Language Drift in Multilingual Retrieval-Augmented Generation: Characterization and Decoding-Time Mitigation

Bo Li[1, 2], **Zhenghua Xu**[1], **Rui Xie**[2*]

[1]State Key Laboratory of Intelligent Power Distribution Equipment and System, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China
[2]National Engineering Research Center for Software Engineering, Peking University, China

Abstract

Multilingual Retrieval-Augmented Generation (RAG) enables large language models (LLMs) to perform knowledge-intensive tasks in multilingual settings by leveraging retrieved documents as external evidence. However, when the retrieved evidence differs in language from the user query and in-context exemplars, the model often exhibits language drift by generating responses in an unintended language. This phenomenon is especially pronounced during reasoning-intensive decoding, such as Chain-of-Thought (CoT) generation, where intermediate steps introduce further language instability. In this paper, we systematically study output language drift in multilingual RAG across multiple datasets, languages, and LLM backbones. Our controlled experiments reveal that the drift results not from comprehension failure but from decoder-level collapse, where dominant token distributions and high-frequency English patterns dominate the intended generation language. We further observe that English serves as a semantic attractor under cross-lingual conditions, emerging as both the strongest interference source and the most frequent fallback language.

To mitigate this, we propose Soft Constrained Decoding (SCD), a lightweight, training-free decoding strategy that gently steers generation toward the target language by penalizing non-target-language tokens. SCD is model-agnostic and can be applied to any generation algorithm without modifying the architecture or requiring additional data. Experiments across three multilingual datasets and multiple typologically diverse languages show that SCD consistently improves language alignment and task performance, providing an effective and generalizable solution in multilingual RAG.

Code and Dataset — https://github.com/pkuserc/SCD

1 Introduction

Recent advances in Retrieval-Augmented Generation (RAG) have significantly enhanced large language models' ability to generate factually grounded answers in open-domain question answering (Xu et al. 2024a; Luo et al. 2024; Fang et al. 2024; Shi et al. 2024). However, in multilingual settings, most existing studies have focused on improving cross-lingual retrieval performance and contextual alignment (Luo et al. 2020; Park and Lee 2025; Chirkova et al. 2024; Wu et al. 2024; Ranaldi, Haddow, and Birch 2025; Ranaldi et al. 2025; Liu et al. 2025; Bland'on et al. 2025), while overlooking a critical issue: the mismatch between the input and output languages.

Figure 1: Illustration of language drift in multilingual RAG. The user query and in-context examples are provided in the target language (e.g., Chinese), while the retrieved context is written in a non-target language (e.g., English). During reasoning, the model mixes languages and ultimately outputs the final answer in a non-target language.

In multilingual RAG settings, the query, instructions, and in-context exemplars are typically written in the target language, aiming to elicit responses in that language. However, due to the predominance of English in open-domain corpora (Xu et al. 2024b; Zeng and Yang 2024; Resnik and Smith 2003), the retrieved context is often in English, even when the query is in another language. This creates a mixed-lingual input scenario where only the retrieved context differs in language. Nevertheless, models frequently generate responses in the language of the retrieved content rather than in the intended target language (Park and Lee 2025; Liu et al. 2025). We refer to this phenomenon as *output language drift*, which poses practical challenges for multilingual applications yet remains underexplored. Our empirical study reveals that such cross-lingual conditions negatively affect both task performance and output language consistency. Notably, English serves as the strongest interference source, significantly degrading output quality in non-English settings, while also serving as the most robust target language when subjected to interference. This issue becomes more severe under few-shot prompting and Chain-of-Thought (CoT) reasoning (Shi et al. 2022; Yu et al. 2025).

Interestingly, we observe that language drift does not nec-

*Corresponding Author.

essarily follow the language of the retrieved context. Instead, models frequently default to English during generation, even when the context passages are in Arabic, Russian, or other non-English languages. This indicates that English plays a dominant role beyond being a common training language: it functions as a semantic attractor in multilingual generation. Our analysis indicates that, under cross-lingual ambiguity, LLMs tend to prefer English over the context language. This fallback tendency further verifies the dominant role of English as the default trajectory in multilingual decoding.

To better understand whether this fallback behavior results from misunderstanding or from generative biases, we conduct human evaluation and reference translation. Interestingly, many of the outputs that drift to the non-target language are still semantically faithful, indicating that the model has accurately understood both the task and the retrieved context. By analyzing intermediate reasoning steps (i.e., CoT traces), we find that the language inconsistency often emerges mid-generation, even when earlier steps remain in the target language. This indicates that the failure stems not from semantic comprehension, but from generation biases favoring frequent English tokens. As a result, the model produces outputs that are structurally fluent but linguistically inconsistent, reflecting a form of language collapse driven by token-level priors rather than task misunderstanding.

