Language-Coupled Reinforcement Learning for Multilingual Retrieval-Augmented Generation

Rui Qi[1,2], Fengran Mo[3], Yufeng Chen[1,2†], Xue Zhang[1,2], Shuo Wang[4], Hongliang Li[1,2]
Jinan Xu[1,2], Meng Jiang[5], Jian-Yun Nie[3], Kaiyu Huang[1,2†]

[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation
（Beijing Jiaotong University), Ministry of Education
[2]School of Computer Science and Technology, Beijing Jiaotong University
[3]University of Montreal [4]Tsinghua University [5]University of Notre Dame
{cherry,kyhuang}@bjtu.edu.cn

Abstract

Multilingual retrieval-augmented generation (MRAG) requires models to effectively acquire and integrate beneficial external knowledge from multilingual collections. However, most existing studies employ a unitive process where queries of equivalent semantics across different languages are processed through a single-turn retrieval and subsequent optimization. Such a "one-size-fits-all" strategy is often suboptimal in multilingual settings, as the models occur to knowledge bias and conflict during the interaction with the search engine. To alleviate the issues, we propose LcRL, a multilingual search-augmented reinforcement learning framework that integrates a language-coupled Group Relative Policy Optimization into the policy and reward models. We adopt the language-coupled group sampling in the rollout module to reduce knowledge bias, and regularize an auxiliary anti-consistency penalty in the reward models to mitigate the knowledge conflict. Experimental results demonstrate that LcRL not only achieves competitive performance but is also appropriate for various practical scenarios such as constrained training data and retrieval over collections encompassing a large number of languages. Our code is available at https://github.com/Cherry-qwq/LcRL-Open.

1 Introduction

Retrieval-augmented generation (RAG) has emerged as a powerful solution to incorporate external knowledge into Large Language Models (LLMs) to enhance the performance via retrieving supportive evidence for knowledge-intensive tasks, which alleviate the hallucinations and lack of knowledge in training corpora (Ni et al., 2024; Ding et al., 2025; Zhang et al., 2025b). In multilingual settings, the training data of LLMs

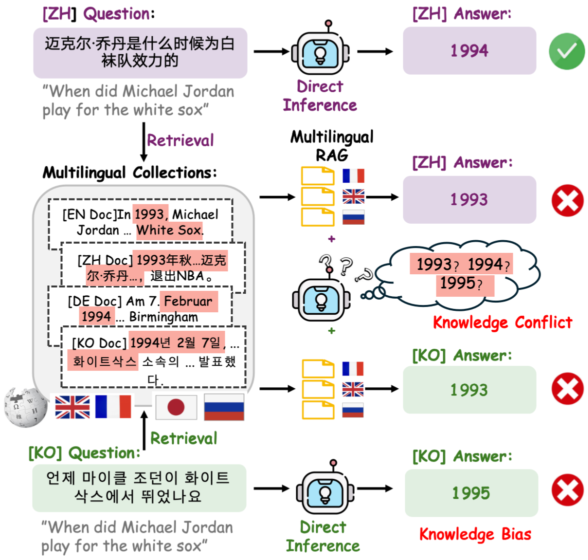


Figure 1: Illustration of knowledge bias and knowledge conflict challenges in the MRAG scenario.

is notoriously imbalanced and culturally biased, leading to significant knowledge disparities across languages (Huang et al., 2024; Ranaldi et al., 2025b; Zhang et al., 2025d; Man et al., 2025). Consequently, RAG remains essential, as it enables LLMs to leverage diverse multilingual collections to bridge these information gaps and enhance the models' capabilities in multiple languages.

Existing approaches in multilingual retrieval-augmented generation (MRAG) typically identify relevant passages from the multilingual collections for a given query as a single-turn retrieval process (Ranaldi et al., 2025a; Wu et al., 2024; Kumar et al., 2025; Ranaldi et al., 2025b; Park and Lee, 2025). Although showing effectiveness, the methods encounter two challenges (Reusens et al., 2023; Chirkova et al., 2024): *knowledge bias* and *knowledge conflict*. As shown in Figure 1, LLMs might generate divergent responses to the identical queries expressed in different languages, due to inherent disparities in knowledge and capabili-

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 14291–14310

ties specific to each language, a phenomenon referred to as knowledge bias. On the other hand, when the retrieval collections contain a wide variety of languages, differences in linguistic expression lead to retrieved documents that are semantically related but factually inconsistent, resulting in knowledge conflicts that hinder the LLM from producing correct responses. Based on these observations, a research question arises: Can LLMs adaptively determine the need for retrieving external knowledge and effectively harmonize contradictory knowledge obtained from different languages, particularly when dealing with a diverse and extensive set of multilingual collections?

To address these issues, the development of reinforcement learning (RL) based post-training provides the possibility to enable the LLMs achieve autonomous query generation and dynamic utilization of retrieved information through optimizing its rollout process (Jin et al., 2025; Wu et al., 2025a; Lupart et al., 2025; Mo et al., 2026a). In these efforts, LLMs can interact with search engines in an interleaved manner and support multi-turn reasoning. Such a paradigm provides the potential idea to dynamically adjust the retrieval and response generation strategies based on the complexity of queries in different languages among MRAG scenarios.

In this study, considering the complementary effect among languages, we propose a language-coupled reinforcement learning (LcRL) method that couples multilingual decision-making and experience reward into the search-augmented RL framework, aiming to mitigate the issues of knowledge bias and knowledge conflict in MRAG. In particular, LcRL adopts Group Relative Policy Optimization (GRPO) (DeepSeek-AI et al., 2025) with a language-coupled rollout mechanism, which mitigates the knowledge bias of multilingual queries through group scores, and breaks the language-independent pattern to address knowledge conflicts arising from factual inconsistencies in single-language queries over multilingual collections. For semantically equivalent queries, the language-coupled sampling generation within the same group focuses on optimizing the model's ability to interact with multilingual collections and perform reasoning in a multi-turn and multilingual paradigm. As a result, queries in different languages can adapt to their respective external retrieval requirements and implicitly learn the transferability from the behaviors of the other languages.

In terms of RL-based model training, we refine the outcome-based reward function by incorporating an n-gram-based recall score, which improves adaptability for multilingual generations of varying lengths. We observe that while the language-coupled GRPO framework exhibits robust cross-lingual transfer capabilities, it encounters significant training instability as the number of supported languages increases. To alleviate this issue, we introduce an auxiliary anti-consistency penalty regularization, which stabilizes the training process by assessing the correctness of the language-coupled response. Experimental results show that our method effectively accommodates increasing language diversity in the retrieval collections and exhibits strong performance and robustness under limited training data conditions. To sum up, our contributions are as follows:

- We propose LcRL, a multilingual search-augmented GRPO framework that is integrated with a language-coupled rollout module and reward modeling, achieving competitive performance on MRAG compared with several baselines.
- Our proposed multi-turn retrieval strategy with the auxiliary anti-consistency penalty regularization provides adaptive control of search needs in multilingual settings and exhibits enhanced performance with increasing language counts.
- Experiments show that LcRL performs strong generalization, which is minimally influenced by the scale and language diversity of training data, successfully transfers to unseen languages, and adapts various LLMs.

2 Related Work

Multilingual RAG. Early research in MRAG focused mainly on the empirical optimization of individual components within the pipeline, such as the introduction of translation strategies (Chirkova et al., 2024; Park and Lee, 2025; Wu et al., 2024; Ranaldi et al., 2025a; Zhang et al., 2025c). The other studies attempt to maximize the efficacy of LLMs across languages through dynamic adaptation (Kumar et al., 2025), including incorporating multi-step dialectic reasoning to alleviate the knowledge conflict during inference (Ranaldi et al., 2025b). However, existing approaches predominantly operate within a fixed pipeline framework,

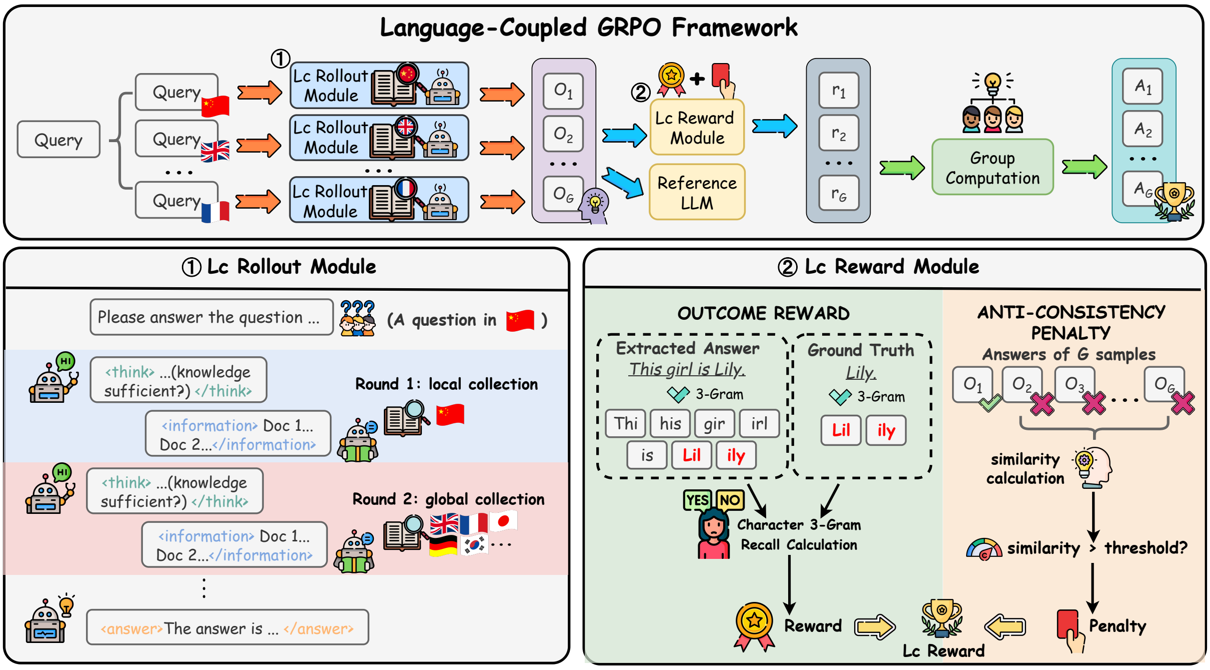


Figure 2: Illustration of our LCRLby integrating two modules (rollout and reward) within the GRPO for language-coupled (Lc) purposes. The Lc rollout is designed with sampling in multiple languages for multilingual scenarios. The Lc reward is incorporated with an anti-consistency mechanism to mitigate training collapse.

