Improving Multilingual Retrieval-Augmented Language Models through Dialectic Reasoning Argumentations

Leonardo Ranaldi (†) **Federico Ranaldi** (‡)
Fabio Massimo Zanzotto (‡) **Barry Haddow**(†) **Alexandra Birch**(†)
(†) School of Informatics, University of Edinburgh, UK
(‡) Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
{first_name.last_name}@ed.ac.uk

Abstract

Retrieval-augmented generation (RAG) is key to improving large language models (LLMs) in systematically accessing richer factual knowledge. Yet, using RAG mechanisms brings intrinsic challenges, as LLMs must deal with conflicting knowledge, especially in multilingual retrieval, where the heterogeneity of knowledge retrieved may deliver different outlooks.

To make RAG more analytical, critical and grounded, we introduce *Dialectic-RAG* (*D*-RAG), a modular approach guided by *Argumentative Explanations*, i.e., structured reasoning process that systematically evaluates retrieved information by comparing, contrasting, and resolving conflicting perspectives. Given a query and a set of multilingual related documents, *D*-RAG selects and exemplifies relevant knowledge for delivering dialectic explanations that, by critically weighing opposing arguments and filtering extraneous content, clearly determine the final response. We show the impact of our framework both as an in-context learning strategy and for constructing demonstrations to instruct smaller models. Our experiments demonstrate that *D*-RAG significantly improves RAG approaches, requiring low-impact computational effort and providing robustness to knowledge perturbations.

1 Introduction

Retrieval-augmented Generation (RAG) has emerged as a promising approach for grounding large language models (LLMs) responses by incorporating relevant knowledge from external sources through structured retrieval mechanisms (Guu et al., 2020; Lewis et al., 2020b). RAG was conceived to handle the limitations of LLMs, such as their inclination towards hallucinations and the lack of knowledge of the specialised domain in their training data (Siriwardhana et al., 2023; Zhang et al., 2023).

Contextualising questions by adding relevant in-context knowledge retrieved from external corpora, such as Wikipedia, effectively reduced inaccurate generation, thereby notably improving accuracies (Gao et al., 2024). Yet, there are still limitations associated with RAGs; recent studies have shown ongoing challenges arising from the retrieved knowledge, where irrelevant or contradictory documents may introduce biases in the models (Menick et al., 2022; Li et al., 2025). These weaknesses arise from the inability of RAG strategies to critically asses the retrieved knowledge (Ranaldi et al., 2025c).

Prior approaches improve the RAG pipeline by incorporating external tools (Li et al., 2023; Yoran et al., 2024) or employing multi-step reasoning strategies (Zhao et al., 2024; Zhang et al., 2024) to determine the relevance of in-context passages. However, these methods may require high computational costs and definitely do not impact smaller-scale LLMs. Recently, Ranaldi et al. (2025c) proposed efficient approaches to enable LLMs to deliver argumentative reasoning trajectories. However, their effort is on English-centric RAG, and this can be a limitation for both the limited variance of retrieved knowledge and the actual operability (Chirkova et al., 2024; Ranaldi et al., 2025a).

In this paper, we present *Dialectic-RAG* (*D*-RAG), a modular framework conceived to enhance multilingual retrieval-augmented language models to follow a *Dialectic Reasoning*, i.e., a structured analytical process that critically examines retrieved knowledge, resolves conflicting perspectives or irrelevant passages, and constructs well-supported responses through structured argumentation (Figure 1). To achieve this, *D*-RAG employs *Argumentative Explanations*, which systematically contrast opposing aspects or filter out irrelevant information, ensuring a coherent and well-grounded final answer. *D*-RAG is designed to enhance the original RAG pipeline by leading the model to leverage knowledge-intensive questions and retrieve sup-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9076–9097
November 4-9, 2025 ©2025 Association for Computational Linguistics

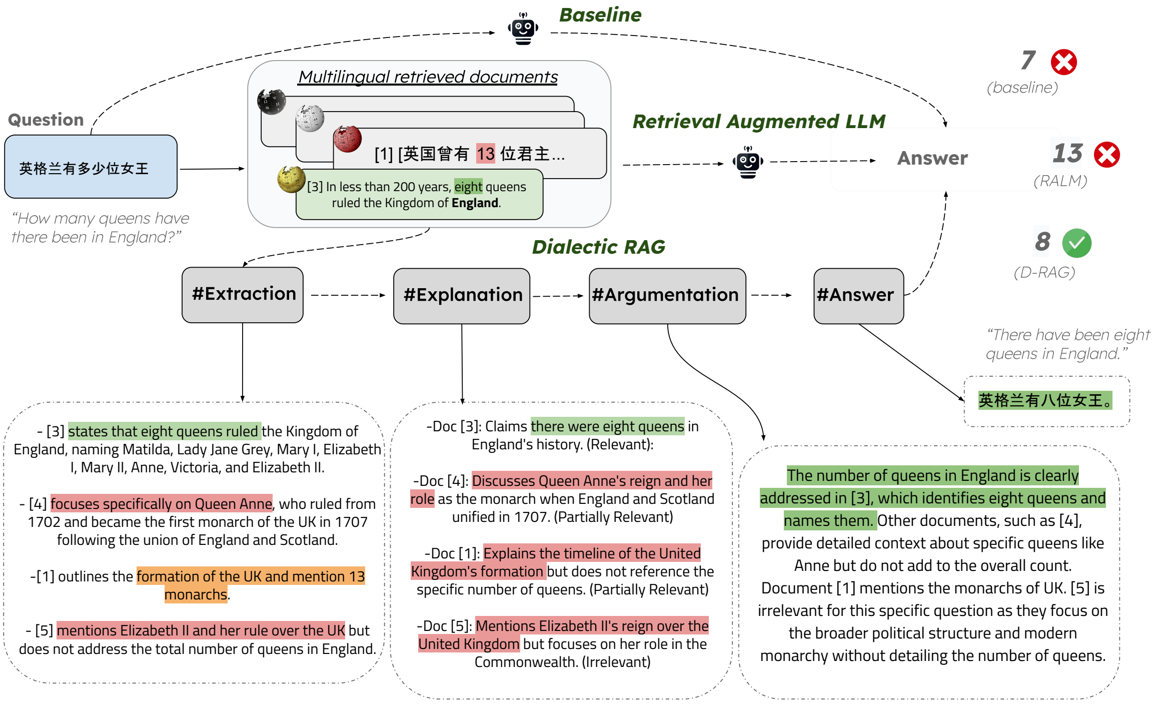

Figure 1: Our *D*-RAG allows LLMs to leverage multilingual knowledge-intensive question answering tasks by delivering argumentative explanations that support the final answer.

porting evidence through step-wise reasoning explanations that, starting from a given query, follow these steps: *(a) extraction*, where a multilingual query and documents are analysed to identify information relevant for answering the query; *(b) explanation*, where the LLMs construct single arguments about the relevance of the extracted passages, highlighting and distinguishing the furnished information; *(c) dialectic argumentation*, where the arguments are consolidated using a neutral perspective into a single final explanation; and *(d) answer*, where a short-form answer is delivered.

To evaluate the efficacy of *D*-RAG, we operate in two different configurations – as an in-context approach to provide explicit instructions for larger and more capable LLMs and as a strategy for constructing synthetic demonstrations to improve the performance and align the reasoning capabilities of smaller LLMs.

Our empirical analyses are carried out in five knowledge-intensive multilingual question-answering (QA) tasks that cover 11 different languages, showing the following results and conclusions:

- *D*-RAG elicits dialectic reasoning trajectories in LLMs by leveraging multilingual knowledge from retrieved documents, significantly outperforming baselines. For instance, on the multilingual knowledge-intensive QA task, when used with GPT-4o, it achieves an absolute average accuracy improvement of 51.6% over no-RAG settings and 12.9% over standard RAG.
- Using *D*-RAG to generate synthetic dialectic multilingual reasoning demonstrations substantially boosts the performance of smaller models. Indeed, on the proposed multilingual knowledge-intensive QA task when applied to Llama3-8B, it results in an average accuracy gain of 9.6% over standard RAG, and 5.5% over instruction-tuned RAG approaches.
- Moreover, *D*-RAG proves effective even in the presence of contradictory evidence. In challenging settings such as the BORDERLINES benchmark (Li et al., 2024), where retrieved content may contain conflicting claims, *D*-RAG supports consistent and grounded outcomes, demonstrating its capacity to maintain answer reliability.
- Finally, we show that *D*-RAG is robust to perturbations that are a limitation for traditional RAG models, including misleading retrieval and misleading reranking (i.e., random shuffling of the retrieved documents).

2 Dialectic Reasoning

Dialectic Reasoning The art of *dialectic* is the ability to reason through critical argument, to examine opposing thoughts, and reach the truth or a principled synthesis between them by providing grounded justifications. In our setting, dialectic is operationalised as a structured contest between *supporting* and *challenging* evidence drawn from heterogeneous sources. An argument advances a claim by adducing premises that are *(i) grounded* in the cited material, *(ii) relevant* to the query at hand, *(iii) sufficient* yet *non-redundant*, and (iv) *mutually consistent*. Counter-arguments are explicitly stated; the procedure then weighs claims against counterclaims under these criteria, retaining the strongest surviving objections and selecting a minimal, adequate supporting set. The outcome is not mere ranking but an *auditable* adjudication that makes the preference relation and its premises transparent.

