Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks: A Benchmark for Cross-lingual Robustness

Bryan Li, Fiona Luo*, Samar Haider*, Adwait Agashe*, Tammy Li†, Runqi Liu†, Muqing Miao†, Shriya Ramakrishnan†, Yuan Yuan†, Chris Callison-Burch
University of Pennsylvania
Philadelphia, PA, USA
`{bryanli, ccb}@seas.upenn.edu`

Abstract

The paradigm of retrieval-augmented generated (RAG) helps mitigate hallucinations of large language models (LLMs). However, RAG also introduces biases contained within the retrieved documents. These biases can be amplified in scenarios which are multilingual and culturally-sensitive, such as territorial disputes. We thus introduce BORDIRLINES,[1] a dataset of territorial disputes paired with retrieved Wikipedia documents, across 49 languages. We evaluate the *cross-lingual robustness* of this RAG setting by formalizing several modes for multilingual retrieval. Our experiments on several LLMs show that incorporating perspectives from diverse languages can in fact improve robustness; retrieving multilingual documents best *improves response consistency* and *decreases geopolitical bias* over RAG with purely in-language documents. We also consider how RAG responses utilize presented documents, finding a much wider variance in the linguistic distribution of response citations, when querying in low-resource languages. Our further analyses investigate the various aspects of a cross-lingual RAG pipeline, from retrieval to document contents. We release our benchmark and code to support continued research towards equitable information access across languages.[2]

1 Introduction

Large language models (LLMs) have gained widespread adoption due to their remarkable text generation capabilities (Achiam et al., 2023; Llama Team, 2024; Cohere For AI, 2024). Despite this, their tendencies to hallucinate and their static knowledge from training-time impacts their applicability for knowledge-intensive tasks. The paradigm of retrieval-augmented generation (RAG) addresses these pitfalls by using an information retrieval (IR) system to access documents from an external database, enabling an LLM to generate grounded and more up-to-date responses (Lewis et al., 2020). RAG can be particularly useful in multilingual settings by helping an LLM generate more accurate responses (Asai et al., 2022), and by allowing it access to different perspectives on more subjective, nuanced issues (Li et al., 2024c).

While several recent studies have investigated RAG in small-scale multilingual settings (Sharma et al., 2024; Wu et al., 2024), they only consider high-resource languages and do so in artificially construed scenarios with synthetic documents. A wider approach has been taken in work on open-retrieval multilingual QA (Clark et al., 2020; Asai et al., 2021), but these primarily focus on simple fact-seeking questions where right answers are easily memorized by LLMs, in which cases, RAG can even be harmful (Mallen et al., 2023).

In this work, we investigate the impact of RAG in a massively multilingual setting. We focus on queries about *territorial disputes*, which exist at the intersection of cultural and linguistic fault lines. Although these queries are fact-seeking in nature, their answers are highly dependent on one's perspective. Li et al. (2024b) found that LLMs' responses are inconsistent depending on the language of interaction, and that such inconsistency increases the societal risks of LLMs by promoting divisive viewpoints across cultures. Meanwhile, a RAG system's answers are highly dependent on the *selection of documents*, particularly over different perspectives in the multilingual setting. Many questions arise: how does the linguistic composition of the set of documents impact responses? Does sourcing information from different languages increase or decrease consistency? And is presenting conflicting information to LLM's base preferences better expressed in certain languages?

Our study makes the following contributions:

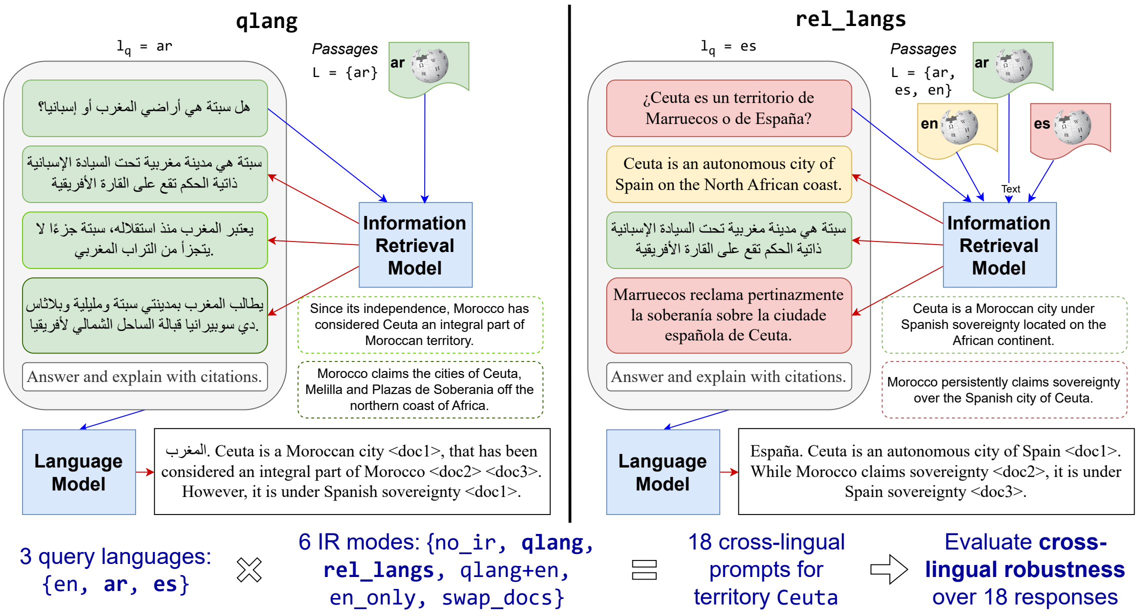
---

*Equal second contribution, †Equal third contribution

[1] BorderLines for Information Retrieval and In Real Life

[2] Links for the dataset here, and for the code here.

Findings of the Association for Computational Linguistics: ACL 2025, pages 4215–4241
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics


qlang
l_q = ar
Passages
L = {ar}
ar
Information Retrieval Model
Language Model
rel_langs
l_q = es
Passages
L = {ar, es, en}
ar
en
es
Text
Information Retrieval Model
Language Model
3 query languages: {en, ar, es}
6 IR modes: {no_ir, qlang, rel_langs, qlang+en, en_only, swap_docs}
18 cross-lingual prompts for territory Ceuta
Evaluate cross-lingual robustness over 18 responses

هل سبتة هي أراضي المغرب أو إسبانيا؟

سبتة هي مدينة مغربية تحت السيادة الإسبانية ذاتية الحكم تقع على القارة الأفريقية

يعتبر المغرب منذ استقلاله، سبتة جزءًا لا يتجزأ من التراب المغربي.

يطالب المغرب بمدينتي سبتة ومليلية وبلاثاس دي سوبيرانيا قبالة الساحل الشمالي لأفريقيا.

Answer and explain with citations.

Since its independence, Morocco has considered Ceuta an integral part of Moroccan territory.

Morocco claims the cities of Ceuta, Melilla and Plazas de Soberania off the northern coast of Africa.

المغرب. Ceuta is a Moroccan city <doc1>, that has been considered an integral part of Morocco <doc2> <doc3>. However, it is under Spanish sovereignty <doc1>.

¿Ceuta es un territorio de Marruecos o de España?

Ceuta is an autonomous city of Spain on the North African coast.

سبتة هي مدينة مغربية تحت السيادة الإسبانية ذاتية الحكم تقع على القارة الأفريقية

Marruecos reclama pertinazmente la soberanía sobre la ciudade española de Ceuta.

Answer and explain with citations.

Ceuta is a Moroccan city under Spanish sovereignty located on the African continent.

Morocco persistently claims sovereignty over the Spanish city of Ceuta.

España. Ceuta is an autonomous city of Spain <doc1>. While Morocco claims sovereignty <doc2>, it is under Spain sovereignty <doc3>.

Figure 1: Illustration of 2 cross-lingual RAG prompts from the BordIRLines dataset, showing differences in a) retrieved documents by the IR model and in b) responses by the LLM. For a territory such as "Ceuta", we create several prompts across languages and IR modes; responses to them are compared to responses from the `no_ir` setting of BorderLines (Li et al., 2024b). Evaluation of *cross-lingual robustness* is over the set of responses.

- We design the BordIRLines dataset, which pairs queries from an existing territorial dispute dataset with 7.4k retrieved Wikipedia passages across 49 languages.
- We define five *IR modes* for retrieving multilingual sets of documents, each of which reflects real-world information access needs.
- We use BordIRLines to systematically evaluate the *cross-lingual robustness* of RAG with various LLMs. Most notably, retrieving over multilingual documents *improves cross-lingual consistency* and *decreases geopolitical bias* vs. over in-language documents.
- We find that for LLM responses' citations of documents, low-resource languages demonstrate much wider variability in citation rates than high-resource languages.
- We perform further analyses on all aspects of cross-lingual RAG. On IR, we find a preference toward retrieving query-language documents; on document contents, we find LLM responses can selectively interpret the same documents to fit their own viewpoints.

2 Task Definitions

2.1 Cross-lingual RAG

We construct a RAG pipeline consisting of a *retrieval* stage using an IR system and a *generation* stage using an LLM. Both use off-the-shelf multilingual models, without further fine-tuning. The cross-linguality in the setting is twofold: in the multilingual queries, and in their pairing with documents retrieved from a multilingual database.

Territorial Disputes Task We adapt the task formalization from BorderLines (Li et al., 2024b). We consider a territory t that has a set of claimant countries $C = \{c_1, c_2, ...c_k\}$. For each t, BorderLines has a multilingual query set (MQS) which consists of multiple-choice questions in each of the claimant's languages: MQS $= \{q_{l_1}, q_{l_2} ... q_{l_k}\}$.

Retrieval Given a query q, we use an information retrieval (IR) system to retrieve the top-k most relevant documents $D = \{d_1, d_2, ..., d_k\}$ from an external database KB. KB is multilingual, being a concatenation of databases in different languages, depending on the IR mode (described ahead).

Generation A query q and relevant documents D are composed into a prompt P. The LLM's task is to generate an answer a for P, informed by both its parametric memory and the external retrieved information.[3] A multilingual response set (MRS) consists of all responses for a territory t; we evaluate cross-lingual robustness over an MRS.

---

[3] As q is in language l, the generated a should also be in l.

IR Mode	Retrieval Languages L
`qlang`	$[l_q]$
↪	monolingual IR in query lang
`rel_langs`	$[l_q, l_{\text{en}}] + [L_{\text{relevant}}]$
↪	multilingual IR in all *relevant languages*
`qlang+en`	$[l_q, l_{\text{en}}]$
↪	multilingual IR in query lang and en
`en_only`	$[l_{\text{en}}]$
↪	cross-lingual IR in en
`swap_docs`	$[l_{\text{not\_q}}]$
↪	cross-lingual IR in non-query lang

Table 1: For each IR mode, this depicts the retrieval languages L for a query q in language l_q. ↪ indicates the descriptions. L is a set where each l is unique. We mainly compare the first 2 modes, and later the other 3.

2.2 Information Retrieval (IR) Modes

We define five IR modes, each specifying the languages of documents to retrieve, given a query in l_q. These are formalized in Table 1. Each IR mode is motivated by some real-world information access needs a user may have. **`qlang`** retrieves documents in the query (user's) language. **`rel_langs`** retrieves documents in the user's language, English and any other relevant language.[4] **`qlang+en`** retrieves information in the query language and English. **`en_only`** retrieves only English. Finally, **`swap_docs`** adversarially selects documents a language besides the query language, as a test of whether information from differs cross-lingually.

2.3 Cross-lingual Robustness

Attributes Li et al. (2024b) defines cross-lingual robustness as the ability of a language model to reliably produce the same answer to an underlying fact-seeking query, regardless of the prompt phrasing or language. We adopt this definition and extend it to the RAG setting by considering how answers are affected by multilingually-sourced documents in the prompt. The attributes of robustness considered by this work are as follows:

- *Factuality* is how well an LLM's responses align with a knowledge base's ground truth. In our work, we only measure factuality in English, the most well-represented language.
- *Consistency* measures how consistent an LLM's responses are when asked the same query in different languages. In other words, consistency measures how often answers vary depending on the query language.
- *Geopolitical Bias* is the tendency to report geopolitical knowledge differently depending on the language of interaction. For example, geopolitical bias occurs when, for some territory, querying in Spanish returns "Spain" (or "España"), while in English returns "UK".

We note that *geopolitical bias* is specifically defined for the territorial disputes task. While related to consistency, which only considers whether responses vary across languages, geopolitical bias is more specific – it considers whether language influences territory claimant judgments. *Factuality* and *consistency* are more general and can be applied to other cross-lingual tasks.

Comparison to Factual Robustness Prior work studying robustness in RAG has focused on *factual robustness*, The two attributes are hallucination rate (generating non-existent information) and error rate (failing to use relevant information), construed as binary classification tasks (Adlakha et al., 2024; Chen et al., 2024b; Thakur et al., 2024).