improving performance through point-wise modifications of isolated modules (Chirkova et al., 2024; Wang et al., 2024b). In contrast to these methods, our approach jointly optimizes retrieval and generation, which dynamically governs the retrieval strategy based on the inherent knowledge bias required for generation across different languages, while effectively mitigating knowledge conflicts in MRAG.

Reinforcement Learning for RAG. The integration of external tools has evolved from static RAG to dynamic, agentic workflows. Such methods, represented by Search-R1 (Jin et al., 2025), allow LLMs to autonomously generate multiple search queries, interact with search engines in a multi-step manner, and perform logical inference purely driven by reward signals (Kumar et al., 2024; Chen et al., 2025; DeepSeek-AI et al., 2025; Man et al., 2024; Zhang et al., 2025a, 2026) Building upon this paradigm, recent studies have significantly expanded the capabilities of search-augmented agents (Zhang et al., 2024; Luo et al., 2025; Song et al., 2025; Lin et al., 2025; Mo et al., 2026b), including efficiency (Xu et al., 2025; Lin et al., 2025) and structural comprehension (Wu et al., 2025a,b). Different from monolingual scenarios, multilingual settings exhibit knowledge conflicts and biases arising from multilingual capabilities. In such situations, current RL-based policies fail to reconcile the contradictory facts yielded by semantically equivalent queries across languages. To address this issue, we propose a language-coupled RL-based framework to solve knowledge bias and knowledge conflict in the rollout and reward phases, respectively.

3 Method

In monolingual scenarios, knowledge bias typically refers to systematic preferences induced by data distribution and model priors, while knowledge conflict arises from mutually inconsistent evidence within the same language corpus. However, the situation is different in RAG. In particular, **knowledge bias** stems from inherent disparities in training data across languages, leading to divergent reasoning capabilities even for semantically equivalent queries. **Knowledge conflict** occurs when top-ranked passages from a multilingual collection present factually inconsistent information, misleading the LLM due to information overload and cross-lingual contradictions. To address specific issues of MRAG, LCRL optimizes reasoning trajectories by coupling multilingual decision-making and experience rewards within a unified RL framework.

Algorithm 1: Multilingual Multi-Turn Retrieval-Augmented Generation

```
Require : Input query q_L, source language L, policy model π_θ, multilingual
          retriever set R_multi = {R_l}_{l∈ℒ}, maximum action budget B.
Ensure  : Final multilingual-informed response y.
Initialize rollout sequence y ← ∅
Initialize action count b ← 0
while b < B do
    Initialize current action rollout sequence y_b ← ∅
    while True do
        Generate response token y_t ∼ π_θ(· | q_L, y + y_b)
        Append y_t to rollout sequence y_b ← y_b + y_t
        if y_t ∈ [</search>, </answer>, <eos>] then
            break
    y ← y + y_b
    if <search> </search> detected in y_b then
        Extract search query
            q ← Parse(y_b, < search >, < /search >)
        Increment action count b ← b + 1
        // Hierarchical Retrieval Strategy
        if b = 1 then
            // Turn 1: Native Language Retrieval
            Retrieve results d = R_L(q)
        else if b = 2 then
            // Turn 2: Global Knowledge Expansion
            Retrieve results d = ⋃_{l∈ℒ\{L}} R_l(q)
        else
            // Turn ≥ 3: High-Resource Anchoring
            Retrieve results d = R_en(q)
        Insert d into rollout y ← y + <information>d</information>
    else if <answer> </answer> detected in y_b then
        return final generated response y
    else
        Ask for rethink
            y ← y + "My action is not correct. Let me rethink."
return final generated response y
```

3.1 Response Generation with Multi-Turn and Multilingual Retrieval

The rollout process for LLM response generation using an interleaved, multi-turn retrieval mechanism designed for the multilingual scenarios is described in Algorithm 1. We define the response generation process is as:

$$y \sim \pi_\theta(\cdot|q_L; \mathcal{R}_{multi}) = \pi_\theta(\cdot|q_L) \otimes \{\mathcal{R}_l\}_{l\in\mathcal{L}}$$

where q_L is the input query with the source language L, and $\mathcal{R}_{multi} = \{\mathcal{R}_l\}_{l\in\mathcal{L}}$ represents the set of all language-specific retrievers for each language l in the multilingual collection $\mathcal{L}$. The model executes a <think> phase to activate the parametric memory. If internal knowledge is insufficient, the model executes a hierarchical retrieval strategy by searching useful information from different language collections.

`Turn 1: Native Language Retrieval.` The model first invokes the retriever $\mathcal{R}_L$ corresponding specifically to the query's native language L, which can capture local context and avoid knowledge conflict by leveraging the culturally and contextually relevant evidence.

`Turn 2: Global Knowledge Expansion.` After processing the native results within a second <think> block, if the model identifies remaining gaps caused by knowledge bias, it triggers a second retrieval turn. The model simultaneously queries all other language collections in the set $\mathcal{L}\setminus\{L\}$, which is formalized as $\mathcal{D}_{global} = \bigcup_{l\in\mathcal{L}\setminus\{L\}} \mathcal{R}_l(q_L)$.

`Turn ≥ 3: High-Resource Anchoring.` For any additional turns, the model is navigated by default to a high-resource retriever (e.g., English $\mathcal{R}_{en}$) to serve as a factual anchor until an <answer> is produced or the maximum action budget B is consumed. This is determined by the broader information coverage and extensive corpus scale in terms of resource (Park and Lee, 2025).

Our multi-turn strategy empowers the model to assess its knowledge adequacy at each step autonomously. By prioritizing the native language in the initial turn, the model seeks to minimize information redundancy and circumvent potential conflicts. The model makes an adaptive decision to expand the search to multilingual, ensuring that cross-lingual complementarity is utilized to address core information deficits, while maintaining the center of native language knowledge to reduce the interference caused by redundant global data. The detailed prompting template used for training and interaction is provided in Appendix C.

3.2 Language-Coupled GRPO with the Search Engine Calling

As shown in the left part of Figure 2, to accommodate multilingual scenarios, we extend the GRPO framework by implementing a language-coupled rollout mechanism. This approach modifies the sampling process to explicitly mitigate linguistic bias and multilingual optimization conflicts.

Multilingual Group Sampling. For a specific question q, we construct a set of semantically equivalent queries across multiple languages: $\mathcal{Q} = \{q_1, q_2, \ldots, q_i, \ldots, q_n\}$, where q_i represents the corresponding language query. The multilingual queries originate from the parallel data expressed in other languages, which are usually unavailable in the datasets but can be generated using translation thinking strategies (Qi et al., 2025). After the interaction with the multilingual and multi-turn retrieval, for a group of size G, we sample responses o_i such that each o_i corresponds to a different language query from $\mathcal{Q}$ as:

$$o_i \sim \pi_\theta(\cdot|q_i; \mathcal{R}), \quad q_i \in \mathcal{Q}$$

Objective Function. Distinct from standard GRPO, which optimizes based on a single-query group, our objective operates over the coupled query set $\mathcal{Q}$. To mitigate the bias in LLMs that leads to divergent responses to equivalent queries due to inherent language-specific differences, the advantage term $\hat{A}_{i,t}^{\text{coupled}}$ is computed over $\mathcal{Q}$, leveraging a shared baseline derived from the entire multilingual group. By normalizing the reward of each language-specific response o_i against the collective performance of the group, the policy is encouraged to bind embeddings from different languages to a unified, shared, and high-quality reasoning path, thereby optimizing trajectories within the same group.