Dialectic Reasoning for RAG Formally, given a query q and a collection of retrieved documents $\mathcal{D} = \{d_i\}_{i=1}^{n}$ expressed in multiple languages $\mathcal{L} = \{\ell_j\}_{j=1}^{k}$, the objective is to construct a grounded *dialectical* argument $\mathcal{A}$ that justifies why a supporting subset $\mathcal{D}_q^+$ is more relevant to answering q than its complement $\mathcal{D}_q^- = \mathcal{D} \setminus \mathcal{D}_q^+$. By *dialectical* we mean a structured weighing of claims and counterclaims drawn from $\mathcal{D}$ in a reasoned preference:

$$\forall\, d^+ \in \mathcal{D}_q^+,\ \forall\, d^- \in \mathcal{D}_q^- :\ \mathrm{Rel}(d^+ \mid q) > \mathrm{Rel}(d^- \mid q),$$

together with explicit premises supporting this imbalance. Dialectic-RAG (*D*-RAG) then produces a natural–language explanation E that articulates the rationale for this preference and the failure modes of competing evidence. In the multilingual setting, E must account for cross-linguistic variation and demonstrate transfer of reasoning across $\mathcal{L}$, thereby yielding robust, language-agnostic justification over heterogeneous evidence:

$$(q, \mathcal{D}, \mathcal{L}) \xrightarrow{\text{Dialectic-RAG}} (\mathcal{A}, E).$$

D-RAG Retrieval-augmented generation enriches data access in large language models (LLMs), but they struggle to critically evaluate retrieved knowledge, handle conflicts, and filter out irrelevant content. Integrating critical reasoning into LLMs is essential to resolve information disputes and ensure more coherent and grounded responses (Xia et al., 2024; Ranaldi et al., 2025c). To instruct an LLM to deliver dialectic multilingual reasoning trajectories in a RAG setting, we propose a modular strategy (Figure 1) formed of: *(a) extraction* (§2.1), where, given a query a set of multilingual documents, the model identify relevant information; *(b) argumentation* (§2.2), where the model delivers argumentative motivations about the extracted information, by displaying and discerning the relevancy about the aspects; *(c) dialectic argumentation* (§2.3), where the arguments constructed in *(b)* are summarised using a dialectic and neutral perspective into a single explanation; *(d) answering* (§2.3), where a final answer to the query is generated adhering to query constraints such as query-language and the compact form of the answer as reported in Appendix B.

We then use *D*-RAG in two scenarios as an in-context learning strategy and a synthetic generator for constructing demonstrations (§2.5), following the experimental settings proposed in (Xia et al., 2024; Ranaldi et al., 2025c) for English. For the in-context learning strategy, we use *D*-RAG to instruct LLMs to follow step-wise dialectic planning that improves the base RAG pipelines (§2.5.1). For the instruction-tuning, we use the synthetic demonstrations to improve smaller LLMs (§2.6) and transfer to them the capability of leveraging the query and the retrieved knowledge for delivering a robust argumentation to reach the answer.

2.1 Extraction

The first step, which we define as α_1 in the proposed pipeline, concerns extracting relevant retrieved knowledge from documents retrieved from a given knowledge base $\mathcal{K}$. Complementary to previous approaches in this paper, we operate in a multilingual retrieval scope (where documents come from knowledge bases in multiple languages as defined in §3.1). We operate via multilingual retriever systems provided by Cohere[1] as the default retriever model $\mathcal{R}$. For completeness, we include the results obtained from a different retrieval system in Appendix P. We then instruct the model to analyse the query, identify the main points from the retrieved documents (i.e., "**#Reference Evidence**") for answering the question, label this phase as "**#Extraction**". Since we work with multilingual queries and documents, this step is crucial to aid the model in planning the reasoning.

[1] Cohere/Cohere-embed-multilingual-v3.0

2.2 Explanations

The second step, defined as α_2, concerns instructing the model to discuss the extracted information and deliver argumentations. Specifically, after identifying and extracting information from the top-k documents, we prompt the model to discuss whether they are actually relevant or irrelevant to the query by clearly citing the passages and labelling this phase as "**#Explanation**".

2.3 Dialectic Reasoning

This step, which we define as α_3, concerns generating a final comprehensive explanatory summary. In particular, for α_3, we leverage the arguments in the previous steps to deliver the final explanation that argues the motivations that support the answer using a dialectic approach, i.e. a critical approach that relies on systematic comparison to arrive at a more articulate and well-founded conclusion. Hence, we instruct the LLM to consider the generated aspects, summarise the main points into a single argumentation, and head this as "**#Dialectic Argumentation:**".

2.4 Final Answer

The last step is defined as α_4 and results in a short-form answer used in the final evaluation. We instruct the model to generate the final answer in this form and in the same language as the query following the pattern "**#Answer:**".

2.5 *D*-RAG Application

2.5.1 *D*-RAG as in-context Learning

We adopt D-RAG as in-context learning strategy by instructing different LLMs to answer knowledge-intensive questions by dealing with retrieved knowledge. D-RAG, in a modular way, identify the most critical information from the retrieved documents (§2.1), arguing the rationale supporting the selection of appropriate points to answer the query by explaining the main passages (§2.2), deliver a single argumentation that best describes the points (§2.3); and finally, generate the final short-form answer in a strict format, to have a more detailed and strict downstream evaluation. Yet, although the sequence of instructions is well-structured and defined, the ability to perform sequential and complex reasoning tasks is limited to larger LLMs (such as GPT-4o, as discussed in the experiments). Hence, we transfer these capabilities to smaller models operating via D-RAG for building synthetic demonstrations as training sets.

2.5.2 *D*-RAG as a Synthetic Annotation

We instruct smaller models via demonstrations produced by high-performing LLMs capable of following structured instructions. In contrast to the methods proposed in (Xia et al., 2024; Asai et al., 2023; Ranaldi et al., 2025c), we use a single prompt composed of a sequence of instructions in a multilingual setting. To filter the quality of generated demonstrations, we follow the method proposed by Ranaldi et al. (2025c), which computes the citation precision for the considered documents as a proxy for the quality of the demonstrations. However, since D-RAG employs a different annotation mechanism, our annotation pipelines firstly filter out the final correct answers through a strict, exact match; then, after the filtering (which removes more than half of the annotated demonstrations), it verifies that the provided instructions have been considered. We detail the description of annotation in Appendix D.

2.6 Tuning Smaller Models

We fine-tune a Language Model θ using the annotations[2] generated via D-RAG. The annotations are augmented with demonstrations α using the standard language modelling objective to maximize the expected log-likelihood:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{(Q,\alpha,Y)\sim\mathcal{D}} \left[\log p_\theta(Y, \alpha \mid Q)\right]$$

where θ^* denotes the optimal model parameters, and $p_\theta(Y, \alpha \mid Q)$ is the joint probability of the output Y and the demonstrations α conditioned on the query Q, learned from the training corpus $\mathcal{D}$ augmented with contrastive reasoning demonstrations. While $\alpha = \alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdot \alpha_4$ is the combination of the multiple reasoning steps performed by the model, "$\cdot$" is the concatenation operator, and α_i are the respective paths generated by the overhead processes. Q is the provided query, and Y is the output, including the intermediate steps and the final answer that compose the training corpus $\mathcal{D}$.

3 Experimental Setup

We evaluate D-RAG on different open-domain question-answering tasks (§3.1). We perform the retrieval and evaluation phases by following standard approaches used to assess the RAG pipeline

[2]we select annotations as described in §2.5.2

(§3.2) and perform the tuning phase by using the setup presented in §3.3.1.

3.1 Tasks & Datasets

We use the following question-answering (QA) tasks: *(i)* MLQA (Lewis et al., 2020a), *(ii)* MKQA (Longpre et al., 2021) and *(iii)* XOR-TyDi QA (Asai et al., 2021) as they best represent multilingual open-ended question-answering tasks. Then, we use BORDERLINES (Li et al., 2024), which contains multilingual questions concerning conflicts over disputed territories (note: we follow the questions and targets delivered by Li et al. (2024)). Finally, we include Natural Questions (NQ) (Kwiatkowski et al., 2019a), as it is a widely used English benchmark for assessing RAG systems, to establish meaningful baselines for comparison. Appendices C and M report the languages and composition of each dataset. Appendix N reports detailed information about BORDERLINES.

3.2 Experimental Settings

Retrieval We employ Wikipedia as the knowledge base $\mathcal{K}$ and Cohere as the retrieval system $\mathcal{R}$. By working through the Wikimedia dump provided by Cohere[3], individual articles are embedded with the embedding model *Cohere_Embed_V3*. This pipeline makes it easy to search Wikipedia for information or to use only specific languages. For each question in the evaluation data, we retrieve the top-5 relevant documents (details Appendix I).

