Enhancing Multilingual RAG Systems with Debiased Language Preference-Guided Query Fusion

Jeonghyun Park, Byeongjeong Kim, Seojin Hwang, Hwanhee Lee[†]
Chung-Ang University, Seoul, Korea
{tom0365, michael97k, swiftie1230, hwanheelee}@cau.ac.kr
https://jeonghyunpark2002.github.io/DELTA_project_page

Abstract

Multilingual Retrieval-Augmented Generation (mRAG) systems often exhibit a perceived preference for high-resource languages, particularly English, resulting in the widespread adoption of English pivoting. While prior studies attribute this advantage to the superior English-centric capabilities of Large Language Models (LLMs), we find that such measurements are significantly distorted by structural priors inherent in evaluation benchmarks. Specifically, we identify *exposure bias* and a *gold availability prior*—both driven by the disproportionate concentration of resources in English—as well as *cultural priors* rooted in topic locality, as factors that hinder accurate assessment of genuine language preference. To address these biases, we propose **DeLP** (**De**biased **L**anguage **P**reference), a calibrated metric designed to explicitly factor out these structural confounds. Our analysis using DeLP reveals that the previously reported English preference is largely a byproduct of evidence distribution rather than an inherent model bias. Instead, we find that retrievers fundamentally favor monolingual alignment between the query and the document language. Building on this insight, we introduce **DELTA** (**DE**biased **L**anguage preference–guided **T**ext **A**ugmentation), a lightweight and efficient mRAG framework that strategically leverages monolingual alignment to optimize cross-lingual retrieval and generation. Experimental results demonstrate that DELTA consistently outperforms English pivoting and mRAG baselines across diverse languages. The Code is available at https://github.com/jeonghyunpark2002/DELTA.git

1 Introduction

Multilingual Retrieval-Augmented Generation (mRAG) (Chirkova et al., 2024) generalizes Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) by retrieving evidence from multilingual knowledge sources. This enables language models to produce responses that are not only factually grounded but also sensitive to the user's language and linguistic context (Rau et al., 2024). In this landscape, mRAG systems frequently exhibit a significant language preference for English (Zhang et al., 2023; Park and Lee, 2025). Consequently, English pivoting—the practice of translating a non-English query into English before retrieval—has emerged as a surprisingly strong heuristic that yields substantial gains across many languages (Chirkova et al., 2024; Ranaldi et al., 2025). Prior works have largely attributed this advantage to the "English-centric" competence of generators, such as superior reasoning in English or reduced translation noise, leading to research that primarily intervenes at the generation stage (Chirkova et al., 2024; Li et al., 2025; Moon et al., 2025).

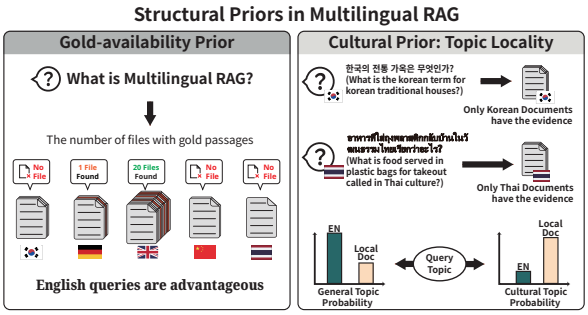

Figure 1: Common causes of language preference of mRAG: gold-availability prior and cultural prior.

However, we find that the perceived effectiveness of English pivoting is primarily driven by retrieval-side structural biases rather than any inherent linguistic preference of the model. As illustrated in Figure 1, we identify two major confounders: a **gold availability prior** and **cultural priors**. Our analysis shows that ground-truth evidence in standard benchmarks is overwhelmingly concentrated in English resources, establishing a dominant gold availability prior. This concentration

[†]Corresponding author.

not only makes English the sole or primary location where correct evidence exists, but also leads to **exposure bias**—the retrieval system's inherent tendency to surface documents from more prevalent languages regardless of the query's intended language—of English documents during retrieval, further amplifying English's advantage. Together, these effects fundamentally distort measured language preference and inflate the apparent superiority of English. In addition, we identify **cultural priors** as an equally critical factor. There are benchmark questions that are tied to specific geographic or cultural contexts and contain native surface forms (e.g., local titles, aliases, and scripts) that act as strong retrieval anchors. When benchmarks overrepresent such locale-specific topics, languages may appear "preferred" due to query–corpus alignment and environmental exposure rather than the model's intrinsic preference. Critically, these structural factors contaminate existing methods (Park and Lee, 2025) for measuring language preference.

To reveal the intrinsic preference of mRAG systems, we propose Debiased Language Preference (**DeLP**), a calibrated measurement that explicitly regresses out these structural confounds. DeLP utilizes a ridge regression framework to predict observed language preference from structural factors (e.g., corpus size, gold availability, cultural prior), treating the residual signal as the true, debiased preference of the model. By applying DeLP, we reveal a qualitatively different landscape: the previously inflated preference for English largely evaporates. Instead, our results show an increased preference for **monolingual alignment**, where the retriever performs most effectively when the query and the target document languages match.

Building on the discovery of monolingual alignment, we introduce DEbiased Language preference–guided Text Augmentation (**DELTA**), a lightweight query-level solution for mRAG. DELTA leverages the debiased preference signals from DeLP to dynamically identify intrinsic model preference for a given query, effectively bridging the gap between the user's query and the languages where the model performs most reliably. By reformulating the query to include these preference-aligned multilingual anchors, DELTA preserves the native script's context while maximizing the benefits of monolingual alignment. DELTA is highly cost-effective, requiring no modifications to the underlying corpus or retriever architecture. Our experiments demonstrate that DELTA outperforms naive English pivoting, proving that accounting for the model's true linguistic preference—rather than following biased environmental cues—is the key to unlocking the true potential of mRAG systems.

2 The Myth of English Preference: Structural Priors in mRAG

A dominant mRAG strategy is *English pivoting*, where non-English queries are translated into English to exploit the perceived superiority of English-centric models. We hypothesize that these gains are not necessarily indicative of model preference but rather reflect a massive exposure bias rooted in the structural distribution of evidence.

2.1 Experimental Setup

Datasets We conduct our analysis on MKQA (Longpre et al., 2021), which provides 10k professionally translated queries. To enable precise measurement of evidence location, we use a 2.7K-example subset that overlaps with KILT NQ [*]. Since MKQA does not provide standardized provenance for each translated instance, using KILT allows us to inherit document-level provenance (i.e., gold Wikipedia passage IDs), which is essential for quantifying gold availability across different linguistic corpora.

Models and Knowledge Sources We employ BGE-m3 (Chen et al., 2024) as the multilingual retriever and re-ranker. For the generation, we use three recently released robust multilingual LLMs: Qwen3-235B (Yang et al., 2025), DeepSeek-v3.1 (Liu et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025). We retrieve top-50 candidate documents per query and apply re-ranking, using the top-5 documents as contexts for generation. In line with previous work (Chirkova et al., 2024; Park and Lee, 2025), we use Wikipedia editions in English and the user's local language to serve as the knowledge sources. Detailed corpus statistics are provided in Appendix D.

2.2 Linguistic Superiority or Data Imbalance?

Following the MKQA protocol anchored to KILT (Longpre et al., 2021), we identify the location of gold passages (WPIDs) within the multilingual Wikipedia datastore. We report this distribution as Gold Availability, as the number of queries whose gold passage WPID is present in that

[*] https://huggingface.co/datasets/facebook/kilt_tasks

	Gold Availability		Retriever Recall		Qwen3-235B-A22B		Gemini-2.5-Flash		DeepSeek-Chat-v3.1	
lang	**#q**	**ratio**	**Base**	**EN**	**Base**	**EN**	**Base**	**EN**	**Base**	**EN**
en	26934	**73.29%**	–	–	70.05(EN)	–	58.26(EN)	–	60.77(EN)	–
ar	214	0.58%	13.36	23.57	47.79	55.14	40.79	48.44	43.64	50.97
de	435	1.18%	21.62	26.40	63.81	60.72	53.52	55.17	54.16	56.92
ja	513	1.40%	16.84	25.83	46.60	59.29	44.26	53.68	44.72	56.52
ko	306	0.83%	15.62	24.81	40.14	54.57	35.97	47.67	34.21	50.49
th	187	0.51%	21.90	27.55	40.73	60.46	31.65	54.86	36.80	57.52
zh	287	0.78%	16.47	26.53	37.52	59.53	30.81	53.59	33.14	56.11

Table 1: Gold availability bias and its impact on multilingual RAG. Gold Availability measures gold-passage coverage per language, and Retriever Recall reports Recall@50. Model columns show end-to-end accuracy. Base denotes native-language queries, and EN denotes English-translated queries.

language's Wikipedia corpus for each language of query. This distribution reflects the extent of corpus-level coverage of gold evidence in each language within the benchmark. We then relate this to retrieval performance, measured by Recall@50—i.e., the fraction of queries whose gold passages appear within the top-50 retrieved candidates—under both native-language queries and English queries (EN). We further evaluate end-to-end mRAG performance using character 3-gram recall between the generated and reference answers. Details are provided in Appendix H.

Our analysis in Table 1 reveals an extreme imbalance in the retrieval environment. English Wikipedia provides substantially higher document density and coverage, introducing a strong **exposure bias**. More critically, for a vast majority of queries, English Wikipedia also serves as the sole repository of ground-truth, inducing a dominant **gold availability prior**. Consequently, English pivoting appears effective not because models prefer English, but because of this structural skew, sustaining the long-standing "myth" of English preference.

ar	de	es	fr	ja	ko	ru	zh
21.43%	16.67%	18.52%	21.74%	14.29%	12.50%	25.00%	6.67%

Table 2: Local-gold coverage by predicted L_{loc}.

2.3 Impact of Cultural Priors

Queries often carry cultural or regional context, whose associated language can naturally align with local-language evidence and be conflated with language preference; we therefore examine where their gold documents reside across Wikipedia languages. We first isolate queries that involve cultural or regional contexts by instructing GPT-4o-mini (Hurst et al., 2024) to predict a single primary locale language L_{loc}, selecting the language that corresponds to the query's main referenced region or culture (Details on the classifier are in Appendix J). Table 2 reports the local-gold rate $p(\text{gold WPID exists in local Wikipedia} \mid L_{\text{loc}})$ for each predicted language. Local evidence is not uniformly absent—across several predicted locale languages, about 20% of these queries have gold pages only in the corresponding local Wikipedia. This distribution introduces a structural bias for retrieval to rely on locale-specific surface-form anchors (e.g., native titles, aliases, scripts) when they exist. As a result, observed language preference can be influenced by locale-tied queries and their local-gold presence, motivating an explicit cultural prior term p_{cult} to avoid conflating topic locality.

