), introducing retrieval reduces F1 for three of five models by 0.096–0.120 points. Only Aya gains (+0.025); LLaMA is near neutral. These drops

Table 4: No-RAG (gold passage) — macro-average over 9 languages.

Model	EM	F1	BS	Hall	LMR
Gemma-2-9B-IT	**0.543**	**0.765**	**0.875**	0.044	0.981
Qwen2.5-7B	0.455	0.688	0.852	0.070	0.980
EuroLLM-9B	0.350	0.600	0.813	**0.033**	0.979
Aya-23-8B	0.310	0.581	0.740	0.050	**0.984**
LLaMA-2-7B	0.081	0.289	0.668	0.370	0.719

represent an **evidence utilisation gap**: high RHR (0.927; Eq. 7) and non-trivial RUR (Eq. 8), yet F1 falls below the clean-passage baseline. The generator implicitly trusts its parametric prior over retrieved spans when the two conflict. This contrasts with the finding of Du et al. (2024) that models are more easily swayed by context for *unfamiliar* entities: here, the models with the *strongest* parametric representations (Gemma, Qwen) suffer the largest retrieval penalties, suggesting that high-resource multilingual training actually *increases* susceptibility to retrieved-context interference, not decreases it. Aya's small gain is the inverse case: its comparatively weaker parametric grounding for several languages means retrieved context supplements rather than conflicts with internal knowledge. Gemma's No-RAG Telugu few-shot score (F1 = 0.881), the single highest language–model result in the study, confirms that the model *possesses* the underlying knowledge; the RAG pipeline fails to resolve conflicts between parametric and retrieved knowledge.

Table 5: RAG vs. No-RAG on TyDiQA. Negative ΔF1 means retrieval hurts.

Model	No-RAG	RAG	ΔF1
Aya-23-8B	0.581	0.606	+0.025
Gemma-2-9B-IT	0.765	0.669	−0.096
Qwen2.5-7B	0.688	0.576	−0.112
EuroLLM-9B	0.600	0.480	−0.120
LLaMA-2-7B	0.289	0.261	−0.028

4.5 Prompt Strategy Effects

We evaluate multiple prompting strategies across both TyDiQA RAG and no-RAG settings. The best-performing strategy varies across models, languages, and retrieval conditions.Under TyDiQA RAG, few-shot prompting is frequently preferred for higher-capacity models, while simpler strategies such as minimal or explicit prompting are more effective for weaker models. In contrast, under the no-RAG setting, minimal prompting is often sufficient, particularly when gold passages are directly provided.A detailed breakdown of the best prompt strategy per language and model for both settings is provided in Appendix C.

4.6 XQuAD No-RAG

Aya achieves the highest performance on XQuAD (macro F1 = 0.757), with strong results on German (0.851), Romanian (0.819), and English (0.873), despite ranking second on TyDiQA under RAG (Table 6). Since XQuAD is evaluated in a *No-RAG*

Table 6: XQuAD No-RAG zero-shot — macro-average over 12 languages.

Model	EM	F1	BS	Hall	Strat.
Aya-23-8B	**0.596**	**0.757**	**0.842**	0.045	strict
Gemma-2-9B-IT	0.466	0.689	0.768	**0.035**	few-shot
EuroLLM-9B	0.406	0.606	0.734	0.084	strict
Qwen2.5-7B	0.305	0.597	0.709	0.080	strict
LLaMA-2-7B	0.148	0.295	0.521	0.395	strict

setting, this contrast highlights that parametric multilingual knowledge and retrieval utilisation are distinct capabilities: Aya excels when answers rely on internal knowledge, but is comparatively less effective at leveraging retrieved context.Furthermore, prior multilingual RAG benchmarks focus largely on Latin-script European languages (Liu et al., 2025), likely underestimating generator difficulty for non-Latin scripts under realistic retrieval conditions.