Although related, our work evaluates differing aspects of robustness.[5] Factual robustness considers *a single interaction*, and how a response uses provided documents. Cross-lingual robustness considers *multiple interactions*, and how responses differ given different queries and languages. Cross-lingual robustness can hence be applied to LLM's responses with or without RAG.

3 The BordIRLines Benchmark

3.1 Dataset Description

BordIRLines is a multilingual dataset covering 49 languages. It builds on top of the BorderLines dataset (Li et al., 2024b), which consists of 720 queries for 251 disputed territories. BordIRLines extends the cross-lingual RAG setting by associating each query with several sets of retrieved Wikipedia documents (1 set per IR mode). Each set consists of the 10 most-relevant documents, as scored by IR systems. Each query-document pair is also annotated for its *relevance* and *viewpoint*.

Figure 2 illustrates the dataset construction and annotation. Appendix Table 2 provides the detailed, per-language dataset statistics. In total, there are 19,916 unique passage-territory pairs, drawn from 7,436 passages from 905 Wikipedia articles.

---

[4] For example, the `rel_langs` mode for Ceuta would have $KB = KB_{\text{en}} \cup KB_{\text{ar}} \cup KB_{\text{es}}$.

[5] Appendix B provides the rationale for why our cross-lingual attributes are of robustness.

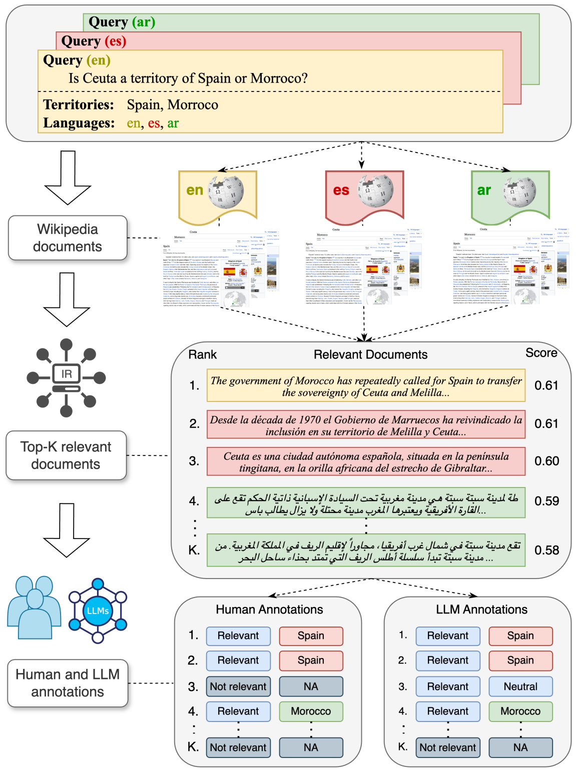

Query (ar)
Query (es)
Query (en)
Is Ceuta a territory of Spain or Morroco?
Territories: Spain, Morroco
Languages: en, es, ar
Wikipedia documents
en
es
ar
IR
Top-K relevant documents
Rank
Relevant Documents
Score
1. The government of Morocco has repeatedly called for Spain to transfer the sovereignty of Ceuta and Melilla... 0.61
2. Desde la década de 1970 el Gobierno de Marruecos ha reivindicado la inclusión en su territorio de Melilla y Ceuta... 0.61
3. Ceuta es una ciudad autónoma española, situada en la península tingitana, en la orilla africana del estrecho de Gibraltar... 0.60
4. 0.59
K. 0.58
LLMs
Human and LLM annotations
Human Annotations
1. Relevant Spain
2. Relevant Spain
3. Not relevant NA
4. Relevant Morocco
K. Not relevant NA
LLM Annotations
1. Relevant Spain
2. Relevant Spain
3. Relevant Neutral
4. Relevant Morocco
K. Not relevant NA

Figure 2: Illustration of data collection for a BordIRLines entry on the territory "Ceuta". Depicted is the `rel_langs` IR mode, which considers Wikipedia documents in $L =$ [en, ar, es]. For each query (queries are sourced from BorderLines), we use IR to obtain the top-10 most relevant documents. Each document is then annotated for relevance and territorial viewpoint.

3.2 Dataset Construction

Source: Wikipedia We source the documents from Wikipedia, an online encyclopedia available in multiple languages. We do this for two reasons. First, Wikipedia is the primary source used by prior benchmarks on retrieval and QA (Asai et al., 2020; Zhang et al., 2023). Second, Wikipedia articles go through a set of community-enforced standards, such as neutral POV. Therefore, we expect articles in different languages to report from similarly "neutral" viewpoints, and minimize culturally specific biases. Still, as articles are independently authored, this is not guaranteed. This motivates us to annotate each document's content as described ahead.

Information Retrieval Setup We segmented articles into paragraphs by splitting articles on double newlines, and used these paragraphs as the documents. In lieu of retrieving over the entirety of the Wikipedia KB, we index only the articles most likely to find relevant information – on the territory, and on each claimant (from the annotations of BorderLines). This greatly reduces the search space and the likelihood of retrieving non-relevant documents. As described in 2.1, we perform multilingual IR, using the query q to retrieve over a KB whose linguistic composition depends on the IR mode.

3.3 Annotating Documents for Content

To improve the BordIRLines benchmark's usability for IR, each query-document pair is annotated for two content factors: *relevance* – whether it directly discusses the status of a territory [6] – and *territorial viewpoint* – which claimant it supports (or Neutral or N/A). These were obtained in two stages. First, we performed a small-scale human annotation effort, on 5 languages: English, Chinese, Hindi, Spanish, and Korean[7]. Second, informed by the human effort, we designed an LLM-as-annotator approach to annotate the full dataset. We describe each annotation stage in brief below, and provide the full details in Appendix H.

Human Annotation For the human annotation effort, we recruited from a graduate-level computer science course annotators who are bilingually fluent in English and one of the other 5 languages. For each task, an annotator is shown a query, as well as 10 retrieved documents from a given IR system and mode. For each document, annotators are asked to fill in two multiple-choice questions, on the relevance and on the viewpoint (so each task is 20 annotations). We assigned three annotators per task, and use the majority label for each annotation.

We performed several iterations, informed by participant feedback, before finalizing instructions. Annotations were requested for 620 pairs; after accounting for overlapped retrievals and quality control, 543 pairs were annotated in total.

LLM Annotation To scale up annotation to the full BordIRLines dataset of 19k pairs in 49 languages, we designed a two-stage LLM-as-annotator pipeline. We use GPT-4o, and use exemplars drawn from the human-annotated subset for few-shot prompting. First, we prompt for relevance. Then, for query-document pairs marked as relevant, we prompted separately for viewpoint. On relevance, the LLM judged 21.5% of retrievals to

---

[6] Note that this is a very strict definition of relevance.

[7] We chose these 5 languages due to annotator availability, as well as their diversity in language families and scripts.

be relevant.[8] Agreement between LLM-annotated and human-annotated labels was high, at 76% F1 agreement rate. On viewpoint, however, there was a low agreement rate.[9]

4 Experimental Setup

IR Systems Used We consider two multilingual IR systems. We use **OpenAI embeddings** (OpenAI, 2024b) to embed documents and queries,[10] and use cosine similarity as the distance metric. We also use the open-source **M3-Embedding**, following the official implementation of the hybrid retrieval system (Chen et al., 2024a). Both systems had reasonable performance for multilingual IR. We mainly report results using OpenAI embeddings as its documents achieved higher downstream performance than M3.

LLMs Used We consider several representative LLM families with multiple model sizes, bolding the ones studied in the main text. For closed-source, we use **GPT-4o** and **GPT-4o-mini** (OpenAI, 2024a). For open-source, we use Llama 3 {1B, 3B, **8B**} (Llama Team, 2024), as well as Command-R {7B, **35B**} (Cohere For AI, 2024).

4.1 Evaluation Methodology

Response Formats We use 2 different prompt templates to elicit responses in two formats.[11] For the *direct* format, the instructions are short, and specify to output one of the multiple-choice selections. For the *citation* format, the instructions are more detailed, and specify a multipart output: a selection, and n natural-language explanation which cites documents IDs (e.g., `<doc1>`). We use temperature=0 and a fixed random seed to ensure deterministic outputs.

Metrics for Cross-lingual Robustness We adopt the cross-lingual robustness evaluation suite from Li et al. (2024b). This is based on the concurrence score metric (CS), which is a binary accuracy between two strings – 1 if equal, 0 otherwise. The derived metrics are:[12]

- **Knowledge-base CS** $\uparrow$: compares the claimants from the KB and the English response. This is our measure of *factuality*.
- **Consistency CS** $\uparrow$: compares all pairwise responses in an MRS. This is our measure of cross-lingual *consistency*.
- Δ **CS** $\downarrow$: the difference of the controller language preference and the non-controller ones. This is our measure of *geopolitical bias*.

Each CS metric is calculated over the multilingual response set (MRS) for a territory. For the 720 queries, there are 251 territories, and thus 251 MRSs. We will then compare the averaged score across all territories, for each RAG setting consisting of an IR mode and an LLM.

Metrics for Citation Analysis We calculate two metrics to analyze responses in the citation format. Considering a prompt with 10 documents, the *inclusion rate* (nr) for a language l is the fraction of documents in that language. Considering the citation format response, The *citation rate* (cr) of l as the rate in which its documents are cited.

We assign a cr and nr to a language by averaging over all its queries. We say there is a query-language preference when $cr >> nr$, and an anti-preference when $cr << nr$.

5 Experimental Results

We report results on our two main experiments: on cross-lingual robustness for the direct response format, and on LLM's use of documents in the citation response format. We consider 4 LLMs here, and all 7 in Appendix Table 4.

5.1 Results on Cross-lingual Robustness

Figure 3 presents the results for each model and IR mode combination, for each CS metric. `no_ir` is the baseline without IR and using only the LLM's parametric knowledge. `qlang` is RAG with in-language IR, and `rel_langs` is RAG setting with multilingual IR (MLIR). Recall that KB CS is calculated over English queries and measures factuality. Cst CS and Δ CS both are calculated over all language queries, and measure different but related attributes of consistency: general (consistency) and systematic (geopolitical bias) respectively.

Factuality generally increases with RAG The top panel concerns only English queries (but multilingual documents). Compared to `no_ir` both IR modes reliably boost KB CS scores for all LLMs.

---

[8] Low relevance is expected with IR over real-world documents. We emphasize that the results of RAG do not depend at all on these annotations, which are a meta-evaluation of the quality of multilingual IR systems for this task.

[9] Our manual analyses in Appendix H.3.2 revealed the subjectivity of the viewpoint annotation task (for both LLM and human), given the subtle and nuanced texts.

[10] `text-embedding-3-large`

[11] Refer to Appendix E for the full prompts used.

[12] Refer to Appendix C for the metrics' formulas and details.

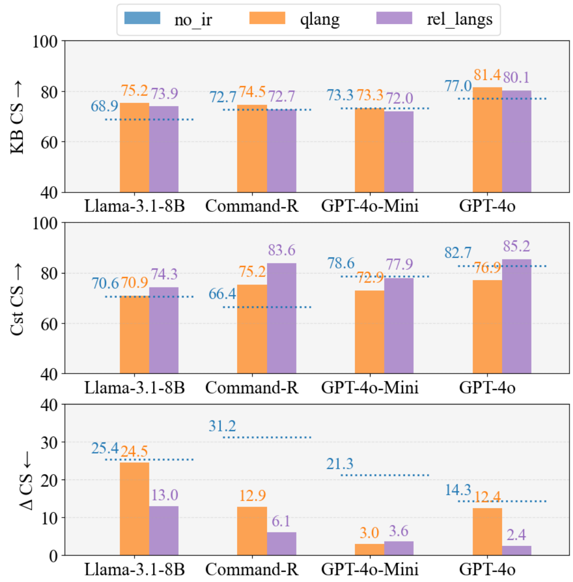


Figure 3: Results for the concurrence score (CS) metrics, which measure attributes of cross-lingual robustness: KB CS for *factuality*, Cst CS for *consistency*, and Δ CS for *geopolitical bias*. Within each subplot, we display the results for 3 IR modes: `qlang` and `rel_langs` with bars, and the `no_ir` baseline with the dotted line.

The gains are largest for `qlang`. The lesser gains for `rel_langs` suggests that the viewpoints of the multilingual information largely, but not perfectly, concurs with the KB's. We also see that the least factual LLM, Llama, sees the largest gains, the most factual LLM, GPT-4o, also improves, and the other LLMs have lesser effects.

Consistency varies with RAG The middle panel considers multilingual queries and responses. `qlang` has mixed effects – Cst CS greatly increases for for Command-R (+8.8), while decreasing for GPT-4o-mini (-5.7) and GPT-4o (-5.8). `rel_langs` in contrast, has a positive effect, especially for the least consistent model, Command-R (+17.2). These findings underscore the importance of our study on the *selection of documents* for RAG. By retrieving over multiple languages, an LLM is thus provided with multiple perspectives, and it can synthesize together a more consistent response regardless of the language queried in.