These findings motivate the need for lightweight decoding-time strategies that maintain output language consistency without compromising reasoning performance. To this end, we introduce Soft Constrained Decoding (SCD), a token-level control mechanism that assigns soft penalties to non-target-language tokens, thereby encouraging target-language generation while preserving fluency. In contrast to hard vocabulary filtering, SCD is a flexible, model-agnostic mechanism compatible with standard decoding algorithms. Extensive experiments across diverse datasets, model backbones, target languages, and context languages demonstrate that SCD improves both output language alignment and answer quality, providing a practical solution to a persistent yet underexplored challenge in multilingual RAG.

Our main contributions are as follows:

- **Multilingual Dataset Construction.** We construct multilingual versions of HotpotQA, MuSiQue, and DuReader by translating and human-verifying all components (queries, answers, prompts, exemplars, and retrieved context), enabling controlled evaluation across four diverse languages.
- **Analysis of Language Drift.** We conduct controlled experiments that vary only the language of retrieved contexts, revealing overlooked patterns in multilingual RAG such as performance degradation, target-language inconsistency, and a strong fallback tendency to English. Chain-of-Thought traces show that drift typically arises mid-generation due to decoding-time biases.
- **Training-free Language Control.** We introduce SCD, a lightweight decoding-time method that softly penalizes non-target tokens. SCD is model-agnostic, requires no training, and improves both output language consistency and task accuracy across datasets and LLMs.

2 Language Drift in Multilingual RAG

In this section, we conduct a comprehensive empirical investigation into the phenomenon of language drift, where model outputs deviate from the intended target language during multilingual RAG generation. To support this study, we construct multilingual variants of several benchmark RAG datasets by translating and aligning all critical components, including queries, answers, prompts, exemplars, and retrieved passages. We then evaluate LLM behavior across a range of controlled conditions. Our findings reveal a set of systematic behaviors that undermine both task accuracy and output language alignment under cross-lingual conditions.

2.1 Multilingual Dataset Construction

To systematically evaluate how multilingual retrieved context influences LLM behavior in RAG, we require datasets in which the language of each input component can be independently controlled. This enables us to isolate the impact of cross-lingual retrieved passages on model reasoning and output consistency. However, no existing benchmark satisfies these constraints while remaining compatible with RAG. To address this gap, we construct multilingual versions of three widely used QA datasets that support retrieval augmentation: HotpotQA (Yang et al. 2018), MuSiQue (Trivedi et al. 2022), and DuReader[1]. These datasets contain high-quality question–answer pairs with human-annotated gold retrieved context, making them well-suited for our purpose. We select four typologically diverse languages, English (`EN`), Chinese (`ZH`), Arabic (`AR`), and Russian (`RU`), to capture a broad range of linguistic variation. Representative data examples and the format used for multilingual annotation are provided in Appendix F.

Each dataset contributes 1,000 samples. For every sample, we prepare five components: a user query, a reference answer, several gold retrieved contexts, a prompt template, and several in-context exemplars. All components are translated into the four languages using GPT-4o, followed by manual verification to ensure semantic fidelity and natural fluency. This multilingual suite enables flexible and language-controlled experimentation across a wide range of configurations.

2.2 Experimental Setup

Based on the multilingual datasets described above, we design a controlled experimental framework to evaluate how the language of the retrieved context influences output behavior in RAG. In our core setup, we fix the language of the query, prompt, and ICL examples to the target language (denoted as the context language), and vary only the language of the retrieved passage to isolate cross-lingual interference effects.

We test across three datasets (HotpotQA, MuSiQue, DuReader) and four target languages (`EN`, `ZH`, `AR`, `RU`), using two instruction-tuned LLMs as backbones: LLaMA3-8B-Instruct (Grattafiori et al. 2024) and Qwen2.5-7B-Instruct (Yang et al. 2024, 2025). All generations are performed using default decoding parameters, and each prompt