3 Measuring Language Preference via Bias Calibration

The structural priors identified in Section 2 suggest that existing metrics, such as MLRS (Park and Lee, 2025), fail to distinguish between a model's intent to use a language and the external necessity imposed by data distribution. To reveal the genuine preference, we introduce **De**biased **L**anguage **P**reference (**DeLP**), a calibrated measurement framework that explicitly regresses out structural confounds.

3.1 Decomposing Structural Bias in mRAG

To isolate intrinsic model preference, we decompose the confounding factors identified in our previous analysis into three primary priors.

Exposure prior (p_{ret}). As observed in the exposure bias of Section 2.2, high-resource corpora—particularly English—dominate the top retrieval results regardless of the encoder's linguistic intent. This prior captures the "popularity bias" of the datastore. A language that appears more frequently in the candidate pool is more likely to be retrieved, potentially leading to a false inflation of preference. Let L_q denote the query language and L_d the language of a retrieved document. We

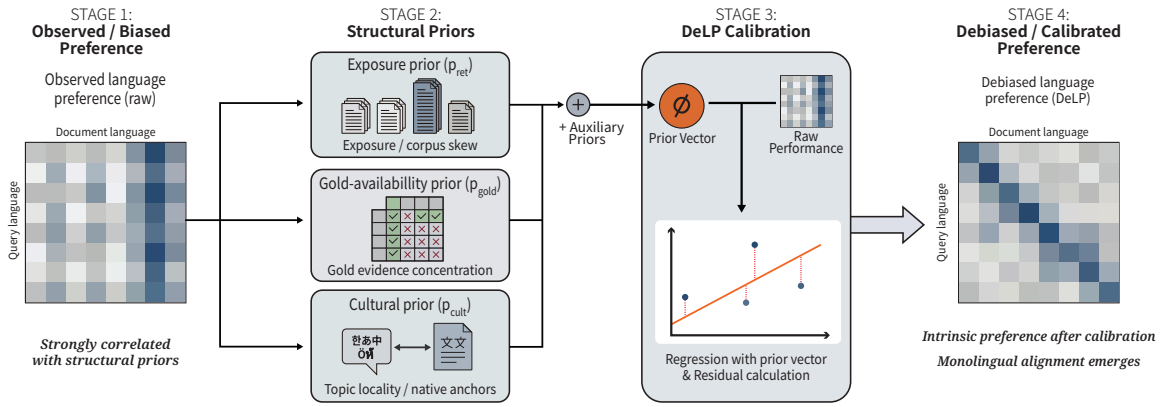

Figure 2: Overview of DeLP, which measures intrinsic language preference in mRAG by regressing out exposure, gold-availability, and cultural priors from raw preference signals.

estimate $p_{\text{ret}}(L_d \mid L_q)$ by calculating the average proportion of document language L_d within the top-50 candidates for queries in L_q.

Gold-availability prior (p_{gold}). Our findings in Table 1 (Section 2.2) demonstrate that retrieval is often forced into English because the gold evidence simply does not exist elsewhere. To prevent such uncontrollable circumstances from being mistaken for model preference, we explicitly model the availability of ground-truth passages. We estimate $p_{\text{gold}}(L_q, L_d)$ on the MKQA–KILT overlap as the empirical fraction of queries in L_q for which gold evidence is present in the L_d corpus.

Cultural prior (p_{cult}). As discussed in Section 2.3, locale-tied queries contain native surface forms that act as structural anchors, naturally pulling retrieval toward the corresponding local-language evidence. This prior captures topic locality; a retriever might show high preference scores for a language simply because the topic is regional. We estimate $p_{\text{cult}}(L_d)$ by identifying the query's associated locale language L_{loc} using *GPT-4o-mini* and computing the fraction of queries where $L_{\text{loc}} = L_d$. The classifier selects the local language of the primary referenced place or culture (e.g., "When did Hong Kong go back to China?" $\rightarrow$ *zh*), reserving en for inherently English-speaking or genuinely global queries. For the details of how we measure cultural prior, please refer to Appendix J.

3.2 Calibration for Genuine Preference

We calibrate the raw language preference scores with respect to the above priors to obtain the residual signal-the component not explained by the priors, which reflects the model's genuine language preference. In addition to the above three main priors, we incorporate two auxiliary structural controls, namely a corpus-size prior p_{db} and a passage-length statistic ℓ, as additional covariates in the prior feature vector $\phi(L_q, L_d)$ (defined below) to account for language-dependent corpus scale and length effects. Let $s_e(L_q, L_d)$ denote the observed language-preference score of the encoder e for the query language L_q, and for the language of evidence L_d. We instantiate s_e by MLRS (Park and Lee, 2025) (Table 10), which measures how often the retriever surfaces evidence in each L_d for a fixed L_q. For each language pair (L_q, L_d), we define a prior feature vector $\phi(L_q, L_d) \in \mathbb{R}^7$:

$$\phi(L_q, L_d) = \begin{bmatrix} 1 \\ \log\big(p_{\text{ret}}(L_d \mid L_q) + \epsilon\big) \\ \log\big(p_{\text{db}}(L_d) + \epsilon\big) \\ \log\big(\ell(L_d) + \epsilon\big) \\ \log\big(p_{\text{gold}}(L_q, L_d) + \epsilon\big) \\ \log\big(p_{\text{cult}}(L_d) + \epsilon\big) \\ \mathbb{I}[L_q = L_d] \end{bmatrix}. \tag{1}$$

where p_{ret} is the exposure prior, p_{db} is the corpus-size prior, ℓ is a passage-length statistic (e.g., median length), p_{gold} is the gold-availability prior, and p_{cult} is the cultural prior, and $\epsilon > 0$ is a small constant added for numerical stability in the log transformation. The vector $\phi(L_q, L_d)$ stacks interpretable covariates that predict s_e without invoking intrinsic model preference. We use log-transformed priors to compress heavy-tailed probabilities and corpus statistics, and make linear effects more reasonable across languages. The indicator $\mathbb{I}[L_q = L_d]$ allows the model to treat same-language retrieval as a special case, ensuring that monolingual matching is not forced to be explained solely by external priors.

Ridge calibration. We fit the regression separately for each encoder e to learn how much of its observed score s_e can be attributed to structural priors. We use ridge regularization to stabilize coefficients under the various priors, preventing any

Query Lang.	$L_q = L_d$	$L_q \neq L_d$							
		en	*ko*	*zh*	*fr*	*ja*	*it*	*pt*	*es*
en	**50.10** (56.79)	–	36.67 (33.94)	40.34 (33.99)	35.57 (37.57)	39.61 (34.18)	35.49 (36.79)	36.46 (36.54)	36.02 (37.49)
ko	43.38 (42.21)	37.69 (44.36)	–	41.75 (35.44)	34.90 (36.84)	**43.59** (38.22)	34.70 (36.00)	35.65 (35.71)	34.77 (36.24)
zh	**50.60** (45.81)	38.68 (45.35)	37.95 (35.06)	–	34.73 (36.73)	41.90 (36.51)	34.87 (36.21)	35.84 (35.91)	35.22 (36.69)
fr	40.05 (43.74)	**41.16** (47.84)	36.77 (34.03)	40.50 (34.16)	–	39.92 (34.50)	36.00 (37.31)	36.69 (36.76)	36.28 (37.76)
ja	**49.19** (45.50)	38.70 (45.37)	38.40 (35.69)	41.56 (35.24)	34.94 (36.94)	–	34.99 (36.29)	35.97 (36.04)	35.23 (36.70)
it	39.05 (41.72)	**40.63** (47.30)	36.85 (34.12)	40.59 (34.25)	36.63 (38.64)	39.86 (34.44)	–	37.01 (37.09)	36.91 (38.39)
pt	**46.08** (39.76)	40.55 (47.23)	36.98 (34.24)	40.63 (34.29)	36.50 (38.52)	40.01 (34.59)	36.48 (37.80)	–	37.73 (39.21)
es	38.19 (41.30)	**40.71** (47.39)	36.86 (34.13)	40.39 (34.04)	36.30 (38.31)	39.76 (34.34)	36.45 (37.76)	37.25 (37.32)	–

Table 3: DeLP for query-document language pairs, averaged over three encoders (raw MLRS in parentheses). Background shading is row-wise min-max scaled (darker = stronger preference); dashes denote $L_q = L_d$. Underline denotes the second-highest per row. The relatively stronger English and Chinese signals are attributable to encoder training-data language imbalance (e.g., BGE-m3 trains on 194 languages with English 43.9% and Chinese 20.5%).

single feature from disproportionately absorbing the preference signal. Let $\mathcal{C}$ be the set of all language pairs used for calibration. For each encoder e, we fit a ridge regression that predicts the raw score from priors:

$$\hat{\beta}_e = \arg\min_{\beta} \mathcal{J}_e(\beta),$$
$$\mathcal{J}_e(\beta) = \sum_{(L_q, L_d) \in \mathcal{C}} \left(s_e(L_q, L_d) - \phi(L_q, L_d)^\top \beta\right)^2 + \lambda \|\beta\|_2^2. \quad (2)$$

where λ is a regularization hyperparameter and ϵ is a small constant for numerical stability.

Debiased preference (DeLP). We define the debiased preference as the residual signal after removing the component explained by structural priors:

$$r_e(L_q, L_d) = s_e(L_q, L_d) - \phi(L_q, L_d)^\top \hat{\beta}_e. \quad (3)$$

The residual $r_e(L_q, L_d)$ represents the portion of the observed score that is independent of structural priors. To keep the overall scale comparable to the raw score, we re-center the residuals by the global mean of raw scores μ_e:

$$\mathrm{DeLP}_e(L_q, L_d) = r_e(L_q, L_d) + \mu_e,$$
$$\mu_e = \frac{1}{|\mathcal{C}|} \sum_{(L_q, L_d) \in \mathcal{C}} s_e(L_q, L_d). \quad (4)$$

By adding back μ_e, DeLP stays on a numeric scale comparable to standard MLRS tables while preserving the relative differences that define the model's intrinsic tendencies. To mitigate potential encoder-specific bias, we apply our calibration procedure independently to each retriever and report all debiased results for three multilingual encoders: BGE-m3 (Chen et al., 2024) and two Sentence-BERT variants (Reimers and Gurevych, 2019), `paraphrase-multilingual-MiniLM-L12-v2` and `paraphrase-multilingual-mpnet-base-v2`. We denote them as `p-mMiniLM` and `p-mMpNet` for compactness in tables.

	p_{ret}		p_{gold}		p_{cult}	
Encoder	**MLRS**	**DeLP**	**MLRS**	**DeLP**	**MLRS**	**DeLP**
`bge-m3`	0.994	0.142	0.914	0.336	0.916	0.335
`p-mMiniLM`	0.997	0.145	0.915	0.321	0.917	0.320
`p-mMpNet`	0.996	0.131	0.917	0.311	0.920	0.310

Table 4: Pearson's r between preference and priors for before (MLRS) and after (DeLP) calibration.