Formally, we optimize the policy π_θ by maximizing the following language-coupled objective:

$$\mathcal{J}_{\text{LcRL}}(\theta) = \mathbb{E}_{\{q_i\}_{i=1}^{G} \sim \mathcal{D}_{\text{couple}}, \{o_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot | \{q_i\}; \mathcal{R})}$$
$$\left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{\sum_{t=1}^{|o_i|} I(o_{i,t})} \sum_{t=1, I(o_{i,t})=1}^{|o_i|} \right.$$
$$\min \left(\frac{\pi_\theta(o_{i,t}|q_i, o_{i,<t}; \mathcal{R})}{\pi_{\text{old}}(o_{i,t}|q_i, o_{i,<t}; \mathcal{R})} \hat{A}_{i,t}^{\text{coupled}}, \text{clip} \right.$$
$$\left. \left(\frac{\pi_\theta(o_{i,t}|q_i, o_{i,<t}; \mathcal{R})}{\pi_{\text{old}}(o_{i,t}|q_i, o_{i,<t}; \mathcal{R})}, 1-\epsilon, 1+\epsilon \right) \hat{A}_{i,t}^{\text{coupled}} \right)$$
$$\left. - \beta \frac{1}{G} \sum_{i=1}^{G} \mathbb{D}_{\text{KL}}[\pi_\theta(\cdot|q_i) || \pi_{\text{ref}}(\cdot|q_i)] \right]$$

where ϵ and β are hyperparameters. Additionally, the indicator function $I(o_{i,t})$ masks retrieved tokens, ensuring that gradients are backpropagated solely through the model's generated reasoning and search actions.

3.3 Language-Coupled Reward Modeling

To guide the optimization process in search-augmented RL, previous methods (e.g., Search-R1) adopt the rule-based outcome reward system using exact matching (EM). However, this discrete approach is suboptimal for multilingual tasks where lexical diversity is prevalent. More critically, recent investigations (Deng et al., 2025) reveal that GRPO-based tool-integrated RL is uniquely vulnerable to training collapse triggered by Lazy Likelihood Displacement (LLD), which is caused by high representational overlap between incorrect and correct trajectories. Thus, we propose the Language-Coupled Reward Modeling system.

Multilingual Outcome Reward. As word-level matching fails to capture such lexical similarities, we replace the binary EM reward with Character 3-gram Recall (r_{ans}) (Chirkova et al., 2024) to assess answer correctness across multilingual generations:

$$r_{\text{ans}}(i) = \text{c3Recall}(\hat{a}_i, a_{\text{gold}})$$

where $\hat{a}_i$ is the extracted answer and a_{gold} is the ground truth. Unlike EM, r_{ans} provides a dense reward signal, which slows down the entry into "high-gradient danger zones" by stabilizing importance ratios, preventing the advantages from being amplified by extreme reward disparities.

Auxiliary Anti-Consistency Penalty. In RL, training instability often arises due to the high variance of the reward signal and the non-stationarity of the environment. To ensure training stability and resolve knowledge conflict, we introduce the Anti-Align Penalty, which acts as a repulsive barrier to break the loop of error mode collapse. We first define the set of "bad samples" B_q for a query q as those failing a correctness threshold τ_{bad}:

$$B_q = \{i \in G_q \mid r_{\text{ans}}(i) < \tau_{\text{bad}}\}$$

For each sample $i \in B_q$, we compute its maximum similarity to other incorrect responses in the group:

$$m_i = \max_{j \in B_q, j \neq i} \text{Sim}_{c3}(\hat{a}_i, \hat{a}_j),$$

where Sim_{c3} represents the similarity calculated by character 3-gram recall. The base penalty p_i is defined with a margin $\gamma = 0.5$ as $p_i = \max(0, m_i - \gamma)$. To ensure stability, we introduce a sample-level weight $w_q = \min(1, \frac{|B_q|}{5})$ and formulate the anti-alignment reward:

$$r_{\text{anti\_align}}(i) = \begin{cases} -p_i \cdot w_q, & i \in B_q \text{ and } |B_q| > 1 \\ 0, & \text{otherwise} \end{cases}$$

By penalizing clusters of similar incorrect answers, the Anti-Consistency Penalty incentivizes the model to critically compare inconsistent evidence across languages, and alleviate the negative gradient issue caused by incorrect mode collapse and representational overlap with correct paths.

Reward Fusion and Normalization. To prevent the penalty from dominating the training, we apply a clipping operation as:

$$\tilde{r}_{\text{anti\_align}}(i) = \text{clip}(r_{\text{anti\_align}}(i), -0.5, 0)$$

Finally, the final reward is defined as:

$$r_{\text{total}}(i) = \max(0, r_{\text{ans}}(i) + \lambda \cdot \tilde{r}_{\text{anti\_align}}(i))$$

This design incentivizes the model to explore the "cross-lingual optimal reasoning path" while maintaining a robust defense against the structural instabilities inherent in tool-integrated RL.

4 Experimental Setup

4.1 Datasets

We utilize MKQA (Longpre et al., 2021) and XOR-TyDi QA (Asai et al., 2021) datasets for evaluation. MKQA consists of 10k queries based on the NQ dataset (Kwiatkowski et al., 2019) translated into non-English languages, which support parallel languages in queries and knowledge. XOR-TyDi QA comprises 40K information-seeking questions in 7 languages. As shown in Table 1, to evaluate the language transferability for the post-training methods, we define seen and unseen languages, where seen languages are used during RL training and unseen languages are only used at evaluation time.

In this paper, we follow Chirkova et al. (2024) and select a subset of queries encompassing 13 languages to maintain the diversity, including both Latin and non-Latin scripts. The details of statistics are provided in Appendix A.1.

Dataset	Type	Languages
MKQA	Training	EN, AR, FI, FR, JA, IT, RU, ZH (8)
	Zero-shot	DE, KO, ES, PT, TH (5)
XOR-TyDi	Training	EN, FI, JA, KO, RU, AR (6)
	Zero-shot	BN, TE (2)

Table 1: Language partition for training and zero-shot evaluation.

4.2 Implementation Details

Base Models. We select three LLMs with different sizes as backbones to implement ours: `Qwen2.5-3B-Instruct`, `Qwen3-4B`, and `Qwen3-8B` (Yang et al., 2025). For the multilingual retriever, we use multilingual E5 (Wang et al., 2024a) and Wikimedia_dump* as the knowledge base for all experiments. To account for variability, the reported performance is averaged over three independent inference runs. More details are in provided in Appendix A.2

Evaluation Metrics. We follow the evaluation settings of Chirkova et al. (2024), which consist of flexible exact match (fEM) and recall on character 3-gram level (c3Recall) (Schick et al., 2023). C3Recall measures the overlap between generated and gold answers at the character 3-gram level, providing a soft matching signal for multilingual evaluation. Furthermore, we measure the correct language rate (CLR), which is the percentage of outputs written in the user language by the models. CLR reflects the usability of multilingual RAG systems by assessing whether the model maintains the intended response language and avoids undesired language switching.

4.3 Baselines

To evaluate the effectiveness of LCRL, we select the following baselines: (1) Direct inference without Retrieval. (2) Inference with Retrieval: native RAG (Lewis et al., 2020), IRCoT (Trivedi et al., 2023), and Search-o1 (Li et al., 2025). (3) Fine-Tuning Methods: Supervised fine-tuning (SFT) (Chung et al., 2024), Argumentative Explanations (D-RAG) (Ranaldi et al., 2025b), and RL-based fine-tuning (Search-R1†) (Jin et al., 2025). These baselines cover a broad spectrum of training-free and post-training RAG approaches, allowing for a comprehensive assessment of LCRL. To make a fair comparison between different methods, we use the same settings of retriever, knowledge corpus, training data, and base LLMs. More details can be found in Appendix A.3.

5 Results

5.1 Main Results

Overall. As shown in Table 2, we investigate the performance of LCRL on two multilingual knowledge-intensive benchmarks. The results demonstrate that our proposed method (LCRL) outperforms several strong baselines in terms of all evaluation metrics for various LLMs. In particular, post-training methods learn to contextualize cross-lingual reasoning processes to minimize their impact on the output language and demonstrate a clear advantage in correct language rate over training-free baselines, which our approach further enhances. We further provide qualitative case studies in Appendix D to illustrate how LcRL resolves knowledge bias and conflict in multilingual settings.

Although the multilingual search-augmented RL method (mSearch-R1) facilitates the knowledge transfer in the multilingual settings compared to other post-training methods such as D-RAG and SFT, the parameters are still hard to optimize,

*https://huggingface.co/datasets/wikimedia/wikipedia

†we reproduce the Search-R1 method in the multilingual setting

Methods	MKQA			XOR-Tydi QA		
	fEM	c3Recall	CLR	fEM	c3Recall	CLR
`Qwen2.5-3B-Instruct`						
Direct Inference	19.7	27.4	72.3	10.1	19.0	70.6
IRCoT	22.4	34.9	65.3	18.5	27.9	69.4
Search-o1	27.7	35.0	70.2	21.1	30.7	67.9
RAG	28.2	34.6	66.3	20.0	29.1	65.2
D-RAG	37.4	43.3	90.2	31.5	38.9	87.1
SFT	35.0	43.0	92.9	28.3	35.5	87.7
Search-R1	22.6	34.8	83.6	18.4	32.0	86.3
mSearch-R1	37.9	53.2	95.6	21.2	35.8	94.1
LcRL (Ours)	**41.2***	**57.0***	**99.1***	**31.7***	**43.9***	**99.2***
`Qwen3-4B`						
Direct Inference	20.3	29.3	65.4	10.0	20.3	66.1
IRCoT	20.3	28.8	68.2	20.7	28.1	69.6
Search-o1	22.1	28.6	72.4	19.6	26.7	73.1
RAG	22.9	29.7	71.0	20.7	28.1	71.7
D-RAG	33.0	39.7	92.7	31.1	38.7	90.3
SFT	21.7	30.6	92.1	28.2	35.6	91.6
Search-R1	24.3	34.7	85.8	17.8	29.8	87.2
mSearch-R1	30.3	40.7	95.0	23.4	36.6	94.6
LcRL (Ours)	**40.6***	**54.9***	**98.9***	**32.4***	**45.7***	**99.4***
`Qwen3-8B`						
Direct Inference	26.2	24.3	66.8	12.3	22.7	67.5
IRCoT	28.7	35.6	70.3	24.0	33.0	71.9
Search-o1	28.6	35.8	73.8	25.2	34.3	74.6
RAG	27.6	34.8	74.2	22.7	31.8	74.9
D-RAG	38.0	46.7	94.1	33.5	42.3	93.5
SFT	29.8	38.3	93.4	30.0	38.9	94.0
Search-R1	29.1	37.4	86.4	16.8	28.1	88.9
mSearch-R1	40.6	50.8	94.8	27.2	38.5	94.2
LcRL (Ours)	**47.6***	**63.2***	**99.4***	**38.3**	**48.9***	**99.3***

Table 2: The overall performance in all languages retrieval collections for MRAG. The highest score based on each base model is highlighted in **bold**. "$*$" denotes significant improvements with t-test at $p < 0.01$ between LcRL and mSearch-R1.