Models & Inference Settings To get a comprehensive evaluation of existing RAG pipelines in the main experiments, we use: GPT-4o (OpenAI, 2023), Llama3-70b-instruct (Grattafiori et al., 2024) and smaller models Llama3-8b-instruct and 1b-instruct[4]. Detailed settings and model versions in Appendix F. We use greedy decoding in all experiments to ensure a deterministic generation process, and we set the temperature to 0 and the generation length to 2048. We observed that these settings deliver better and deterministic performances.

3.3 Evaluation Metrics

We use flexible exact-match accuracy following Schick et al. (2023), which is based on whether or not ground-truth answers are included in the generated answers provided by the models instead of a strict exact match. Moreover, our prompting pipelines instruct the models to use **'#Answer'** as a final label (see Appendix A) to elicit a conclusive generation that contains a short-form answer.

3.3.1 Training Setting

To evaluate the impact of *D*-RAG reasoning demonstrations on smaller models (§2), we employ the annotations produced following the *D*-RAG strategy (§2.5.2). Further, for a fair comparison, we deliver annotations using Llama-3-SFT, where Llama is tuned on training samples without *D*-RAG (annotation generated using same query, retrieved documents and the prompt in Table 5). We fine-tune the models for three epochs with a batch size of 32 and a learning rate equal to 1e-5 with a 0.001 weight decay. We use the cosine learning rate scheduler with a warmup ratio of 0.03. We conducted our experiments on a workstation with four Nvidia RTX A6000 and 48GB of VRAM.

3.4 Evaluated Methods

We propose the following settings:

Baseline - without RAG We evaluate the baseline capabilities of selected models in a zero-shot way without introducing any documents (without RAG) using the instruction (prompt) in Table 4.

Retrieval Augmented LLM (RAG) We assess the impact of retrieved knowledge by instructing the evaluated models to consider the *top*-5 retrieved documents. We use the retrievers in §3.2.
→ **ICL** As baseline settings we use the instruction in Table 5.
→ ***D*-RAG (ICL)** To complete the RAG-based settings, we use *D*-RAG as an in-context learning strategy as in Table 6.
→ **fine-tuning** Finally, we tune Llama models using *SFT* and *D*-RAG as presented in §3.3.1 and prompt using RAG instruction (Table 5).

4 Results

Table 1 shows that *D*-RAG aids the models in leveraging retrieved documents for multilingual QA tasks, displaying the impact of dialectic argumentations on RAG (complete results in Appendices Q, R, W). We found that *D*-RAG is effective as an in-context learning (ICL) approach in larger LLMs and is helpful as a demonstration strategy to improve the performance of smaller models, achieving solid results compared to fine-tuning approaches. To this end, the following sections

[3]Cohere/wikipedia-2023-11-embed-multilingual-v3

[4]to simplify notation we omit *instruct* for the rest of the paper

Models	MKQA	MLQA	X.TyDi	Avg
Baseline				
Llama3-1B	32.5	33.7	27.3	31.2
Llama3-8B	38.9	43.4	34.5	38.6
Llama3-70B	40.7	43.9	36.5	40.4
GPT-4o	44.8	46.9	36.7	42.8
RAG				
Llama3-1B	50.6	48.6	41.7	46.9
Llama3-8B	57.3	54.5	48.1	53.1
Llama3-70B	60.1	56.6	49.2	55.3
GPT-4o	61.4	58.6	51.2	57.4
RAG → ***D*-RAG** as ICL				
Llama3-1B	48.6	48.0	38.3	45.0
Llama3-8B	56.7	53.5	48.1	52.8
Llama3-70B	67.3	62.4	55.8	62.4
GPT-4o	**68.2**	**65.5**	**60.7**	**64.8**
RAG → tuning via *SFT* and *D*-RAG				
Llama3-1B$_{SFT}$	52.1	50.0	41.3	47.8
Llama3-8B$_{SFT}$	60.3	56.3	48.5	55.0
Llama3-1B$_{D\text{-RAG}}$	55.8	53.7	46.6	51.9
Llama3-8B$_{D\text{-RAG}}$	**63.6**	**59.3**	**52.7**	**58.5**

Table 1: Average results on multilingual QA tasks (§3.1). Models instructed as detailed in §3.4. In bold, best performances of ICL and fine-tuned models.

analyse the impact of *D*-RAG when adopted as both an ICL (§4.1) and as a framework for generating annotations to instruct LLMs (§4.2). Then, in §4.4, we study a practical application on BORDERLINES (Li et al., 2024). Finally, we investigate the role of the argumentative explanations (§4.3) and revealed evidence of robustness on perturbations and functionality in low-resource settings (§4.5).

4.1 *D*-RAG in-context learning

Table 1 reports the results of *D*-RAG when adopted as an in-context learning (ICL) strategy for different models. We observe an overall improvement over the baseline models without retrieved documents (average improvements of +22 points for GPT-4o and for Llama3-70B, +14.2 points for Llama3-8B and +13.8 points for Llama3-1B); however, the results show that the impact of *D*-RAG in a RAG setting emerges for GPT-4o and Llama3-70B where *D*-RAG achieves a general improvement of +7.4 and +8.9 average points respecting to RAG. In contrast, for Llama3-8B and Llama3-1B, we observe a decrease in performance compared to the RAG pipeline, suggesting that these smaller models cannot deliver the dialectic reasoning explanations required to support their responses.

4.2 *D*-RAG Annotation Approach

Table 1 reports the impacts of *D*-RAG used as an annotation strategy for different smaller models (denoted as **RAG** → tuning via *SFT* and *D*-RAG). *D*-RAG effectively enhances the performance of smaller models when employed to deliver reasoning demonstrations via GPT-4o. We found that *D*-RAG outperform SFT approaches for both model versions. It emerges that both tuning strategies work well and outperform the baseline RAG approaches—for instance, Llama3-8B improves 52.8 → 55.0 average accuracy comparing RAG and SFT versions. Yet, the models tuned via *D*-RAG annotations consistently surpass the SFT (Llama3-1B improves +3.1 and Llama3-8B +2.9 average points). These results indicate the benefits provided by *D*-RAG demonstrations and their ability to efficiently elicit argumentations in smaller LMs (as shown in an inference example in Appendix **??**). Finally, Appendix V compares *D*-RAG with related work focusing on English, showing that although tuning is multilingual, *D*-RAG achieves sustainable performance; instead, the others underperform in multilingual QA task.

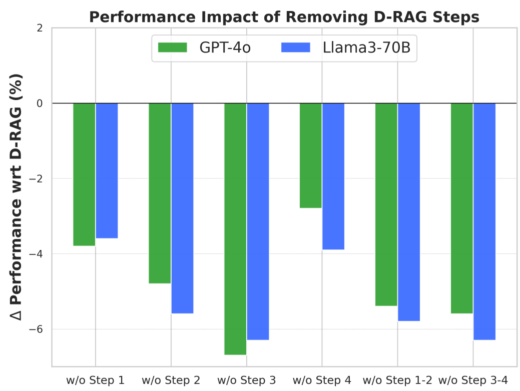

Figure 2: Performance differences (Δ) for GPT-4-o and Llama3-70B. We analyse the impact of each component on MKQA by eliminating (w/o) the *D*-RAG steps.

4.3 The Role of *D*-RAG Components

Figure 2, Table 23, and Appendix O evaluate the impact of *D*-RAG components on the final performance. The results in Figure 2 demonstrate the importance of each phase of the process introduced in §2. For GPT-4o and Llama3-70B, we observe the highest decrease in performance when removing the second and third steps. In particular, removing the second step (w/o Step 2), also defined as α_2, which is concerned with arguing and breaking down relevant points of retrieved documents to answer the given query, it is possible to observe an average decrease of -5.2% compared to *D*-RAG. Removing Step 3, which is responsible for delivering the argumentation, we observe an average reduction of -6.5% compared to *D*-RAG. These results demonstrate the crucial impact of each passage of *D*-RAG for eliciting dialectic explanations from the model. The impact of steps for ICL operation affects the tuning as well. As reported in detail in Appendix U the models tuned via modified *D*-RAG or randomly mixed steps negatively impact performance (the crucial points are Steps 2 and 3 as in the case of *D*-RAG as ICL). Finally, Appendix O argues that the error propagation rate of each component is on average around 10%.

Model	%Agreement English (En)	%Agreement X,Y,En
GPT-4o	75%	66.6%
+RAG	85%	81.6%
+*D*-RAG	**100%**	**100%**
Llama3-8B$_{ICL}$	35%	43.3%
+RAG$_{ICL}$	50%	51.6%
+*D*-RAG $_{ICL}$	65%	68.3%
Llama3-8B$_{SFT}$	65%	70%
Llama3-8B$_{D\text{-RAG}}$	**95%**	**98.3%**

Table 2: Agreement rate with controller in BORDERLINES dataset (Li et al., 2024). Details in Appendix N.