4.7 Script-Level Bias and Per-Language Patterns

Errors are strongly language- and script-dependent, with three illustrative cases. **Telugu** is the hardest under RAG (F1: 0.162–0.357): abugida tokenization fragments retrieved chunks, degrading generation even when the answer is present.This fragmentation directly widens the evidence utilization gap: Telugu shows near-perfect EV (1.0) yet the lowest F1, confirming that morpheme-level fragmentation causes incoherent span copying rather than meaningful extraction. Gemma's No-RAG score (F1 = 0.812) confirms the bottleneck lies in retrieval, not model knowledge. **Finnish** shows low F1 despite perfect script fidelity (LMR = 1.00; Eq. 4): agglutinative morphology means a minor suffix variations incur large token-level penalties. chrF scores (Gemma: 0.678, EuroLLM: 0.635) reveal partially correct predictions, exposing a metric artifact rather than the true failure . **Russian** illustrates script-switching: Qwen produced 44% Latin characters in v4 (LMR = 0.565; Eq. 4), flagged by TR (Eq. 6).A Cyrillic enforcement clause in the prompt restored LMR to 0.984 without retraining, showing that SC (Eq. 5) and TR (Eq. 6) provide actionable inference-time diagnostics. For Korean, Latin script leakage (LMR = 0.148) was successfully detected by the TR metric. For Arabic, negation-mismatch conflicts (CONF 0.50–0.60) were isolated via our conflict detection module; resolution via constrained decoding remains future work. Hindi maintained strong script fidelity (LMR = 0.930), with errors attributable to comprehension rather than script failure.

Table 7: Language-specific knowledge conflicts and hallucination under RAG (Aya).

Lang	**Type**	**HALL**	**CONF**
Arabic	Negation	0.030	0.570
Russian	Semantic drift	0.100	0.615
Indonesian	Prompt-induced	0.070 / 0.133	0.515

4.8 Knowledge Conflicts and Hallucination

Three distinct conflict types emerge across languages (Table 7). Arabic shows the highest conflict rate (CONF $\approx$ 0.50–0.60) across all models: negated facts in retrieved passages are consistently rendered as affirmatives, a script-and-syntax-specific conflict type absent from prior work focused on entity substitution in English (Xu et al., 2024). Russian exhibits the highest hallucination among non-LLaMA models (Aya HALL = 0.100–0.140; Eq. 3) even after the Cyrillic fix, indicating residual *semantic* drift that prompt engineering cannot resolve. Indonesian represents a third type: prompt-induced conflict, where explicit instructions raised HALL to 0.133 while minimal prompting reduced it below 0.075 universally. Minimal prompting serves as an effective mitigation, identified through pilot experiments and applied as a deliberate language-specific intervention (Appendix B.3).

4.9 Summary

We highlight four key findings. (1) **Retrieval is not the primary bottleneck**: high RHR = 0.927 demonstrates that failures arise predominantly from poor generator utilisation rather than missing evidence. While our study uses controlled sample sizes (200 per language), the consistency of this pattern across 21 languages and 5 models of varying capacity suggests it is systematic rather than an artefact of our experimental setup. (2) **Stronger models are more vulnerable to retrieval**: models with richer parametric knowledge (Gemma, Qwen) suffer larger RAG drops, suggesting increased interference between parametric and retrieved knowledge. (3) **High overlap does not imply correctness**: despite higher RUR in non-Latin languages, F1 remains lower, indicating that surface copying does not guarantee semantic accuracy. (4) **Errors are language-specific**: failures vary by script and linguistic properties (e.g., negation, morphology, tokenization), and require targeted interventions;

SC/TR metrics help identify these patterns.

5 Conclusion

In this study, we provide multilingual analysis, We observed that retrieval coverage is not the primary bottleneck in multilingual RAG: despite high RHR (0.927), performance drops in three of five models due to parametric–retrieved conflict rather than missing evidence. Stronger parametric models (Gemma, Qwen) are more susceptible to this interference. We identify script fidelity as an additional failure factor in non-Latin languages and introduce LMR, SC, and TR as lightweight diagnostics. A simple Cyrillic fix (LMR: 0.565→0.984) highlights their practical value without retraining. Overall, our findings emphasize generator-side grounding and script-aware conflict resolution over improving retrieval coverage.