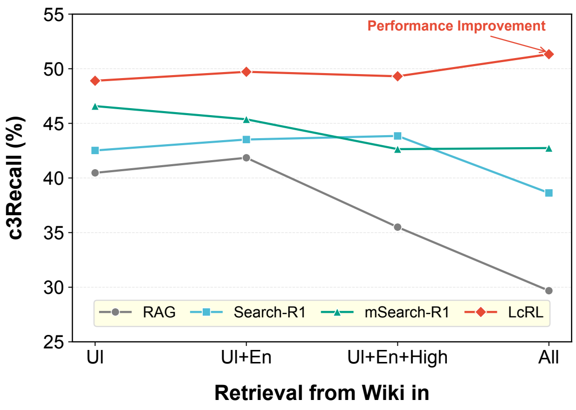


Figure 3: Performance with different retrieval options. Ul denotes the language of the user query, and High represents a randomly selected high-resource language.

	fEM	c3Recall	CLR
Our Full Model	41.2	57.0	99.1
w/o Lc Reward	30.8	42.2	98.1
w/o c3Recall Reward	18.0	20.2	97.8
w/o Lc Rollout	30.4	45.7	98.7
w/o multi-language Rollout	27.9	38.5	84.6
Replace by PPO	15.5	21.7	72.3

Table 3: Results of LcRL with algorithms and components based on `Qwen2.5-3B-Instruct` for MKQA.

and its performance has not been significantly improved. The limitations arise from optimization instability and knowledge conflicts, in which deep, multi-turn retrieval over multilingual collections is incorporated into the search engine calling process. We address these issues by incorporating a language-coupled mechanism into the rollout and reward modules, enabling stable training and effective multilingual knowledge transfer within groups, which achieves significant performance and successfully transfers to unseen languages. In addition, we also conduct experiments on unseen languages that were not included in the training data. LcRL also outperforms other methods. Results show that LcRL demonstrates superior generalization performance compared to all baseline methods. The results and further explanations for each language are provided in Appendix B.

Performance on Language Count in Multilingual Collections. As shown in Figure 3, we investigate the impact of increasing language diversity in the retrieval collections on different RAG methods. The results show that only our method exhibits a steady improvement, whereas all other approaches suffer a sharp performance degradation when the number of languages in the retrieval collections exceeds two. This degradation is attributable to knowledge conflicts arising from factual inconsistencies across languages in the collections. Our method explicitly addresses this issue through an anti-consistency penalty in the reward model. Moreover, detailed per-language results and analysis are provided in Appendix B, including performance comparisons across both seen and unseen languages, as well as additional analyses on data scaling and language coverage. The analysis shows that the language-coupled approach prevents negative transfer from low-resource languages where the multilingual curse is prevalent in high-resource languages.

5.2 Ablation Studies

As shown in Table 3, we further investigate the effectiveness of our method with different algorithms and components. *First*, we replace the outcome-based reward function. The EM objective provides a sparse optimization signal for MRAG. By introducing n-gram preference optimization, we observe a 22% and a 12.8% gain with respect to c3Recall

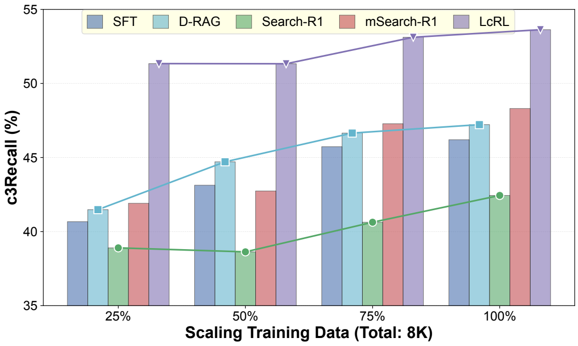


Figure 4: Results of post-training methods with scaling utilized data.

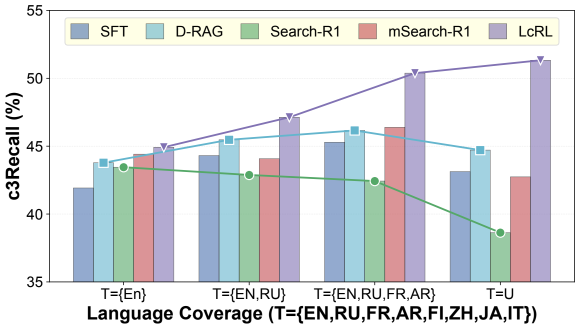


Figure 5: Performance of post-training methods with different language coverages.

and fEM, respectively. The proposed language-coupled mechanism delivers further performance gains by suppressing training collapse arising from language-consistency error modes. *Second*, we replace the search engine calling strategy in the rollout module. The results show that multi-turn retrieval not only improves overall performance but also reduces language error rates, and our proposed language-coupled multi-turn retrieval further strengthens the interaction between LLMs and the search engine in multilingual scenarios. *Finally*, we compare different RL algorithms (GRPO v.s. PPO). The results demonstrate the superiority of GRPO in multilingual settings. It is attributed to the group-based learning mechanism that facilitates cross-lingual generalization, and the limitations of value LLMs in multilingual training.

6 Analysis

6.1 Effects of Training Data

To further investigate the generalization and cross-lingual transferability of post-training methods, we systematically vary both the scale and the language coverage of the training corpus. Overall, our method exhibits positive robustness when handling a larger number of languages.

Data Scale. Figure 4 presents a comparison of post-training methods under training corpus scaling. SFT-based methods perform poorly in low-resource settings but show substantial performance gains as the data scale increases. In contrast, RL-based methods achieve competitive performance even with limited data, with scaling providing only marginal improvements. However, once the training corpus reaches a sufficient size, RL-based methods experience significant performance gains.

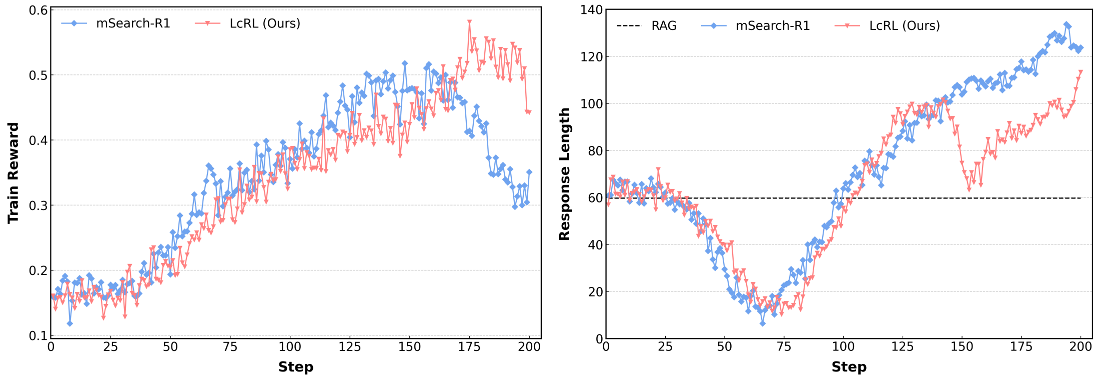
Language Coverage. Figure 5 shows that our method achieves superior performance when the training corpus covers a large number of languages, indicating that it effectively mitigates the "multilingual curse" and highlighting the distinct effectiveness in multilingual scenarios. Although performance decreases as the number of languages is reduced, our approach consistently outperforms competing methods. When the multilingual setting is absent (i.e., the training corpus contains only a single language), our method exhibits performance comparable to other approaches.

6.2 Response Length and Training Stability

As illustrated in Figure 6a, we analyze the training dynamics to verify the stability of LcRL, and mSearch-R1 exhibits a sharp "reward collapse" starting around step 160. In contrast, LcRL maintains a robust upward trajectory throughout the training process, stabilizing at higher reward levels without succumbing to performance degradation. This indicates that our language-coupled reward modeling effectively regularizes the policy against error mode collapse. Furthermore, Figure 6b highlights the efficiency of the generated reasoning trajectories. While mSearch-R1 suffers from uncontrolled length expansion, which is a common symptom of reasoning loops in RL. LcRL effectively regulates the generation process. In the later training stages, LcRL achieves a response length reduction of approximately 12% compared to mSearch-R1, indicating that LcRL does not rely on redundant token generation to improve performance and optimize for more concise multilingual reasoning paths.