4.4 Dialectic Reasoning in BORDERLINES

To investigate the impact of our *D*-RAG in real contexts, we used BORDERLINES (Li et al., 2024). This resource provides questions concerning disputed territories as detailed in Appendix N. These questions are in English and in two additional languages, which are the land disputants (defined as **X** and **Y**). Finally, a target or controller value indicates the country that controls the territory[5]. To study the consistency and dialectic capabilities of our *D*-RAG, we then conducted a retrieval phase and evaluated GPT-4o and Llama3-8B (tuned and not) with the questions in the specific languages and English using the prompts defined in Appendices A and B. Then, setting the controller as **X**, we estimated the percentage of times the answer provided by the models prompted in English matched with the target or named controller (denoted as **%Agreement English**), and the percentage when the models prompted via queries in three languages matches among them and with the controller.

Table 2 shows that the consistency percentage increases when *D*-RAG is used. In particular, in GPT-4o, there is a 15% and 19.6% increase when *D*-RAG is compared with RAG. Similarly, it occurs between Llama3-8B instructed via *D*-RAG. Finally, Llama3-8B tuned with *D*-RAG has the most robust constancy.

4.5 Additional Analysis

Robustness To test the robustness of the proposed framework and avoid the possible performance bias obtained from noisy or misleading retrieval, we follow the methodology used in previous works. We shuffled the order of the retrieved documents *(Random Shuffle)* and inserted misleading and irrelevant documents *(Random Noise)*. Figure 3 and Appendix X report the results. *D*-RAG outperforms the baseline model with RAG as ICL and annotation strategy. The random shuffling of retrieved documents minimally impacts performance, demonstrating the permutation invariance property of *D*-RAG. Moreover, when noisy documents are added, all the evaluated models suffer a higher performance drop. The drop for *D*-RAG is typically lower than the standard RAG approach, which shows that the proposed method is more robust even when dealing with noisier results.

Quantity of Instructions Figure 4 shows the behaviour of *D*-RAG when scaling-up the number of training examples. While we found that the quantity of the demonstrations used in *D*-RAG is important in determining the final performance, we found that *D*-RAG can outperform the baseline RAG models with only 50% of training demonstrations, also achieving superior training performance when compared to the fine-tuned SFT model (i.e., the model fine-tuned without *D*-RAG demonstrations as explained in §3). This further highlights the quality of the training signal provided by the constructed synthetic demonstrations.

Quality of Generation Table 3 shows the tendency to generate answers in the same query language and follow the provided instructions at inference time (details in Appendix K). The requirements that our framework must satisfy are *i)* all instructions given in the prompt must be followed, and *ii)* in the multilingual task, the answer must be in the same query language. We show that the GPT4-o and Llama3-70B are consistent with the requirements. Instead, the Llama3 models do not

[5]in some cases, there are no defined places that we do not consider in our analysis.

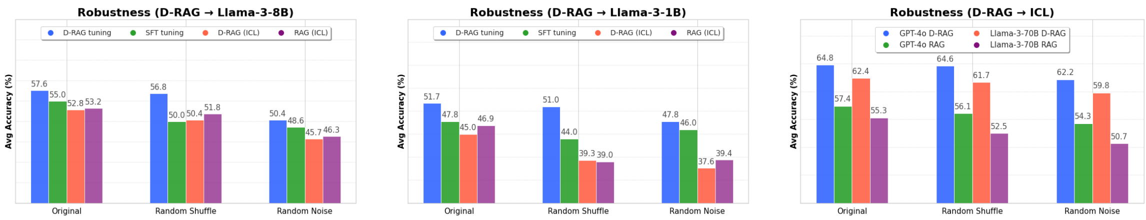

Figure 3: Robustness results on QA datasets (§3.1). We provide retrieved documents by randomly shuffling them (Random Shuffle) and introducing irrelevant documents (Random Noise). Additional experiments in Appendix X.

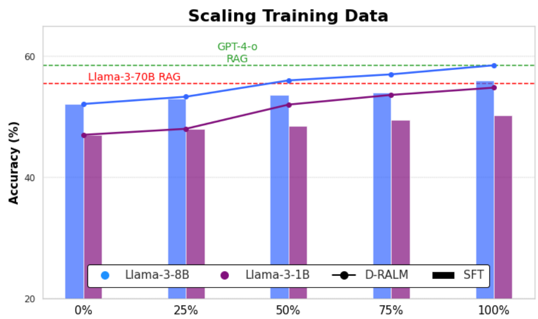

Figure 4: Performances assessment of Llama3-8B and -1B by scaling *D*-RAG (lines) and SFT (bars) tuning demonstrations on ablation set (Appendix E).

follow the instructions, but when tuned, using *D*-RAG demonstrations, they become consistent.

Models	IF	CL	***LR*-IF**	***LR*-CL**
GPT-4o	-	85.6%	-	72.2%
+ *D*-RAG $_{ICL}$	90.5%	**94.8%**	83.6%	**86.4%**
Llama3-70B	-	65.2%	-	63.8%
+ *D*-RAG $_{ICL}$	83.5%	**79.4%**	77.4%	**70.2%**
Llama3-8B	-	65.9%	-	46.0%
+ SFT	-	72.8%	-	64.6%
+ *D*-RAG $_{ICL}$	58.4%	66.2%	45.5%	44.0%
+ *D*-RAG $_{FT}$	**78.3%**	**72.0%**	**67.1%**	**69.6%**
Llama3-1B	-	57.2%	-	30.4%
+ SFT	-	66.3%	-	48.8%
+ *D*-RAG $_{ICL}$	40.0%	53.3%	40.7%	32.2%
+ *D*-RAG $_{FT}$	**60.4%**	**69.5%**	**45.3%**	**59.9%**

Table 3: Percentage (%) of answers that follow the prompt instructions (IF) and generate the final answer in the correct language (CL). FT indicates fine-tuned models via *D*-RAG. ***LR*** indicates the results for low-resource languages considering the MKQA answers.

***D*-RAG Settings & Comparisons** We provide evidence for the robustness of the *D*-RAG by proposing three experiments. Firstly, Appendix L shows that decomposing our *D*-RAG into different prompts delivers benefits which are minimal compared to the cost of increasing the number of prompts (four prompts against a single one). Then, in Appendix J, we analyse the impact of internal argumentation in the query language. As shown in Table 13, argumentation in a language other than English (a language in which the models used in this work are more profitable) leads to a drop in performance that will definitely be a matter of future investigation. In Appendix V, we show that *D*-RAG perform well even in monolingual tasks (English). In contrast, related methods achieve lower performance in multilingual tasks. Finally, in Appendix P we show that the proposed framework is not related to the proposed multilingual retrieval system (cohere) but works well with other systems as well.

5 Applicability & Future Work

We propose a method to improve RAG capabilities in multilingual scenarios by eliciting LLMs to consider heterogeneous sources of knowledge and argue the reasons that support the answer in a dialectic manner. Our work applications are related to: *(i)* improving the answering of questions that involve a retrieval in a setting with unbalanced resource availability, e.g., in the case of Wikipedia, where the number of documents differs across languages (Table 12). *(ii)* improving the argumentation in scenarios where there is an information overlap on retrieved statements (§4.4). *(iii)* Transferring the capabilities of delivering dialectic explanations to smaller LLMs by teaching them via synthetic demonstrations. We plan to analyse the role different languages can play in delivering reasoning and how much the multilingual proficiency of LLMs can influence this task.

6 Related Work

Lewis et al. (2020b) investigated the advantages of augmenting LLMs with retrieved knowledge, a technique known as Retrieval-augmented Language Models (RAG). Shi et al. (2023) demonstrated that the benefits of RAG could be undermined by noisy retrieval. Several studies have enhanced RAG through in-context solutions, tuning, or retriever interventions (Menick et al., 2022; Jiang et al., 2023; Gao et al., 2023; Sawarkar et al., 2024). While effective, in-context learning only partially mitigates retrieval bias, and tuning remains costly (Asai et al., 2023). Xia et al. (2024) proposed low-impact reasoning techniques, later enhanced via contrastive reasoning by Ranaldi et al. (2025c). Unlike these English-centric approaches, we focus on multilingual knowledge-intensive tasks. Complementing (Zhang et al., 2022), we study the inference phase and enrich the work proposed by Chirkova et al. (2024), we propose a framework that allows the LLMs to leverage the different knowledge, reason about them, and deliver argumentative explanations by using a dialectic approach. Our effort aims to improve the limitations of multilingual RAG, bias towards language, information disparity (Sharma et al., 2024) or conflicting knowledge (Li et al., 2024).

7 Conclusion

RAG has demonstrated its potential to improve LLM performances in knowledge-intensive tasks; yet, a major limitation lies in handling heterogeneous retrieved data, especially in multilingual cases. To address this, we propose *Dialectic-RAG* (*D*-RAG) to improve retrieval-based reasoning via argumentative explanations. We show that *D*-RAG significantly improves multilingual retrieval-augmented inference, enhancing both in-context learning and demonstration-based instruction for smaller models. Structuring reasoning over retrieved knowledge mitigates misleading inferences and improves response consistency, reinforcing the importance of dialectic reasoning for reliable multilingual RAG applications.