[1]https://github.com/baidu/DuReader

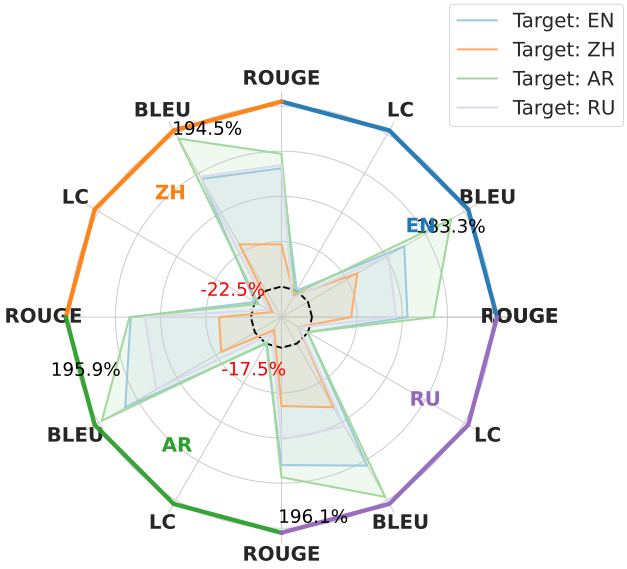

Figure 2: Relative performance gap between with-ICL and without-ICL settings across different target and context language combinations on the HotpotQA dataset, visualized as a polar radar chart. Each axis corresponds to one evaluation metric (ROUGE, BLEU, or LC) under a specific context language (EN, ZH, AR, or RU), totaling 12 axes. Solid lines represent different target languages, color-coded accordingly. Values indicate the percentage difference between ICL and non-ICL performance under each configuration. The black dashed ring at 0% denotes no change. Labels mark the highest gains and the most severe LC degradations. The chart reveals that ICL consistently improves BLEU and ROUGE, but often reduces language consistency, especially under ZH and RU contexts.

includes four ICL exemplars in the same language as the query.

For evaluation, we report standard BLEU-1/2/3 and ROUGE-1/2/L scores, along with their averaged variants (**BLEU** and **ROUGE**), using reference answers in the target language as the gold standard. To further quantify language fidelity, we introduce a **Language Consistency (LC)** metric, which measures the proportion of generated responses written in the expected target language. This comprehensive metric suite allows us to jointly evaluate reasoning accuracy and language control in multilingual RAG settings.

2.3 ICL Improves Performance but Undermines Consistency

To investigate how multilingual retrieved context and ICL jointly affect RAG performance, we conduct controlled experiments across various target–context language pairs. Specifically, we fix the query, prompt, and exemplars in the target language (EN, ZH, AR, or RU) and vary only the language of the retrieved context. For each configuration, we compare model outputs with and without ICL exemplars, allowing us to isolate the effects of ICL under multilingual interference.

Figure 2 summarizes these effects using a radar chart on the HotpotQA dataset with `LLaMA3-8B-Instruct` as the backbone. Each colored line represents a fixed target language, while each radial group corresponds to one context language (EN, ZH, AR, RU), covering three evaluation metrics: ROUGE, BLEU, and LC. The plotted values represent the *relative percentage change* introduced by ICL compared to the non-ICL baseline under the same configuration. Positive values indicate improvements, whereas negative values reflect degradation. As a concrete example, the green point within the orange ZH-labeled frame represents the ZH-AR condition. It shows that ICL increases BLEU significantly but reduces LC, reflecting the common pattern where richer reasoning comes at the cost of linguistic fidelity under cross-lingual retrieved context. Due to space constraints, we report the radar plot results only for HotpotQA, which is representative of the broader trends. Similar patterns are observed across other datasets, languages, and backbone models; detailed results are included in Appendix D. Our results in Figure 2 reveal two key findings:

- **Multilingual interference degrades both performance and consistency.** When the retrieved context is in a language different from the target, both task performance (measured by BLEU and ROUGE) and output language consistency decline significantly. Notably, we observe that **English acts as the strongest interfering language**: when used as cross-lingual retrieved context, it induces the most severe performance degradation across non-English targets. For example, in the ICL setting with ZH as the target language, language consistency drops from 92.0% to 68.4% when switching retrieved contexts from ZH to EN retrieved context, with a drop in average BLEU score from 0.212 to 0.086. In contrast, EN exhibits the **strongest resistance to interference** when serving as the target language, while ZH shows the **greatest sensitivity** across all datasets.

- **In-context learning improves performance but worsens consistency.** Adding ICL examples consistently improves generation quality across all datasets and models. However, it also intensifies output language drift, leading the model to deviate further from the expected target language. For example, with RU as the target language, the average ROUGE increases from 0.193 to 0.373 after adding ICL, while language consistency drops from 0.991 to 0.895. Similar trends are observed when the context language differs from the target language: ICL improves accuracy but significantly reduces alignment with the expected output language.

These findings indicate that while ICL improves semantic fidelity, it also increases vulnerability to language drift due to extended reasoning and exposure to non-target-language tokens. Since ICL reflects real-world usage and consistently improves performance, we adopt it as the default in all experiments, with prompts explicitly instructing the model to generate in the target language.

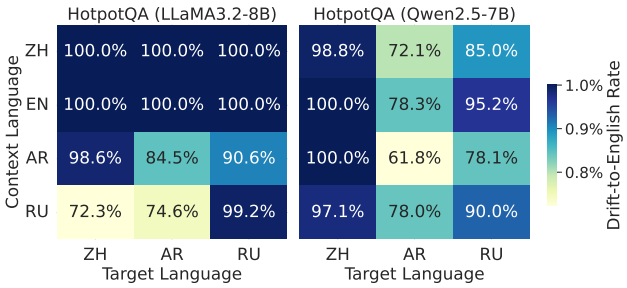

Figure 3: Language drift patterns on the HotpotQA dataset for `LLaMA3-8B` and `Qwen2.5-7B` models. Each cell shows the percentage of inconsistent outputs that are generated in English (`EN`). Both models exhibit a strong fallback tendency toward English across all cross-lingual settings.