6 Acknowledgments

We thank our mentors and collaborators for their guidance and support throughout this work. We also acknowledge the use of open-source datasets and models, including TyDiQA, XQuAD, and publicly available instruction-tuned LLMs, which made this study possible.

7 Limitations

We identify five limitations. First, evaluations use up to 200 samples per language (Bengali: 130); the Bengali sample size reflects the full size of the TyDiQA Bengali evaluation split — a benchmark constraint, not a sampling choice. Bengali results are directionally consistent with other non-Latin abugida scripts: RAG reduces F1 for four of five models (e.g., Gemma: 0.874→0.739), script compliance remains high for stronger models (SC: 0.89–0.99), and the EV–F1 gap pattern holds across all five models, corroborating the paper's core surface-copying claim. These consistent patterns across 21 languages and 5 models suggest conclusions are systematic rather than an artefact of sample size. Second, all evaluated models are sub-10B parameters (7B–9B); whether and how findings generalize to larger models (14B, 30B, 70B+) remains an open question. The direction of this effect is non-trivial: while script fidelity issues, such as those observed in Russian, Korean, and Telugu, would likely improve with scale due to stronger multilingual coverage and instruction-following ability, the evidence utilization gap may persist or intensify, as larger models possess stronger parametric priors that are harder to override with retrieved evidence. This makes generator-side grounding interventions even more critical at larger scales, and evaluating this regime is an important direction for future work. Third, retrieval errors from BGE-M3 may propagate into generation, meaning some observed hallucinations may reflect retrieval failures rather than generator behaviour. Fourth, prompting strategies are manually designed rather than exhaustively optimized; automated prompt search may yield stronger baselines. Fifth, findings are scoped to extractive QA and may not generalize to abstractive or reasoning-heavy tasks.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, and 1 others. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of EMNLP*.

Mikel Artetxe and 1 others. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of ICLR*.

Jian Chen and 1 others. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Nadezhda Chirkova and 1 others. 2024. Retrieval-augmented generation in multilingual settings. *arXiv preprint arXiv:2407.01463*.

Jonathan Clark and 1 others. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024.

Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Google DeepMind. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

P. Hrudya and 1 others. 2025. A multilayered approach to identifying social media events using LLM. In *Information Systems for Intelligent Systems*.

Junjie Hu and 1 others. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. In *IEEE Transactions on Big Data*. ArXiv preprint arXiv:1702.08734.

Vishnu Kanjirangat and 1 others. 2025. Tokenization and representation biases in multilingual models on dialectal NLP tasks. In *Proceedings of EMNLP*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergei Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

Dongho Ki and 1 others. 2025. Linguistic nepotism: Trading-off quality for language preference in multilingual RAG. *arXiv preprint arXiv:2509.13930*.

G. S. Kohli and 1 others. 2024. Building a LLaMA2-finetuned LLM for Odia language utilizing domain knowledge instruction set. In *Proceedings of AI-ML Systems*.

Patrick Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xi Lin and 1 others. 2022. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2203.08300*.

Pengfei Liu and 1 others. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. 2025. XRAG: Cross-lingual retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of EMNLP*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of ACL*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*.

A. R. Nair and D. Gupta. 2024. Evaluating performance and accuracy of large language models in translating code-mixed Hindi to English: A comparative study. In *Proceedings of INDICON*.

A. Padmavilochanan and 1 others. 2025. Multimetric evaluation of LLMs on major Indian language translation tasks. In *Proceedings of ASIANCON*.

Edoardo Ponti and 1 others. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00206*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

N. Pradeesh and 1 others. 2025. Retrieval-augmented generation for multiple-choice questions and answers generation. *Procedia Computer Science*.