8 Future Work

The study of LLMs' reasoning capabilities in non-English settings is an emerging research domain. Multiple studies have proposed techniques to increase (Ranaldi et al., 2024b,c,d), transfer (Ranaldi and Pucci, 2023), or align (Ranaldi et al., 2024a) reasoning capabilities beyond English. Although our contribution has focused on the benefits of instruction models to follow a reasoning trace to maximise the understanding and generation capabilities, we are interested in continuing our studies. In particular, we aim to investigate the impacts of further tuning in the proposed experimental settings, using synthetic data produced by models of the same family (Ranaldi and Freitas, 2024a) or self-generated (Ranaldi and Freitas, 2024b), adopting reinforcement learning in language and multimodal spaces as initiated in our parallel works (Ranaldi and Pucci, 2025; Ranaldi et al., 2025b), as well as instructing the model to abstract reasoning passages (Ranaldi et al., 2025d).

Acknowledgements

This work is funded by EU Horizon Europe (HE) Research and Innovation programme grant No 101070631, and UK Research and Innovation under the UK HE funding grant No 10039436

Limitations

Due to the limitations imposed by the evaluation benchmarks and the cost of the closed-source models, we conducted tests on three tasks in different languages, which only scratches the surface of the world's vast array of languages. In the future, it will be appropriate to study the generality of our model compared to other large language models, both closed-source and open-source. In addition, it will be of interest to us to analyse in detail the argumentative capabilities of LLMs in specific languages, exemplifying the benefits and biases that can be provided.

Ethics Statement

We did not address ethical considerations in our work. The data employed for our experiments were sourced exclusively from open, publicly available benchmarks detailed reported in the appendices. Additionally, official and respected sources obtained statistics related to language differences in commonly used pre-training datasets. Our analysis and methodology were conducted with sensitivity and care, ensuring that no aspects concerning gender, sex, or race were considered or affected.

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.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Common Crawl. 2021. Common crawl 2021. Web. Accessed: 2023-12-12.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. 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. 2019a. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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. 2019b. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Tammy Li, Runqi Liu, Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

OpenAI. 2023. Gpt-4 technical report.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Leonardo Ranaldi and Andre Freitas. 2024a. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi and Andre Freitas. 2024b. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2347, Miami, Florida, USA. Association for Computational Linguistics.

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

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, and Andrè Freitas. 2024a. Does the *Order* matter? Curriculum learning over languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5212–5220, Torino, Italia. ELRA and ICCL.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024b. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. 2024c. Empowering multi-step reasoning across languages via program-aided language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187, Miami, Florida, USA. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024d. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241, Mexico City, Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Federico Ranaldi, and Giulia Pucci. 2025b. R2-MultiOmnia: Leading multilingual multimodal reasoning via self-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8220–8234, Vienna, Austria. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025c. Eliciting critical reasoning in retrieval-augmented generation via contrastive explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11168–11183, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025d. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17222–17240, Vienna, Austria. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and Andrè Freitas. 2024e. Eliciting critical reasoning in retrieval-augmented language models via contrastive explanations.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

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

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

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2024. Improving retrieval augmented language model with self-reasoning.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yuetong Zhao, Hongyu Cao, Xianyu Zhao, and Zhijian Ou. 2024. An empirical study of retrieval augmented generation with chain-of-thought.

A Prompting Approaches

Baseline Prompt Template (no RAG)

#Role
Please answer the question by following the provided instructions.

#Instructions:
Answer the question as clearly as possible based on your knowledge following the format "**#Answer:**"
Note: answer in the query language.

#Question:
{`question`}

Table 4: Baseline prompting template.

Baseline RAG Prompt Template

#Role
Please answer the question by following the provided instructions.

#Instructions:
Answer the question as clearly as possible using the provided ***Reference Evidence*** *and follow the format* "**#Answer:**" *Note: answer in the query language*.

#Reference Evidence:
[1] {$\texttt{Document}_1$}
[2] {$\texttt{Document}_2$}
[3] {$\texttt{Document}_3$}
[4] {$\texttt{Document}_4$}
[5] {$\texttt{Document}_5$}

#Question:
{`question`}

Table 5: RAG prompting example.

B *D*-RAG prompting Template

D-RAG Prompt

#Role
You are helpful assistant. Please answer the question by following the provided instructions.

#Requirements:
Answer the question as clearly as possible using the provided ***#Reference Evidence*** *and follow the* ***#Instructions***.

#Reference Evidence
[1] {`Document`$_1$}
[2] {`Document`$_2$}
[3] {`Document`$_3$}
[4] {`Document`$_4$}
[5] {`Document`$_5$}

#Instructions

1) Consider the provided documents labelled "**#Reference Evidence**", identify and understand the main points. Follow the directions in detail and use only the information in the documents, exemplifying which points are most relevant for answering the question **#Question**.
Note: Ensure all documents are considered and provide a precise and well-structured response using English as the shared language. Name this passage "**#Extraction:**".

2) For each document, extract the most relevant information for answering the **#Question** discussing whether they are actually relevant or irrelevant.
To ensure clarity, include the exact passages from each supporting document and reference their document numbers. Organise your argumentation as follows:" Document [1] claims [specific argument], whereas passage [4] claims.... Name this passage as "**#Explaination:"**.

3) Please consider the step **2)** in detail, ensure they are correct. Then, provide a single argumentative explanation that considers the passages and their supporting motivations from a *neutral* perspective, as concern argumentative passages.
*Note: To enhance clarity, present your detailed explanation under the heading "****#Dialectic Argumentation:****"*

4) Finally, to facilitate the final evaluation, deliver a short-form answer by labelling it as "**#Answer:**"
Note: answer in the query language.

#Question
{`question`}

Table 6: The Dialectic RAG (*D*-RAG) framework instructs the model to deliver multi-step reasoning paths that lead the models to solve the task by explaining the perspectives that have emerged.

C Data Composition

In our experiments, we use three knowledge-intensive question-answering task: *(i)* MLQA (Lewis et al., 2020a), *(ii)* MKQA (Longpre et al., 2021) and *(iii)* XOR-TyDi QA (Asai et al., 2021) as they best represent multilingual open-ended question-answering tasks. MLQA is manually translated from SQuAD v1.1 (Rajpurkar et al., 2016), MKQA and XOR-TyDi QA are machine translated and manually controlled by Natural Questions (Kwiatkowski et al., 2019b) and TyDi QA (Clark et al., 2020), respectively.
We use test sets in the languages in Table 15. For each language, we used the same questions and, consequently, the same number of questions to avoid any imbalance in double-checking by retrieving the corresponding ids. Details on the number of instances are in Table 7. In addition, since the experimental setting of our work requires a subset of examples to conduct the annotation phase (§2.5), we used instances defined in Table 8 (not present in the evaluation set) and annotated them as described in Appendix D.

D Data Annotation

We use *D*-RAG annotations to fine-tune smaller models to leverage knowledge-intensive tasks using retrieved documents (§2.6). To ensure the quality of the annotations firstly, we use an exact-match as the first filter then we use GPT-4o-mini as annotator. HThen, after ensuring that the final answer matches the target, we systematically instruct the GPT-4o-mini using the *D*-RAG (Table 6). This double-check assess the accuracy of the outcomes delivered. Hence, we prompt the model as follows:

> **#Role:**
> You are an experienced expert skilled in answering complex problems through logical reasoning and structured analysis.
> **#Task:**
> Given the following sentences, you are a decision maker who decides whether the 'Response' provides the 'Target' as the final outcome and follows the given 'Instructions'. If the output doesn't align with the target answer and doesn't not follow the instructions, respond with '0', whereas if it's correct, then respond with '1'. *Please, ensure that all criteria are complied with the requests and do not provide any other answer beyond '0' or '1'.*
> **#Senteces:**
> #Response: {model_result}
> #Target: {target_answer}.
> #Instructions: {D-RAG_template}.

E Splitting Informations

As in §2.5.2 and in Appendix C, we conducted an evaluation phase on equally distributed portions of the data on the languages (Table 7). In addition, we annotated a set of samples (Table 8) equally distributed among the languages in Table 9. The annotation data were filtered, and some questions are repeated for different languages (by task and dataset construction), the arguments are different because the documents retrieved are different.

Testing Sets

Dataset	# per lang available	# per lang used	#Tot. used	#Tot. ablation
MLQA	1.5k	0.8k	7.2k	1.8k
MKQA	2k	1.0k	6.0k	1.0k
XOR-TyDi	0.6k	0.4k	2.4k	0.6k

Table 7: Number (#) of instances for evaluation (test/ablation) phases which are equally distributed among the languages in Table 15. (k denotes 1000 instances)

Training Sets

Dataset	#example	#example correct	#Total used
MLQA	3500	1920	1920
MKQA	2000	1128	920
XOR-TyDi	800	556	200
Total	6.3k	3.6k	3.02k

Table 8: Number of datasets used for evaluation phases which are equally distributed among the languages in Table 15. (k denotes 1000 instances)

Language used for training

Dataset	Languages
MKQA	English, Spanish, German, Russian, Chinese, Finnish, Arabic
MLQA	English, Chinese, Arabic, German, Spanish
XORTyDi QA	English, Chinese, Arabic, Finnish

Table 9: Languages annotation.