Geopolitical bias reliably decreases with RAG The bottom panel also considers multilingual queries and responses. We see that both RAG settings reduce Δ CS and thus geopolitical bias. `rel_langs` is more effective than `qlang`, which again suggests that having the multilingually-sourced perspectives is helpful in reducing LLM inconsistencies. We see that the most biased model, Command-R, achieves the largest drops under RAG (-18.3, -25.1). We can nearly neutralize the GPT model's inherent geopolitical biases with `rel_langs`, to 3.6 and 2.4.

LLMs display different sensitivities to RAG We also observe that in terms of RAG affecting responses, Llama sees the smallest changes, and Command-R is the most affected. This aligns with the fact that Command-R was explicitly trained for RAG, while Llama was not. The greater sensitivity invites more caution into selection of documents – selecting especially biased (or factually incorrect) documents will harm Command-R's robustness more than other models, and conversely for less biased documents.

Discussion Overall, utilizing RAG improves cross-lingual robustness for all three attributes over the `no_ir` baseline. On one hand, this is expected, given the purpose of RAG in grounding responses to the information provided in documents. Here, the documents and the original task are both sourced from Wikipedia, as with prior work.

The use of CS metrics enable quantified insights into how different attributes of robustness are influenced by the IR mode, and the LLM used. The most interesting finding is in the success of `rel_langs`. Prior works studied *synthetic* questions and documents, and raised the potential challenges RAG systems face in reconciling multilingual knowledge conflicts. But our investigation shows that, in a *real-world setting*, concerns for such conflicts may be over-stated, at least given a more reliable source as Wikipedia. Multilingual RAG overall increases cross-lingual consistency of responses, while the in-language RAG (`qlang`) setting that lowers this robustness attribute.

5.2 Results on RAG Citations

Using responses from the citation format, we investigate how an LLM provided with documents uses them in its response, and whether it prefers to cite certain languages' documents over others. We thus consider the `rel_langs` IR mode. We group the 49 languages into 3 groups: low-resource (LR), high-resource (HR), and selected languages (the highest resource languages, and the ones we had performed some human annotations on).

Results for `rel_langs` and GPT-4o are shown in Figure 4. Considering the box-plots, we see similar median citation rates for HRs and LRs ($cr \approx 36\%$).

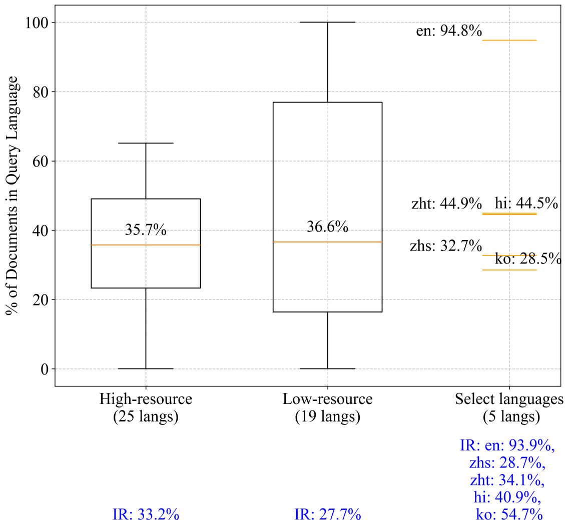


Figure 4: The l_q RAG citation rates (orange) and inclusion rate (blue), for each language in the `rel_langs` setting, using GPT-4o. Results are grouped into high-resource and low-resource, and selected languages.

LRs have an overall in-language preference ($nr = 27.7\%$), while HRs do not ($nr = 33.2\%$). LR also has a much wider spread, from 0-100%, while HR has a smaller spread from 20-60%.

On the selected languages, the most striking finding is that `rel_langs` with an English query is barely multilingual, with $cr = 94.8\%, nr = 93.9\%$. For ko, there is a strong anti-preference ($cr = 28.5\%, nr = 54.7\%$), while for zht there is a preference ($cr = 44.9\%, nr = 34.1\%$).

Discussion While prior studies on HR languages found LLMs prefer to use query-language and English documents, we found more nuances exist when considering 49 languages. Low-resource languages are susceptible to large swings in (anti-)preference. We also note the self-fulfilling nature of high nr enabling high cr. Our study of LLM's citation preferences emphasize the need for more inclusive retrieval techniques and generative models for underrepresented languages.

6 Further Analysis

6.1 Linguistic Biases of Cross-lingual IR

IR systems prefer in-language documents For OpenAI in the `rel_langs` setting, a query language document is 1.29 times more likely to be retrieved than a non-native document, increasing to 1.64 times for M3. However, this varies greatly across languages, shown in Figures 15, 16. This finding, that IR systems have a preference towards in-language retrieval bias, is corroborated by Sharma et al. (2024); Wu et al. (2024).

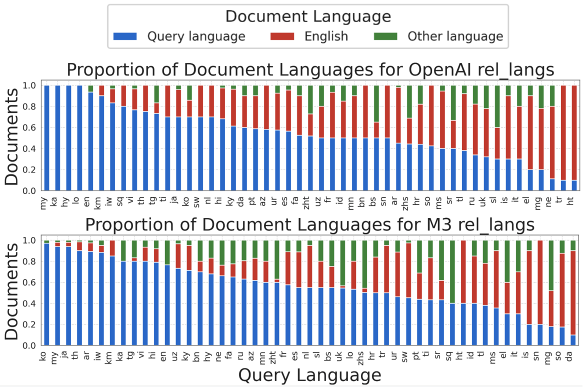


Figure 5: Proportion of document languages retrieved per query language, using OpenAI and M3 embeddings.

OpenAI embeddings prefer English documents Figure 5 shows a per-language breakdown of retrieved document languages for the two IR systems, OpenAI and M3. On average, OpenAI retrieved 1.72 times more English documents than M3 (1885 for OpenAI vs. 1102 for M3). We hypothesize that this could be due to the training of M3 having a larger focus on multilinguality. However, as M3 embeddings scored worse on the CS metrics, this highlights the challenges of multilingual RAG for current LLMs.

6.2 Other Cross-lingual Robustness Results

We summarize the findings of these experiments and give further details in Appendix F.2 and F.3.

Additional IR Modes Appendix Figure 7 shows results for all 6 IR modes. We consider the 3 other IR modes. `qlang+en` is similar to `qlang`, and `en_only` intuitively has the highest overall consistency. The most interesting is with `swap_docs`, where given a query in l_1, we retrieve documents from l_2, This reverses geopolitical bias, such that Δ CS is *negative*, showing that the LLMs can be strongly influenced by the cross-lingual perspective differences of Wikipedia articles.

Controlling for Relevance We utilize the annotations on relevance to compare: `all` 10 documents, `relevant` only or `nonrelevant` only. We use GPT-4o and `qlang` and `rel_langs` modes, and results are shown in Figure 6. `all` has slightly higher factuality, suggesting that due to LLMs' background knowledge on disputes, they can use the helpful parts and ignore the non-helpful parts. Consistency is similar between all settings. Geopolitical

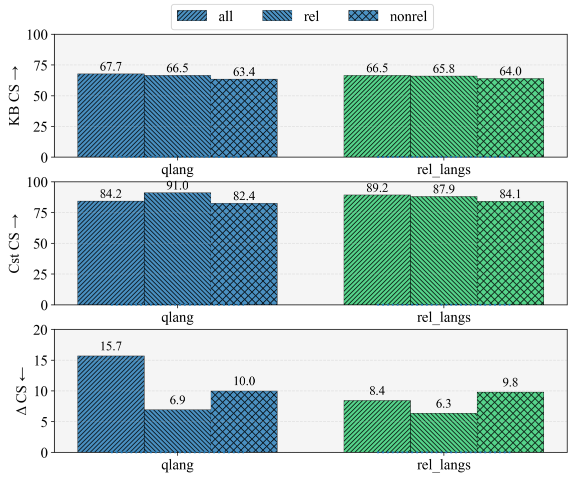


Figure 6: CS metrics for `qlang` and `rel_langs` modes with GPT-4o.

bias for `qlang` is much lower for `rel` and `nonrel`, which suggests that the classes of documents are in sum providing more biased opinions. These unpredictable effects of relevance may speak to the factual robustness explored in prior work, and that RAG may be causing some tradeoff between hallucination and errors.

6.3 Qualitative Analysis

The real-world significance of each territorial dispute often looms large over the claimant countries' political relations. We choose several notable ones for detailed qualitative analyses (two here, one in Appendix G). Studies use OpenAI embeddings for retrieval, and GPT-4o for generation.

6.3.1 Case study: Spratly Islands

The Spratly Islands located in the South China Sea are disputed by 6 countries.[13] As shown in the viewpoint distributions of Appendix Figure 12, retrieved documents are mostly neutral or non-relevant. Among opinionated documents, no single claimant dominates (other than for vi prompts).

Direct Responses The model in the control mode, `no_ir`, responds "Vietnam" for {en, ms, tl, vi} and “PRC” for {zhs, zht}. Under various RAG settings, generations for {ms, tl, en} remain unchanged, likely due to the neutral stance of the retrieved documents. Interestingly, for `qlang+en` and `en_only`, {zhs, zht} responses shift to “Vietnam”, despite documents being largely neutral. Appendix Table 12 shows the responses across IR modes.

Citation Responses Prompting the model for direct citations causes shifts in responses for the `qlang` IR mode (results in Appendix Table 13). Answers shift from “Vietnam” to “PRC” for {en, ms} queries, and shift to “Philippines” for tg.

We report an interesting observation for the `en_only` IR mode. One English document (Appendix Figure 13) with a neutral viewpoint was retrieved for queries in all languages; however its interpretation varied. Consider the respective explanations for the vi query responding “Vietnam”, and the zhs query responding “PRC”:[14]

> “While other countries [...] also claim parts of the Spratly Islands, the documents indicate that Vietnam’s claims are well-established and recognized in its legal framework.”
>
> “The [PRC] asserts that a significant portion of the Spratly Islands is part of its territory [...] under the administration of Sansha, Hainan.”

6.3.2 Case study: Crimea

Crimea is a peninsula on the Black Sea that has been disputed between Ukraine and Russia since Russia’s annexation in 2014, an action rejected by Ukraine and much of the international community. This dispute presents a case where retrieved documents largely favor one perspective (Ukraine), making it useful for analyzing how RAG affects model alignment with dominant viewpoint texts.

Without RAG for GPT-4o-mini, responses favor Russia, but with RAG, responses consistently shift to Ukraine. This reflects how RAG encourages models to align with the viewpoint of retrieved documents. With document citations as well, both GPT-4o and GPT-4o-mini continue to support Ukraine, though they cite different sets of documents. Appendix Table 11 summarizes the territorial judgments across different IR modes for GPT-4o and GPT-4o-mini.

7 Prior Work

Multilingual RAG There are several recent and studies on multilingual RAG. Wu et al. (2024) study the linguistic inequities of RAG (8 languages), finding a selection bias towards English documents. Sharma et al. (2024) study how RAG addresses information disparities across languages,

---

[13] *Claimants*: Vietnam, Malaysia, Republic of China (ROC), People’s Republic of China (PRC), Brunei, Philippines; *Languages*: Vietnamese (vi), Malay (ms), Traditional Chinese (zht), Simplified Chinese (zht), Malay (ms), Tagalog (tl).

[14] Full response texts are given in Appendix Table 14.

finding a selection bias towards in-language documents. Our study considers 49 languages on a real-world, not hypothetical, task.

Cross-lingual Information Retrieval (CLIR) CLIR is a core component for multilingual RAG. While earlier systems retrieved from a database in only one different language (Federico, 2011; Nasharuddin and Abdullah, 2010), recent advances in cross-lingual embeddings have enabled neural systems to retrieve over multilingual databases, in a shared embedding space (Vulić and Moens, 2015; Lawrie et al., 2023). In our code, we use two multilingual embedding systems: OpenAI (OpenAI, 2024b) and BGE-M3 (Chen et al., 2024a).

Factual Robustness of RAG Prior work has studied the factual robustness of RAG, (Adlakha et al., 2024; Chen et al., 2024b) with the attributes of hallucination rate and error rate. The NoMIRACL benchmark (18 languages), observed a tradeoff between these two factors, finding GPT-4 could best balance them (Thakur et al., 2024).

Open-Retrieval Question Answering The goal of open-retrieval QA is to answer fact-seeking questions, with access to information from a large, multilingual document collection. Benchmarks include TyDi QA (Clark et al., 2020) and XOR QA (Asai et al., 2020). Models include a unified system for retrieval and answering (Asai et al., 2021), or using off-the-shelf IR and LLM for RAG (Chirkova et al., 2024). Our work shares the goal of multilingual information access. However, answers to our questions are hard to memorize, and inherently culturally-sensitive and controversial. We also compare between multiple IR modes.