Emergence of Monolingual Alignment. After calibration, we find that the preference landscape shifts qualitatively from the raw preference as in Table 3. The previously dominant English preference largely disappears, and the strongest signal consistently moves to the diagonal ($L_q = L_d$). This reveals that retrievers fundamentally favor monolingual alignment—the matching of query and document in the same language. We also observe that queries favor the linguistically or regionally related languages, such as Korean with Japanese. Overall, the DeLP score suggests that much of the apparent English preference in prior protocols was induced by structural priors, while the residual preference signal is dominated by query-language alignment and interpretable related-language effects. For a more detailed DeLP score, refer to Appendix G.

Correlation Analysis. To validate DeLP, we compute the correlation between preference scores and priors before and after calibration as shown in Table 4. Raw scores (MLRS) are highly correlated with all three priors (exposure, gold-availability, and cultural), suggesting that existing language-

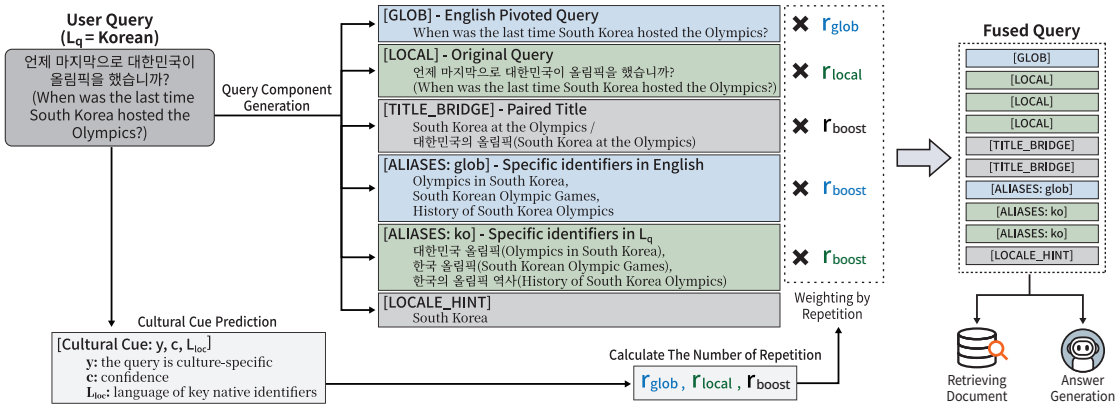

Figure 3: Overview of DELTA query fusion. DELTA fuses global and local query segments into a single preference-aligned query using lightweight repetition-based weighting.

preference measurements largely reflect prior-driven preference rather than intrinsic model preference. After calibration, these correlations drop sharply after applying DeLP. This confirms that DeLP effectively decouples intrinsic model tendencies from the structural signals.

4 Debiasing mRAG through Preference-Aligned Augmentation

Monolingual alignment in Section 3 reveals that *retrievers intrinsically perform best* when the *query language matches the document language*. This suggests that while English pivoting provides coverage (due to gold availability), it often sacrifices the retrieval anchors present in the user's native tongue. Motivated by this point, we propose DELTA (**DE**biased **L**anguage preference guided **T**ext **A**ugmentation), a lightweight query reformulation strategy that injects preference-aligned cues into a single fused query.

4.1 Query Fusion with Native Anchors

DELTA aims to maximize the benefits of both global coverage and local discriminative matching. As illustrated in Figure 3, given a local question q_{local}, DELTA constructs an English pivot q_{glob} and extracts a set of cultural identifiers—canonical titles $(t_{\text{glob}}, t_{\text{loc}})$, aliases, and a regional hint—using a frozen LLM (Instruction in Appendix N). These elements are then concatenated into a single query Q_{fused} composed of five segments, each optionally weighted by cultural cues' confidence score:

- `[GLOB]`: The English pivot q_{glob}.
- `[LOCAL]`: The original query q_{local} to leverage monolingual alignment.
- `[TITLE_BRIDGE]`: Paired titles $(t_{\text{glob}}, t_{\text{loc}})$ to facilitate cross-lingual mapping.
- `[ALIASES]` and `[LOCALE_HINT]`: Specific identifiers that serve as stable retrieval anchors.

4.2 Repetition-based Weighting

To implement the debiased preference control, DELTA utilizes a repetition-based weighting policy(Wang et al., 2023). We first predict a cultural cue (y, c, L_{loc}), where $y \in \{0, 1\}$ indicates whether the query is culture-specific, $c \in [0, 1]$ is the confidence score, and L_{loc} is the language of the key native identifiers. This cue determines how strongly we upweight locale-specific blocks versus the global English back-off when forming the fused query Q_{fused}. We map the confidence score c into three discrete repetition levels using two thresholds τ_{low} and τ_{high}. We then set repetition counts for the local block `[LOCAL:`L_{loc}`]` and the global pivot block `[GLOB]` as:

$$r_{\text{local}} = \begin{cases} 1 + \mathbb{I}[c \geq \tau_{\text{low}}] + \mathbb{I}[c \geq \tau_{\text{high}}], & y = 1 \\ 1, & y = 0 \end{cases}$$
$$r_{\text{glob}} = \begin{cases} 1 + \mathbb{I}[c < \tau_{\text{low}}], & y = 1 \\ 2, & y = 0 \end{cases} \quad (5)$$

$c < \tau_{\text{low}}$ triggers no additional upweighting, $\tau_{\text{low}} \leq c < \tau_{\text{high}}$ adds one extra repetition, and $c \geq \tau_{\text{high}}$ adds two, yielding $r_{\text{local}} \in \{1, 2, 3\}$ and $r_{\text{glob}} \in \{1, 2\}$. Intuitively, we upweight high-confidence culture-specific queries toward the local expression to preserve culturally grounded identifiers, while non-culture-specific queries mildly favor the global pivot for robust back-off. In addition, when y=1 and $c \geq \tau_{\text{boost}}$, we duplicate local-side disambiguation anchors (i.e., `[TITLE_BRIDGE]` and `[ALIASES]`) once more to further emphasize native surface-form anchors and reduce entity ambiguity. Overall, DELTA realizes preference control via text-only weighting, and all concrete hyperparameter values are reported in Appendix K.

Method	*en*	*ar*	*es*	*zh*	*ja*	*de*	*ko*	*th*	AVG ↑	Latency ↓
Qwen3-235b-a22b-2507										
Document Level										
MultiRAG	70.05	47.79	**63.76**	37.52	46.60	**63.81**	40.14	40.73	51.30	1.38
CrossRAG	68.21	43.95	61.14	37.81	44.75	60.16	38.13	42.87	49.63	1.29
DKM-RAG	69.13	42.69	62.12	35.13	43.90	61.13	39.49	38.88	49.06	3.80
QTT-RAG	**70.11**	46.44	63.02	37.68	46.94	62.79	44.13	42.12	51.65	1.80
Query Level										
English Translation	-	55.14	61.94	59.53	59.29	60.72	54.57	60.46	58.81	1.17
DELTA (ours)	63.85	**62.55**	63.03	**62.59**	**62.38**	62.86	**63.26**	**62.51**	**62.88**	**1.13**
Gemini-2.5-flash										
Document Level										
MultiRAG	58.26	40.79	55.11	30.81	44.26	53.52	35.97	31.65	43.80	1.53
CrossRAG	63.40	41.87	57.24	29.74	44.14	56.80	36.09	32.49	45.22	2.60
DKM-RAG	64.21	39.41	**59.26**	31.34	43.45	**57.74**	37.26	33.64	45.79	5.63
QTT-RAG	**65.32**	42.64	57.81	31.56	45.18	56.27	40.65	35.97	46.93	5.55
Query Level										
English Translation	-	48.44	55.84	53.59	53.68	55.17	47.67	54.86	52.75	1.55
DELTA (ours)	56.97	**56.45**	55.95	**55.83**	**56.18**	55.98	**56.44**	**56.45**	**56.28**	**1.48**
Deepseek-chat-v3.1										
Document Level										
MultiRAG	60.77	43.64	56.22	33.14	44.72	54.16	34.21	36.80	45.46	2.56
CrossRAG	67.83	48.34	62.24	39.05	49.27	61.33	39.85	45.70	51.70	2.64
DKM-RAG	67.84	44.07	**62.49**	37.63	45.66	**61.65**	40.30	40.38	50.00	2.39
QTT-RAG	**68.28**	46.13	61.81	37.24	47.36	60.48	41.06	41.29	50.46	1.93
Query Level										
English Translation	-	50.97	58.32	56.11	56.52	56.92	50.49	**57.52**	55.26	2.05
DELTA (ours)	59.85	**59.46**	58.61	**59.67**	**59.02**	59.25	**53.51**	56.45	**58.23**	**1.13**

Table 5: Main results (end-to-end mRAG performance). We use bge-m3 (Chen et al., 2024) for retrieval, and evaluate with character 3-gram recall (Chirkova et al., 2024). **Best**, second-best AVG (mean) are computed per generator and language. We report generation time as latency.

5 Experiments

5.1 Experimental Setup

Baselines. (1) **MultiRAG**: Retrieve and re-rank from multilingual datastores using the original MKQA query, then generate the answer in the query language. (Chirkova et al., 2024) (2) **CrossRAG**: Run the same multilingual retrieval as MultiRAG, translate the retrieved passages into a single pivot language (English). (Ranaldi et al., 2025) (3) **DKM-RAG**: Translate the retrieved passages into the query language, use an LLM to produce multiple refined passages. (Park and Lee, 2025) (4) **QTT-RAG**: Translate retrieved passages and attach translation-quality tags so the generator can decide which contexts to trust. (Moon et al., 2025) (5) **English Translation**: Translate the original query into the global pivot language.

5.2 Results and Analysis

Main Results. Table 5 shows that DELTA achieves the best average performance for each generator and is comparable to, or better than, document-level frameworks that require substantially higher cost due to the document's long context length. The gains are particularly pronounced on non-English queries, indicating that preference-aligned query augmentation is more effective than relying on document-side transformations in multilingual settings. DELTA provides little benefit on English queries (the en column in Table 5) because the local query and the global pivot become nearly identical when $L_q = \text{en}$. As a result, DELTA injects redundant segments with repeated, overlapping content, which unnecessarily lengthens the query and can dilute the useful signal for retrieval, resulting in no gain or even a slight degradation.

Statistic	Value
Queries (N)	16,828
Newly recovered queries (N_{new})	1,235
Gold best-rank (mean)	10.39
Gold best-rank (median)	5
Top-10 rate	66.23%
Rank in [10,49] rate	35.14%
Rank in [40,50] rate	4.62%

Table 6: Retrieval rank analysis, which reports the rank distribution of gold passages newly recovered by DELTA relative to English-pivot.