7 Conclusion

In this work, we propose a language-coupled GRPO method for MRAG, which leverages multilingual and multi-turn decision-making with an



(a) Train reward.

(b) Response length.

Figure 6: Comparisons of train reward and response length between mSearch-R1 and ours.

anti-consistency reward. It can adaptively determine the need for retrieving external knowledge and effectively harmonize contradictory knowledge obtained from different languages, mitigating the issue of knowledge bias and conflicts. Experimental results demonstrate that our method exhibits strong robustness under resource-constrained conditions and is effective in dealing with an increasing number of language counts in retrieval collections.

Limitation

Although the effectiveness shown in our proposed method, some potential limitations should be addressed in future studies.

First, current evaluations typically reuse existing QA datasets combined with generic Wikipedia dumps. Due to the lack of dedicated query-document relevance judgments annotations in multilingual RAG scenarios, we are unable to systematically evaluate the correlation between the search results evaluated by traditional IR metrics and the answer results produced by generator models. Second, due to cost constraints, we evaluate our method on three representative multilingual LLMs, which only scratches the surface of the world's vast array of open-resourced models. Furthermore, as this study prioritizes the interaction between LLMs with multilingual retrieved evidence over the optimization of retrieval algorithms themselves, we utilize multilingual-e5-base as our fixed retriever. This model represents a widely adopted and computationally efficient baseline for multilingual dense retrieval. While a more expansive experimental configuration might yield higher absolute performance, the current setup provides a robust foundation for evaluating our core contributions.

Ethics Statement

We did not explicitly address ethical considerations in this work. All data used in our experiments were obtained exclusively from open and publicly available benchmarks, as detailed in the appendices. Statistics related to language distributions in commonly used pre-training datasets were sourced from official open-resourced technical report. Our analysis and methodology were conducted with sensitivity, and do not involve or affect attributes related to gender, sex, or race.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities of China under Grant 2024JBGP008 and the National Natural Science Foundation of China (No. 62406018, 62476023, 61976016, 62376019), and the authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. In particular, Bowen Jin (University of Illinois Urbana-Champaign, USA) as the first author of search-R1, has provided valuable thoughts and contributes part of the implementation of the multilingual strategies based on search-R1.

References

Akari Asai, Jungo Kasai, Jonathan 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, Online. Association for Computational Linguistics.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze

Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. Research: Learning to reason with search for llms via reinforcement learning. *Preprint*, arXiv:2503.19470.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Wenlong Deng, Yushu Li, Boying Gong, Yi Ren, Christos Thrampoulidis, and Xiaoxiao Li. 2025. On grpo collapse in search-r1: The lazy likelihood-displacement death spiral. *Preprint*, arXiv:2512.04220.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in llms. *Preprint*, arXiv:2402.10612.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Sunayana Sitaram, Kalika Bali, Tanuja Ganu, and Akshay Nambi. 2025. Bridging the language gap: Dynamic learning strategies for improving multilingual performance in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9209–9223, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lyu. 2025. Resisting contextual interference in rag via parametric-knowledge reinforcement. *Preprint*, arXiv:2506.05154.

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.

Qi Luo, Xiaonan Li, Yuxin Wang, Tingshuo Fan, Yuan Li, Xinchi Chen, and Xipeng Qiu. 2025. Marag-r1: Beyond single retriever via reinforcement-learned multi-tool agentic retrieval. *Preprint*, arXiv:2510.27569.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. Chatr1: Reinforcement learning for conversational reasoning and retrieval augmented question answering. *arXiv preprint arXiv:2510.13312*.

Zhibo Man, Yuanmeng Chen, Yujie Zhang, and Jinan Xu. 2025. DMDTEval: An evaluation and analysis of LLMs on disambiguation in multi-domain translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6065–6082, Suzhou, China. Association for Computational Linguistics.

Zhibo Man, Kaiyu Huang, Yujie Zhang, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024. ICL: Iterative continual learning for multi-domain neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7732–7743, Miami, Florida, USA. Association for Computational Linguistics.

Fengran Mo, Yifan Gao, Sha Li, Hansi Zeng, Xin Liu, Zhaoxuan Tan, Xian Li, Jianshu Chen, Dakuo Wang, and Meng Jiang. 2026a. Agentic conversational search with contextualized reasoning via reinforcement learning. *Preprint*, arXiv:2601.13115.

Fengran Mo, Zhan Su, Yuchen Hui, Jinghan Zhang, Jia Ao Sun, Zheyuan Liu, Chao Zhang, Tetsuya Sakai, and Jian-Yun Nie. 2026b. Opendecoder: Open large language model decoding to incorporate document quality in rag. *arXiv preprint arXiv:2601.09028*.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do LLMs need retrieval augmentation? mitigating LLMs' overconfidence helps retrieval augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11375–11388, Bangkok, Thailand. Association for Computational Linguistics.

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.

Rui Qi, Zhibo Man, Yufeng Chen, Fengran Mo, Jinan Xu, and Kaiyu Huang. 2025. Sot: Structured-of-thought prompting guides multilingual reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11024–11039.

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

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025b. Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9075–9096, Suzhou, China. Association for Computational Linguistics.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *Preprint*, arXiv:2505.17005.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenmin Wang, Peilin Zhang, Ge Liu, Ruihua Wu, and Guixiang Song. 2024b. Investigating on the external knowledge in rag for zero-shot cross-language transfer. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 1479–1484.

Junlin Wu, Xianrui Zhong, Jiashuo Sun, Bolian Li, Bowen Jin, Jiawei Han, and Qingkai Zeng. 2025a. Structure-r1: Dynamically leveraging structural knowledge in llm reasoning through reinforcement learning. *Preprint*, arXiv:2510.15191.

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.

Yicong Wu, Guangyue Lu, Yuan Zuo, Huarong Zhang, and Junjie Wu. 2025b. Graph-r1: Incentivizing the zero-shot graph learning capability in LLMs via explicit reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23920–23938, Suzhou, China. Association for Computational Linguistics.

Zhichao Xu, Minheng Wang, Yawei Wang, Wenqian Ye, Yuntao Du, Yunpu Ma, and Yijun Tian. 2025. Recon: Reasoning with condensation for efficient retrieval-augmented generation. *Preprint*, arXiv:2510.10448.

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*.

Jinghan Zhang, Fengran Mo, Xiting Wang, and Kunpeng Liu. 2024. Blind spot navigation in llm reasoning with thought space explorer. *arXiv preprint arXiv:2410.24155*.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025a. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025b. CM-align: Consistency-based multilingual alignment for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25689–25702, Suzhou, China. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025c. Less, but better: Efficient multilingual expansion for LLMs via layer-wise mixture-of-experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17948–17963, Vienna, Austria. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025d. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2026. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. *Preprint*, arXiv:2510.07300.

Appendix

A Details in Experiments

A.1 Datasets

The details of training and evaluation data are presented in Table 1 and 4. For MKQA, we initially select a subset of 2k instances per language for the test set. For XOR-TyDi QA, we select the already segmented validation set in 8 languages and ground questions in Wikipedia in the same language as the question.

Dataset	Training Set			Test Set		
	# Lang	# / Lang	**Total**	# Lang	# / Lang	**Total**
MKQA	8	1,000	8,000	13	2000	16,601
XOR-TyDi	6	1,000	6,000	7	U*	2,560

***Note:** XOR-TyDi counts in group U: fi(615), ja(433), ko(371), ru(568), ar(708), bn(427), te(351).

Table 4: Number (#) of training and evaluation instances.

A.2 Retriever and Generator

Retriever. We pick a strong and publicly available Multilingual-E5-base[‡] as our multilingual retriever, and Wikimedia_dump as the knowledge base for all experiments. We split the Wikipedia article into chunks of 100 words (or 100 Unicode characters for non-whitespace separated languages) and prepending the article title to each chunk. Following Search-R1 (Jin et al., 2025), we set the number of retrieved passages to 3 across all retrieval-based methods.

Generator. We use various strong LLM, including recently released Qwen3-4B[§] and Qwen3-8B[¶], as well as widely-used LLM Qwen2.5-3B-Instruct[‖].

Training is performed on a single node with 4 A800 GPUs. We use a total batch size of 512, with a mini-batch size of 256 and a micro-batch size of 16. For GRPO training, we set the policy LLM learning rate to 1e-6 and sample 3 responses per prompt. For our auxiliary anti-consistency penalty introduced in section 3.3, λ is set to 0.02.

The group size G is treated as a hyperparameter and is not necessarily equal to the number of available languages N. When $G < N$, we always include the original-language query and randomly sample the remaining $G - 1$ queries from other seen languages. We search G in the range of 3 to 8 and select $G = 5$, which is consistent with the optimal setting reported in Search-R1.