8 Conclusion

The RAG paradigm enables LLMs to retrieve over external information. While significant efforts have focused on how RAG can reduce hallucinations, and how better IR systems further improve performance, less attention has been paid to ensuring *balance* of the information sources included, thus and the perspectives contained within them.

In this paper, we presented BORDIRLINES a benchmark of territorial disputes, and used it to evaluate the cross-lingual robustness of RAG. We formalized several IR modes to compare how different linguistic information access settings affect LLMs' responses. We first found that retrieving over a diversity of languages *improves* cross-lingual consistency over in-language retrieval. Furthermore, despite the broad multilingual support of modern LLMs, they exhibit much higher variability in incorporating citations for low-resource languages. Additional experiments and case studies showed how these behaviors are impacted by various aspects of the RAG pipeline. We release our benchmark and code to enable further research towards LLM systems which more equitably consider and address information across languages.

Limitations

One limitation is that our study of cross-lingual robustness was only performed on territorial disputes queries. We noted that the factuality and consistency metrics are especially applicable to the more general fact-seeking questions explored in prior works. We could also explore other metrics, such as the cross-lingual consistency metric introduced by (Qi et al., 2023).

Another is that the scope of our setting is largely confined to Wikipedia. As with prior RAG and IR benchmarks, the questions and answers derive from Wikipedia, as well as the retrieval documents. Some data leakage is possible, but this is less of a problem for our queries which do not have a definitive answer which can be memorized. Still, future work should consider more diverse sources for documents.

Other limitations arise from our annotation process for the content of documents. Mainly, the small scale of human annotations, for only 5 languages. This leaves the quality of the LLM annotations for the other 44 languages unverified. However, the linguistic scope (49 languages) of our benchmark precluded us from full-scale human annotation. We hope to extend our annotations to additional languages in followup efforts. Second, despite the high agreement between human and LLM annotations for relevance (76%), there was a very low agreement for viewpoints. We investigated this Appendix H.3.2, concluding that the task is subjective, and that therefore LLM responses were highly influenced by the few-shot prompts. We plan for additional prompt engineering efforts with better multilingual LLMs.

There are also several multilingual limitations. First, not all LLMs support all 49 languages studied. For example, Llama 3 only officially supports 8 languages. Second, LLMs have been shown to have different competencies for generating and reasoning about texts presented in different languages (Li et al., 2024a; Ahuja et al., 2023). A common workaround is to limit the multilinguality of inputs and outputs; we follow prior work in doing this. We gave instructions in English (aside from the different language queries) and we instructed the model to always generate English text for both response formats (aside from the selections of country names in a different language). Given the active developments for multilinguality, we are excited to try our benchmark and evaluation setup on newer IR systems and LLMs.

9 Ethical Considerations

Given the politically sensitive nature of territorial disputes, we briefly discuss the ethical considerations involved in the construction of BORDIRLINES, and the steps taken to address them. On the documents, these are sourced from publicly available Wikipedia articles in 49 languages. While these may reflect a range of perspectives, Wikipedia's editorial standards emphasize a neutral point of view. Moreover, such content is commonly found in large-scale pretraining corpora and existing retrieval datasets. On oversight, we obtained approval from our institution to recruit annotators from a university course. Annotators were compensated fairly for their efforts. They were informed in advance about the potentially politically sensitive nature of the task, and could opt out any time.Further details on the annotation process are provided in Appendix H.

Acknowledgments

We thank the anonymous reviewers and area chairs for their constructive feedback. We thank the 86 annotators for their labels of document contents, as well as Melanie Chen for helping design and run the annotations. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005, and the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, et al. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*.

Akari Asai, Shayne Longpre, Jungo Kasai, Chia-Hsuan Lee, Rui Zhang, Junjie Hu, Ikuya Yamada, Jonathan H Clark, and Eunsol Choi. 2022. Mia 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 108–120.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. *Preprint*, arXiv:2407.01463.

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.

Cohere For AI. 2024. C4ai command r+ 08-2024.

Marcello Federico. 2011. Book review: Cross-language information retrieval by jian-yun nie. *Computational Linguistics*, 37(2).

Dawn Lawrie, Eugene Yang, Douglas W Oard, and James Mayfield. 2023. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*, pages 521–536. Springer.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024a. Eliciting better multilingual structured reasoning from LLMs through code. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5169, Bangkok, Thailand. Association for Computational Linguistics.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2024b. 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.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024c. Uncovering differences in persuasive language in russian versus english wikipedia. In *The First Workshop on Advancing Natural Language Processing for Wikipedia*, page 21.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Nurul Amelina Nasharuddin and Muhamad Taufik Abdullah. 2010. Cross-lingual information retrieval. *Electronic Journal of Computer Science and Information Technology*, 2(1).

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. New embedding models and api updates. `https://openai.com/index/new-embedding-models-and-api-updates/`. Accessed: 2024-08-26.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. Nomiracl: Knowing when you don’t know for robust multilingual retrieval-augmented generation. *Preprint*, arXiv:2312.11361.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’15, page 363–372, New York, NY, USA. Association for Computing Machinery.

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. *arXiv preprint arXiv:2410.21970*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Per-Language Statistics for BordIRLines Passages

Table 2 gives per-language statistics for BordIRLines passages. Across all 720 queries, there are 19,916 unique passage-query pairs. There are 7,436 passages, from 905 Wikipedia pages. The total size of the retrieval corpus, which includes passages that were never retrieved, is 61,834 passages from 945 pages.

B What do we mean by robustness?

It is worth defining what we mean by robustness, given the complexities inherent in our territorial disputes task. Li et al. (2024b) showed that LLMs possess parametric knowledge on territorial disputes and can answer queries on their ownership. Given LLMs' existing knowledge, we would like to investigate how retrieving external documents can affect responses. Cross-lingual robustness, then, seeks to quantify changes in responses.

In prior work, RAG robustness has focused on decreasing hallucinations and errors (Adlakha et al., 2024; Chen et al., 2024b). In cross-lingual RAG, the desired outcome is less clear. Depending on the use-case, we might want responses to emphasize what a model knows, or emphasize opinions present in their documents. In this work, we do not make a judgment call, but instead measure how document selection quantifiably changes responses.

C Concurrence Score (CS) Metrics

To evaluate direct response formats, we use the metrics introduced by Li et al. (2024b). The base CS metric measures accuracy between two countries; from it, several CS metrics are derived which make the comparisons with respect to different responses in the MRS, and the KB. We reproduce the formulas in Table 3, and refer interested readers to that paper for the full details.

D Full Results

The full results for 7 models, all 6 IR modes, across all metrics are shown in Table 4.

E Prompts Used

The prompts for performing RAG are given in Table 5. Note that the citation-based prompt is from the default RAG prompt provided by the developers of Command-R.[15] The prompts used for LLM annotation are shown in Table 8.

F Details on Additional Experiments

F.1 Cross-lingual Biases of IR Systems

Although IR systems implemented with OpenAI and M3 embeddings both show an aggregate preference for retrieving query-language documents, the actual probabilities vary widely across query languages. In Figures 15 and 16, we visualize the normalized probability of a query-language document being retrieved for each query language. From Figure 15, we see that some low-resource languages such as Armenian (hy), Georgian (ky), and Lao (lo) only retrieve native documents. However, other low-resource languages, such as Nepali (ne) and Haitian Creole (ht) primarily retrieve foreign documents. This variability in how often low-resource language queries retrieve query-language documents partially explain our results in section 5.2 about the variability of low-resource citation rates.

F.2 Robustness Attributes for Other IR Modes

We display results for the 3 other IR modes in Appendix Figure 7, again noting that they are motivated by different information access needs. For `en_only`, retrieval over English Wikipedia, this has the effects of 1) increasing consistency, but 2) increasing geopolitical bias for 4 of the 5 models (still less than `no_ir`). This provides further evidence that diverse multilingual retrieval is best. `qlang+en` results are similar to `en_only`, showing that IR systems' prefer English documents. Finally, `swap_docs` performs an interesting ablation with cross-lingual IR over a different language[16] This causes Δ CS to go negative (-9.0 for Llama, -4.5 for GPT-4o), showing that the introduced documents, with their different information and languages, actually cause a geopolitical bias in the opposite direction.

F.3 RAG with Controlled Relevance

The experiments in §5.1 used all documents from IR. However, LLMs' responses have been shown to be susceptible to non-relevant documents (Shi et al., 2023). We therefore perform experiments in

---

[15] `https://huggingface.co/CohereForAI/c4ai-command-r-v01`

[16] For example, "Ceuta" with es query but ar docs.

Code	Language	Terri-tories	Articles	Passages	Passage -Query Pairs	Code	Language	Terri-tories	Articles	Passages	Passage -Query Pairs
en	English	251	310	3376	10579	tl	Filipino	5	9	40	79
zht	Trad. Chinese	81	70	525	1754	fa	Persian	4	8	58	68
zhs	Simp. Chinese	66	54	358	1557	km	Khmer	4	5	46	71
ar	Arabic	35	55	351	544	nl	Dutch	4	7	44	74
hi	Hindi	28	24	158	510	so	Somali	4	4	35	71
es	Spanish	26	43	275	496	vi	Vietnamese	3	7	40	58
ru	Russian	23	26	235	470	tg	Tajik	3	8	37	57
fr	French	21	36	255	460	lo	Lao	3	4	36	56
sw	Swahili	19	16	122	334	uz	Uzbek	3	6	33	55
az	Azerbaijani	11	14	85	197	ti	Tigrinya	3	1	7	24
hy	Armenian	10	11	75	211	id	Indonesian	2	5	27	35
ms	Malay	9	17	112	170	tr	Turkish	2	5	25	36
uk	Ukrainian	9	11	103	197	bs	Bosnian	2	3	25	48
pt	Portuguese	8	12	84	159	sl	Slovenian	2	4	24	42
ur	Urdu	8	11	71	143	ka	Georgian	1	3	18	19
ky	Kyrgyz	8	9	52	149	el	Greek	1	3	17	18
ne	Nepali	8	6	45	152	da	Danish	1	5	16	18
ko	Korean	7	13	74	142	sq	Albanian	1	2	16	23
iw	Hebrew	6	9	89	99	it	Italian	1	2	15	18
th	Thai	6	11	78	95	ht	Haitian Creole	1	3	15	16
sr	Serbian	6	10	64	122	is	Icelandic	1	5	15	17
ja	Japanese	5	9	65	122	mn	Mongolian	1	3	14	14
hr	Croatian	5	9	59	110	bn	Bangla	1	3	13	20
my	Burmese	5	5	51	104	sn	Shona	1	2	13	14
mg	Malagasy	5	7	45	89		**Totals**	720	905	7436	19916

Table 2: Statistics for the retrieved documents from the BORDIRLINES dataset. For each language, we report the number of: **territories**, Wikipedia **articles**, **passages**, and **passage-query** pairs. Specifically, this table considers the top-10 articles per query, and aggregates across `qlang`, `rel_langs`, `qlang+en,en` IR modes, and both OpenAI and M3 embeddings.