Analyzing Gold Passage Recall and Ranking. To investigate how DELTA recovers missing gold evidence to improve the overall mRAG system, we compare its retrieval performance against English-

pivot retrieval across seven languages (ar, de, es, ja, ko, th, zh), totaling 16,828 queries. As shown in Table 6, while English pivoting provides coverage due to English-heavy gold availability, it often degrades native surface-form anchors—such as titles, aliases, and original scripts—that are critical for precise entity matching. DELTA restores these anchors, facilitating better alignment with English gold documents. Among 1,235 newly recovered queries, DELTA achieves a mean best rank of 10.39 (median 5) with a 66.2% Top-10 entry rate. This indicates that DELTA does not merely rescue missed gold pages near the cutoff but significantly elevates them to high-ranking, actionable positions.

Method	ar	de	es	ja	ko	th	zh	Avg
Orig	63.68	62.46	63.26	63.26	62.84	63.37	63.04	63.13
+Global	71.42	72.02	71.37	71.51	71.77	71.70	71.57	71.62
+Title	68.17	67.83	67.75	67.78	67.93	68.17	68.34	68.00
+Aliases	68.14	67.46	67.43	67.38	68.61	68.15	68.13	67.90
+Locale	67.57	67.63	67.78	67.89	67.81	67.65	67.44	67.68
All cues	**72.99**	**73.01**	**72.48**	**73.26**	**72.88**	**72.93**	**72.70**	**72.89**

Table 7: Cue ablations for DELTA with fixed evidence. We incrementally add query-side cues and report end-to-end generation accuracy.

Impact of Cues on Evidence Interpretation. To isolate generation-stage effects from retrieval-stage effects, we conduct a cue ablation study under a fixed-evidence setting, which is reported in Table 7. Specifically, we first retrieve passages and re-rank, then hold those retrieved passages constant while modifying only the query-side cues at generation time. This design ensures that any observed performance differences stem solely from how the generator interprets the fixed evidence under different query formulations, not from changes in retrieved content. Under this setup, the global pivot cue significantly outperforms the original query, indicating that concise global English paraphrasing aids the generator in aligning evidence. Bridge cues also provide independent gains, showing that even when the retrieved evidence context is held fixed, varying only the query-side cues at generation time improves the model's ability to select precise evidence spans. The best performance across all languages is achieved by combining all cues, suggesting that bridge cues offer critical disambiguation and entity grounding.

Latency Analysis. To assess the efficiency of DELTA, we report average end-to-end latency (wall-clock time in seconds per query, averaged over all test queries) in the rightmost column of Table 5. We provide detailed per-language latency measurements in Appendix I. DELTA maintains high efficiency by generating a single fused query and avoiding document translation. It can even be faster than English Translation; by incorporating local cues and disambiguation anchors, DELTA enables direct retrieval, reducing the overhead of processing overly generic English-only signals.

6 Related Works

6.1 Multilingual RAG

Prior work in mRAG has explored how performance varies with the query language (Ranaldi et al., 2025; Longpre et al., 2021), the language of relevant or irrelevant evidence (Qi et al., 2025; Wu et al., 2024), as well as document ordering and prompting strategies that affect how models consume multilingual contexts (Sharma et al., 2024; Wu et al., 2024; Shankar et al., 2024; Ki et al., 2025). A common and effective heuristic is pivot translation, where non-English queries are translated into English before retrieval, often producing large gains (Asai et al., 2021; Ranaldi et al., 2025). However, much of the existing analysis of why pivot translation helps centers on the generation stage (e.g., English-centric generation competence, translation noise, and cross-lingual drift), which motivates generator-side interventions such as translation-aware prompting or decoding-time control (Sharma et al., 2024; Moon et al., 2025). In contrast, our work focuses on a retrieval-side explanation: we empirically show that gold evidence is structurally skewed toward English corpora.

6.2 Language Preference

In mRAG, language preference is shown both in retrieval (over-retrieving high-resource languages) and in generation (differentially using evidence by language even under *matched relevance*—a setting where the gold passage or core supporting evidence is correctly retrieved across all compared conditions), degrading consistency and downstream quality (Park and Lee, 2025). Existing measurements of language preference in mRAG commonly rely on behavioral proxies, such as comparing outputs across query languages via information overlap (Sharma et al., 2024) or embedding similarity to references (Park and Lee, 2025), and, in more controlled settings, analyzing citation or attribution behavior as evidence that language varies while

other variables are fixed (Ki et al., 2025; Qi et al., 2025). While prior approaches offer useful signals, they miss a key confound in mRAG: structural priors can dominate preference scores. We therefore debias preference by regressing out these priors and using the residual as the preference signal.

7 Conclusion

We demonstrate that gains from English pivoting in mRAG stem from retrieval-side evidence imbalance, which biases preference measurements. We address this with DeLP, a debiased metric that calibrates structural priors to reveal preference shifts toward the query language. Leveraging DeLP, we introduce DELTA, a lightweight query reformulation strategy that fuses global and local cues into a single query, consistently outperforming baselines.

Limitations

First, our debiasing targets retriever-level preference, while generator-level preference can still remain. Therefore, extending debiasing to how generators consume multilingual evidence is an important direction for future work. Second, our conclusions are drawn from a Wikipedia-based mRAG setup. Evaluating DeLP and DELTA on broader, domain-specific multilingual corpora is therefore necessary to assess their generalizability. Third, DELTA controls the balance between global and local signals using simple repetition, which is coarse. More precise and principled weighting or adaptive control logic could further improve effectiveness and stability.

Ethics Statement

We conduct our experiments using publicly available multilingual datasets, knowledge sources, and models that are widely used in the research community and released under established data-sharing and licensing guidelines. We follow the usage protocols and license agreements specified by the original providers. While these resources are designed to reduce harmful biases and inappropriate content, they may still contain artifacts of data imbalance and may not fully represent the diversity of languages, dialects, and cultural contexts. Our work analyzes and mitigates retrieval-side evidence imbalance and does not involve human subject data, user interaction logs, or the collection of personally identifiable information. We encourage future deployments to consider downstream risks such as uneven coverage across languages and potential disparities in answer quality for under-resourced communities.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25494299). This research was supported by the Chung-Ang University Graduate Research Scholarship in 2025.

References

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. 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*, pages 547–564.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh. 2025. Linguistic nepotism: Trading-off quality for language preference in multilingual rag. *arXiv preprint arXiv:2509.13930*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Bo Li, Zhenghua Xu, and Rui Xie. 2025. Language drift in multilingual retrieval-augmented generation: Characterization and decoding-time mitigation. *arXiv preprint arXiv:2511.09984*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Hoyeon Moon, Byeolhee Kim, and Nikhil Verma. 2025. Quality-aware translation tagging in multilingual rag system. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 161–177.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual RAG systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5647–5675, Vienna, Austria. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2025. On the consistency of multilingual context utilization in retrieval-augmented generation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 199–225, Suzhuo, China. Association for Computational Linguistics.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Bergen: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

Nikhil Sharma, Kenton Murray, and Ziang Xiao. 2024. Faux polyglot: A study on information disparity in multilingual large language models. *Preprint*, arXiv:2407.05502.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *Preprint*, arXiv:2410.21970.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatgpt when your question is not in english: a study of multilingual abilities and types of llms. *arXiv preprint arXiv:2305.16339*.

Appendix

A The Use of Large Language Models

We write the manuscript ourselves, and an LLM (ChatGPT-5.2) is used solely for refinement—style, clarity, and grammar. It is not used for ideation or content generation.

B Implementation Details

We adopt the multilingual retrieval baseline of Bergen (Chirkova et al., 2024), which retrieves evidence from a datastore spanning all languages. For generations, we follow Bergen's prompting setup and use the basic_translated_langspec template (Figure 4) to produce the final mRAG response. Building on this standardized pipeline, we conduct a series of experiments to quantify language preference in mRAG under the Bergen framework, which systematically examines the key components and practical adjustments necessary for a robust multilingual RAG baseline. We use a robust multilingual LLM, qwen3-235b-a22b-2507, to translate the user's local question into the global pivot language, English. We instruct GPT-4o-mini (Hurst et al., 2024) to generate a compact lexical bundle that supplies candidate titles, aliases, and a short disambiguation hint for the global and local segments. All LLM calls are made using the OpenRouter API. We conduct our experiments using an AMD EPYC 7313 CPU (3.0 GHz) paired with four NVIDIA RTX 6000 Ada GPUs. We use Python 3.11.5 and PyTorch 2.3.1 for the software environment.

C Language Notation

We use standard ISO 639-1 language codes to denote the languages in our experiments. Specifically, en denotes English, ar represents Arabic, es corresponds to Spanish, zh refers to Chinese (Simplified), ja indicates Japanese, de stands for German, ko denotes Korean, and th denotes Thai. These concise codes facilitate consistent identification and processing of language-specific data across datasets and models in multilingual NLP research.

D Dataset Details & Statistics

Wikipedia is a widely adopted knowledge source in both monolingual RAG and mRAG systems, as it provides broad topical coverage and is commonly used to benchmark RAG pipelines. In most experiments, we retrieve from various linguistic data sources from (i) the KILT snapshot of English Wikipedia[†] and (ii) the Wikipedia edition in the user's local language[‡]. This two-source design reflects a standard and practical mRAG setting where English serves as a high-coverage reference corpus while local-language Wikipedia captures language-specific evidence and terminology.

We report summary statistics for the data resources used in our experiments in Table 14. MKQA is our primary evaluation dataset, and we provide the number of examples along with the median lengths of questions and answers. We also use Wikipedia as the external corpus for the retriever datastore; its statistics, including the number of passages and their median lengths, are likewise presented in Table 14. These statistics provide an overview of the datasets and corpora underlying our experimental setup.

E Raw Language Preference Score

MLRS. Following the standard MultiLingual-RankShift (MLRS) protocol, we quantify retriever-level language preference by measuring how much the ranks of non-query-language documents improve after being translated into the query language (Park and Lee, 2025). For each query q with language L_q, we first retrieve a ranked list D_q from a multilingual datastore, assigning each document $d \in D_q$ an initial rank r_d^{init}. We then translate documents with $L_d \neq L_q$ into L_q and re-rank them using the same retriever, obtaining $r_d^{\text{re-rank}}$. The (non-negative) rank gain is computed as $\Delta r_d = \max\big(r_d^{\text{init}} - r_d^{\text{re-rank}}, 0\big)$, and aggregated per query as $\Delta r_q = \sum_d \Delta r_d$. Normalizing by the maximum possible gain $\Delta r_q^{\max} = \sum_d (r_d^{\text{init}} - 1)$ yields the query-level score $\text{MLRS}_q = \frac{\Delta r_q}{\Delta r_q^{\max}} \times 100$ (or 0 if $\Delta r_q^{\max} = 0$), and the final MLRS is the average over queries.