F Models Version

We used the following models on the architectures reported in §3 or via API (we spent $250).

Model	Version
GPT-4o	OpenAI API (gpt-4-o)
Llama3-70B	meta-llama/Meta-Llama-3-70B-Instruct
Llama3-8B	meta-llama/Meta-Llama-3-8B-Instruct
Llama3-1B	meta-llama/Meta-Llama-3.2-1B-Instruct

Table 10: Models (huggingface.co). We used the configurations described in §3 in the repositories for each model *(access verified on 25 Jan 2025).

G Difference between High- and Low-resource Languages

In this work, we define the differences between high-resource (HR) and low-resource (LR) using the consideration already taken in previous works (Chirkova et al., 2024; Qin et al., 2023). We report two tables: Table 11 reports the language distribution of CommonCrawl, and Table 12 the number of documents in the Wikipedia dump used in our work (§3).

Language	Percentage
English (en)	46.3%
Russian (ru)	6.0%
German (de)	5.4%
Chinese (zh)	5.3%
French (fr)	4.4%
Japanese (ja)	4.3%
Spanish (es)	4.2%
Other	23.1%

Table 11: Language distribution of CommonCrawl (Common Crawl, 2021).

H Documents in Wikimedia_Dump

Language	Percentage
English (en)	41,488k
Russian (ru)	13,784k
German (de)	20,772k
Chinese (zh)	7,875k
Italian (it)	10,462k
French (fr)	17,813k
Japanese (ja)	6,626k
Spanish (es)	12,865k
Portuguese (pt)	5,637k
Bengali (bn)	767k
Finnish (fn)	272k
Arabic (ar)	1,050k
Thai (th)	876k
Vietnamese (vi)	2,067k
Telogu (te)	124k

Table 12: Language distribution of Wikimedia Dump introduced in §3.

I Retrieval Details

Retrieval We use Cohere as the retrieval system and Wikimedia_dump as the knowledge base $\mathcal{K}$ for all experiments. We use $\mathcal{K}$ provided by Cohere *wikipedia-2023-11-embed-multilingual-v3* (available on huggingface). They provide individual documents embedded with multilingual embedding model *Cohere_Embed_V3* (in Table 12 are reported the dump composition). For each question in the evaluation data, we retrieve 10 relevant documents and then filter the top-5 most relevant ones as done in the related repository (dot score between query embedding and document embeddings).

J Ablation Argumentation Language

D-RAG is instructed to use an English argumentation (see Table 6). In this experiment, we instruct the model to operate in Chinese, Arabic and German and report the differences with the original *D*-RAG, which is in English.

Models+*D*-RAG	ΔDE	ΔZH	ΔAR
GPT-4o	-2.4	-6.3	-8.6
Llama3-70B	-6.8	-9.5	-12.6
Llama3-8B	-8.1	-9.3	-14.6
Llama3-1B	**-12.8**	**-16.6**	**-18.4**

Table 13: Ablation on argumentation language impacts on *D*-RAG using MKQAs' ablation set.

K Ablation Output Analysis

To control the quality of the generations, we defined two different metrics: Instruction Following (IF) and Correct Language (CL). The role of IF is to investigate whether the models followed the instructions given in the prompt. The role of CL, on the other hand, is to analyse whether the language of the final response is the same as that of the query (note that this requirement was well defined in the prompt. In order to have a robust result, we conducted these two analyses using GPT-4o-mini as an instructed evaluator, using the prompt in Appendix D and avoiding the target part in the case of IF. We computed the CL using OpenLID framework (Burchell et al., 2023). For both values, we reported the percentage of correctness (accuracy).

L Ablation number of Steps

D-RAG operates via a single instruction. To observe the impact of instruction splitting on the final performances, we apply the same prompt shown in Table 6 by giving the model one step at a time.

Models	MKQA	MLQA	XoR TyDi
GPT-4o			
Single Step	**68.6**	65.8	61.3
4 Steps	68.4	66.9	63.0
Llama3-70B			
Single Step	67.0	62.9	56.2
4 Steps	**67.5**	**63.4**	**56.0**
Llama3-8B tuned via *D*-RAG			
Single Step	62.4	59.8	52.1
4 Steps	**63.5**	**60.9**	**53.6**
Llama3-8B tuned via *D*-RAG			
Single Step	55.9	53.2	46.4
4 Steps	**57.4**	**55.3**	**48.9**

Table 14: *D*-RAG using Single Step prompting (traditional approach) and breaking the steps into single phases on ablation set of proposed QA tasks.

M Proposed Task

Dataset	Languages	#Languages
MKQA	`English, Spanish, German, Russian, Chinese, Finnish, Arabic,` **`Italian, Korean`**	9
MLQA	`English, Chinese, Arabic, German, Spanish,` **`Hindi`**	6
XORTyDi QA	`English, Chinese, Arabic, Finnish,` **`Korean, Telugu`**	6

Table 15: Languages present in datasets used in this work. *In **bold**, the languages are used only for evaluation as described in Appendix C.

N Experiment on BORDERLINES

To investigate the impact of our *D*-RAG in real contexts, we used examples from the BORDERLINES (Li et al., 2024). This resource has questions concerning disputed territories between two nations that bureaucratically belong to a specific country. The questions have the form `Is` **`Place P`** `a territory of` **`A) Country X`** `or` **`B) Country Y`**`?`. These questions are in English, language **`X`** and **`Y`** (are the languages spoken in the countries). Finally, a target or controller value indicates the country that controls the **`P`**. (in some cases, there are no defined places that we do not consider in our analysis)
To study the consistency and dialectic capabilities of our *D*-RAG, we selected a small set consisting of 60 instances (20 questions in English, l20 language **`X`** and 20 **`Y`**). We then conducted a retrieval phase and prompt GPT-4o with the questions in the specific languages and English using the prompts defined in Appendices A and B. Then, we set the controller as **`X`** we estimated the percentage of times the answer provided by the model prompted in English matched with the controller (denoted as **%Agreement English**, and the percentage when the models prompted via queries in three languages matches among them and with the controller.

O Error Propagation & Document Consideration in *D*-RAG

We provide a detailed analysis of error propagation across the four stages proposed in *D*-RAG. We quantifying the error rates attributed to each subcomponent, recognising that every stage performs a distinct function (as proposed in §2). The analysis was conducted on an ablation subset combining the MKQA dataset and GPT-4o outputs. The error rate for each step was independently assessed using GPT-4o as an initial annotator, supplemented by a rigorous double manual verification process to ensure reliability. The cumulative error increases across stages due to the dependency of each step on the previous one. The total failure rate across the full pipeline is **34%**.

Stage	Description	Error Rate
Extraction	Extraction of relevant information	11%
Explanation	Explanation relevance	9%
Dialectic Reasoning	Argumentation about output of previous steps	8%
Answering	Generation of the final solution	6%

Step	Isolated Error	Cumulative Error
Extraction	11%	11%
Explanation	9%	20%
Dialectic Reasoning	8%	28%
Answering	6%	34%

Table 16: Error rates per stage and cumulative error analysis in the D-RAG framework. Each stage contributes independently and sequentially to the overall error rate.

P Ablation Retrieval System

D-RAG is independent of the retrieval system used. In the initial setup to facilitate reproducibility we used cohere (explained in detail in Appendix I). To demonstrate complete independence, we conducted the same experimental setting by replacing cohere with BGE-m3 (Chen et al., 2024). We then reported the results in Table P and the overlap of the retrieved documents in Table 18.

Dataset	English	Spanish	German	Russian	Chinese	Finnish	Arabic	Italian	Korean	Avg
baseline (cohere)	69.2	56.0	54.0	43.4	40.2	31.3	30.0	57.2	26.0	45.2
baseline (BGE-m)	68.2	54.6	55.4	43.8	38.8	31.8	31.0	56.4	24.9	45.0
D-RAG (cohere)	79.0	72.8	76.9	68.9	65.0	70.7	60.3	75.7	55.8	68.2
D-RAG (BGE-m)	78.6	72.0	75.6	70.0	64.2	68.5	60.6	76.0	55.8	69.0

Table 17: Comparison between D-RAG performed by cohere and BGE-m on MKQA using GPT-4o across different languages.

Dataset	English	Spanish	German	Russian	Chinese	Finnish	Arabic	Italian	Korean
MKQA	92%	82%	86%	84%	86%	82%	84%	88%	85%
MLQA	95%	86%	84%	–	79%	–	86%	–	–

Table 18: Overlap of documents retrieved using Cohere and BGE-m3 on the ablation set.