Jian Qi and 1 others. 2025. On the consistency of multilingual context utilisation in retrieval-augmented generation. *arXiv preprint arXiv:2504.00597*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

J. R. and 1 others. 2024. A few-shot multi-accented speech classification for Indian languages using transformers and LLM fine-tuning approaches. In *Proceedings of the DravidianLangTech Workshop*.

A. Rajagopalan Nair, D. Gupta, B. Paul, and J. Siva Bhavani. 2026. Enhancing low-resource Indian language machine translation using large language models with preference optimization and hypergeometric-gamma reward. *IEEE Access*, 14:1641–1665.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *Proceedings of ICML*.

Freda Shi, Mirac Suzgun, Markus Freitag, and 1 others. 2023b. Language models are multilingual chain-of-thought reasoners. In *Proceedings of ICLR*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, and 1 others. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of NAACL*.

Jayasimha Talur and 1 others. 2025. MEMERAG: A multilingual end-to-end meta-evaluation benchmark for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hugo Touvron and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Utter Project. 2024. EuroLLM: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Shijie Wu and 1 others. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *arXiv preprint arXiv:2410.21970*.

Rongwu Xu and 1 others. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of EMNLP*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR*.

Ahmet Üstün and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Implementation Details

Experiments are implemented using Hugging Face Transformers and Datasets. Inference is performed on NVIDIA A100 GPUs with automatic device mapping (`device_map="auto"`).

All models use greedy decoding (`do_sample=False`) with a maximum of 40 new tokens. Models are loaded in `float16` precision. Where VRAM is constrained, 4-bit NF4 double quantization is applied.

Evaluation is conducted in two phases: (i) answer generation and metric computation, and (ii) BERTScore evaluation after releasing the generative model from GPU memory to avoid out-of-memory issues.

We evaluate up to 200 samples per language per condition and report macro-averaged results across languages. Best-strategy selection by F1 is reported separately as an oracle upper bound.

B Per-Language Analysis

Per-language performance varies notably across models, as shown in Table 9. While script compliance (SC) remains consistently high, semantic overlap (SEM) differs significantly across languages. Gemma-2-9B and Aya-23-8B achieve strong multilingual performance, whereas Qwen2.5-7B remains relatively balanced. EuroLLM-9B shows moderate but stable results, while LLaMA-2-7B underperforms, particularly on non-English languages.

C Prompting Strategies

We evaluate five prompting strategies across all datasets and configurations. Each strategy controls the level of instruction specificity and grounding constraints imposed on the model.

C.1 Prompt Types

Minimal A basic extractive QA prompt with no explicit constraints beyond answering from the context.

Explicit (English Instruction) Adds explicit instruction in English to use exact words from the context.

```
  Context:
{context}

  Question: {question}

  Answer in {language} using EXACT words
from the context. No extra text.
Answer:
```

Explicit (Native Instruction) Provides instructions in the target language to enforce language and script consistency.

```
  Context:
{context}

  Question: {question}

  [Instruction in target language: answer using
exact words from context]
Answer:
```

Strict Imposes hard constraints on answer format (length and extractiveness).

```
  Context:
{context}

  Question: {question}

  RULE: Copy the answer phrase exactly from the
context.
Maximum 5-6 words. No explanation.
Answer:
```

Few-shot Includes a demonstration example before the query.

```
  Example:
Context: Paris is the capital of France.
Question: What is the capital of France?
Answer: Paris

  Now answer the following:

  Context:
{context}

  Question: {question}
Answer:
```

C.2 Dataset-Specific Variations

XQuAD (No-RAG) Prompts follow a standard context-question-answer format using the gold passage. In English, a question-first format is used:

```
  Question: {question}

  Context:
{context}

  Copy the answer EXACTLY from the context.
Answer:
```

TyDiQA (No-RAG) Prompts are identical to the RAG setup but use the full gold passage as context without retrieval. Language-specific constraints (e.g., script enforcement for Russian, native instructions) are applied.