Results. Table 10 reports the retriever's language preference scores before calibration with DeLP. Overall, we observe three consistent patterns. First, cross-lingual retrieval ($L_q \neq L_d$) generally yields lower MLRS than monolingual retrieval, indicating that cross-lingual matching is less preferred in most cases. Second, English emerges as a dominant target language: when the retrieved document language L_d is English, the retriever attains near-maximum preference scores and often even

[†] https://huggingface.co/datasets/facebook/kilt_wikipedia

[‡] https://huggingface.co/datasets/wikimedia/wikipedia

Generator	Level	Method	*en*	*ar*	*es*	zh	*ja*	*de*	*ko*	*th*
Deepseek-chat-v3.1	Document	MultiRAG	1.925	3.093	2.456	2.683	2.816	2.456	2.362	2.716
		CrossRAG	2.005	2.897	2.706	2.822	2.968	2.328	2.778	2.653
		DKM-RAG	1.955	2.710	2.262	2.572	2.841	2.227	1.733	2.843
		QTT-RAG	1.671	2.347	2.054	1.348	2.475	1.663	**1.383**	2.525
	Query	English Translation	–	2.170	1.992	1.800	1.857	2.224	2.179	2.110
		Ours	**0.851**	**0.988**	**1.496**	**1.288**	**1.004**	**0.853**	1.637	**0.955**
Gemini	Document	MultiRAG	1.524	1.518	1.502	1.546	1.595	1.494	1.519	1.581
		CrossRAG	1.688	4.821	**0.780**	1.511	2.632	3.234	3.883	2.277
		DKM-RAG	4.308	7.293	5.301	**0.765**	6.261	5.559	6.187	9.346
		QTT-RAG	2.511	7.355	5.306	1.177	6.581	5.639	6.274	9.553
	Query	English Translation	–	**1.484**	1.510	1.528	1.802	1.481	**1.478**	1.539
		Ours	**1.483**	1.485	1.469	1.484	**1.499**	**1.473**	1.490	**1.481**
Qwen	Document	MultiRAG	**1.067**	1.905	1.367	1.657	1.147	1.343	1.141	1.410
		CrossRAG	1.082	1.533	1.357	2.475	**1.033**	**0.630**	0.693	1.543
		DKM-RAG	4.573	1.368	3.794	5.171	5.721	0.764	**0.627**	8.364
		QTT-RAG	1.401	1.813	1.698	3.263	1.494	1.350	1.037	2.352
	Query	English Translation	–	1.086	1.347	1.141	1.095	1.108	1.214	1.226
		Ours	1.069	**1.036**	**1.272**	**1.128**	1.045	1.111	1.123	**1.225**

Table 8: Average generation time (sec/query; lower is better). **Bold** = lowest, underline = second-lowest *across both Document-Level and Query-Level rows* for each (generator, language).

surpasses monolingual settings, consistent with English-heavy pretraining and stronger English representations. Third, cross-lingual preference is partly modulated by linguistic relatedness: closely related Romance languages (fr/it/pt/es) preserve relatively high cross-lingual scores, while East Asian pairs (ko/ja/zh) show moderate but noticeable drops compared to monolingual baselines.

F Language Distribution of Retrieved Documents

Table 13 reports the language composition of the top-50 documents retrieved for each MKQA query-language split. Across nearly all query languages, the retrieved evidence is heavily concentrated in English, and English often remains the most frequently retrieved language even for non-English queries. This trend is also reflected in the aggregated distribution (**mkqa_avg**), indicating that the observed language preference in standard mRAG pipelines largely mirrors structural priors of the retrieval setup rather than purely intrinsic model preference.

G Calibration Details

Table 9 reports the detailed DeLP scores for each of the three encoders. After removing the variance explained by the structural priors, the calibrated matrices exhibit a consistent pattern across encoders: the strongest preference concentrates on the diagonal (L_q=L_d), indicating a robust shift toward query–document language alignment rather than an English-dominant bias. Residual cross-lingual preferences remain comparatively mild and structured, reflecting interpretable related-language effects instead of exposure- or coverage-driven artifacts.

H Gold Passage Counting Protocol

We compute Table 1 on the 2,827-question subset that overlaps with KILT NQ provenance. Because prior work (Park and Lee, 2025) provides the same underlying question translated into 13 query languages, the unit counted in Table 1 is not the number of unique questions but the number of question$\times$query-language instances. Hence the total number of instances is $2{,}827 \times 13 = 36{,}751$, and values such as "#q = 26,934 (73.29%)" can legitimately exceed 2,827; the ratio is computed as $26{,}934/36{,}751$. Gold labels originate from KILT's provenance, which is anchored to English Wikipedia page IDs (WPIDs). All 13 translations of the same question share the same gold WPID set. We then assess gold availability in each Wikipedia language edition by mapping each English WPID to a corresponding page in language ℓ using Wikipedia/Wikidata interlanguage links (sitelinks), and checking whether the mapped page exists in the Wikipedia dump used to build our

Query Lang.	Encoder	$L_q = L_d$	$L_q \neq L_d$							
			en	ko	zh	fr	ja	it	pt	es
en	bge-m3	**49.25**	–	35.87 (-13.37)	39.48 (-9.77)	34.58 (-14.67)	38.79 (-10.46)	34.59 (-14.66)	35.81 (-13.43)	35.13 (-14.12)
	p-mMiniLM	**50.26**	–	37.00 (-13.27)	40.94 (-9.32)	36.18 (-14.08)	39.96 (-10.31)	35.83 (-14.43)	36.62 (-13.64)	36.49 (-13.78)
	p-mMpNet	**50.80**	–	37.15 (-13.65)	40.61 (-10.19)	35.94 (-14.86)	40.09 (-10.71)	36.04 (-14.76)	36.96 (-13.84)	36.43 (-14.37)
ko	bge-m3	42.34	36.76 (-5.58)	–	40.69 (-1.65)	34.41 (-7.93)	**42.45** (+0.11)	34.44 (-7.90)	35.29 (-7.05)	34.46 (-7.87)
	p-mMiniLM	**44.09**	38.03 (-6.06)	–	42.37 (-1.72)	35.09 (-9.00)	43.90 (-0.19)	34.75 (-9.34)	36.07 (-8.02)	34.98 (-9.11)
	p-mMpNet	43.72	38.29 (-5.43)	–	42.19 (-1.53)	35.20 (-8.52)	**44.43** (+0.72)	34.91 (-8.80)	35.59 (-8.13)	34.87 (-8.84)
zh	bge-m3	**49.67**	38.52 (-11.15)	37.32 (-12.36)	–	34.33 (-15.34)	41.36 (-8.32)	34.58 (-15.09)	35.71 (-13.96)	34.98 (-14.69)
	p-mMiniLM	**51.01**	38.80 (-12.21)	38.11 (-12.90)	–	34.99 (-16.03)	42.20 (-8.81)	35.06 (-15.95)	35.94 (-15.07)	35.38 (-15.64)
	p-mMpNet	**51.11**	38.72 (-12.39)	37.91 (-13.20)	–	34.87 (-16.24)	42.13 (-8.98)	34.98 (-16.13)	35.88 (-15.23)	35.31 (-15.80)
fr	bge-m3	39.45	**40.48** (+1.02)	36.15 (-3.30)	39.95 (+0.49)	–	39.47 (+0.01)	35.39 (-4.06)	36.25 (-3.20)	35.75 (-3.70)
	p-mMiniLM	40.42	**41.56** (+1.14)	37.20 (-3.22)	40.85 (+0.43)	–	40.27 (-0.15)	36.33 (-4.09)	36.94 (-3.48)	36.56 (-3.86)
	p-mMpNet	40.28	**41.45** (+1.17)	36.95 (-3.33)	40.71 (+0.43)	–	40.03 (-0.25)	36.29 (-3.98)	36.87 (-3.41)	36.54 (-3.74)
ja	bge-m3	**48.61**	38.44 (-10.16)	38.21 (-10.39)	41.15 (-7.46)	34.70 (-13.91)	–	34.83 (-13.78)	35.86 (-12.75)	35.09 (-13.52)
	p-mMiniLM	**49.56**	38.95 (-10.61)	38.55 (-11.01)	41.90 (-7.66)	35.19 (-14.37)	–	35.21 (-14.35)	36.14 (-13.42)	35.44 (-14.12)
	p-mMpNet	**49.41**	38.70 (-10.72)	38.43 (-10.98)	41.64 (-7.78)	34.94 (-14.47)	–	34.94 (-14.47)	35.92 (-13.49)	35.15 (-14.26)
it	bge-m3	38.38	**39.88** (+1.50)	36.15 (-2.23)	39.83 (+1.45)	35.87 (-2.51)	39.26 (+0.88)	–	36.39 (-2.00)	36.17 (-2.21)
	p-mMiniLM	39.44	**41.10** (+1.67)	37.23 (-2.21)	40.92 (+1.48)	37.08 (-2.36)	40.24 (+0.80)	–	37.44 (-2.00)	37.36 (-2.08)
	p-mMpNet	39.33	40.90 (+1.57)	37.18 (-2.15)	**41.02** (+1.69)	36.94 (-2.39)	40.09 (+0.77)	–	37.21 (-2.12)	37.20 (-2.12)
pt	bge-m3	**45.38**	39.89 (-5.49)	36.22 (-9.17)	39.82 (-5.56)	35.78 (-9.60)	39.41 (-5.97)	35.81 (-9.57)	–	37.09 (-8.29)
	p-mMiniLM	**46.52**	41.16 (-5.36)	37.33 (-9.19)	41.24 (-5.28)	37.03 (-9.49)	40.47 (-6.05)	36.93 (-9.59)	–	38.21 (-8.31)
	p-mMpNet	**46.33**	40.61 (-5.72)	37.38 (-8.95)	40.84 (-5.49)	36.70 (-9.63)	40.14 (-6.19)	36.71 (-9.61)	–	37.88 (-8.45)
es	bge-m3	37.64	**40.18** (+2.54)	36.21 (-1.43)	39.78 (+2.15)	35.68 (-1.95)	39.28 (+1.64)	35.90 (-1.74)	36.82 (-0.82)	–
	p-mMiniLM	38.71	**41.31** (+2.60)	37.29 (-1.42)	40.85 (+2.14)	36.87 (-1.84)	40.20 (+1.49)	37.01 (-1.70)	37.73 (-0.98)	–
	p-mMpNet	38.23	**40.65** (+2.42)	37.09 (-1.14)	40.53 (+2.30)	36.34 (-1.89)	39.81 (+1.58)	36.43 (-1.80)	37.19 (-1.04)	–

Table 9: Language preference measured by DeLP. Each cell reports the debiased preference score and its delta from the matching-language baseline ($L_q = L_d$). Background shading is row-wise min–max scaled (including the diagonal cell); Darker cells indicate a stronger preference for the document language.

corpus.