Q Results GPT-4-o

	-	RALM	RALM	D-RAG
Lang.	-	($\mathcal{R}$ *from* $\mathbf{W}_{SL}$)	($\mathcal{R}$ *from* $\mathbf{W}_{ML}$)	
		MKQA		
English	60.9	71.9	73.8	78.8
Tot.Avg	44.8	49.2	61.4	68.2
		MLQA		
English	68.8	75.1	75.1	76.8
Tot.Avg	46.9	54.2	58.6	65.5
		XOR TyDi QA		
English	54.7	68.3	68.0	71.8
Tot.Avg	36.7	47.2	51.2	60.7

Table 19: Performance GPT-4 (exact-match %) across different languages and retrieval settings (retrieval $\mathcal{R}$ from Wikipedia in Specific Language denoted as $\mathbf{W}_{SL}$ and Multiple Languages denoted as $\mathbf{W}_{ML}$).

R Results Llama3-70B

		RALM	RALM	D-RAG
Language	-	($\mathcal{R}$ *from* $\mathbf{W}_{SL}$)	($\mathcal{R}$ *from* $\mathbf{W}_{SL}$)	
		MKQA		
English	58.5	69.3	72.8	76.1
Tot Avg	40.7	51.4	60.1	67.3
		MLQA		
English	68.0	73.4	74.1	75.6
Tot Avg	43.9	52.9	56.6	62.4
		XOR TyDi QA		
English	53.6	66.8	67.3	71.6
Tot Avg	36.5	45.8	49.2	55.8

Table 20: Performance Llama3-70B (exact-match %) across different languages and retrieval settings (retrieval $\mathcal{R}$ from Wikipedia in Specific Language denoted as $\mathbf{W}_{SL}$ and Multiple Languages denoted as $\mathbf{W}_{ML}$).

S Results Llama-3-8b

		RALM	RALM	+SFT	*D*-RAG (ICL)	*D*-RAG (SFT)	*D*-RAG (SFT+ICL)
Language	-	($\mathcal{R}$ *from* $\mathbf{W}_{SL}$)	($\mathcal{R}$ *from* $\mathbf{W}_{ALL}$)				
MKQA							
English	57.4	68.8	70.7	72.0	70.3	75.8	76.0
Avg	41.3	51.4	60.3	62.7	58.8	65.4	66.0
Total Avg	38.9	49.7	57.3	60.3	56.7	63.6	64.0
MLQA							
English	66.4	70.2	72.7	74.9	72.8	75.0	75.0
Avg	47.6	55.8	58.7	61.1	57.6	63.4	63.8
Total Avg	43.4	51.5	54.5	56.3	53.5	59.3	59.6
XOR TyDi QA							
English	53.0	65.6	65.4	68.4	66.0	68.8	70.0
Arabic	38.6	52.3	56.2	58.9	54.8	59.7	60.0
Russian	41.2	51.4	53.0	53.5	50.0	53.8	53.6
Finnish	31.5	49.2	51.3	50.9	49.2	52.0	52.4
*Korean	28.7	34.3	40.0	38.2	36.0	39.7	39.5
*Telugu	14.3	16.2	22.3	22.3	19.8	21.9	22.3
Avg	41.0	54.5	56.6	57.8	55.0	58.7	60.0
Total Avg	34.5	44.6	48.1	48.5	45.9	52.7	52.9

Table 21: Performance Llama3-8B (exact-match %) across different languages and retrieval settings (retrieval $\mathcal{R}$ from Wikipedia in Specific Language denoted as $\mathbf{W}_{SL}$ and Multiple Languages denoted as $\mathbf{W}_{ML}$). We denote with * languages that are not part of the tuning set as described in Appendix E.

T Languages of Retrieved Documents

	$\mathcal{R}$ `from` $\mathtt{W}_{ALL}$		
Question Lang.	**%En**	**%SL**	**%Oth**
MKQA			
English	98.9%	-	1.1%
German	10.2%	86.3%	3.1%
Italian	11.8%	85.8%	2.4%
Spanish	11.4%	86.0%	2.8%
Finnish	22.6%	67.1%	10.3%
Russian	22.2%	65.2%	12.6%
Chinese	14.4%	81.2%	4.4%
Arabic	24.3%	66.2%	9.5%
Korean	24.0%	65.5%	10.5%
MLQA			
English	99.2%	-	0.8%
Chinese	15.3%	83.5%	2.2%
Arabic	20.8%	70.0%	9.2%
German	13.0%	85.5%	1.5%
Spanish	11.4%	86.0%	2.8%
Hindi	32.6%	58.8%	9.2%
XORTyDi QA			
English	98.4%	-	1.6%
Arabic	16.3%	76.6%	7.1%
Korean	31.2%	59.2%	9.8%
Russian	19.8%	68.4%	11.8%
Finnish	19.8%	70.8%	9.4%
Telogu	42.0%	45.6%	12.4%

Table 22: Percentage of the languages of retrieved documents. We retrieve the documents using $\mathcal{R}$ system from the Wikipedia dump (detailed in §3) considering all languages analysed in the task $\mathtt{W}_{ALL}$). The languages are double-checked using OpenLID framework (Burchell et al., 2023).

U Ablation on *D*-RAG Components

Models	MKQA	MLQA	XoR TyDi
FULL	**63.6**	**59.3**	**52.7**
w/o *(2)*	53.6	50.9	42.8
w/o *(3)*	49.2	48.6	42.1
w/o *(4)*	58.6	55.8	49.7
RANDOM	50.4	41.5	39.3

Table 23: Evaluation of impacts of each component on evaluation sets of proposed QA tasks with Llama3-8B$_{D\text{-RAG}}$. We eliminate (w/o) or RANDOM shuffling the four defined steps (§2).

V *D*-RAG on monolingual task

Models	MKQA	NQ
Self-RAG (Asai et al., 2023)	38.7	39.2
C-RAG (Ranaldi et al., 2024e)	41.5	**40.2**
D-RAG	**49.6**	39.4

Table 24: Accuracies of *D*-RAG on Llama-2-7b on MKQA and NQ (Kwiatkowski et al., 2019a).

W Results Llama3-1B

Language	-	RALM ($\mathcal{R}$ *from* $\mathbf{W}_{SL}$)	RALM	+SFT	*D*-RAG (ICL)	*D*-RAG (SFT)
			($\mathcal{R}$ *from* $\mathbf{W}_{ALL}$)			
			MKQA			
English	52.0	63.5	65.4	66.2	63.9	69.4
Avg	33.8	44.0	52.3	54.6	50.4	57.7
Total Avg	32.5	42.2	50.6	52.1	48.6	55.3
			MLQA			
English	58.2	63.6	64.0	67.3	68.2	70.3
Avg	36.8	49.4	53.0	54.6	50.3	48.2
Total Avg	33.7	45.5	48.6	50.0	48.0	53.7
			XOR TyDi QA			
English	47.3	60.0	59.7	62.8	60.4	62.7
Avg	33.8	48.0	50.2	51.4	47.5	53.5
Total Avg	27.3	40.6	41.7	41.3	38.3	46.6

Table 25: Performance Llama3-1B (exact-match %) across different languages and retrieval settings (retrieval $\mathcal{R}$ from Wikipedia in Specific Language denoted as $\mathbf{W}_{SL}$ and Multiple Languages denoted as $\mathbf{W}_{ML}$). We denote with * languages that are not part of the tuning set as described in Appendix E.

X Ablation top-k retrieved

In the experiments shown in Figure 3 and the ablation settings proposed in §4.5 we reproduced the same one proposed in (Xia et al., 2024; Ranaldi et al., 2025c). To strengthen these results, we have added further documents and reproduce the same settings. In Table 26 is possible to observe that the benefits of *D*-RAG are confirmed with more documents.

		Ablation Retrieval Experiments Llama-3-8B			
Model	**top-5**	**random shuffle**	**random bias**	**top-3**	**top-10**
RAG	53.2	51.8	46.3	44.2	52.3
D-RAG (ICL)	52.8	50.4	45.7	45.8	50.2
D-RAG (SFT)	**57.6**	**56.8**	**50.4**	**50.4**	**57.6**
		Ablation Retrieval Experiments GPT-4o			
Model	**top-5**	**random shuffle**	**random bias**	**top-3**	**top-10**
RAG	57.4	56.1	54.3	50.2	57.0
D-RAG (ICL)	**64.8**	**64.6**	**62.2**	**61.8**	**64.8**

Table 26: Ablation of retrieval strategies across Llama-3-8B and GPT-4o, showing top-k retrieval performance under different perturbations.

Y *D*-RAG and GPT-4o on BORDERLINES

- $\mathcal{Q}_{En}$: *Is Sixty-Four Villages East of the River a territory of A) Russia or B) Republic of China?*
- $\mathcal{Q}_{Ru}$: Являющя ли Шестьдесят четыре деревни к востоку от реки территорией А) России Б) Республики Китай?
- $\mathcal{Q}_{Zh}$: 河东六十四村是A）俄罗斯还是B）中华民国的领土?
- **Target:** A) *Russia* (as reported in BORDERLINES (Li et al., 2024))

Baseline Answer to $\mathcal{Q}_{En}$: I couldn't find any information about 'Sixty-Four Villages East of the River'.
Answer to $\mathcal{Q}_{Zh}$: 河六十四是 B) 中民的土。
Answer to $\mathcal{Q}_{Ru}$: Правильный ответ: А) Россия.