TyDiQA (RAG) Prompts are constructed over retrieved evidence chunks. The context is formed by concatenating top-ranked passages retrieved using a multilingual dense retriever.

Language-specific adaptations include:

- Script enforcement (e.g., Cyrillic-only responses for Russian)
- Language-specific few-shot examples
- Minimal prompting for languages prone to hallucination

```
   Context:
{retrieved_passages}

   Question: {question}

   Answer using exact words from the context.
Answer:
```

C.3 Language-Specific Prompt Adaptations

In addition to shared prompt templates, several language-specific adaptations are applied to improve generation quality and ensure script consistency.

Script Enforcement For languages with distinct writing systems (e.g., Russian, Greek, Hindi), prompts explicitly enforce output in the correct script (e.g., Cyrillic for Russian, Devanagari for Hindi). This reduces cross-script leakage and improves language fidelity metrics such as Language Match Rate.

Native-Language Instructions For the explicit-native strategy, instructions are provided entirely in the target language rather than translated templates. This improves alignment with the input context and encourages more accurate extraction.

English Prompt Structure For English, a question-first format is used ("Question–Context–Answer"), while other languages follow a context-first format. This aligns with standard English QA benchmarks and improves extraction accuracy.

Indonesian Simplification For Indonesian, all prompt strategies are simplified to a minimal instruction format. This design choice mitigates hallucination observed with more complex prompting strategies in preliminary experiments.

Language-Specific Few-Shot Examples Few-shot prompts include examples written in the target language and script, ensuring that demonstrations reflect the expected answer format and linguistic structure.

These adaptations enable controlled evaluation of prompting effects across languages while accounting for script, structure, and dataset-specific characteristics. To quantify the impact of these prompting strategies across models and languages, Table 9 reports the best-performing strategy and corresponding F1 score under both RAG and No-RAG settings for TyDiQA.

D Evaluation Details

Normalization We apply language-specific normalization including Arabic alef unification, English article removal, Finnish stemming, and Vietnamese diacritic normalization.

E Retrieval Effects and EU Analysis

This section presents additional analysis of retrieval effects and evidence utilisation across models. Figure 2a illustrates the impact of retrieval on performance, while Figure 2b highlights the evidence utilisation gap across models.