A key source of confusion is that our corpus is passage-based: each Wikipedia page is split into multiple chunks, so a single WPID may correspond to multiple passages. Moreover, for a given question, KILT may provide multiple gold provenances that map to different passages within the same WPID. In Table 1, we use a WPID-level convention: when multiple gold passages correspond to the same WPID for a query, we treat them as a single gold item rather than counting them multiple times. Out of the 2,827 questions in our KILT-overlap subset, 2,404 questions have at least one available gold WPID (i.e., a mapped gold page exists in our Wikipedia dumps). Accordingly, the number of questions with gold evidence satisfies `only_en` + `both` = 2,404 in Table 1.

I Detailed Latency

Table 8 reports detailed latency measured as average generation time (sec/query; lower is better) for each generator across languages. DELTA remains consistently efficient because it produces a single fused query and avoids document translation overhead; in several settings, it is even faster than English Translation, since the fused query retains local cues and disambiguation anchors that help retrieval focus earlier and reduce wasted computation on overly generic English-only signals.

J Cultural Prior Measurement

To model whether a query is intrinsically tied to a particular cultural or regional context (independent of corpus size or retrieval exposure), we construct a *cultural prior* using an LLM-based classifier. Starting from the English version of each query (MKQA-en), we assign exactly one *cultural database language* from a fixed set of 13 languages (*en, ar, es, de, ja, ko, th, zh, fr, it, pt, ru, fi*). We use GPT-4o mini (via OpenRouter) with constrained JSON output to enforce a single-label decision. We instruct the classifier to choose the local language of the primary place/culture the query is about (e.g., France $\rightarrow$ *fr*, Hong Kong/China $\rightarrow$ *zh*), and to select *en* only when the cultural context is inherently English-speaking (e.g., US/UK-specific) or when the query is genuinely global/multi-country and not tied to a single locale.

In addition to the cultural-language label, we record lightweight *cultural metadata* for analysis and filtering: (i) `country_or_region` (a single primary place/region), (ii) `is_culture_specific` (whether the question is judged to be culture/locale-specific), (iii) `confidence` (0–1), and (iv) a short

rationale. These fields are used only to characterize the dataset and to support qualitative inspection; our core metric relies on the language label.

Finally, we define the cultural prior $p_{\text{cult}}(\ell)$ as the empirical probability that a query's predicted cultural language equals ℓ, i.e., the normalized frequency of the single-label assignments over the evaluation set. This prior captures *where evidence should exist* in a fair localized setting, and is incorporated as a structural factor alongside other priors (e.g., exposure and gold availability) in our calibration analysis. We use those cultural prior and metadata for calibration and DELTA.

K Repetition-based weighting for DELTA

Query construction with repetition. To control cue influence without changing the retriever or learning parameters, we apply a deterministic repetition policy while constructing the fused query string Q_{fused}. We use the same notation as in Eq. 5: $y \in \{0, 1\}$ indicates whether the query is culture-specific and $c \in [0, 1]$ is the confidence score. We repeat the local block [LOCAL:L_q] r_{local} times and the global pivot block [GLOB] r_{glob} times, and concatenate all segments with a delimiter (" | ") to form a single retrieval query.

Length control. To keep retrieval budgets comparable across methods, we truncate the final Q_{fused} to a fixed maximum length (e.g., 900 characters) after concatenation.

Deduplication. We apply conservative deduplication to avoid redundant anchors: (i) if the global and local titles are identical, we keep only a single [TITLE_BRIDGE]; (ii) if alias sets match across languages, we keep only the global alias block; and (iii) when the query language is not English, we always include [LOCAL:L_q] at least once.

L Thresholds and fixed hyperparameters.

We do not exhaustively tune $(\tau_{\text{low}}, \tau_{\text{high}}, \tau_{\text{boost}})$ because a full sweep is combinatorial and would couple these knobs to expensive end-to-end RAG runs. Instead, we instantiate the confidence thresholds with three goals: (i) discretize the continuous confidence c into a small number of stable intervals, (ii) keep the query-length increase bounded, and (iii) reserve upweighting for only the most reliable culture-specific cases. Concretely, we set two cutoffs $\tau_{\text{low}} < \tau_{\text{high}}$ to map c into three repetition levels for the local block, $r_{\text{local}} \in \{1, 2, 3\}$, where τ_{low} marks the onset of *reliable* culture-specificity and τ_{high} indicates *high-confidence* cases that warrant the strongest local emphasis. In our implementation, we use $\tau_{\text{low}} = 0.6$ and $\tau_{\text{high}} = 0.85$, which empirically balance coverage (triggering local upweighting for sufficiently confident cases) and conservativeness (avoiding frequent over-repetition under noisy cue predictions).

For auxiliary local boosting, we use a separate threshold τ_{boost} that applies only to the disambiguation anchors ([TITLE_BRIDGE] and [ALIASES]), not the full [LOCAL] query text. Specifically, for culture-specific queries ($y = 1$), we set $b = \mathbb{I}[c \geq \tau_{\text{boost}}]$ and, when $b = 1$, duplicate [TITLE_BRIDGE] and [ALIASES] once to strengthen culturally grounded anchoring and reduce entity ambiguity

We set $\tau_{\text{boost}} = 0.7$ so that anchor duplication is enabled for moderately-to-high confidence culture-specific queries, providing extra entity anchoring/disambiguation without incurring the larger length increase of repeating the entire local block.

For ridge calibration, we likewise keep a single regularization strength λ across all encoders; this choice is motivated by the small calibration design ($|\mathcal{C}|$ language pairs with a low-dimensional prior vector) where ridge mainly stabilizes coefficients against correlated priors rather than serving as a performance-tuned knob.

M Case Study

DELTA. Table 15 illustrates DELTA on a Korean query asking "when was the last time South Korea had the Olympics." DELTA forms the global pivot q_{glob} and emits it as [GLOB], and places the original Korean surface form as [LOCAL:ko]. It then injects multilingual anchors: [TITLE_BRIDGE] contains paired Wikipedia-style titles, while [ALIASES:GLOB] and [ALIASES:ko] provide short alias cues in the global and local languages, respectively. Finally, [LOCALE_HINT] adds a brief region hint with minimal disambiguation to bias retrieval toward region-appropriate evidence.

Crucially, DELTA controls the balance between global and local signals purely through repetition. Because this query is labeled culture-specific (y=1) with high confidence c=0.93, the policy sets r_{local}=3 (since $c \geq 0.85$) while keeping r_{glob}=1 (since $c \geq 0.6$), yielding three copies of [LOCAL:ko] but only one copy of [GLOB]. Moreover, the auxiliary local-

Query Lang.	Encoder	$L_q = L_d$	$L_q \neq L_d$							
			en	ko	zh	fr	ja	it	pt	es
en	bge-m3	**56.03**	–	33.02 (-23.01)	33.10 (-22.93)	36.61 (-19.42)	33.36 (-22.67)	35.89 (-20.14)	35.86 (-20.17)	36.62 (-19.41)
	p-mMiniLM	**56.85**	–	34.34 (-22.51)	34.61 (-22.24)	38.17 (-18.68)	34.52 (-22.33)	37.15 (-19.70)	36.73 (-20.12)	37.96 (-18.89)
	p-mMpNet	**57.49**	–	34.45 (-23.04)	34.27 (-23.22)	37.94 (-19.55)	34.67 (-22.82)	37.34 (-20.15)	37.02 (-20.47)	37.90 (-19.59)
ko	bge-m3	41.15	**43.49** (+2.34)	–	34.42 (-6.73)	36.42 (-4.73)	37.18 (-3.97)	35.72 (-5.43)	35.30 (-5.85)	35.93 (-5.22)
	p-mMiniLM	42.95	**44.62** (+1.67)	–	36.04 (-6.91)	37.08 (-5.87)	38.47 (-4.48)	36.07 (-6.88)	36.18 (-6.77)	36.45 (-6.50)
	p-mMpNet	42.53	**44.98** (+2.45)	–	35.85 (-6.68)	37.20 (-5.33)	39.01 (-3.52)	36.21 (-6.32)	35.65 (-6.88)	36.34 (-6.19)
zh	bge-m3	44.98	**45.26** (+0.28)	34.52 (-10.46)	–	36.34 (-8.64)	36.05 (-8.93)	35.86 (-9.12)	35.73 (-9.25)	36.45 (-8.53)
	p-mMiniLM	**46.18**	45.39 (-0.79)	35.46 (-10.72)	–	36.98 (-9.20)	36.77 (-9.41)	36.38 (-9.80)	36.05 (-10.13)	36.85 (-9.33)
	p-mMpNet	**46.27**	45.41 (-0.86)	35.21 (-11.06)	–	36.87 (-9.40)	36.71 (-9.56)	36.28 (-9.99)	35.94 (-10.33)	36.78 (-9.49)
fr	bge-m3	43.18	**47.23** (+4.05)	33.29 (-9.89)	33.58 (-9.60)	–	34.07 (-9.11)	36.70 (-6.48)	36.30 (-6.88)	37.25 (-5.93)
	p-mMiniLM	44.09	**48.15** (+4.06)	34.54 (-9.55)	34.52 (-9.57)	–	34.83 (-9.26)	37.65 (-6.44)	37.05 (-7.04)	38.03 (-6.06)
	p-mMpNet	43.96	**48.14** (+4.18)	34.25 (-9.71)	34.37 (-9.59)	–	34.61 (-9.35)	37.59 (-6.37)	36.93 (-7.03)	38.01 (-5.95)
ja	bge-m3	45.03	**45.18** (+0.15)	35.45 (-9.58)	34.86 (-10.17)	36.71 (-8.32)	–	36.11 (-8.92)	35.88 (-9.15)	36.56 (-8.47)
	p-mMiniLM	**45.80**	45.54 (-0.26)	35.90 (-9.90)	35.57 (-10.23)	37.18 (-8.62)	–	36.53 (-9.27)	36.25 (-9.55)	36.91 (-8.89)
	p-mMpNet	**45.67**	45.39 (-0.28)	35.73 (-9.94)	35.30 (-10.37)	36.94 (-8.73)	–	36.24 (-9.43)	35.98 (-9.69)	36.62 (-9.05)
it	bge-m3	41.06	**46.63** (+5.57)	33.30 (-7.76)	33.47 (-7.59)	37.92 (-3.14)	33.86 (-7.20)	–	36.44 (-4.62)	37.68 (-3.38)
	p-mMiniLM	42.11	**47.69** (+5.58)	34.57 (-7.54)	34.59 (-7.52)	39.07 (-3.04)	34.80 (-7.31)	–	37.55 (-4.56)	38.83 (-3.28)
	p-mMpNet	41.98	**47.59** (+5.61)	34.48 (-7.50)	34.68 (-7.30)	38.94 (-3.04)	34.67 (-7.31)	–	37.27 (-4.71)	38.67 (-3.31)
pt	bge-m3	39.19	**46.64** (+7.45)	33.37 (-5.82)	33.46 (-5.73)	37.83 (-1.36)	34.02 (-5.17)	37.13 (-2.06)	–	38.61 (-0.58)
	p-mMiniLM	40.17	**47.75** (+7.58)	34.67 (-5.50)	34.91 (-5.26)	39.02 (-1.15)	35.03 (-5.14)	38.25 (-1.92)	–	39.68 (-0.49)
	p-mMpNet	39.91	**47.30** (+7.39)	34.68 (-5.23)	34.50 (-5.41)	38.70 (-1.21)	34.72 (-5.19)	38.01 (-1.90)	–	39.35 (-0.56)
es	bge-m3	40.76	**46.93** (+6.17)	33.36 (-7.40)	33.42 (-7.34)	37.73 (-3.03)	33.87 (-6.89)	37.22 (-3.54)	36.88 (-3.88)	–
	p-mMiniLM	41.81	**47.90** (+6.09)	34.63 (-7.18)	34.52 (-7.29)	38.86 (-2.95)	34.76 (-7.05)	38.33 (-3.48)	37.84 (-3.97)	–
	p-mMpNet	41.33	**47.34** (+6.01)	34.39 (-6.94)	34.19 (-7.14)	38.34 (-2.99)	34.39 (-6.94)	37.73 (-3.60)	37.25 (-4.08)	–