2.4 English as the Default Fallback Language

While previous results show that cross-lingual interference reduces output consistency, we further investigate *which language the model tends to generate* when it fails to remain in the target language. Specifically, we analyze all inconsistent outputs and identify their actual output language. Strikingly, we observe that in the majority of drift cases across all target languages and datasets, the model defaults to generating in `EN` regardless of whether the retrieved context is `EN`, as shown in Figure 3. Due to the space limitations, additional results with similar conclusions are provided in the Appendix A.

This fallback behavior suggests that `EN` plays a dominant role not only in training but also during decoding. Rather than aligning with the context language, the model often defaults to `EN` when facing ambiguity, a tendency driven by structural biases such as the over-representation of English tokens during pretraining and the concentration of factual knowledge in `EN`. Our experiments further confirm that even when both the target languages and context are non-English, misaligned outputs predominantly appear in `EN`, indicating that language drift is not random but guided by `EN` acting as a default semantic attractor.

2.5 Language Collapse During Decoding

To assess whether the observed language drift arises from comprehension failure or unstable decoding behavior, we conduct a semantic agreement analysis. As shown in Table 1, we compare three evaluation metrics under cross-lingual settings: (1) Standard ROUGE between the model output and the target-language reference; (2) ROUGE after translating drifted outputs back into the target language and recomputing scores against the original reference (denoted as ROUGE(T)); (3) Semantic Match Rate, scored by GPT-4o, which evaluates whether the model output is factually aligned with the reference regardless of surface language. We observe that translation leads to a *significant* improvement in ROUGE scores. For example, ROUGE increases from 0.182 to 0.263 for `ZH`, and from 0.333 to 0.388 for `RU`, indicating that the original outputs are semantically aligned despite being expressed in the wrong language. Moreover, the Semantic Match Rate further confirms that even when ROUGE is low, the match rate often exceeds 60% for `RU` and over 50% for `ZH` and `AR`, demonstrating strong task understanding. These findings suggest that language drift stems not from comprehension failure but from decoder-level instability. Additional results on other datasets (see Appendix B) show similar patterns across models and languages.

This pattern suggests a language collapse during decoding, where the LLM correctly processes the input and understands the intended task but fails to maintain the target language throughout generation. We hypothesize that this issue arises from token-level priors learned during pretraining, as English tokens tend to dominate due to their higher frequency, more stable syntactic structures, and richer factual coverage. During multi-step reasoning, especially under CoT prompting, such biases can override explicit language instructions and gradually shift the generation toward English. The drift typically unfolds over time, with the generation beginning in the target language but progressively deviating into English. This highlights a fundamental limitation in multilingual LLMs: **strong semantic reasoning does not guarantee stable language control during generation.**

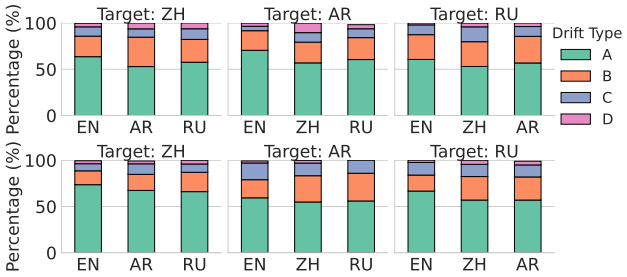

Figure 4: Distribution of four language drift types across different target–context language pairs in the HotpotQA dataset. Each subplot corresponds to a fixed target language (`ZH`, `AR`, `RU`), with the x-axis denoting the context language. The top row displays results for `LLaMA3-8B`, and the bottom row for `Qwen2.5-7B`.

2.6 Types of Language Drift Behaviors

To better understand how language drift manifests in multilingual reasoning, we categorize drifted outputs into four distinct behavioral types based on multilingual generations. We randomly sampled 1,000 language-inconsistent outputs and had them manually annotated by three trained reviewers with backgrounds in linguistics or multilingual NLP. The taxonomy includes: **Type A:** *Named Entity Representation Divergence*, where inconsistent transliteration or spelling results in mismatches despite semantic equivalence; **Type B:** *Answer Target Shift*, where the model alters answer granularity or is misled by context-language cues, leading to an incorrect sub-answer; **Type C:** *Reasoning Chain Misalignment*, where the CoT path becomes structurally disrupted due to language mixing or code-switching; and **Type D:** *Conceptual Reference Shift*, where cultural or semantic biases embedded in the dominant language (such as English)