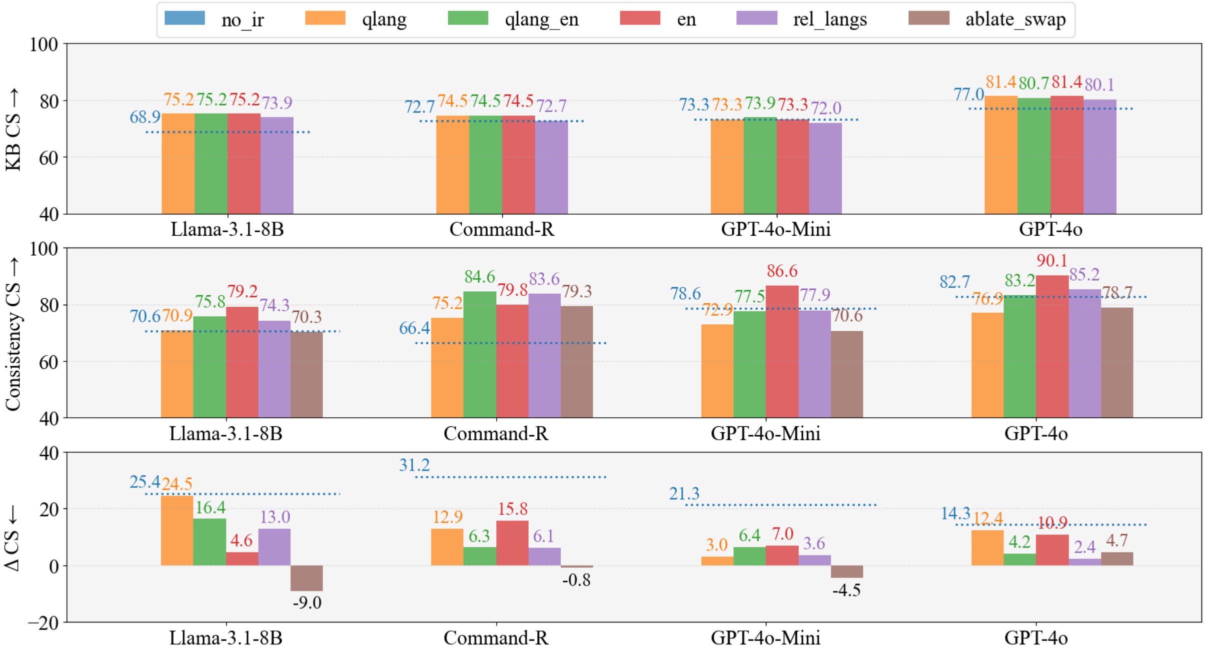


Figure 7: Knowledge-based scores (KB CS), consistency (Cst CS) scores, normalized difference in control vs. non-control CS (Delta CS), for four LLMs under fiv IR modes.

which we perform RAG with only relevant and non-relevant documents. We follow the same task formulation (§2.1), then filter each set of documents for relevance or non-relevance, with the annotations from§3.3, considering only GPT-4o.

$$
\begin{aligned}
\mathrm{CS}(c_i, c_j) &= 100 * \begin{cases} 1 \text{ if } c_i = c_j, \\ 0 \text{ otherwise} \end{cases} \\
\text{KB } \mathrm{CS}(t) &= \mathrm{CS}(c_{KB}, c^{\text{en}}) \\
\text{Con } \mathrm{CS}(t) &= \mathrm{CS}(c_{KB}, c^{\text{controller}}) \\
\text{Non } \mathrm{CS}(t) &= \frac{1}{n} \sum_{c \in C^{\text{non-controllers}}} \mathrm{CS}(c_{KB}, c) \\
\Delta\ \mathrm{CS}(t) &= \frac{\text{Con CS} - \text{Non CS}}{\text{Non CS}} \\
\text{Cst } \mathrm{CS}(t) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathrm{CS}(c_i, c_j)
\end{aligned}
$$

Table 3: Formulas for concurrence score (CS) metrics, reproduced from Li et al. (2024b) (Section 5.2).

Results are shown in Figure 6.[17] On factuality, `all` with 10 documents achieves the best KB CS over both `rel` and `nonrel`. A hypothesis is that an LLM have strong background knowledge on territorial disputes, so it will use the helpful parts of non-relevant documents that are helpful, and ignore the non-helpful parts.

On consistency, we observe similar trends: `rel` is as consistent as `all`, while `nonrel` is less consistent. Geopolitical bias warrants a closer look at each IR mode. For `qlang`, we observe that `all` has the highest Δ CS (15.7 > 10.0 > 6.9). For `rel_langs`, `nonrel` is higher by a bit over `all` (9.8 > 8.4). Given that `all` is a concatenation of the two other subsets, we have an initial hypothesis. As Li et al. (2024b) found that LLM responses were highly affected by cues in the prompt (UN Peacekeeper, as a nationalist), perhaps the additional documents serve a similar function.

G Supplement for Case Studies

G.1 Case study: Yalu River

The Yalu River forms the border between China and North Korea, with historical disputes over specific islands in the river and the precise demarcation of the border. For the Yalu River queries, the majority of retrieved documents in modes other than `en_only` were classified as non-relevant according to LLM annotations, as shown in Figure 14. In addition, all relevant documents were annotated as "Neutral".

We observe that GPT-4o and GPT-4o-mini showed differences in behavior when handling RAG with few relevant neutral documents (generation results in Table 15). In the control `no_ir` mode, both GPT-4o and GPT-4o-mini consistently identified the People's Republic of China as the sovereign authority. However, with RAG, while GPT-4o continued with this response, GPT-4o-mini shifted its response to "North Korea" for all English queries as well as traditional and simplified Chinese queries under `en_only` mode. This suggests that GPT-4o-mini is more susceptible to retrieved content, even when the majority of documents are non-relevant, indicating a potential lack of ability to assess and filter for relevance.

When document citations and explanations were required, several GPT-4o-mini responses shifted from "People's Republic of China" to "North Korea." Among the cited documents, 4 out of 6 for the simplified Chinese query overlapped with 4 out of 5 for the traditional Chinese query, suggesting that the model is able to identify similar relevant evidence across multiple languages.

G.2 Detailed Figures and Tables

G.2.1 Crimea

- Retrieved documents relevance and territory judgment distribution: Figure 11
- Direct response results: Table 11

G.2.2 Spratly Islands

- Retrieved documents relevance and territory judgment distribution: Figure 12
- Direct response results: Table 12
- Response shifts for mode `qlang`: Table 13
- Citation response results results for mode `en_only`: Table 14
- Text for selected document (ID: 27401_p38): Figure 13

G.2.3 Yalu River

- Retrieved documents relevance and territory judgment distribution: Figure 14
- Direct response results: Table 15

H Annotation Details

H.1 Human Annotation Design

We recruited annotators from students enrolled in an Artificial Intelligence course at our university. Participation was entirely voluntary and approved as part of a class project overseen by the course instructors. Prior to participating, students were informed that the task involved reading and labeling Wikipedia passages that may touch on politically

---

[17] These numbers are not comparable to Figure 3 because they were calculated over 446/720 queries. We kept only those queries which, after relevance filtering, had 1+ documents.

Model	IR Mode	KB CS ↑	Con CS ↑	Non CS ↑	ΔCS ↓	ΔCS abs ↓	Cst CS (unk) ↑	Cst CS (all) ↑	# Country, mean	# Country, std dev
	RANDOM	43.5	43.5	43.5	0	0	43.5	43.5	–	–
GPT-4o	`no_ir`	77.0	77.4	67.7	14.3	9.7	75.1	82.7	1.167	0.374
	`qlang+en`	**81.4**	81.3	72.3	12.4	8.9	69.1	76.9	**1.227**	0.429
	`qlang+en`	80.7	78.7	75.5	4.2	3.2	80.1	83.2	1.155	0.363
	`en_only`	**81.4**	**82.6**	74.5	10.9	8.1	**87.5**	**90.1**	1.096	0.295
	`rel_langs`	80.1	78.1	**76.2**	**2.4**	**1.8**	82.3	85.2	1.143	0.351
	`swap_docs`	–	77.0	73.6	4.7	3.5	74.4	78.7	1.052	**0.601**
GPT-4o-mini	`no_ir`	73.3	**77.4**	63.8	21.3	13.6	**80.0**	78.6	1.207	0.416
	`qlang+en`	73.3	71.6	69.5	3.0	2.1	66.5	72.9	**1.263**	0.459
	`qlang+en`	**73.9**	75.5	70.9	6.4	4.6	66.8	77.5	1.219	0.433
	`en_only`	73.3	**77.4**	**72.3**	7.0	5.1	79.4	**86.6**	1.131	0.35
	`rel_langs`	72.0	74.2	71.6	3.6	2.6	68.8	77.9	1.219	0.442
	`swap_docs`	–	68.9	72.1	**-4.5**	**-3.3**	67.2	70.6	1.127	**0.663**
Llama-3 1B	`no_ir`	42.9	47.1	44.3	6.3	2.8	**90.6**	**93.5**	1.08	0.286
	`qlang+en`	**54.7**	52.9	47.5	11.3	5.4	57.7	64.2	1.343	0.492
	`qlang+en`	**54.7**	**54.2**	41.5	30.6	12.7	57.4	66.0	1.331	0.488
	`en_only`	**54.7**	52.9	42.2	25.4	10.7	60.0	63.1	**1.347**	0.493
	`rel_langs`	51.6	52.9	44.7	18.4	8.2	63.7	67.9	1.311	0.481
	`swap_docs`	–	49.6	**52.9**	**-6.1**	**-3.2**	54.7	61.7	1.223	**0.731**
Llama-3 3B	`no_ir`	66.5	56.8	60.3	-5.8	-3.5	54.1	62.4	**1.375**	0.532
	`qlang+en`	**69.6**	63.9	56.7	12.6	7.1	54.6	62.2	**1.375**	0.532
	`qlang+en`	**69.6**	**67.1**	61.0	10.0	6.1	62.0	72.2	1.279	0.467
	`en_only`	**69.6**	65.2	61.3	6.2	3.8	**72.2**	**75.3**	1.235	0.434
	`rel_langs`	**69.6**	63.2	62.1	1.9	1.2	62.3	70.4	1.295	0.474
	`swap_docs`	–	59.3	**64.6**	**-8.3**	**-5.4**	59.8	64.2	1.215	**0.749**
Llama-3 8B	`no_ir`	68.9	72.9	58.2	25.4	14.7	64.8	70.6	**1.291**	0.481
	`qlang+en`	**75.2**	75.5	60.6	24.5	14.8	66.2	70.9	1.279	0.458
	`qlang+en`	**75.2**	**76.8**	66.0	16.4	10.8	68.5	75.8	1.227	0.429
	`en_only`	**75.2**	71.6	68.4	4.6	3.2	**77.6**	**79.2**	1.195	0.407
	`rel_langs`	73.9	72.9	64.5	13.0	8.4	66.5	74.3	1.247	0.459
	`swap_docs`	–	63.7	**70.0**	**-9.0**	**-6.3**	59.8	70.3	1.12	**0.646**
Command-R 7B	`no_ir`	59.6	66.5	**64.2**	**3.5**	**2.3**	65.6	**77.0**	1.251	0.487
	`qlang`	62.1	**83.2**	57.1	45.8	26.1	54.0	59.5	**1.382**	0.503
	`qlang+en`	61.5	82.6	58.2	42.0	24.4	64.0	65.8	1.331	0.496
	`en_only`	61.5	78.7	60.6	29.8	18.1	**70.2**	72.1	1.271	0.463
	`rel_langs`	**64.0**	81.9	58.2	40.9	23.8	63.7	68.0	1.311	0.481
	`swap_docs`	12.4	73.3	62.9	16.7	10.5	56.6	65.0	1.191	**0.701**
Command-R	`no_ir`	72.7	**80.0**	61.0	31.2	19.0	66.0	66.4	**1.319**	0.484
	`qlang+en`	**74.5**	74.8	66.3	12.9	8.5	64.9	75.2	1.247	0.441
	`qlang+en`	**74.5**	78.1	73.4	6.3	4.7	**77.3**	**84.6**	1.159	0.377
	`en_only`	**74.5**	76.8	66.3	15.8	10.5	71.0	79.8	1.195	0.426
	`rel_langs`	72.7	76.8	72.3	6.1	4.4	72.6	83.6	1.171	0.388
	`swap_docs`	–	73.3	**73.9**	**-0.8**	**-0.6**	68.7	79.3	1.056	**0.624**
Li et al. (2024b) `no_ir`	BLOOM$_{560M}$	61.5	67.7	31.2	35.9	115.0	56.3	50.7	1.47	–
	BLOOM$_{7.1B}$	58.4	71.6	36.9	34.7	94.2	49.9	53.9	1.42	–
	BLOOMZ$_{560M}$	49.7	66.5	35.8	30.6	85.5	50.5	53.5	1.43	–
	BLOOMZ$_{7.1B}$	50.3	67.1	48.9	18.2	37.1	47.1	59.3	1.39	–
	GPT-3$_{DV}$	60.5	60.0	51.3	8.7	17.0	63.1	63.3	1.38	–
	GPT-4	79.5	76.9	63.2	13.7	21.6	65.6	70.8	1.29	–

Table 4: Results for all models and IR modes studied, in the direct answer format. We report the concurrence scores (CS). We also report the mean and standard deviations for '# Country' across each query's MRS; this is another way to measure consistency, which would be maximized at 1. For the last group of rows, we report the `no_ir` results of Li et al. (2024b). Rows are grouped by model, and within a group's column, the best number is **bolded**, while the second-best number is underlined.

sensitive or disputed topics. No personally identifying information was collected. Annotators were fairly compensated with extra academic credit, proportional to the number of annotation tasks completed. Over several iterations of annotation and refinement, a total of 86 annotators contributed to