[‡] https://huggingface.co/intfloat/multilingual-e5-base

[§] https://huggingface.co/Qwen/Qwen3-4B

[¶] https://huggingface.co/Qwen/Qwen3-8B

[‖] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

Methods	**Language**							
	AR	**FI**	**JA**	**KO**	**RU**	**BN**	**TE**	**AVE**
Qwen2.5-3B-Instruct								
Direct Inference	26.9	18.8	11.1	17.4	21.0	10.2	7.1	16.1
IRCoT	39.4	35.7	22.7	9.6	32.3	9.3	7.0	22.3
Search-o1	42.5	40.4	21.8	12.5	36.3	9.7	6.1	24.2
RAG	40.1	39.1	21.2	10.8	34.3	10.2	6.0	23.1
D-RAG	50.4	48.7	31.9	20.6	43.1	9.7	6.7	30.1
SFT	48.0	45.5	32.2	12.4	39.3	5.7	6.1	27.0
Search-R1	37.5	41.5	20.7	21.7	38.7	**12.7**	6.5	25.6
mSearch-R1	44.7	45.6	21.1	23.7	44.0	12.5	7.0	28.4
LcRL (Ours)	**53.4**	**57.6**	**32.4**	**30.3**	**45.8**	11.5	**8.6**	**34.2**
Qwen3-4B								
Direct Inference	27.1	22.6	12.3	17.7	21.7	**12.9**	11.2	17.9
IRCoT	31.9	43.9	18.3	13.0	26.6	5.4	6.8	20.9
Search-o1	34.2	48.2	18.8	13.4	27.7	5.0	9.8	22.4
RAG	32.7	46.1	20.2	12.4	29.0	4.8	7.8	21.8
D-RAG	44.8	56.2	33.6	17.3	41.6	9.8	10.5	30.5
SFT	41.4	55.0	29.7	14.1	38.0	5.9	8.7	27.5
Search-R1	39.6	36.6	19.8	24.4	28.7	8.9	4.5	23.2
mSearch-R1	44.4	40.9	25.9	**32.6**	39.3	10.0	6.1	28.4
LcRL (Ours)	**56.1**	**59.5**	**34.2**	29.9	**48.5**	11.5	**11.4**	**35.9**
Qwen3-8B								
Direct Inference	29.1	27.0	13.1	21.3	22.9	**17.8**	**16.7**	21.1
IRCoT	42.7	47.2	22.4	15.2	37.6	13.1	12.0	27.2
Search-o1	43.6	47.3	25.6	14.2	40.7	13.4	10.4	27.9
RAG	41.1	44.1	23.8	13.2	36.6	13.9	11.3	26.3
D-RAG	52.3	53.9	36.7	22.8	45.9	13.3	12.6	33.9
SFT	49.7	50.0	33.3	19.0	42.7	13.2	10.9	31.3
Search-R1	38.9	32.2	20.7	19.1	29.7	9.6	5.0	22.2
mSearch-R1	46.0	45.4	26.0	33.1	43.2	12.9	8.0	30.5
LcRL (Ours)	**59.1**	**63.1**	**39.0**	**35.0**	**48.2**	15.8	11.0	**38.8**

Table 5: Main results (%) of XOR-Tydi QA. Background colors distinguish between seen languages and unseen languages. The highest score on each base model is highlighted in **bold**. The second highest score on each base model is highlighted in underline.

A.3 Baselines

We provide a brief introduction to the baseline methods used in our experiments.

- Direct Inference. The model generates answers directly based on its internal parametric knowledge without any external retrieval.
- RAG (Lewis et al., 2020). The traditional Retrieval-Augmented Generation framework. Documents are retrieved in a single step to provide context for the final generation.
- IRCoT (Trivedi et al., 2023). Interleaving Retrieval with Chain-of-Thought Reasoning. It guides the retrieval process by using the intermediate steps of a CoT as queries.
- Search-o1 (Li et al., 2025). An agentic search-enhanced framework that allows the model

Methods	Language													
	En	Fr	It	Ru	Fi	Zh	Ja	Ar	Es	Pt	Ko	Th	De	Ave
Qwen2.5-3B-Instruct														
Direct Inference	51.8	39.8	36.9	21.5	23.6	13.2	14.9	17.5	39.3	37.8	8.8	18.0	39.7	27.9
IRCoT	61.8	50.1	45.6	33.9	31.7	15.6	17.1	23.3	57.3	55.3	12.7	20.0	48.1	36.3
Search-o1	60.9	43.2	50.8	27.4	37.6	18.9	25.1	16.3	58.7	58.9	14.2	22.6	57.5	37.9
RAG	64.1	46.4	46.4	30.7	33.1	14.7	20.6	20.9	54.9	52.2	10.6	18.8	51.9	35.8
D-RAG	72.6	54.9	55.1	38.8	42.3	22.9	29.4	30.6	62.8	59.7	18.9	26.1	61.3	44.3
SFT	60.1	55.8	52.8	44.9	41.0	23.3	**33.5**	32.6	56.9	53.4	16.5	28.0	54.7	36.3
Search-R1	59.7	45.9	47.9	36.9	36.5	16.4	15.2	20.2	50.3	43.2	12.2	19.8	44.8	34.5
mSearch-R1	74.6	73.9	72.5	53.6	63.5	26.6	28.2	33.0	67.4	72.8	18.4	32.9	73.3	53.1
LcRL (Ours)	**81.9**	**75.6**	**76.0**	**57.6**	**68.7**	**27.8**	30.8	**37.9**	**78.2**	**74.4**	**25.0**	**40.3**	**74.9**	**57.6**
Qwen3-4B														
Direct Inference	51.9	37.8	38.6	26.4	24.6	18.9	16.4	20.1	39.4	39.2	10.2	19.5	38.6	29.4
IRCoT	41.9	36.4	24.7	33.3	34.2	22.7	17.4	20.0	40.3	30.0	9.8	20.9	47.5	29.1
Search-o1	43.9	34.1	25.4	31.0	37.3	21.0	18.0	17.8	42.1	27.7	11.0	18.4	48.4	28.9
RAG	46.6	32.9	29.9	29.9	39.9	19.3	22.5	16.6	45.9	26.1	13.7	17.1	51.6	30.1
D-RAG	62.7	48.2	45.6	33.9	45.4	24.9	27.2	29.4	52.2	31.7	18.9	23.5	58.0	38.6
SFT	53.2	38.6	39.4	28.4	26.1	19.2	17.5	22.0	48.1	32.7	19.9	24.0	54.4	32.6
Search-R1	58.5	47.5	46.1	31.7	32.5	21.6	19.3	20.5	33.8	33.0	13.9	20.5	28.5	31.3
mSearch-R1	64.7	53.2	52.8	38.1	39.2	27.0	24.9	25.9	40.4	38.4	20.1	26.9	34.9	37.4
LcRL (Ours)	**73.1**	**60.9**	**73.5**	**59.5**	**69.1**	**29.9**	**33.4**	**39.8**	**64.9**	**54.5**	**21.6**	**29.2**	**60.0**	**51.5**
Qwen3-8B														
Direct Inference	59.5	46.5	45.4	29.0	34.3	21.6	18.7	19.4	47.9	44.9	12.5	22.3	46.6	34.5
IRCoT	54.2	38.5	36.9	33.1	46.0	22.8	25.9	27.4	58.7	29.7	18.9	17.8	52.1	35.5
Search-o1	53.3	40.7	35.6	35.3	44.9	22.2	28.3	26.1	57.5	31.7	18.5	20.4	51.6	35.9
RAG	52.1	39.9	34.1	34.7	43.8	21.3	27.0	25.4	56.5	30.9	17.1	19.8	50.4	34.8
D-RAG	68.8	56.4	53.6	44.1	53.0	28.3	35.1	34.6	57.4	40.7	19.1	24.5	59.3	44.2
SFT	55.4	42.9	36.0	38.9	47.1	25.6	30.5	29.7	60.9	34.6	19.7	22.4	55.3	38.4
Search-R1	56.7	42.3	33.7	39.9	45.7	22.5	31.2	27.6	52.4	33.4	15.8	20.5	52.3	36.5
mSearch-R1	71.0	70.5	69.4	47.5	59.8	23.8	31.3	33.3	74.7	73.2	20.1	33.3	72.5	52.3
LcRL (Ours)	**88.3**	**78.9**	**78.2**	**63.5**	**75.0**	**35.4**	**39.3**	**46.9**	**84.5**	**80.0**	**33.8**	**48.6**	**80.6**	**64.1**

Table 6: Main results (%) of MKQA. Background colors distinguish between seen languages and unseen languages. The highest score on each base model is highlighted in **bold**. The second-highest score on each base model is highlighted in underline.

to autonomously plan, execute search queries, and refine its internal thought process before producing the final answer.

- D-RAG (Ranaldi et al., 2025b). A dialectic reasoning argumentation approach that systematically evaluates retrieved information by comparing, contrasting, and resolving conflicting perspectives.
- SFT (Chung et al., 2024). Supervised Fine-Tuning. LLMs are trained by queries and retrieved documents.
- Search-R1 (Jin et al., 2025). A framework that trains LLMs to reason with search through Reinforcement Learning (RL). We also extend the Search-R1 to the Multilingual Search-augmented RL method (mSearch-R1) with just Multi-Turn and Multilingual Retrieval introduced in section 3.1.