Targe Language	Context Language	LLaMA3-8B ROUGE	LLaMA3-8B ROUGE(T)	LLaMA3-8B Semantic Match Rate	Qwen2.5-7B ROUGE	Qwen2.5-7B ROUGE(T)	Qwen2.5-7B Semantic Match Rate
ZH	EN	0.182	0.263	54.7%	0.331	0.352	62.7%
	AR	0.211	0.258	46.2%	0.342	0.366	55.3%
	RU	0.209	0.261	49.4%	0.337	0.359	54.2%
AR	EN	0.294	0.331	53.4%	0.201	0.221	46.4%
	ZH	0.265	0.288	48.3%	0.187	0.202	42.2%
	RU	0.280	0.303	50.0%	0.206	0.220	43.0%
RU	EN	0.333	0.388	62.8%	0.240	0.262	60.1%
	ZH	0.335	0.367	59.1%	0.248	0.257	56.1%
	AR	0.339	0.361	62.9%	0.248	0.252	59.0%

Table 1: Performance under cross-lingual retrieved context for non-English target languages (ZH, AR, RU) using LLaMA3-8B and Qwen2.5-7B on HotpotQA. We report standard ROUGE, ROUGE after translating the model output to the target language (ROUGE(T)), and Semantic Match Rate assessed by GPT. Despite language drift, many outputs remain semantically correct, highlighting decoder-level instability rather than comprehension failure.

trigger unintended knowledge concepts. Full category definitions and examples are provided in Appendix C.

We use GPT-4o to classify a representative set of drifted outputs into the four categories defined in our taxonomy, followed by manual verification to ensure label quality. As shown in Figure 4, the most common behavior across both models and all target languages is *Named Entity Representation Divergence* (Type A), which accounts for approximately 55% to 74% of drifted cases on average. This is followed by *Answer Target Shift* (Type B), occurring in roughly 17% to 31% of cases, with greater variation across context languages. *Reasoning Chain Misalignment* (Type C) is less frequent, comprising around 9% to 18%, while *Conceptual Reference Shift* (Type D) remains rare, often below 5%.

These findings suggest that most drift cases arise from surface-level inconsistencies, such as entity formatting or answer phrasing, rather than from deeper reasoning failures. Recognizing how such drift emerges during the later stages of CoT decoding can inform more targeted control strategies, including applying penalties for answer-level deviations or reinforcing consistency in entity representation.

3 Soft-Constrained Decoding

3.1 Soft-Constrained Decoding (SCD)

To mitigate output language drift in multilingual generation, we propose **Soft-Constrained Decoding (SCD)**, a lightweight decoding-time control strategy that incorporates token-level language awareness into the generation process. Instead of applying rigid vocabulary restrictions, SCD subtly adjusts the token probability distribution to favor the target language, while preserving open-ended reasoning capabilities and fluent output.

Token Categorization. Let $\mathcal{V}$ denote the model vocabulary, and we partition $\mathcal{V}$ into three disjoint sets:

- $\mathcal{V}_{\text{target}}$: tokens associated with the *target language*,
- $\mathcal{V}_{\text{neutral}}$: *neutral tokens* such as punctuation, digits, and shared symbols,
- $\mathcal{V}_{\text{distractor}}$: tokens linked to *non-target languages*.

This categorization is performed via Unicode ranges or tokenizer-based heuristics and cached prior to generation.

Logits Adjustment. Let $\mathbf{z}^{(t)} \in \mathbb{R}^{|\mathcal{V}|}$ be the raw logits output at decoding step t. SCD adjusts $\mathbf{z}^{(t)}$ before softmax as follows:

$$\tilde{z}_i^{(t)} = \begin{cases} \alpha z_i^{(t)}, & \text{if } i \in \mathcal{V}_{\text{target}} \\ z_i^{(t)}, & \text{if } i \in \mathcal{V}_{\text{neutral}} \\ \beta z_i^{(t)}, & \text{if } i \in \mathcal{V}_{\text{distractor}} \end{cases}$$

Here, $\alpha > 1.0$ is a soft boost to target-language tokens, and $\beta < 1.0$ is a penalty for distractor-language tokens. This modification biases generation while preserving flexibility.

Cold Start Smoothing. Multilingual LLMs, especially in low-resource languages, often generate unstable initial outputs such as repeated prompts or template fragments. To minimize such disruptions, we introduce a *warm-up period* by delaying the activation of language constraints until decoding step T_{start}. This design ensures a fluent transition into reasoning before language control is applied.

Integration. SCD is *model-agnostic* and fully compatible with standard decoding algorithms. It requires no additional training or architectural changes.

SCD operates as a lightweight decoding-time strategy that gently discourages the selection of non-target language tokens without eliminating them entirely. By incorporating language awareness directly into the token selection process, SCD guides the model to favor tokens in the target language while retaining the flexibility needed for open-ended reasoning.