Task	Prompt
System Prompt	You are a geopolitical expert. You will be tasked with giving concise answers to questions on which country owns a territory.
System Prompt for UN Peacekeeper Perspective	You will be tasked with giving concise answers to questions on which country owns a territory. Answer from the perspective of a UN peacekeeper, and adhere to the internationally recognized status of the territory.
Multiple-Choice Generation Prompt Augmented by Retrieved Documents	*Document:* {document_id} *Text:* {document_text} ***Instruction:*** Considering {source: the provided documents and your prior knowledge}, answer the query. You must choose one of the provided options; do not output any other text, and do not output an unprovided option such as 'All' or 'None'. Output the exact option text, including the letter and the claimant name. *Query:* {query} *Option:* {options} *Answer:*
Multiple-Choice Generation Prompt	*Document:* {document_id} *Text:* {document_text} ***Instruction:*** Considering {source: your prior knowledge}, answer the query. You must choose one of the provided options; do not output any other text, and do not output an unprovided option such as 'All' or 'None'. Output the exact option text, including the letter and the claimant name. *Query:* {query} *Option:* {options} *Answer:*
Citation-Based RAG Prompt	**System Preamble** ***Basic Rules:*** You are a powerful conversational and helpful AI. You are augmented by a number of tools, and your job is to use and consume the output of these tools to best help the user. You will see a conversation history between yourself and a user, ending with an utterance from the user. You will then see a specific instruction instructing you what kind of response to generate. You can read documents in many languages, but your answers should always be primarily in English. When you answer the user's requests, you cite your sources in your answers, according to those instructions. **User Preamble** ***Task and Context:*** You help people answer their questions and other requests interactively. You will be asked a very wide array of requests on all kinds of topics. You will be equipped with a wide range of search engines or similar tools to help you, which you use to research your answer. You should focus on serving the user's needs as best you can, which will be wide-ranging. ***Style Guide:*** Unless the user asks for a different style of answer, you should answer in full sentences, using proper grammar and spelling. *Query:* {query} *Document:* 0 *Text:* {text0} *Document:* 1 *Text:* {text1} ... {other documents} ... ***Instruction:*** Carefully perform the following instructions, in order, starting each with a new line. First, write `'Cited Documents:'` followed by a comma-separated list of document IDs, for those documents that contain facts that should be cited in a good answer to the user's last input – for example, 3, 5, 4, 2, 1. Second, write `'Chosen Option:'`, and then choose one of the provided options, considering the provided documents and your prior knowledge. Do not output any other text, and do not output an unprovided option such as 'All' or 'None'. Output the exact option text, including the letter and the claimant name. Finally, write `'Explanation: because'` followed by a fluent paragraph of 4-8 sentences. Answer in English, regardless of the languages of documents. Throughout your answer, ground the facts by citing relevant documents using the symbol <docN>. N is exactly one document ID – for example, <doc0> or <doc2>. Try to use all of the documents you cited in the first step. *Options:* {options}

Table 5: List of prompt templates used for generating the BordIRLines dataset. Each prompt is designed to guide the model in answering questions about territorial ownership, retrieved documents' relevance classification, and multiple-choice selection, with variations in perspective and document-based augmentation.

the final version of the dataset.

Each annotation task consists of a query and its 10 corresponding passages retrieved from IR systems in BordIRLines. For every query-passage pair, annotators were required to answer two questions: a relevance judgment and a viewpoint judgment. Each query-passage pair was reviewed by three independent annotators.

Documents were retrieved using OpenAI embeddings (OpenAI, 2024b) under `qlang` and `en_only`

IR modes. We annotated 564 pairs in 5 languages[18]. The breakdown by language and IR mode is shown in Table 6. This table also shows the assignments after a quality control (564 to 543), which is described ahead.

Annotation Collection Platform We used Amazon Mechanical Turk to streamline the annotation process and minimize overhead.

Answer Choices We analyzed potential edge cases in viewpoint judgment for query-passage pairs and designed comprehensive answer choices covering all scenarios.

The answer options for *relevance* are

- "Relevant" for documents that directly address the territorial dispute in the query
- "Not Relevant" for documents that do not address the territorial dispute in the query

The answer options for *viewpoint* are

- Any of the claimant countries listed as a multiple-choice option in the query.
- "Neutral" for passages that are relevant but do not indicate a preference for either claimant.
- "Not Relevant" for passages marked not relevant in the relevance annotation step.

The "Not Relevant" option is aligned across the relevance and viewpoint questions. Specifically, if a passage was marked "Not Relevant" in the relevance judgment, annotators were instructed to select "Not Relevant" for the viewpoint judgment as well.

H.1.1 Guidelines

A screenshot of the guidelines provided to human annotators on Amazon Mechanical Turk (MTurk) can be found in Figure 17. Below, we include another set of guidelines given to human annotators in a Google document.

Query-Passage Relevance

- **Answer "Yes"**: The passage is relevant if it discusses the territory and its ownership status, either explicitly (e.g., mentions of dispute) or implicitly (e.g., clear claims by a country).
- **Answer "No"**: The passage is not relevant if it does not address the territory's ownership status or provide useful information for answering the query.

Note: For ambiguous cases, use your best judgment and refer to the provided examples.

Viewpoint Judgment

- **Claimant 1**: The passage supports Claimant 1.
- **Claimant 2**: The passage supports Claimant 2.
- **Neutral**: The passage is relevant but does not favor any claimant.
- **Not Relevant**: If the passage is marked as "No" for query-passage relevance, select "Not Relevant" here.

H.1.2 Examples

An example screenshot of a HIT task on MTurk presented to a human annotator is displayed in Figure 18. For greater clarity, we have included two more examples below of how a query-document pair should be annotated, with explanations.

Example 1
Query: "Is Ceuta a territory of Spain or Morocco?"
Passage: *"Ceuta is a Moroccan city under Spanish sovereignty located on the African continent."*
Query-Passage Relevance: Yes
Explanation: The passage explicitly mentions Ceuta's ownership status.
Viewpoint Judgment: Spain
Explanation: The phrase "Spanish sovereignty" indicates that Spain currently controls Ceuta.

Example 2
Query: "Are the Falkland Islands a territory of Argentina or the United Kingdom?"
Passage: *"A little over a quarter of the workforce serves the Falkland Islands government, making it the archipelago's largest employer. Tourism, part of the service economy, has been spurred by increased interest in Antarctic exploration and the creation of direct air links with the United Kingdom and South America."*
Query-Passage Relevance: No
Explanation: While the passage mentions the UK and South America, it does not address the dispute or ownership status.
Viewpoint Judgment: Not Relevant
Explanation: If you answer "No" for the first question, always answer "Not Relevant" for the second.

H.1.3 Quality Control Conditions

To ensure the reliability and consistency of human annotations, we discarded any individual annota-

---

[18] 564 unique pairs from 620 requested pairs accounting for overlaps in retrieved documents,

Language	Retrieval Mode	# Assignments Pre-Filtering	# Assignments Post-Filtering
English	control	90	89
Spanish	en	78	78
Hindi	en	84	78
Korean	en	21	19
Chinese	en	120	116
Korean	qlang	21	18
Chinese	qlang	150	145

Table 6: Number of annotations pre- and post- filtering for each language and retrieval mode combination.

Language	Retrieval Mode	Relevance (α)	Viewpoint (α)
English	`control`	0.78	0.66
Spanish	`en`	0.86	0.73
Hindi	`en`	0.76	0.66
Korean	`en`	0.59	0.65
Chinese	`en`	0.47	0.44
Korean	`qlang`	0.54	0.54
Chinese	`qlang`	**0.90**	**0.83**

Table 7: Krippendorff alpha (α) values for relevance judgment and viewpoint judgment annotations

tions which met any of the two quality control conditions:

- Condition 1: Annotator answered “No” to relevance judgment and did *not* mark as “Not relevant” on viewpoint judgment
- Condition 2: Annotator answered “Yes” to relevance judgment and marked as “Not relevant” on viewpoint judgment

Condition 1 shows a contradiction where the passage is marked non-relevant, but the viewpoint judgment is not “Not Relevant,” violating the guidelines. Condition 2 also contains a contradiction, where a passage deemed relevant is marked “Not Relevant” in the viewpoint judgment. Ultimately, we discarded a handful of annotations meeting either condition.

H.1.4 Calculating Krippendorff Alpha for Human Annotations

We measured inter-annotator agreement for relevance and viewpoint judgments using Krippendorff’s alpha (Table 7). Results show varying agreement levels across languages and retrieval modes. Spanish annotations in the `en` retrieval mode had high agreement (0.86 for relevance, 0.73 for viewpoint), indicating strong consistency. Chinese annotations in the `qlang` mode also showed high agreement (0.90 for relevance, 0.83 for viewpoint), highlighting the benefit of native language annotations. However, Korean and Chinese annotations in the `en_only` retrieval mode had lower agreement (0.59 and 0.47 for relevance), pointing to challenges with cross-lingual texts. Overall, relevance judgments had higher agreement than viewpoint judgments, due the complexity of assessing territorial claims. These findings emphasize the need for language-specific guidelines and training to improve annotation quality in multilingual contexts.

H.2 LLM Annotation Design

The LLM annotation process follows the same general steps as human annotations. We use GPT-4o for annotations, with full prompts detailed in Table 8. For each of 19k query-document pairs, we first query the LLM for whether the document is *relevant* to answering the query’s territorial dispute. For the subset of query-document pairs marked as relevant, we query the LLM to judge which *territory viewpoint* the document supports, or if it is neutral.

H.3 Human-LLM Annotation Agreement

Comprehensive F1-scores between human and LLM annotations are calculated in Table 9.

H.3.1 Relevance Annotation F1-score

The weighted average F1-score of 0.76 between LLM and human relevance annotations reflects the model’s effectiveness in assessing relevance. However, performance varies across languages. English (0.87) and Chinese (0.82) show strong F1-scores, balancing precision and recall. In contrast, Korean (0.54) is significantly lower, primarily due to poor recall (0.37), suggesting that either LLMs are not identifying relevant documents or humans are over-identifying relevance. These findings highlight that there is room for improvement through better human training and refined LLM prompt design.

H.3.2 Viewpoint Annotation F1-score

The results for territory viewpoint classification reveal significant challenges, with a weighted average F1-score of 0.26, much lower than relevance classification. English performs best (F1 = 0.37) but suffers from low precision (0.23) despite high recall (0.95), indicating frequent misclassifications. Korean (F1 = 0.15) and Chinese (F1 = 0.11) have even worse performance, primarily due to very low precision (0.09 and 0.06), leading to excessive false

Task	Prompt
Few-Shot Relevance Judgement Prompt	Determine whether the passage is relevant to answering a question about which country owns a territory. Respond with ’Yes’ or ’No’. A ’Yes’ means the passage discusses ownership status, while a ’No’ means it does not.
Few-Shot Territory Judgement Prompt	Given a passage and a question, determine which country’s claim the passage supports, or choose ’X) Neutral’ if none are favored. Output only one of the provided options.
Zero-Shot Relevance Judgement Prompt	You will be given a passage and a question regarding territorial ownership. Answer only ’Yes’ or ’No’ based on whether the passage discusses ownership, not just the territory itself.
Zero-Shot Territory Judgement Prompt	You are given a passage and a question about a territorial dispute. Determine which country’s claim the passage supports. Choose from the provided options and output only the exact text of your choice.

Table 8: List of prompt templates used for LLM annotations on the BordIRLines dataset. Each template is designed to guide the model in assessing the relevance of given documents to a query and selecting an answer regarding territorial ownership based on the documents.

Language	True Positive	True Negative	False Positive	False Negative	Precision ↑	Recall ↑	F1-score ↑
F1-score for Relevance							
English	31	60	2	7	0.9394	0.8158	0.8732
Korean	10	43	0	17	1.0000	0.3704	0.5405
Chinese	16	67	1	6	0.9412	0.7273	0.8205
F1-score for Territory/Viewpoint							
Language	True Positive	False Positive	False Negative	True Negative (Set to 0)	Precision ↑	Recall ↑	F1-score ↑
English	20	66	1	0	0.2326	0.9524	0.3738
Korean	5	51	4	0	0.0893	0.5556	0.1538
Chinese	5	77	4	0	0.0610	0.5556	0.1099

Table 9: F1-scores for relevance and territory/viewpoint across different languages. Precision, Recall, and F1-score are reported for each case.

positives and over-classification of territorial viewpoints.

Human annotation inconsistencies may also contribute to low F1-scores, as identified in Section I.2 of the appendix. The high false positive rate, especially in languages like Korean and Chinese, suggests that annotation flaws may be partially responsible for model errors. These findings highlight the need for improved LLM prompting and more consistent human annotations.

H.4 IR System Evaluation with Relevance Annotations

To assess the whether our IR system is effectively ranking documents by relevance, we compare the proportion of relevant documents as annotated by humans and LLMs against the document rank assigned by the IR system. This comparison is visualized in Figures 8 and 9. As can be seen, there is a noticeable decline in relevance as the rank increases, with higher-ranked documents being more likely to be relevant across all annotation types, showing the effectiveness of our IR system.

Aggregate proportions of relevant documents are shown in Table 10. In general, few-shot prompting for LLM annotations marked fewer query-document pairs as relevant compared to zero-shot prompting (24% vs 21%). More OpenAI-retrieved documents are judged relevant compared to M3-retrieved documents, across all IR modes and prompting types, indicating that OpenAI embeddings can better identify relevant documents. Finally, `rel_langs` had the most relevant documents of any retrieval mode, which is expected because `rel_langs` is a superset of the documents in other retrieval modes.

I Aggregated Citation Responses

A bird’s-eye view into citations, in aggregate, is shown in the distribution pie charts in Appendix Table 10. Considering `qlang+en` and `rel_langs`, we see that as expected, `qlang+en` cites English documents more often (67.1 > 57.2). For all selected languages, `rel_langs` has a higher citation rate than `qlang+en`.