B Details on Experimental Results

Details on Overall Performance. Table 5 and Table 6 show the specific performance of LcRL in all languages on MKQA and XOR-Tydi QA. Results show that methods with retrieval consistently outperform methods with direct inference across nearly all languages. Notably, the performance on Qwen2.5-3B-Instruct generally surpasses that of Qwen3-4B, as the instruct-tuned variant exhibits superior instruction-following capabilities and more robust tag adherence compared to the base models (Jin et al., 2025). Specifically, Latin-script languages exhibit higher baseline scores and stable performance trajectories, where unseen languages such as Es, Pt and De perform comparably to seen languages like Fr and It, indicating robust intra-family cross-lingual transfer capabilities in current models. Conversely, non-Latin scripts manifest significant performance volatility and pronounced

System Prompt:
Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information.
After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search>, and it will return the top searched results between <information> and </information>.

Hierarchical Search Logic:
You can search at least 2 times:
1. First search in **{lang_name}** knowledge base.
2. Second search in **English and other languages** (e.g., French) to provide comprehensive information and resolve potential conflicts.
If needed, continue searching in English.

Output Constraints:
If no further external knowledge is needed, provide the answer inside <answer> and </answer> without detailed illustrations.
Note: The answer **must** be in **{lang_name}**.
For example: <answer> {example_answer} </answer>.

Question: {question}

Table 7: Multilingual Reasoning and Hierarchical Retrieval Template for LCRL. The structure enforces a "local-first, global-supplement" strategy to resolve cross-lingual knowledge conflicts.

"performance gaps" with languages like ZH, JA and TH, frequently experiencing performance degradation in several baseline models. In extreme low-resource scenarios such as TE and BN, models face a severe "low-resource dilemma" where direct inference often remains nearly optimal, identifying TE as the most challenging language across all evaluated methods. Notably, while KO and TH exhibits poor performance in unseen languages, it shows substantial improvement potential through retrieval-augmented strategies. Furthermore, RL-based baselines prove highly unstable when generalizing to unseen non-Latin languages, highlighting the inherent difficulty of applying standard RL policies to diverse and complex linguistic scripts.

Model Sensitivity to Data Scale. As shown in figure 7, by maintaining a fixed number of languages N and scaling the samples per language M from 25% to 100% of the total training budget, we observe distinct behavioral patterns: supervised baselines like SFT and D-RAG exhibit steep linear growth and a heavy dependency on data volume. In contrast, the baseline RL method displays "unstable" behavior characterized by performance regressions around the 50% data scale in languages such as FI, IT, and ZH. Our proposed LCRL demonstrates an "efficient and robust" trajectory with only 25% of the training data, abd it consistently surpasses all baselines trained on the full dataset. On high-resource languages such as EN, FR, and RU, performance across all evaluated methods remains relatively stable. In contrast, on linguistically complex or low-resource languages like AR, ZH, JA, and FI, baseline models exhibit increased volatility, for instance, Search-R1 shows a notable performance decline on Finnish (FI) as data scales. LCRL maintains high performance starting from the 25% data volume and exhibits a steady upward trend with the addition of training samples, demonstrating consistent behavior across diverse linguistic settings.

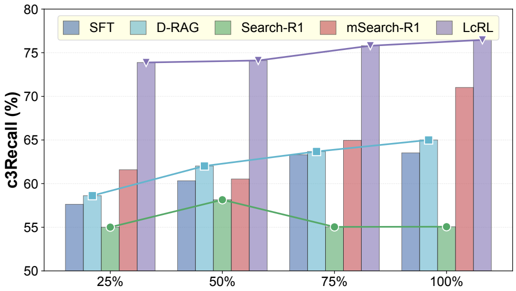
Model Sensitivity to Language Coverage. As illustrated in Figure 8, baseline methods typically exhibit an inverted U-shaped or downward performance trajectory as the number of training languages N increases. This decline is particularly pronounced in high-resource Latin languages like EN, FR, and IT, where the inclusion of more languages appears to serve as a hindrance to native performance. Conversely, non-Latin and complex script languages such as ZH, JA, and AR maintain relatively stable trends, suggesting a lower susceptibility to cross-lingual interference from additional training data. In unseen language scenarios, baselines often stagnatie or even decline with increasing N. Furthermore, RL-based approaches like Search-R1 demonstrate greater sensitivity to data distribution changes with larger performance fluctuations, whereas fine-tuning methods remain comparatively stable. Notably, models consistently achieve effective results even on languages excluded from their specific training sets, underscoring the robust cross-lingual generalization capabilities inherent in these multilingual frameworks.

C Training Template

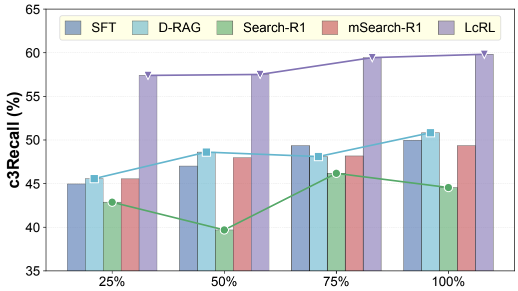
To enable the policy model to autonomously navigate the multilingual knowledge space, we employ a structured template that establishes a clear interaction protocol between the LLM and the search environment. As illustrated in Table 7, the template partitions the model's trajectory into three iterative phases: (1) internal reasoning within <think> tags, (2) tool-calling via <search> tags, and (3) final answer synthesis within <answer> tags. Unlike monolingual reasoning prompts, our template explicitly guides the model to adopt a multi-turn, multi-language retrieval strategy.

D Case Study

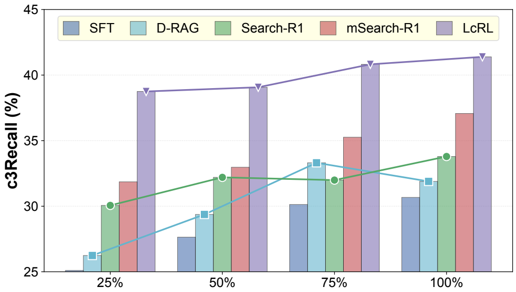
As shown in Figures 9 and 10, we conduct a case study on two scenarios involving knowledge bias and knowledge conflict. In the first case (Fig. 9), when asked about the 50th U.S. state, the model recognizes its lack of internal knowledge and initially searches the local Chinese collection. However, it fails to find relevant facts due to knowledge bias, and subsequently triggers a second-round search in global collections. By successfully synthesizing information from English and Japanese documents, the model correctly identifies "Hawaii" as the final answer. In the second case (Fig. 10), when asked about the release time of "Power of Love", the model encountered severe knowledge conflict because of multiple distracting dates retrieved from collections in multiple languages. Despite this interference, the model demonstrates robust reasoning by filtering out irrelevant noise and accurately outputting the golden answer (1984). These examples highlight the model's effectiveness in cross-lingual information synthesis and its resilience against contradictory information.



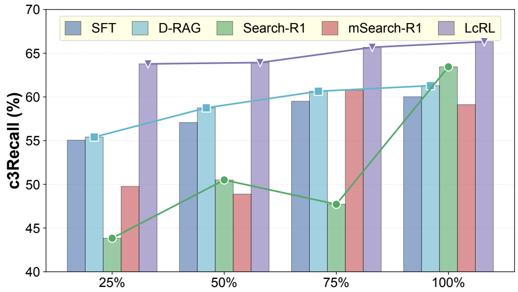
(a) Performance on Language EN.



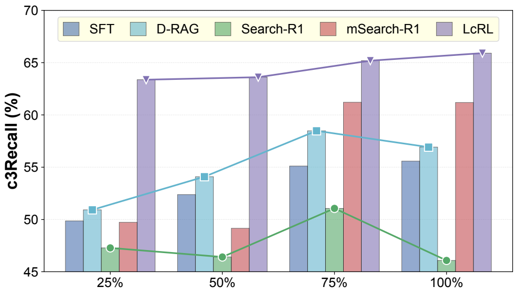
(b) Performance on Language RU.



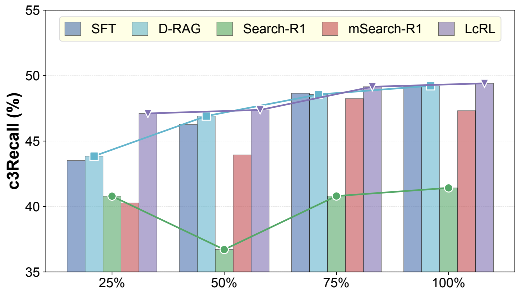
(c) Performance on Language AR.



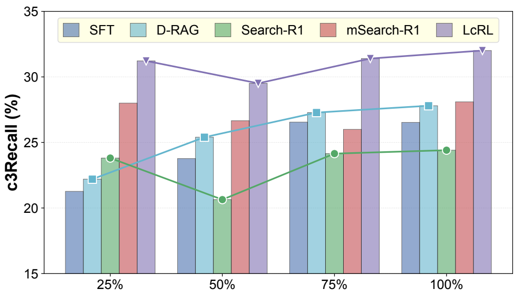
(d) Performance on Language FR.



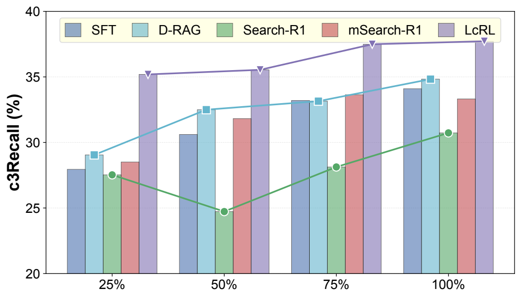
(e) Performance on Language IT.