3.2 Experimental Setup and Baselines

We evaluate our proposed SCD on three multilingual retrieval-augmented QA datasets, i.e., HotpotQA, MuSiQue, and DuReader, which are described in Section 2.1. Experiments are conducted using two instruction-tuned LLMs:

Targe Language	Context Language	HotpotQA			Musique			DuReader		
		ROUGE	BLEU	LC	ROUGE	BLEU	LC	ROUGE	BLEU	LC
Prompted Language Instruction										
ZH	EN	0.182	0.086	68.4%	0.187	0.097	63.9%	0.339	0.166	84.2%
	AR	0.211	0.106	77.7%	0.181	0.089	76.5%	0.358	0.175	90.1%
	RU	0.209	0.107	79.5%	0.169	0.087	64.5%	0.343	0.168	83.1%
AR	EN	0.294	0.162	85.4%	0.144	0.080	90.0%	0.209	0.099	88.2%
	ZH	0.265	0.143	88.4%	0.120	0.057	89.2%	0.193	0.080	87.0%
	RU	0.280	0.151	88.6%	0.121	0.061	89.8%	0.186	0.077	89.5%
RU	EN	0.333	0.177	80.2%	0.218	0.119	81.9%	0.285	0.150	84.3%
	ZH	0.335	0.172	85.1%	0.206	0.102	90.2%	0.296	0.149	85.8%
	AR	0.339	0.179	86.8%	0.214	0.109	92.5%	0.288	0.143	90.9%
Translation-Based Evaluation										
ZH	EN	0.263	0.135	100.0%	0.257	0.142	100.0%	0.366	0.178	100.0%
	AR	0.258	0.132	100.0%	0.214	0.105	100.0%	0.364	0.177	100.0%
	RU	0.261	0.136	100.0%	0.235	0.124	100.0%	0.365	0.175	100.0%
AR	EN	0.331	0.183	100.0%	0.168	0.095	100.0%	0.231	0.114	100.0%
	ZH	0.288	0.156	100.0%	0.135	0.066	100.0%	0.202	0.087	100.0%
	RU	0.303	0.165	100.0%	0.140	0.074	100.0%	0.195	0.083	100.0%
RU	EN	0.388	0.218	100.0%	0.258	0.148	100.0%	0.314	0.167	100.0%
	ZH	0.367	0.196	100.0%	0.215	0.109	100.0%	0.309	0.156	100.0%
	AR	0.361	0.196	100.0%	0.217	0.114	100.0%	0.293	0.148	100.0%
Soft-Constrained Decoding (Ours)										
ZH	EN	0.306	0.155	90.6%	0.276	0.146	91.8%	0.403	0.190	95.2%
	AR	0.283	0.146	93.9%	0.234	0.118	94.8%	0.408	0.195	96.6%
	RU	0.293	0.156	92.5%	0.243	0.130	92.3%	0.404	0.190	95.7%
AR	EN	0.352	0.197	96.4%	0.187	0.106	98.8%	0.241	0.113	96.7%
	ZH	0.312	0.170	95.5%	0.157	0.079	97.6%	0.236	0.104	94.1%
	RU	0.326	0.183	96.3%	0.152	0.080	98.0%	0.220	0.092	95.4%
RU	EN	0.422	0.238	95.4%	0.270	0.162	94.1%	0.334	0.174	94.4%
	ZH	0.400	0.216	94.1%	0.230	0.126	94.7%	0.335	0.165	94.3%
	AR	0.392	0.216	94.0%	0.232	0.128	94.3%	0.317	0.155	94.7%

Table 2: Performance comparison across three language control strategies: Prompted Language Instruction, Translation-Based Evaluation, and SCD on three multilingual RAG datasets. We report results for `LLaMA3-8B`, where SCD consistently improves both LC and content metrics across datasets compared to strong baselines. Results for `Qwen2.5-7B` are provided in Appendix E due to space constraints.

LLaMA3-8B-Instruct and Qwen2.5-7B-Instruct. We empirically find moderate settings ($\alpha = 1.1$, $\beta = 0.9$, $T_{\text{start}} = 5$) to balance language fidelity and semantic fluency in SCD.

To benchmark SCD against other lightweight language control strategies, we compare it with the following decoding-time baselines: (1) **Prompted Language Instruction**: Explicitly appending an instruction in the prompt that requests answers to be generated in the target language; (2) **Translation-Based Evaluation**: Evaluating drifted outputs by translating them back into the target language using the same LLM, before computing BLEU/ROUGE scores; (3) **Vocabulary Restriction Decoding**: Restricting the decoding space to tokens belonging to the target language only, effectively applying a hard constraint on generation.

We evaluate all methods using three complementary metrics: (1) BLEU (mean of BLEU-1/2/3), (2) ROUGE (mean of ROUGE-1/2/L), and (3) language consistency (LC), defined as the percentage of outputs generated in the correct target language. All decoding parameters follow the default settings of each model, and no task-specific or model-specific fine-tuning is applied. Additional performance improvements may be obtained by tuning decoding parameters, we leave this for future work. All reported scores are averaged over five independent runs to reduce randomness.