Table 10: Raw language preference measured by MLRS with different re-ranking encoders for various query–document language pairs. The $L_q = L_d$ column shows scores for matching query and document languages, while the remaining columns represent cross-lingual scenarios. Parentheses indicate the change from the $L_q = L_d$ column (positive for improvement, negative for decline). The highest score per row is in bold, and the second highest is underlined.

boost flag triggers at $c \geq 0.7$, duplicating the local-side anchors once more, which explains why `[TITLE_BRIDGE]` and `[ALIASES:ko]` appear twice, whereas `[ALIASES:GLOB]` remains single-copy. Overall, this design realizes a global back-off (`[GLOB]`) with preference-aligned local emphasis (`[LOCAL]`, `[TITLE_BRIDGE]`, `[ALIASES:ko]`) within a single Q_{fused}, without modifying the retriever or adding model parameters.

Success Case. Table 16 presents a representative top-1 retrieval example comparing DELTA with a simple English-translation query for the question "언제 마지막으로 대한민국이 올림픽을 했었나요. (When was the last time South Korea had the Olympics?)." Although both methods use the same retriever and multilingual datastore, the retrieved evidence differs markedly: DELTA's fused query contains explicit host-oriented cues (local surface form, title/alias anchors, and a locale hint), which increases lexical alignment with passages that describe Olympics *held in* Korea (e.g., 개최, 서울 1988, 평창 2018). In contrast, the English-translation query is more underspecified and can drift to participation-centric passages that match broad entities ("South Korea", "Olympics") but do not emphasize hosting-related facts. As a result, the DELTA top-1 passage provides the necessary host evidence for inferring the most recent domestically held Olympics, enabling the generator to produce the correct answer, while the translation-based pipeline is more likely to miss the hosting signal and return an incorrect year/event.

Failure Case. Table 17 shows a representative failure where the question "who is the president during the Korean War" is underspecified: "president" can plausibly refer to the U.S. president overseeing U.S. involvement (gold: Harry S. Truman and Dwight D. Eisenhower) or to the South Korean president during the same period (Syngman Rhee). In this example, DELTA's cultural/locale cues and title bridge steer the query toward *South Korean leadership*, effectively resolving the ambiguity in the wrong direction. Consequently, the top-1 retrieved passage focuses on Syngman Rhee and contains strong lexical overlap with the localized cues (e.g., "대한민국 대통령", "이승만", "한국 전쟁"), making the generator likely to output Syngman Rhee despite the dataset's gold reference targeting

U.S. presidents. This failure highlights a limitation of repetition- and cue-based weighting: when the underlying intent is ambiguous, aggressively injecting locale-specific anchors can over-localize retrieval and suppress globally relevant evidence, suggesting the need for ambiguity-aware safeguards (e.g., intent disambiguation or controlled locale injection) for such queries.

N Prompts

As shown in Figure 4, we provide the exact prompt templates used throughout our pipeline. Prompt (A) specifies the RAG answer-generation instruction, with two variants depending on whether retrieved documents are provided, enforcing concise English outputs and (when available) conditioning answers on the supplied evidence. Prompt (B) defines our cultural-context annotation step, where an LLM assigns a single cultural database language from a fixed set under strict locality-oriented rules and returns lightweight metadata (region, culture-specificity, confidence, and a brief rationale) in a structured JSON format. Prompt (C) is used by DELTA to produce retrieval anchors—English and local Wikipedia-style titles, alias lists, and a short disambiguation hint—which are then assembled into a fused query; this prompt enforces a fixed JSON schema and language constraints to keep the generated anchors consistent and directly usable for retrieval. Finally, in Prompt (D), we provide the prompts used for English translation.

O Validation of the DeLP

We validate DeLP by constructing a controlled experiment that directly tests whether the metric remains stable under artificial shifts in gold-answer distribution—a scenario where a robust metric should reflect consistent model preference regardless of structural changes in the corpus.

Experimental Setup. We fix the underlying model preference and artificially vary the gold-answer language ratio between Korean and English from 0 (all gold answers in Korean) to 1 (all gold answers in English) in incremental steps. Under this setup, a metric that disentangles structural priors from intrinsic preference should remain stable across the spectrum, whereas a metric that conflates the two should exhibit high sensitivity.

Results. Table 11 reports the range and standard deviation of each metric across all ratio configurations. We observe that MLRS consistently exhibits larger variability (range 10.35, std 3.27) as the gold-language ratio shifts, confirming that raw MLRS scores are heavily confounded by gold-distribution bias. In contrast, DeLP remains more stable across the same configurations (range 9.56, std 3.02). We further confirm this difference via paired permutation tests, obtaining $p = 0.00016$ for range and $p = 0.00010$ for standard deviation, both indicating statistical significance. In terms of effect size, DeLP reduces variability by 7.6% relative to MLRS, supporting its robustness against structural biases caused by gold-language fluctuations.

Furthermore, as reported in Table 4, we validate DeLP by measuring the correlation between preference scores and structural priors before and after calibration. Raw MLRS scores exhibit near-perfect correlation with the exposure prior ($r > 0.99$), gold-availability prior ($r > 0.91$), and cultural prior ($r > 0.91$) across all encoders. After applying DeLP calibration, these correlations drop sharply, confirming that DeLP effectively decouples intrinsic model preference from the structural signals that contaminate standard benchmarks.

Metric	Range	Std
MLRS (raw)	10.35	3.27
DeLP (ours)	**9.56**	**3.02**

Table 11: Sensitivity of MLRS and DeLP to gold-language ratio shifts (Korean $\rightarrow$ English). Lower range and std indicate greater robustness against gold-language distribution shifts.

P RAG Utility and Sanity Check

We clarify that Gold Availability in Table 1 measures whether the KILT gold provenance page exists in the target language's Wikipedia via inter-language mapping—it does not directly measure answerability. Consequently, a Gold Availability of $\sim$1% does not imply that 99% of questions are unanswerable or that the *Base* column primarily reflects parametric memory.

To empirically verify the actual utility of retrieval under these conditions, we conduct a sanity check by comparing mRAG performance with and without retrieval augmentation across multiple non-English query languages and models. As shown in Table 12, we observe consistent performance gains from RAG across all tested languages and models. Notably, the improvements are substantial

in several cases—for example, Qwen3-235B gains 14.97 points on Arabic and 14.75 points on Korean when retrieval is enabled. These results demonstrate that even when the exact gold provenance page is absent in the local Wikipedia, retrieval still surfaces relevant supporting evidence distributed across multilingual Wikipedia corpora, yielding consistent and meaningful performance improvements over relying solely on parametric memory.

Model	Lang.	RAG	No RAG
Gemini-2.5-Flash	ar	43.60	38.42
	es	58.76	57.13
	ko	38.83	34.66
	th	31.55	30.20
	zh	30.87	30.65
DeepSeek-Chat-v3.1	ar	48.37	38.38
	es	61.82	59.35
	ko	41.56	32.81
	th	35.42	33.28
	zh	38.99	37.89
Qwen3-235B	ar	45.58	30.61
	es	64.21	61.92
	ko	42.29	27.54
	th	43.86	32.37
	zh	38.67	33.85

Table 12: mRAG performance comparison with and without retrieval augmentation across non-English query languages. RAG consistently outperforms No RAG, demonstrating that multilingual Wikipedia provides relevant evidence beyond the exact gold provenance page.

Q Explanation of Low Gold Availability

We clarify that the >99% figure does not indicate that non-English questions are unanswerable. Gold Availability measures whether the KILT gold provenance page *exists* in the target language's Wikipedia via interlanguage mapping—it does not directly measure whether a question can be answered from that language's corpus.

Why does Gold Availability appear so low? As detailed in Appendix H, KILT provenance is constructed on English Wikipedia and mapped to other languages via interlanguage links. If the corresponding page is absent or the interlanguage mapping is incomplete, Gold Availability becomes zero even when partially relevant content exists in that language. Furthermore, we aggregate at the WPID level, treating multiple gold passages from the same page as a single item, which further reduces the reported ratio. For instance, 17,667 Korean gold passages correspond to a smaller number of distinct page IDs, making the percentage appear lower than the raw passage count would suggest.

The absolute scale is not negligible. We compute statistics over 36,751 expanded samples (2,827 questions $\times$ 13 query languages). Even at 1%, this corresponds to approximately 367 gold page IDs—an absolute scale that is by no means trivial. We also note that this analysis is conducted on the widely adopted MKQA/KILT benchmark, so the observed distribution reflects a standard setting rather than an artifact of an unusual corpus.