(f) Performance on Language FI.

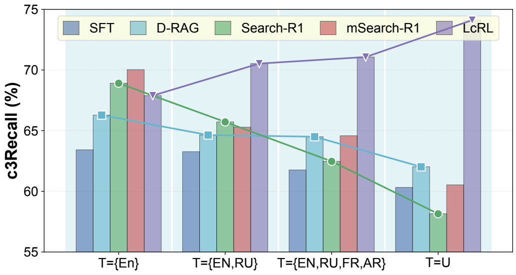


(g) Performance on Language ZH.

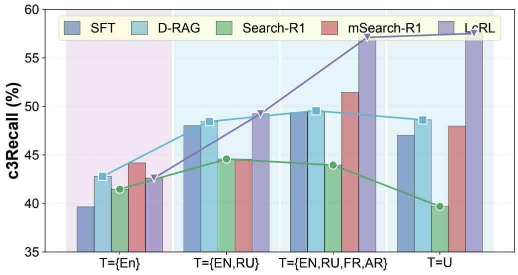


(h) Performance on Language JA.

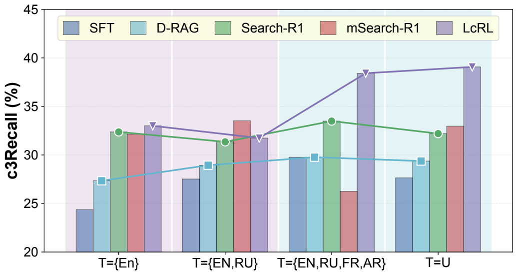
Figure 7: Results of post-training methods on each language with scaling utilized data.



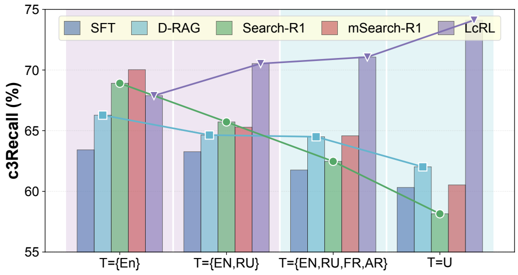
(a) Performance on Language EN.



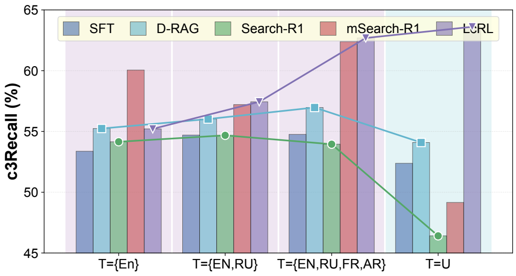
(b) Performance on Language RU.



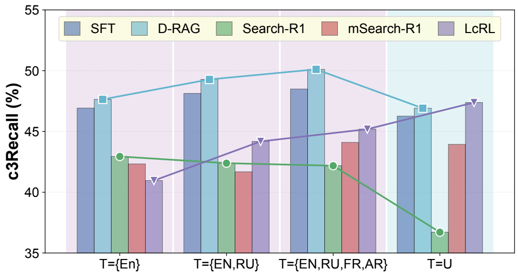
(c) Performance on Language AR.



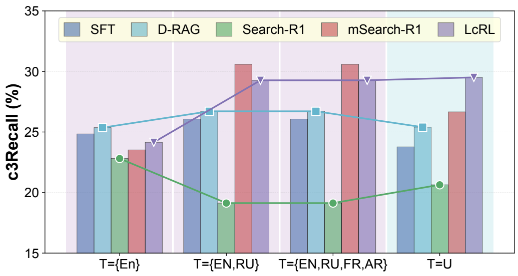
(d) Performance on Language FR.



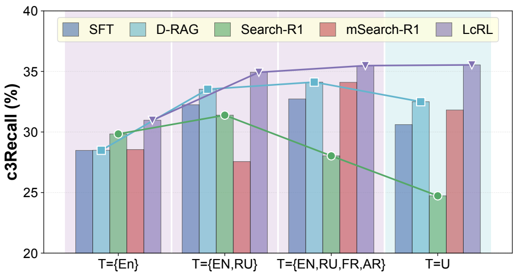
(e) Performance on Language IT.



(f) Performance on Language FI.



(g) Performance on Language ZH.



(h) Performance on Language JA.

Figure 8: Results of post-training methods on each language with different language coverages. Background colors distinguish between seen languages and unseen languages .

Question: 美国的第50个州是哪一个?
Golden Answer: 夏威夷州

<think>美国的第50个州是哪个，我暂时没有这方面的信息，我先进行搜索。</think>

First search in Chinese collection.

Round 1
<search>美国的第50个州是哪个? </search>

<information>
[ZH] Doc 1(Title: 美国50个州的名称源自不同的语言。)
[ZH] Doc 2(Title: 50号国道总共穿过了12个州，分别是加利福尼亚州、内华达州、犹他州...)
[ZH] Doc 3(Title: 各州同时认可美国宪法，因此其部分的主权被转移给了联邦政府，也就形成由50个共享主权的小国家形成的国家，即美国。)
</information>

<think>给出的信息中没有包含"美国的第50个州是哪个? "的相关信息。我需要再次进行搜索。</think>

Next search in global collections.

Round 2
<search>美国的第50个州是哪个? </search>

<information>
[lang=en]
[EN] Doc 1(Title: In other fields) Fifty is: There are 50 states in the United States of America. The TV show Hawaii Five-O and its reimagined version, Hawaii Five-0, are so called because Hawaii is the last (50th) of...
[EN] Doc 2(Title: Route 50S (Ohio–West Virginia) (former)) Alabama State Route 50
Arizona State Route 50 (never built)
Arkansas Highway 50...
[EN] Doc 3(Title: Route 50) U.S.

[lang=ja]
[JA] Doc 1(Title: アメリカ合国の州は、邦共和国であるアメリカ合国を成する「国家」で在50州がある,各州と邦は国家主を共有している。)
[JA] Doc 2(Title: 五十州：アメリカ合国の州の数。アメリカ合国の50番目の州はハワイ州である。ロマ数字では50 はL で表される。50年祭には、金が念品として使われたり、形容とされることが多い。)
[JA] Doc 3(Title: アメリカ合国の50の州は邦政府とは主を共有しながらも独立した国家治体である。
......
</information>

<answer>美国的第50个州是夏威夷州（Hawaii）。美国一共有50个州，而夏威夷是在1959年正式成为美国联邦的一部分的，因此它是美国最后一个加入联邦的州。</answer>

Figure 9: Case study for knowledge bias. Green block indicates correct information, and red block indicates relative but incorrect information.

Question: frankie goes to hollywood power of love 是哪一年的?
Golden Answer: 1984

<think>frankie goes to hollywood power of love 是哪一年的?
我需要通过搜索引擎来查找相关信息。</think>

First search in Chinese collection.

Round 1
<search>frankie goes to hollywood power of love 发行时间</search>

<information>
[ZH] Doc 1(Title: 影片改期到2021年9月3日在美国院线发行，海外发行于2021年8月25日。)
[ZH] Doc 2(Title: 据报道，发行商焦点影业将该片计划于2018年8月17日在美国上映。之后，美国上映日被改至2019年3月29日；不久后，上映日又提前为2019年3月15日。)
[ZH] Doc 3(Title: 影片于2022年7月15日由派拉蒙影在北美地区发行，当地原定档期为2017年4月14日，后来推迟到2017年8月4日。)
</information>

<think>我需要搜索更多信息来给出答案。</think>

Next search in global collections.

Round 2
<search>when was the song power of love by frankie released</search>

<information>
[lang=en]
[EN] Doc 1(Title: It also topped the charts in several other European countries...)
[EN] Doc 2(Title: "The Power of Love" is a song originally recorded and released by British band Frankie Goes to Hollywood. It was written by...)
[EN] Doc 3(Title: "Bat Out of Hell" originally peaked outside the top ten at number 15 upon its initial release in 1979. "The Power of Love" originally peaked at number-one upon its initial release in 1984.)

[lang=it]
[IT] Doc 1(Title: The Power of Love è il terzo singolo estratto dall'album Welcome to the Pleasuredome dei Frankie Goes to Hollywood, pubblicato nel 1984.)
[IT] Doc 2(Title: Power of Love è il settimo album in studio del cantautore statunitense Luther Vandross, pubblicato nel 1991. Tracce)
[IT] Doc 3(Title: Nel giugno 1985, The Power of Love venne pubblicato come singolo nel Regno Unito, dove scalò la classifica britannica, rimanendo in prima posizione per cinque settimane nell'ottobre 1985 e diventando...)
......
</information>

<think> 根据搜索结果，Frankie Goes to Hollywood 的歌曲《Power of Love》最初是在1984年发行的。这首歌在美国取得了巨大的成功，并且在英国排行榜上也表现优异，最终在1985年达到了榜首位置。虽然有报道称该歌曲计划在2018年和2019年在美国上映，但这些信息似乎与歌曲的发行时间无关。因此，《Power of Love》最准确的发行年份是1984年。</think>
<answer>《Power of Love》最准确的发行年份是1984年。</answer>

Figure 10: Case study for knowledge conflict. Green block indicates correct information. Red block indicates relative but incorrect information. Purple block indicates no relative information.