3.3 Effectiveness of Soft-Constrained Decoding

As shown in Table 2, SCD consistently outperforms existing language control methods, achieving notable improvements in both *language consistency* (LC) and *semantic generation quality*, as measured by average BLEU and ROUGE scores. These results support our central hypothesis that maintaining alignment with the target language can reinforce, rather than hinder, the coherence and accuracy of reasoning paths.

Across all datasets and language configurations, SCD consistently improves both language consistency and content quality compared to the Prompted Language Instruction

Targe Lang.	Context Lang.	ROUGE			CoT Length		
		PLI	VRD	SCD	PLI	VRD	SCD
ZH	EN	0.182	0.155	0.306	104.0	38.6	134.9
	AR	0.211	0.184	0.283	103.5	40.2	142.4
	RU	0.209	0.173	0.293	77.6	42.8	143.1
AR	EN	0.294	0.295	0.352	77.0	50.5	90.2
	ZH	0.265	0.266	0.312	86.4	49.9	92.9
	RU	0.280	0.281	0.326	86.3	57.4	100.2
RU	EN	0.333	0.343	0.422	85.6	58.8	111.8
	ZH	0.335	0.339	0.400	89.4	56.1	111.0
	AR	0.339	0.341	0.392	89.4	56.6	111.8

Table 3: Comparison of three decoding strategies on HotpotQA across ROUGE score and average CoT length.

baseline. For instance, under the challenging ZH−EN condition on HotpotQA, SCD increases LC from 68.4% to 90.6%, while also boosting BLEU from 0.086 to 0.155 and ROUGE from 0.182 to 0.306. Similar trends are observed for other target languages such as AR and RU, with LC improvements ranging from 10 to 22 percentage points.

While the translation-based method trivially achieves 100% LC by converting drifted outputs into the target language after generation, it often underperforms SCD in BLEU and ROUGE. This outcome is expected, as translation does not recover the original reasoning trajectory but merely reformulates its surface form. Moreover, translation-based evaluation adds additional complexity, increases inference cost, and may amplify noise when the original outputs are incomplete or syntactically broken.

The above results demonstrate that SCD is a practical, lightweight, and model-agnostic decoding-time intervention. It requires no additional training or architectural modifications, and can be seamlessly integrated into standard decoding workflows (e.g., greedy, sampling, top-p). Across models, languages, and datasets, SCD provides consistent and substantial improvements in both linguistic alignment and semantic quality, making it a strong candidate for real-world multilingual RAG and generation-based applications.

3.4 Should Multilingual Generation Be Fully Language Isolated?

To examine the trade-offs of different language control strategies in multilingual generation, we compare three decoding methods: Prompted Language Instruction (PLI), Vocabulary-Restricted Decoding (VRD), and our proposed SCD. PLI uses explicit prompts to enforce the target language; VRD imposes hard constraints by restricting generation to target-language tokens; and SCD softly penalizes non-target tokens while maintaining generation flexibility.

As shown in Table 3, SCD consistently achieves the highest ROUGE scores across all target–context language pairs on HotpotQA. For instance, in the ZH-EN setting, SCD reaches 0.306 ROUGE, compared to 0.155 under VRD and 0.182 under PLI. Similar trends are observed for AR and RU targets. Interestingly, VRD often underperforms PLI, suggesting that overly strict language filtering can suppress useful multilingual cues and degrade output quality, despite improving consistency.

To assess generation dynamics, we compare the average length of generated CoT responses. VRD consistently yields the shortest outputs, e.g., only 38.6 tokens in ZH-EN, compared to 104.0 with PLI and 134.9 with SCD—indicating that hard constraints truncate reasoning. In contrast, SCD preserves longer and more complete reasoning chains by allowing controlled cross-lingual flexibility. We further analyze how reasoning length affects language drift and control effectiveness in Appendix G, where SCD demonstrates robust performance across various CoT trajectories.

These results suggest that *effective multilingual generation does not require full language isolation*. Allowing limited access to non-target tokens during reasoning, while softly guiding the output toward the desired language, improves both language consistency and semantic fidelity.

4 Related Work

Multilingual RAG has received increasing attention as a means to enhance LLMs with access to cross-lingual knowledge. Prior research has primarily focused on improving the quality of multilingual retrieval (Liu et al. 2025; Chirkova et al. 2024; Ranaldi, Haddow, and Birch 2025), aligning retrieved passages with user queries across languages (Ranaldi et al. 2025; Bland'on et al. 2025), and adapting RAG pipelines to typologically diverse settings (Wu et al. 2024; Zeng and Yang 2024). These efforts have significantly advanced retrieval-stage effectiveness in non-English tasks and established multilingual evaluation protocols. Some recent works have further explored language preferences in RAG models (Park and Lee 2025; Shi et al. 2022; Yu et al. 2025), highlighting accuracy disparities across languages. However, most of these works either evaluate generation outcomes at the answer level or focus on upstream retrieval modules, without deeply investigating how language behavior evolves throughout the decoding process.