Retrieved documents [1] The Sixty-Four Villages East of the River were inhabited villages by Manchu, Daur, and Han, located on the left (north) bank of the Amur River opposite Heihe and the Blagoveshchensk. Among Russian historians, the district occupied by the villages referred to as Zazeysky [...] (from $\mathbf{W}_{En}$)
[2] The Treaty of Aigun of 1858, signed by the Chinese Qing Dynasty and the Russian Empire, ceded the lands north of the Amur to Russia. Manchu residents who lived on the left bank of the Amur received the right to stay and live on Russian territory. During the Boxer Rebellion of 1900, Russia sent its troops to occupy 64 villages and forced the people to cross the Amur River and settle in China. Currently, most of this territory belongs to the Blagoveshchensky and Tambov districts of the Amur region. (from $\mathbf{W}_{Ru}$)
[3] The 'Treaty of Peking' signed between the Qing Dynasty and the Russian Empire stated: 'The land mentioned above is open land. Wherever the Chinese live, or the Chinese occupy the fishing and hunting land, Russia is not allowed to occupy it. The Chinese may still occupy it by fishing and hunting as usual. The Boxer Rebellion took place in 1900, and the Qing government had no time to consider the situation in the northeast. Russia used the destruction of the Dongqing railway by the Boxer Rebellion as an excuse to send troops [...] (from $\mathbf{W}_{Zh}$)
[4] The 1900 Amur anti-Chinese pogroms were a series of ethnic killings and reprisals undertaken by the Russian Empire against subjects of the Qing dynasty of various ethnicities. They took place in the Sixty-Four Villages East of the River in the Amur region during the same time as the Boxer Rebellion. The events resulted in thousands of deaths, the loss of residency for Chinese subjects living in the villages, and increased Russian control over the region.[...] (from $\mathbf{W}_{En}$)
[5] The Boxer Rebellion, also known as the Boxer Uprising, was an anti-foreign, imperialist, and Christian uprising in North China between 1899-1901, towards the end of the Qing dynasty, by the Society of Righteous and [...] (from $\mathbf{W}_{En}$)

RAG Answer to $\mathcal{Q}_{En}$: The Sixty-Four Villages East of the River The villages are currently a territory of **A) Russia**.
Answer to $\mathcal{Q}_{Zh}$: 河东六十四村历史上属于清朝（中华民国前身）领土，根据《北京条约》规定，俄罗斯不得占据中国居民捕鱼狩猎的土地。[...] 尽管该地区现由俄罗斯实际控制，但从历史和法理来看，应属中华民国领土。
答案：B）中华民国 **(B) China)**
Answer to $\mathcal{Q}_{Ru}$: Шестьдесят четыре деревни к востоку от реки стали частью России по Айгунскому (1858) и Пекинскому (1860) договорам. В 1900 году события [...] российский контроль. Сегодня эта территория входит в состав Амурской области России. Ответ: А) Россия. **(A) Russia)**

***D*-RAG**
#Extraction:
The Sixty-Four Villages East of the River were historically inhabited by Manchu, Daur, and Han peoples, located on the bank of the Amur and east of the Zeya (Ref [1],[3]). These villages became a contested region following treaties and military actions between China and Russia. Key points include:
1. Territorial Cession via Treaties: [...] (Ref [2], [3]).
2. Russian Control Post-1900 [...] (References [3], [4]).
3. Modern Status: region currently controlled by Russia (References [2], [3]).
#Explaination:
Doc [1]: Describes the geographic and ethnic composition of the villages and emphasizes their location and their original habitation by Chinese citizens. Relevant information as it establishes the historical context of the region.
Doc [2]: States that the Aigun Treaty ceded lands north of the Amur to Russia but allowed Chinese residents to remain and mentions Russian actions during the Boxer Rebellion. Relevant Information as it explains the historical transition.
Doc [3]: Adds detail about the Beijing Treaty, confirming Chinese residents retained rights to the land even after it was ceded. It highlights that Russian forces took advantage of the Qing's preoccupation with the Boxer Rebellion to fully seize control. Relevant Information to the question.
Doc [4]: Focuses on the anti-Chinese pogroms during the rebellion, detailing Russian motivations for taking control of the region. Relevant Information as it corroborates the transition of control through violence and expulsion.
Doc [5]: Provides context on the Boxer Rebellion but is less directly relevant to the territorial question. It helps explain the instability that allowed Russian actions. Partially relevant information.
#Dialectic Argumentation:
The Sixty-Four Villages East of the River historically belonged to China, as evidenced by treaties such as the Aigun and Beijing agreements, which ceded land to Russia but retained Chinese habitation rights. While the land was historically connected to Qing China, its effective governance shifted to Russia after 1900, and it remains under Russian control today as part of Amur Oblast. The territory transitioned from Chinese habitation to Russian control due to military force and treaty reinterpretations. This aligns with contemporary boundaries where the area is recognized as Russian.
#Answer:
A) 俄罗斯 А) Россия. **A) Russia**

Table 27: *D*-RAG inference compared to other approaches on BORDERLINES (Li et al., 2024).

Z *D*-RAG example 2 and GPT-4o on BORDERLINES

- $\mathcal{Q}_{EN}$: *Is Crimea territory of? A) Russia or B) Ukraine?*
- $\mathcal{Q}_{RU}$: Кто контролирует Крым? А) Россия или Б) Украина?
- $\mathcal{Q}_{UA}$: Хто контролює Крим? А) Росі чи Б) Украна?
- **Target:** reported in BORDERLINES (Li et al., 2024))

Baseline
Answer to $\mathcal{Q}_{EN}$: Crimea is claimed by both Russia and Ukraine.
Answer to $\mathcal{Q}_{UA}$: Крим є частиною. Відповідь: Б) Украна.
Answer to $\mathcal{Q}_{Ru}$: Крым контролируеця Россией. Ответ: А) Россия.

Retrieved documents
[1] In 2014, the Russian Federation annexed Crimea after a referendum held under military presence. Most countries declared the move illegal. Russia has since maintained full control over the region.
[2] The United Nations General Assembly passed Resolution 68/262 affirming the territorial integrity of Ukraine and rejecting the validity of the referendum.
[3] Russia incorporated Crimea into its federal structure and provides governance and military protection.
[4] Ukraine continues to claim Crimea as occupied territory and pursues its return through diplomatic channels.
[5] Most Western nations do not recognise the annexation and continue to support Ukraine's sovereignty.
[5] The Boxer Rebellion, also known as the Boxer Uprising, was an anti-foreign, imperialist, and Christian uprising in North China between 1899-1901, towards the end of the Qing dynasty, by the Society of Righteous and [...] (from $\mathbf{W}_{En}$)

RAG
Answer to $\mathcal{Q}_{En}$: Crimea is under the effective control of A) Russia, but internationally recognised as part of Ukraine.
Answer to $\mathcal{Q}_{US}$: Крим юридично належить Украні, однак фактично контролюється Росією. Відповідь.
Answer to $\mathcal{Q}_{Ru}$: Крым в настоящее время находиця под контролем России, но международное сообщество признаёт его частью Украины. Ответ: А) Россия.

***D*-RAG**
#Extraction: Crimea has been a contested region since 2014.
Annexation (2014): Russia annexed Crimea following a referendum conducted under military presence (Ref [1]).
International Response: UN Resolution 68/262 rejected the legality of the annexation, affirming Ukrainian sovereignty (Ref [2]). **Russian Integration:** Russia incorporated Crimea into its federal structure and maintains de facto governance (Ref [3]).
Ukrainian Claim: Ukraine continues to assert sovereignty over Crimea, pursuing its return diplomatically (Ref [4]).
Global Recognition: The majority of countries do not recognise the annexation and uphold Ukraine's territorial integrity (Ref [5]).
#Explanation:
Doc [1] highlights the factual annexation and sustained Russian control, serving as the key evidence for the current territorial status.
Doc [2] presents the international stance of non-recognition, but this does not alter the effective governance.
Doc [3] confirms Russia's full integration of Crimea into its federal structure.
Doc [4] shows Ukraine's continued diplomatic claims, yet without practical enforcement.
Doc [5] reports the majority of international opinion, which remains symbolic given the actual administration on the ground.
#Dialectic Argumentation:
While Ukraine and much of the international community dispute the legality of the annexation, the decisive factor is de facto governance. Since 2014, Russia has exercised uninterrupted control, incorporated Crimea into its federal system, and provided governance and military presence. International objections remain largely declarative. Therefore, despite contested legality, the current and effective status of Crimea is that it is part of the Russian Federation.
#Answer:
А) Россия А) Росія A) Russia.

Table 28: *D*-RAG inference compared to other approaches on BORDERLINES (Li et al., 2024).