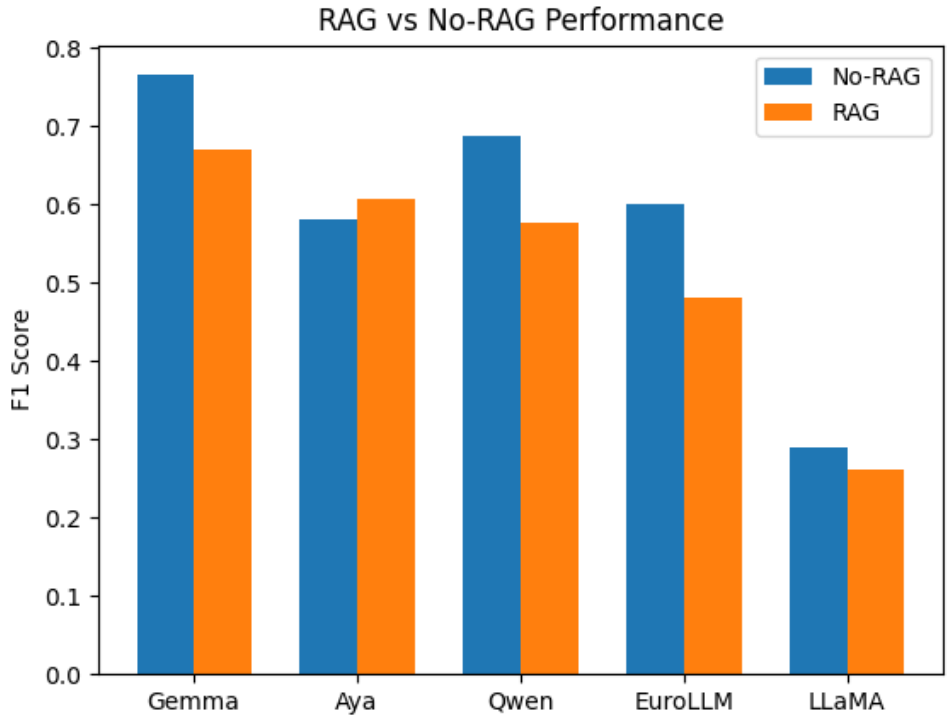
RAG			
Model	**Lang**	**Strat**	**F1**
Aya	Ar	min	0.761
	Bn	exp_na	0.652
	En	exp_na	0.611
	Fi	exp_na	0.484
	Id	min	0.671
	Ko	exp_en	0.795
	Ru	min	0.633
	Sw	fs	0.493
	Te	exp_na	0.357
Qwen	Ar	exp_en	0.469
	Bn	exp_en	0.679
	En	exp_en	0.273
	Fi	exp_en	0.431
	Id	min	0.344
	Ko	exp_en	0.569
	Ru	exp_en	0.431
	Sw	exp_en	0.413
	Te	exp_en	0.581
Gemma	Ar	fs	0.705
	Bn	fs	0.739
	En	fs	0.549
	Fi	fs	0.609
	Id	min	0.616
	Ko	fs	0.696
	Ru	min	0.565
	Sw	fs	0.732
	Te	fs	0.812
EuroLLM	Ar	fs	0.587
	Bn	fs	0.441
	En	fs	0.377
	Fi	fs	0.498
	Id	min	0.487
	Ko	fs	0.697
	Ru	fs	0.417
	Sw	fs	0.380
	Te	fs	0.190
LLaMA	Ar	str	0.224
	Bn	str	0.268
	En	str	0.335
	Fi	str	0.413
	Id	min	0.205
	Ko	str	0.253
	Ru	str	0.282
	Sw	str	0.205
	Te	min	0.162

No-RAG			
Model	**Lang**	**Strat**	**F1**
Aya	Ar	fs	0.513
	Bn	min	0.516
	En	min	0.746
	Fi	min	0.602
	Id	min	0.683
	Ko	min	0.663
	Ru	min	0.642
	Sw	min	0.465
	Te	str	0.311
Qwen	Ar	fs	0.613
	Bn	fs	0.761
	En	fs	0.445
	Fi	fs	0.494
	Id	fs	0.517
	Ko	fs	0.672
	Ru	fs	0.534
	Sw	fs	0.536
	Te	str	0.609
Gemma	Ar	fs	0.725
	Bn	fs	0.874
	En	fs	0.731
	Fi	fs	0.715
	Id	min	0.768
	Ko	fs	0.807
	Ru	fs	0.619
	Sw	fs	0.792
	Te	fs	0.881
EuroLLM	Ar	fs	0.678
	Bn	str	0.578
	En	str	0.714
	Fi	str	0.609
	Id	min	0.658
	Ko	fs	0.744
	Ru	str	0.622
	Sw	fs	0.580
	Te	fs	0.395
LLaMA	Ar	min	0.111
	Bn	str	0.268
	En	str	0.335
	Fi	min	0.296
	Id	min	0.205
	Ko	min	0.100
	Ru	min	0.226
	Sw	min	0.145
	Te	min	0.162