In contrast, we focus on the overlooked issue of *language drift* in multilingual RAG, where model outputs shift away from the target language during reasoning. We demonstrate that this drift arises during decoding, with English acting as a default fallback. To mitigate it, we propose a lightweight decoding-time strategy that improves language alignment without requiring model retraining.

5 Conclusion

This work addresses a key challenge in multilingual RAG: large language models often generate outputs in unintended languages when reasoning over cross-lingual evidence. Through controlled experiments and CoT analysis, we find that such drift arises from decoder-stage biases rather than comprehension failure. To mitigate this, we introduce SCD, a lightweight, model-agnostic strategy that softly penalizes non-target-language tokens. SCD consistently enhances both language consistency and task performance across models, languages, and datasets. These findings underscore the value of decoding-time control for building more robust and controllable multilingual RAG systems.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62276089), the Natural Science Foundation of Tianjin (Grant No. 24JCJQJC00200 and Grant No. 24JCQNJC01230), the Natural Science Foundation of Hebei Province (Grant No. F2024202064), the Science Research Project of Hebei Education Department (Grant No. BJ2025004), the Ministry of Human Resources and Social Security of China (Grant No. RSTH-2023-135-1), and the Science and Technology Program of Hebei Province (Grant No. 24464401D).

References

Bland'on, M. A. C.; Talur, J.; Charron, B.; Liu, D.; Mansour, S.; and Federico, M. 2025. MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation. *ArXiv*, abs/2502.17163.

Chirkova, N.; Rau, D.; D'ejean, H.; Formal, T.; Clinchant, S.; and Nikoulina, V. 2024. Retrieval-augmented generation in multilingual settings. *ArXiv*, abs/2407.01463.

Fang, F.; Bai, Y.; Ni, S.; Yang, M.; Chen, X.; and Xu, R. 2024. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. In *Annual Meeting of the Association for Computational Linguistics*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liu, W.; Trenous, S.; Ribeiro, L. F. R.; Byrne, B.; and Hieber, F. 2025. XRAG: Cross-lingual Retrieval-Augmented Generation.

Luo, F.; Wang, W.; Liu, J.; Liu, Y.; Bi, B.; Huang, S.; Huang, F.; and Si, L. 2020. VECO: Variable and Flexible Cross-lingual Pre-training for Language Understanding and Generation. In *Annual Meeting of the Association for Computational Linguistics*.

Luo, K.; Liu, Z.; Xiao, S.; Zhou, T.; Chen, Y.; Zhao, J.; and Liu, K. 2024. Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models. In *Annual Meeting of the Association for Computational Linguistics*.

Park, J.; and Lee, H. 2025. Investigating Language Preference of Multilingual RAG Systems. *ArXiv*, abs/2502.11175.

Ranaldi, L.; Haddow, B.; and Birch, A. 2025. Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task. *ArXiv*, abs/2504.03616.

Ranaldi, L.; Ranaldi, F.; Zanzotto, F. M.; Haddow, B.; and Birch, A. 2025. Improving Multilingual Retrieval-Augmented Language Models through Dialectic Reasoning Argumentations. *ArXiv*, abs/2504.04771.

Resnik, P.; and Smith, N. A. 2003. The Web as a Parallel Corpus. *Computational Linguistics*, 29: 349–380.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; Das, D.; and Wei, J. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. *ArXiv*, abs/2210.03057.

Shi, Z.; Sun, W.; Gao, S.; Ren, P.; Chen, Z.; and Ren, Z. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. *arXiv preprint arXiv:2406.14891*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*.

Wu, S.; Tang, J.; Yang, B.; Wang, A.; Jia, K.; Yu, J.; Yao, J.; and Su, J. 2024. Not All Languages are Equal: Insights into Multilingual Retrieval-Augmented Generation. *ArXiv*, abs/2410.21970.

Xu, S.; Pang, L.; Yu, M.; Meng, F.; Shen, H.; Cheng, X.; and Zhou, J. 2024a. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. *ArXiv*, abs/2402.18150.

Xu, Y.; Hu, L.; Zhao, J.; Qiu, Z.; Ye, Y.; and Gu, H. 2024b. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *Frontiers Comput. Sci.*, 19: 1911362.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing*.

Yu, Z.; Li, T.; Wang, C.; Chen, H.; and Zhou, L. 2025. Cross-Lingual Consistency: A Novel Inference Framework for Advancing Reasoning in Large Language Models. *ArXiv*, abs/2504.01857.

Zeng, J.; and Yang, J. 2024. English language hegemony: retrospect and prospect. *Humanities and Social Sciences Communications*, 11: 1–9.