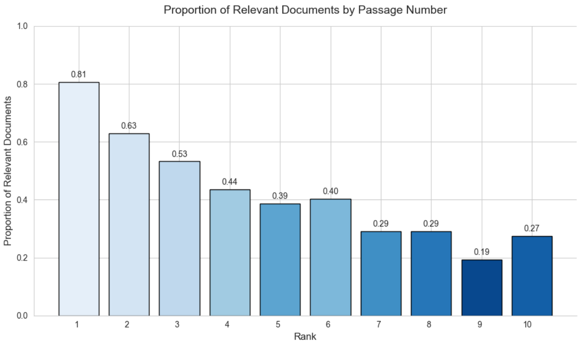


Figure 8: Human annotations for the proportion of relevant documents vs. the document rank assigned by the OpenAI IR system. Top ranked documents are more likely to be relevant, with a drop-off as rank increases.

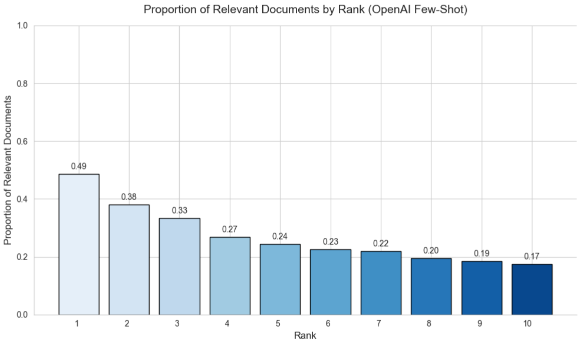


Figure 9: LLM annotations (gpt-4o, few-shot prompting) for relevant document proportions vs. rank by the OpenAI IR system. Similar to Figure 8, relevance decreases monotonically with rank. The lower proportions compared to Figure 8 reflect human annotations being done on a subset of the most popular languages, whereas LLM annotations cover all documents.

	OpenAI		**BGE-M3**	
	Zero-shot	Few-shot	Zero-shot	Few-shot
`qlang`	*0.2876*	0.2580	0.2844	0.2515
`qlang+en`	*0.3556*	0.3207	0.3297	0.2938
`en_only`	*0.3229*	0.2946	0.3049	0.2767
`rel_langs`	***0.3864***	**0.3497**	**0.3739**	**0.3373**

Table 10: Zero-shot and few-shot LLM annotation relevance proportions across IR Modes, for the sets of documents retrieved by two separate IR systems (OpenAI and M3 embeddings).

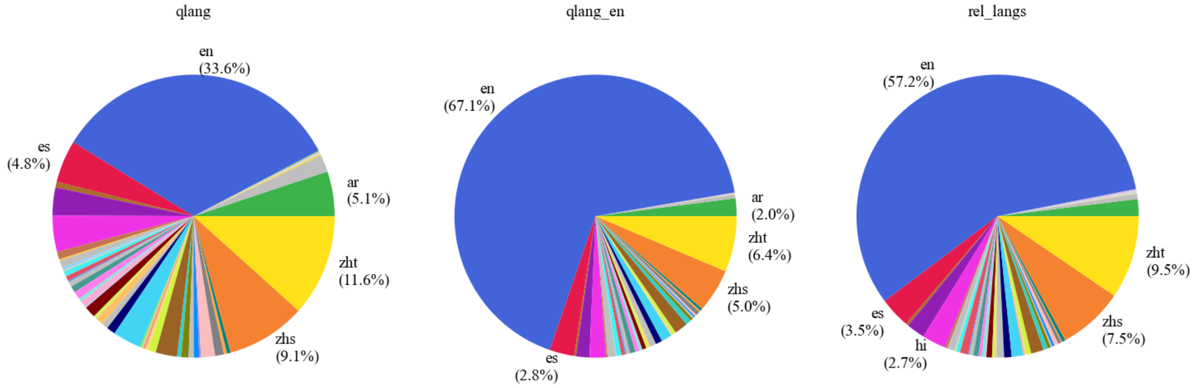


Figure 10: Citation distribution across three RAG modes. Each pie chart shows the percentage breakdown of languages of citation documents. `qlang` simply reflects the linguistic distribution of the 720 queries in BordIRLines; it can be considered a denominator for normalization of the other two charts. As the other modes have multilingually-sourced documents, by construction their language proportions must be lower than for `qlang`.

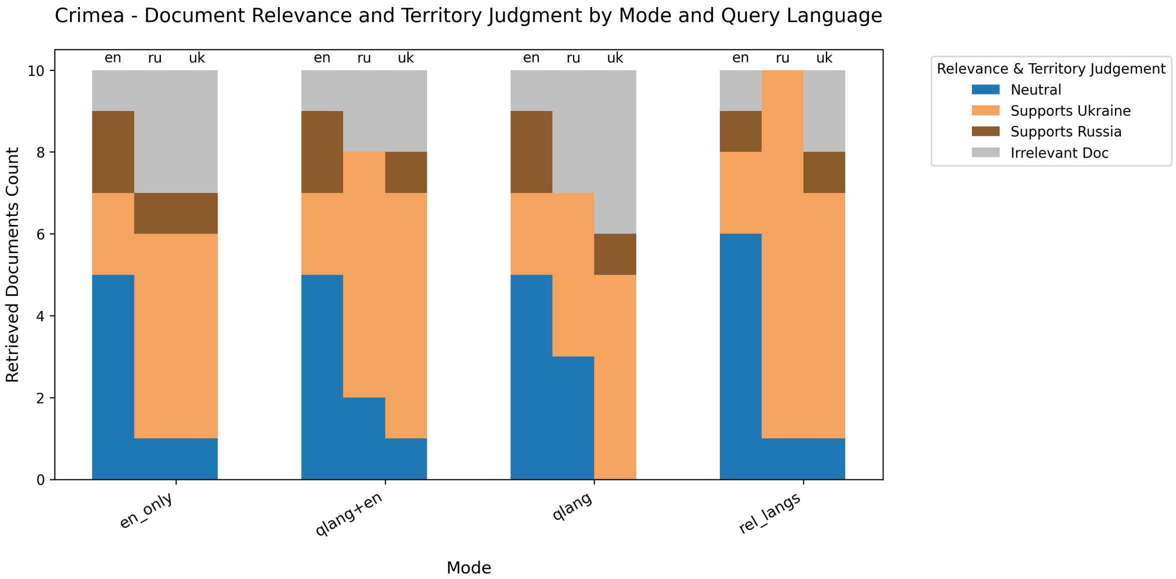


Figure 11: Document relevance and territorial judgment in the Crimea case study, grouped by query language and retrieval mode. Stacked bars show the distribution of retrieved documents *Supports Ukraine*, *Supports Russia*, *Neutral*, or *Irrelevant Doc*. The label at the top of each bar indicates the query language: en (English), ru (Russian), uk (Ukrainian).

Model	Mode	Crimea_en	Crimea_ru	Crimea_uk
GPT-4o	no_ir	Ukraine	Ukraine	Ukraine
	en_only	Ukraine	Ukraine	Ukraine
	qlang+en	Ukraine	Ukraine	Ukraine
	qlang	Ukraine	Ukraine	Ukraine
	rel_langs	Ukraine	Ukraine	Ukraine
GPT-4o-mini	no_ir	Russia	Russia	Russia
	en_only	Ukraine	Ukraine	Ukraine
	qlang+en	Ukraine	Ukraine	Ukraine
	qlang	Ukraine	Ukraine	Ukraine
	rel_langs	Ukraine	Ukraine	Ukraine

Table 11: Crimea territorial judgment results for GPT-4o-mini and GPT-4o across different retrieval modes. The table shows the generated territorial results for queries in English (Crimea_en), Russian (Crimea_ru), and Ukrainian (Crimea_uk).

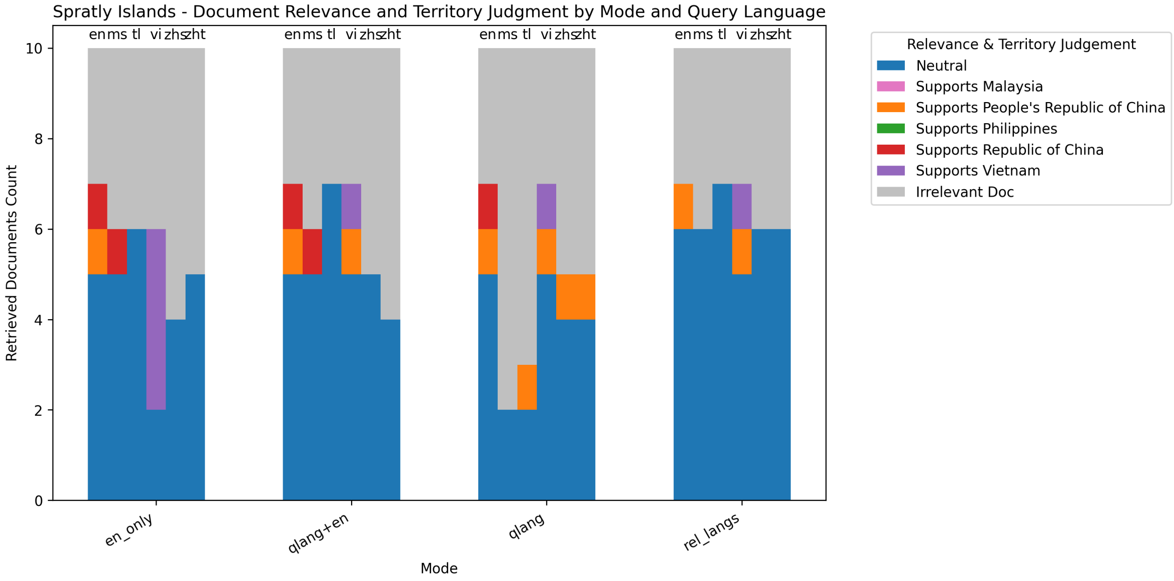


Figure 12: Document relevance and territorial judgment in the Spratly Islands case study, grouped by query language and retrieval mode. Stacked bars show the distribution of retrieved documents *Supports Malaysia*, *Supports People's Republic of China*, *Supports Philippines*, *Supports Republic of China*, *Supports Vietnam*, *Neutral*, or *Irrelevant Doc*. The label at the top of each bar indicates the query language: en (English), ms (Malay), tl (Tagalog), vi (Vietnamese), zhs (Simplified Chinese), zht (Traditional Chinese).

Model	Mode	Spratly_en	Spratly_ms	Spratly_tl	Spratly_vi	Spratly_zhs	Spratly_zht
GPT-4o	no_ir	Vietnam	Vietnam	Vietnam	Vietnam	PRC	PRC
	en_only	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam
	qlang+en	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam
	qlang	Vietnam	Vietnam	Vietnam	Vietnam	PRC	PRC
	rel_langs	Vietnam	Vietnam	Vietnam	Vietnam	Vietnam	PRC

Table 12: Spratly Islands territorial judgment results for GPT-4o across different retrieval modes. The table shows the generated territorial results for queries in English (Spratly_en), Malay (Spratly_mi), Tagalog (Spratly_tl), Vietnamese (Spratly_vi), simplified Chinese (Spratly_zhs), and traditional Chinese (Spratly_zht).

Response Mode	en	Spratly_ms	Spratly_tl	Spratly_vi	Spratly_zhs	Spratly_zht
Direct	Vietnam	Vietnam	PRC	Vietnam	PRC	PRC
Citation	PRC	PRC	Philippines	Vietnam	PRC	PRC

Table 13: Change in Spratly Islands territorial judgment results for mode `qlang` using model *GPT-4o-mini*. Note that the 'Citation Response' in this table does not include document selection or justification.

> The following are political divisions for the Spratly Islands claimed by various area nations (in alphabetical order): Brunei: Part of Brunei's Exclusive Economic Zone, China: Part of Sansha, Hainan, Malaysia: Part of Sabah state, Philippines: Part of Kalayaan, Palawan province, Taiwan: Part of Kaohsiung municipality, Vietnam: Part of Trường Sa, Khánh Hòa Province.

Figure 13: A retrieved document for the Spratly Islands (**doc_id: 27401_p38**).