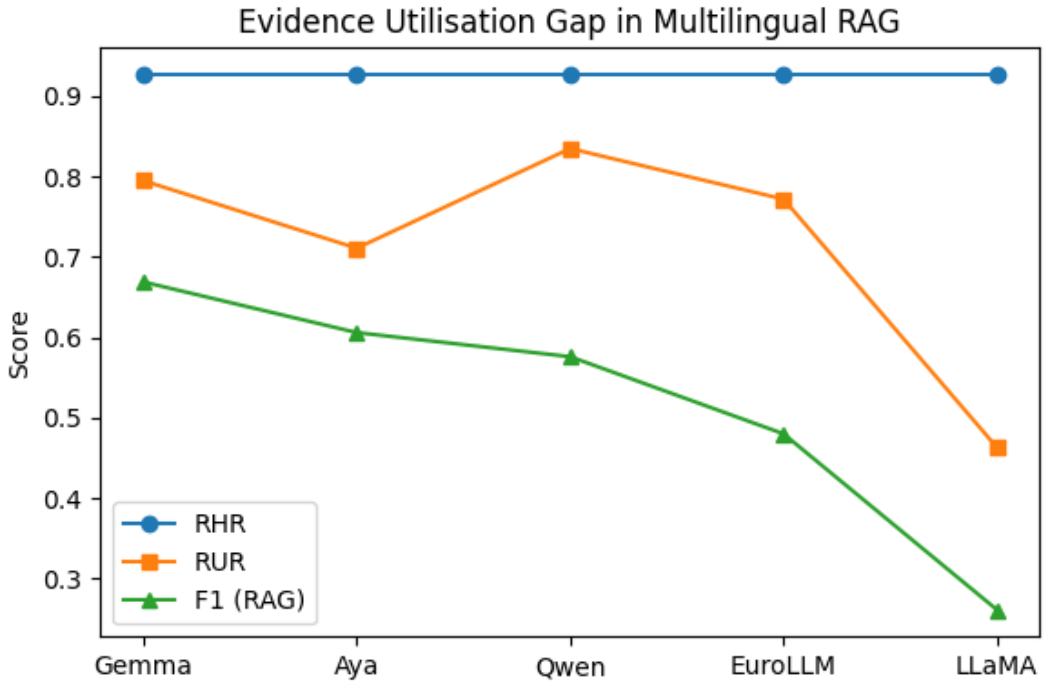
Table 8: Best prompt strategy and F1 score per language under RAG and No-RAG.

Gemma-2-9B / Aya-23-8B				Qwen / EuroLLM / LLaMA			
Model	**Lang**	**SEM**	**SC**	**Model**	**Lang**	**SEM**	**SC**
Gemma-2-9B	Ar	0.75	0.99	Qwen2.5-7B	Ar	0.70	0.99
	Bn	0.77	0.89		Bn	0.81	0.97
	En	0.58	1.00		En	0.50	1.00
	Fi	0.61	1.00		Fi	0.51	0.99
	Id	0.68	1.00		Id	0.63	1.00
	Ko	0.71	0.88		Ko	0.68	0.90
	Ru	0.58	0.99		Ru	0.58	0.99
	Sw	0.76	1.00		Sw	0.58	1.00
	Te	0.83	0.90		Te	0.70	0.99
Aya-23-8B	Ar	0.77	1.00	EuroLLM-9B	Ar	0.655	0.758
	Bn	0.70	0.99		Bn	0.589	0.778
	En	0.63	1.00		En	0.645	0.755
	Fi	0.50	1.00		Fi	0.649	0.818
	Id	0.67	1.00		Id	0.660	0.797
	Ko	0.80	0.96		Ko	0.621	0.655
	Ru	0.66	0.99		Ru	0.626	0.767
	Sw	0.50	1.00		Sw	0.615	0.778
	Te	0.44	0.99		Te	0.543	0.596
				LLaMA-2-7B	Ar	0.31	0.67
					Bn	0.39	0.62
					En	0.46	1.00
					Fi	0.53	1.00
					Id	0.31	1.00
					Ko	0.30	0.35
					Ru	0.47	0.77
					Sw	0.32	1.00
					Te	0.16	0.39

Table 9: Per-language Semantic Overlap (SEM) and Script Compliance (SC) under best prompting strategy (TyDiQA RAG).



(a) Introducing retrieval reduces performance for most models, indicating interference between parametric and retrieved knowledge.



(b) RHR, RUR, and F1 across models, showing an evidence utilisation gap. RHR is model-independent and shown as a reference.

Figure 2: Comparison of RAG performance and evidence utilisation across models.