What does this result actually tell us? The key implication of this finding is not about QA difficulty. Rather, it demonstrates that standard benchmarks carry a systematic English-centric provenance prior, which causes gold evidence to concentrate overwhelmingly in English corpora. We argue that this structural skew—not any intrinsic linguistic superiority of English—drives the apparent advantage of English pivoting in mRAG systems, which is the central motivation for our DeLP calibration framework.

	en	ko	ar	zh	fi	fr	de	ja	it	pt	ru	es	th
mkqa_en	44.12	1.60	1.19	1.30	2.54	10.03	6.90	1.44	8.32	7.67	4.85	9.90	0.13
mkqa_ko	23.07	17.35	1.99	4.81	2.04	7.90	5.96	10.36	6.16	5.06	6.85	6.85	1.58
mkqa_ar	24.93	3.30	15.29	4.07	2.10	8.30	6.53	6.64	6.80	5.71	7.78	7.65	0.89
mkqa_zh	24.70	3.17	1.76	23.22	2.01	7.47	6.17	6.27	6.08	5.24	6.37	7.27	0.27
mkqa_fi	30.32	2.27	1.63	2.33	7.92	11.11	8.20	3.78	8.77	7.18	6.51	9.42	0.58
mkqa_fr	29.90	1.48	1.25	1.55	2.50	21.44	6.96	2.06	9.40	7.96	4.77	10.55	0.19
mkqa_de	32.54	1.46	1.17	1.44	2.96	11.40	15.12	1.89	9.09	7.69	4.83	10.17	0.24
mkqa_ja	24.56	4.80	1.69	3.99	2.19	7.97	5.99	22.55	6.38	5.66	6.49	7.45	0.28
mkqa_it	28.72	1.59	1.30	1.58	2.52	12.30	6.97	1.95	17.46	8.47	5.26	11.70	0.17
mkqa_pt	28.82	1.71	1.40	1.63	2.60	11.92	6.74	2.23	10.24	13.78	5.38	13.33	0.24
mkqa_ru	27.02	2.53	1.92	1.98	2.45	8.83	6.44	2.71	7.36	6.24	23.83	8.43	0.26
mkqa_es	29.45	1.73	1.27	1.60	2.66	11.85	6.93	1.83	10.55	9.33	5.27	17.36	0.16
mkqa_th	32.39	3.10	2.10	2.96	2.53	10.00	7.40	4.43	8.06	7.43	6.80	9.70	3.10
mkqa_avg	**29.27**	**3.55**	**2.61**	**4.04**	**2.85**	**10.81**	**7.41**	**5.24**	**8.82**	**7.49**	**7.31**	**9.98**	**0.62**

Table 13: Language distribution of retrieved documents for each MKQA query-language split. Each row corresponds to the query language (dataset), and each column indicates the language of the retrieved passages; values are shown as percentages (without the %). The final row (**mkqa_avg**) reports the average retrieved-language distribution across all query languages.

Dataset	**en**	**ar**	**es**	**fi**	**fr**	**de**	**ja**	**it**	**ko**	**pt**	**ru**	**zh**	**th**
MKQA													
# examples	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
len question.	43	38	48	46	49	47	26	48	22	45	42	16	41
len answer.	11	10	11	11	11	11	8	11	6	11	12	6	12
Wikipedia													
# ex. (M)	25	3.3	10	1.5	13	14	27	8.2	1.6	4.7	8.6	11	3.7
len passage.	624	585	619	833	627	720	208	650	431	619	721	206	217

Table 14: Statistics of the datasets used in our experiments. MKQA Number of examples and median lengths of questions and answers (in Unicode characters). Wikipedia: Number of passages (in millions) and their median lengths.

DELTA segment	**Instantiated content (case study)**	**Rep.**
`[GLOB]`	when was the last time south korea had the olympics	1
`[LOCAL:ko]`	언제 마지막으로 대한민국이 올림픽을 했었나요	3
`[TITLE_BRIDGE]`	South Korea at the Olympics / 대한민국의 올림픽	2
`[ALIASES:ko]`	대한민국 올림픽, 한국 올림픽, 한국의 올림픽 역사	2
`[ALIASES:GLOB]`	Olympics in South Korea, South Korean Olympic Games, History of South Korea Olympics	1
`[LOCALE_HINT]`	South Korea + Last Olympic Games in South Korea	1

Table 15: DELTA case study. A Korean culture-specific query (c=0.93) is converted into a single fused query Q_{fused} by concatenating labeled segments. Repetition counts follow Eq. 5, which upweights local cues while maintaining a global back-off.

Item	Content
DELTA	[GLOB] when was the last time south korea had the olympics |[LOCAL:ko] 언제 마지막으로 대한민국이 올림픽을 했었나요 |[TITLE_BRIDGE] South Korea at the Olympics / 대한민국의 올림픽 |[ALIASES:ko] 대한민국 올림픽, 한국 올림픽, 한국의 올림픽 역사 |[ALIASES:GLOB] Olympics in South Korea, South Korean Olympic Games, History of South Korea Olympics |[LOCALE_HINT] South Korea Last Olympic Games in South Korea
English Translation	When was the last time south korea had the olympics
Top-1 passage (DELTA)	대한민국에서 열린 올림픽으로는 1988년 서울 하계 올림픽과 2018년 평창 동계 올림픽이 널리 알려져 있다. 서울 대회는 20세기 후반 대한민국의 국제 스포츠 행사 유치와 관련해 자주 언급되며, 주요 경기장은 서울 및 인근 지역에 분산되어 운영되었다. 평창 대회는 강원 지역을 중심으로 동계 종목이 진행되었고, 개폐회식과 일부 경기장이 평창 및 주변 권역에 배치되었다. 두 대회 모두 대한민국 내에서 개최된 사례로 정리되며, 대회의 성격(하계/동계)과 개최 지역(서울/평창)이 함께 기술되는 경우가 많다.
Top-1 passage (English Translation)	대한민국(South Korea)은 근대 올림픽(Olympics)에 지속적으로 참가해 왔으며, 여러 종목에서 의미 있는 성과를 거두었다. 이 문서는 연도별 참가 개요, 선수단 규모, 주요 종목에서의 메달 기록과 같은 정보를 중심으로 구성된다. 예를 들어 양궁, 태권도, 쇼트트랙 등에서의 성과가 요약되고, 대회별 대표 선수나 주목할 만한 기록이 덧붙여지기도 한다. 또한 특정 대회에서의 종합 순위 변화나 메달 수 추이처럼 참가 및 성과를 설명하는 통계적 서술이 포함될 수 있다.
Gold answer	2018년 평창 동계 올림픽. (*The 2018 PyeongChang Winter Olympics.*)
English translation answer	1988년 서울 하계 올림픽. (*The 1988 Seoul Summer Olympics.*)
DELTA answer	대한민국에서 개최된 올림픽으로 1988년 서울 하계 올림픽과 2018년 평창 동계 올림픽이 언급되므로, 질문에서 묻는 "마지막으로" 개최된 올림픽은 2018년 평창 동계 올림픽이다. (*The 2018 PyeongChang Winter Olympics.*)

Table 16: Case study: DELTA vs. English translation (top-1 retrieval).

Item	Content
DELTA (misled)	[GLOB] who is the president during the korean war |[TITLE_BRIDGE] President of South Korea during the Korean War / 한국 전쟁 중 대한민국 대통령 |[ALIASES:ko] 이승만, 이승만 대통령, 대통령 이승만 |[ALIASES:GLOB] Syngman Rhee, Rhee Syngman, President Rhee, Rhee |[LOCALE_HINT] Korea (Korean Peninsula) President during Korean War era
Top-1 passage (DELTA)	이승만(Syngman Rhee)은 1948년부터 1960년까지 대한민국의 대통령으로 재임한 정치인이다. 대한민국 정부 수립 이후 초대 대통령으로 선출되었으며, 냉전 초기 한반도의 분단 체제 속에서 정부 운영을 주도했다. 재임 기간에는 한국 전쟁(1950–1953) 시기가 포함되며, 전쟁 전후의 정치적 갈등과 대외 관계가 함께 언급된다. 관련 문서들은 대체로 이승만의 생애, 대통령 재임 기간, 당대의 국내 정치 상황과 외교적 맥락을 중심으로 개괄한다.
Gold answer	해리 S. 트루먼; 드와이트 D. 아이젠하워 (Harry S. Truman; Dwight D. Eisenhower)
DELTA answer	이승만 (Syngman Rhee)

Table 17: Failure case (top-1 retrieval).

```
(A) RAG Answer Generation

Goal: Answer as concisely as possible in {lang}.

With Documents:
  System: Extract relevant information from provided documents and answer briefly. Reply in {lang}.
  User: Background: {docs} \n\nQuestion: {question}

Without Documents:
  System: Answer briefly. Reply in {lang}.
  User: Question: {question}


(B) Cultural Language Classifier

System (instruction):
  You are annotating a FAIR multilingual retrieval setup.
  Given an English query, decide the SINGLE most appropriate "cultural database language"
  where the relevant evidence SHOULD exist in a fair, localized setting.

CRITICAL RULES:
  - You MUST choose exactly ONE language from this fixed set:
    {en, ar, es, de, ja, ko, th, zh, fr, it, pt, ru, fi}
  - Prefer the LOCAL language of the primary place/culture the query is about.
  - Do NOT choose 'en' just because the query text is English.
  - Choose 'en' only if the query's primary cultural context is inherently English-speaking
    (e.g., US/UK-specific) OR the query is truly global / multi-country / not place-specific.
  - If the query mentions a place that maps to one of the non-English languages,
    pick that non-English language.

Examples:
  - "when did hong kong go back to china" -> cultural_language="zh"
  - "what is the capital of france" -> cultural_language="fr"
  - "who was the first president of the united states" -> cultural_language="en"
  - "compare gdp of france and germany" -> cultural_language="en" (multi-country/global)

Output (JSON only; no extra text):
  {
    "country_or_region": string (SINGLE primary place/region),
    "cultural_language": string (exactly one from the set),
    "is_culture_specific": boolean,
    "confidence": number in [0,1],
    "rationale": short string
  }

Input:
  User: Query: {query_en}
```

Figure 4: Prompt templates used in our pipelines: RAG generation and cultural-language classification.

```
(C) DELTA Bundle Generator

Goal: Produce title/alias anchors and a short disambiguation hint for fused query construction.

Return Format:
  - SINGLE-LINE JSON object only (no markdown, no explanation).
  - Keys must be EXACTLY:
      en_title, local_title, aliases_en, aliases_local, extra_disambig

Constraints:
  - aliases_en / aliases_local: 0..K items each
  - Titles: plausible Wikipedia page titles; use null if unsure
  - extra_disambig: <= 8 words
  - local_title & aliases_local MUST be in {query_lang}; English fields MUST be English
  - Do not add new keys

Input (User JSON):
  {
    "q_en": "{q_en}",
    "q_orig": "{q_orig}",
    "query_lang": "{query_lang}",
    "country_or_region": "{country_or_region}",
    "cultural_language": "{cultural_language}",
    "is_culture_specific": {is_culture_specific},
    "confidence": {confidence}
  }


(D) English Translation

Goal: Translate the question from {lang_name} to fluent, natural English while preserving
      the original meaning as much as possible.

Rules:
  - Keep named entities as appropriate English forms.
  - Do not add explanations or extra information.
  - Return STRICT JSON with a single key "translation".

System:
  You are a professional translator from {lang_name} to English.
  You receive a question in the source language and must translate it into fluent,
  natural English while preserving the original meaning as much as possible.
  - Keep named entities as appropriate English forms.
  - Do not add explanations or extra information.
  Return STRICT JSON with a single key "translation".

User:
  Question in {lang_name}:
  {query}

  Return only:
  {"translation": "<the question translated into English>"}
```

Figure 5: Prompt templates used in our pipeline: DELTA bundle generation and English translation.