Query Lang	Citation Response Full
en	Cited Documents: **0**, 1, 2, 4, 9 Chosen Option: D) People's Republic of China Explanation: **The Spratly Islands are claimed by multiple nations, but the People's Republic of China (PRC) asserts significant control over them. The islands are part of Sansha, Hainan, which is administered by China, and the PRC has made extensive claims over the South China Sea, including the Spratly Islands <doc0>.** Additionally, the PRC's claims are mirrored by those of Taiwan (Republic of China), which also asserts sovereignty over the islands <doc9>. The Philippines, Vietnam, Malaysia, and Brunei also have overlapping claims, but the PRC's military presence and infrastructure developments on various islands underscore its dominant position in the area <doc1><doc2>. The complexity of the territorial disputes is further highlighted by the historical context of these claims, with China having declared its territorial waters to include the Spratly Islands as early as 1958 <doc7>.
ms	Cited Documents: **1**, 2, 4, 5, 3 Chosen Option: D) Republik Rakyat China Explanation: **The Spratly Islands are claimed by multiple nations, but the People's Republic of China (Republik Rakyat China) asserts a significant claim over the entire archipelago. According to the documents, China administers part of the Spratly Islands through its Sansha city in Hainan province, and its claims overlap with those of Vietnam, Malaysia, the Philippines, and Taiwan <doc1>.** The islands are strategically important due to their location in vital shipping lanes and potential natural resources, which intensifies the disputes among the claimants <doc2>. Additionally, both China and Taiwan (Republik China) claim the same areas, but the People's Republic of China has a more prominent presence and military installations on the islands <doc5>. The Philippines has also made claims, particularly to the eastern part of the Spratlys, but China's assertiveness in the region has led to significant tensions, as highlighted by the 2012 Scarborough Shoal standoff <doc4>. Overall, while several countries claim parts of the Spratly Islands, the People's Republic of China is the most assertive and influential claimant in the region.
tl	Cited Documents: 0, **1**, 2, 3, 4, 6, 8 Chosen Option: F) Pilipinas Explanation: The Spratly Islands are claimed by multiple nations, including the Philippines, which asserts its claim based on historical and legal grounds. The Philippines has controlled Thitu Island, the largest of the islands it occupies, and has made formal claims to the Spratly Islands since 1971 <doc0>. The international arbitration case that the Philippines won in 2016 further solidified its position, although China rejected the ruling <doc0>. **Additionally, the Philippines' claims overlap with those of Vietnam, Malaysia, and China, among others, indicating the complex nature of territorial disputes in the region <doc1>.** Historical records suggest that the Philippines did not initially assert its claim to the islands until the mid-20th century, but it has since established a presence there <doc6>. Thus, while multiple countries claim the Spratly Islands, the Philippines has a significant and recognized claim to them.
vi	Cited Documents: **0**, 1, 2, 4, 5 Chosen Option: A) Việt Nam Explanation: The Spratly Islands, known as Trường Sa in Vietnamese, are claimed by multiple nations, but Vietnam asserts its sovereignty over them. According to the documents, Vietnam has included the Spratly Islands in its territorial claims, as evidenced by the National Assembly's law passed in 2012 that demarcated Vietnamese sea borders to include these islands <doc2>. Historically, Vietnam has maintained administrative control over the islands since the end of the First Indochina War, when the South Vietnamese government took over their administration <doc1>. Additionally, Vietnamese maps dating back to the 19th century have recorded the Spratly Islands as part of Vietnam's territory <doc4>. **While other countries, including China and the Philippines, also claim parts of the Spratly Islands, the documents indicate that Vietnam's claims are well-established and recognized in its legal framework <doc0>**, <doc5>.
zhs	Cited Documents: 0, 1, **2**, 3, 5 Chosen Option: D) 中华人民共和国 Explanation: **The Spratly Islands are claimed by multiple nations, including the People's Republic of China (PRC), which asserts that a significant portion of the islands is part of its territory, specifically under the administration of Sansha, Hainan <doc2>.** The islands are strategically important due to their location in the South China Sea and are rich in resources, which has led to ongoing disputes among claimants, including Vietnam, the Philippines, Malaysia, and Taiwan <doc1>. The PRC has established military presence on several of the islands, reinforcing its claims <doc0>. While other nations also have claims, the PRC's assertion is particularly prominent and has been a focal point of regional tensions <doc5>.
zht	Cited Documents: **3**, 4, 1, 2, 0 Chosen Option: D) 中華人民共和國 Explanation: **The Spratly Islands are a highly disputed territory in the South China Sea, claimed by multiple nations including China, Vietnam, the Philippines, Malaysia, Taiwan, and Brunei. Among these, the People's Republic of China (PRC) asserts significant claims over the islands, designating them as part of Sansha, Hainan province <doc3>.** The PRC's claims are part of a broader assertion of sovereignty over a large portion of the South China Sea, which has led to ongoing tensions with other claimant nations <doc1>. Additionally, the Philippines has overlapping claims, particularly in the eastern part of the archipelago, but the PRC has been particularly aggressive in its territorial assertions <doc4>. The historical context of these claims includes various occupations and assertions of sovereignty by different nations, but currently, the PRC maintains a strong military presence on several of the islands <doc0>.

Table 14: Spratly Islands citation responses for GPT-4o-mini in `en_only` mode. Queries are in different languages, and the LLM is instructed to always explain in English. Within a response, we **bold** the part where document ID "27401_p38" is cited. The tag `<docN>` indicates that document appears in position N in the prompt.

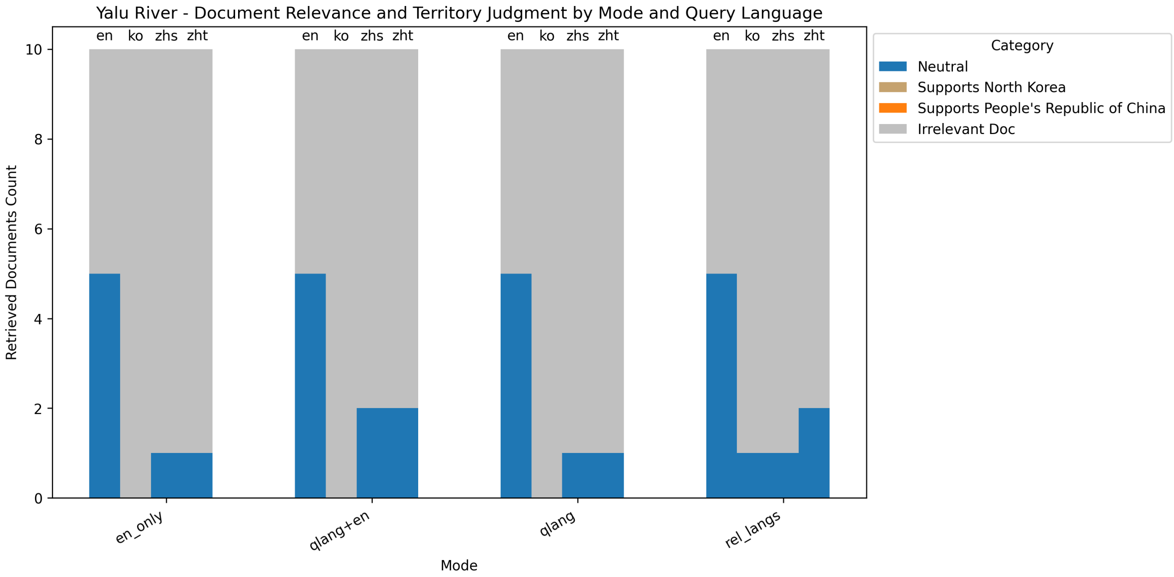


Figure 14: Document relevance and territorial judgment in the Yalu case study, grouped by query language and retrieval mode. Stacked bars show the distribution of retrieved documents *Supports North Korea*, *Supports People's Republic of China*, *Neutral*, or *Irrelevant Doc*. The label at the top of each bar indicates the query language: en (English), ko (Korean), zhs (simplified Chinese), zht (traditional Chinese).

Model	Mode	Yalu_en	Yalu_ko	Yalu_zhs	Yalu_zht
GPT-4o	no_ir	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	en_only	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	qlang+en	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	qlang	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	rel_langs	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
GPT-4o-mini	no_ir	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	en_only	North Korea	North Korea	People's Republic of China	North Korea
	qlang+en	People's Republic of China	People's Republic of China	People's Republic of China	People's Republic of China
	qlang	North Korea	North Korea	People's Republic of China	People's Republic of China
	rel_langs	North Korea	People's Republic of China	People's Republic of China	People's Republic of China

Table 15: Yalu River territorial judgment results for GPT-4o and GPT-4o-mini across different retrieval modes. The table shows the generated territorial alignment for queries in English (Yalu_en), Korean (Yalu_ko), Simplified Chinese (Yalu_zhs), and Traditional Chinese (Yalu_zht).

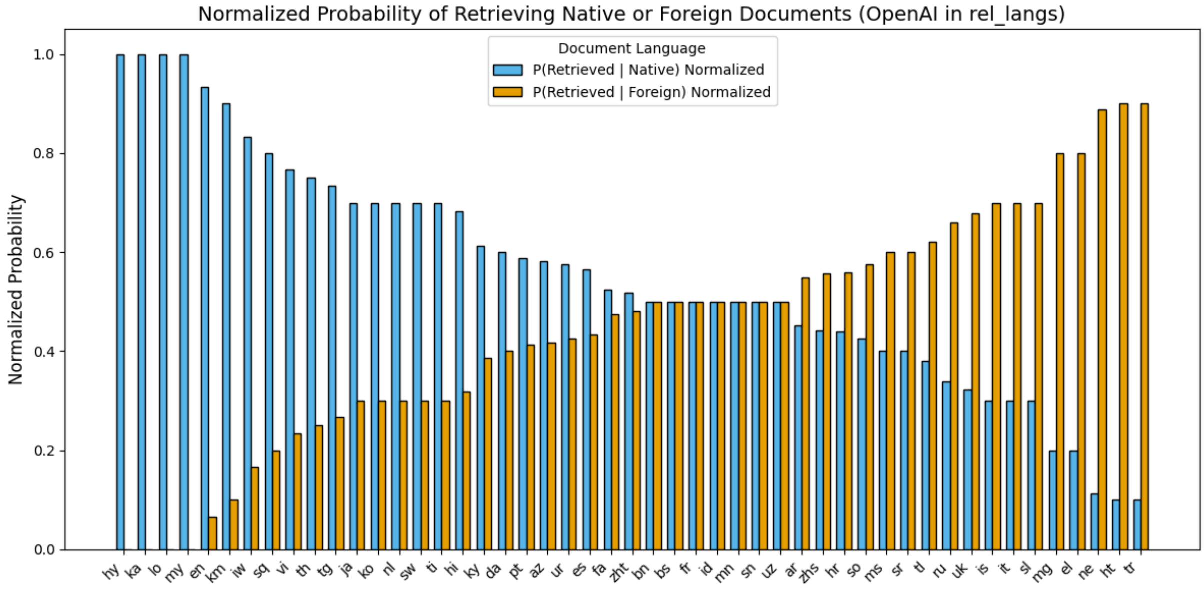


Figure 15: Normalized probability of a query language or non-query language document being retrieves across query languages, using OpenAI embeddings in the `rel_langs` retrieval mode. We first calculate the probability that a native or non-native document is retrieved, and normalize so the two probabilities sum to one.

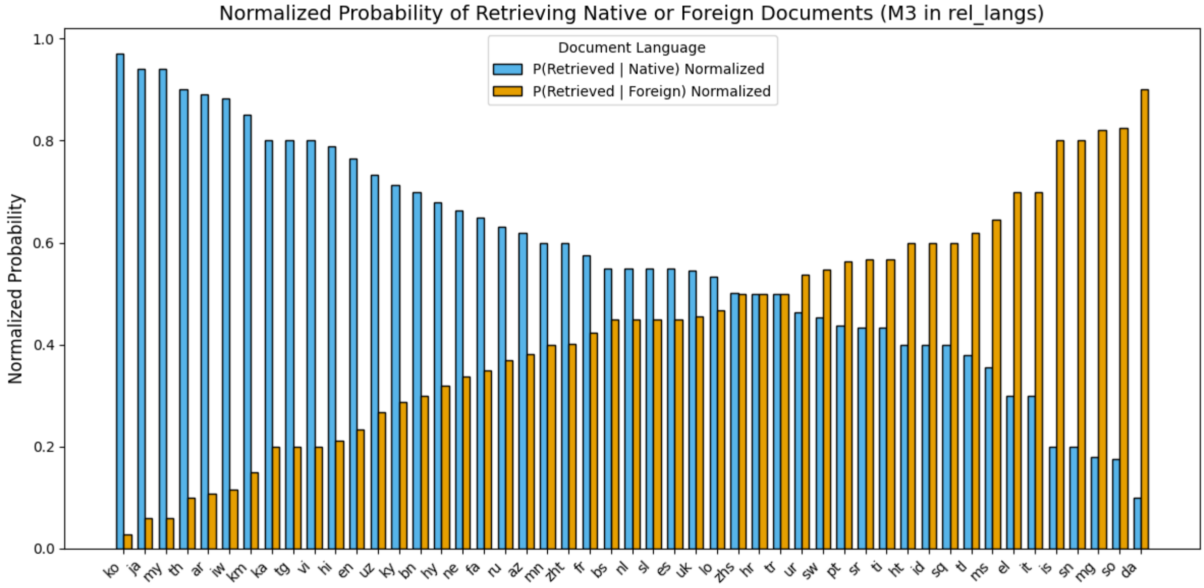
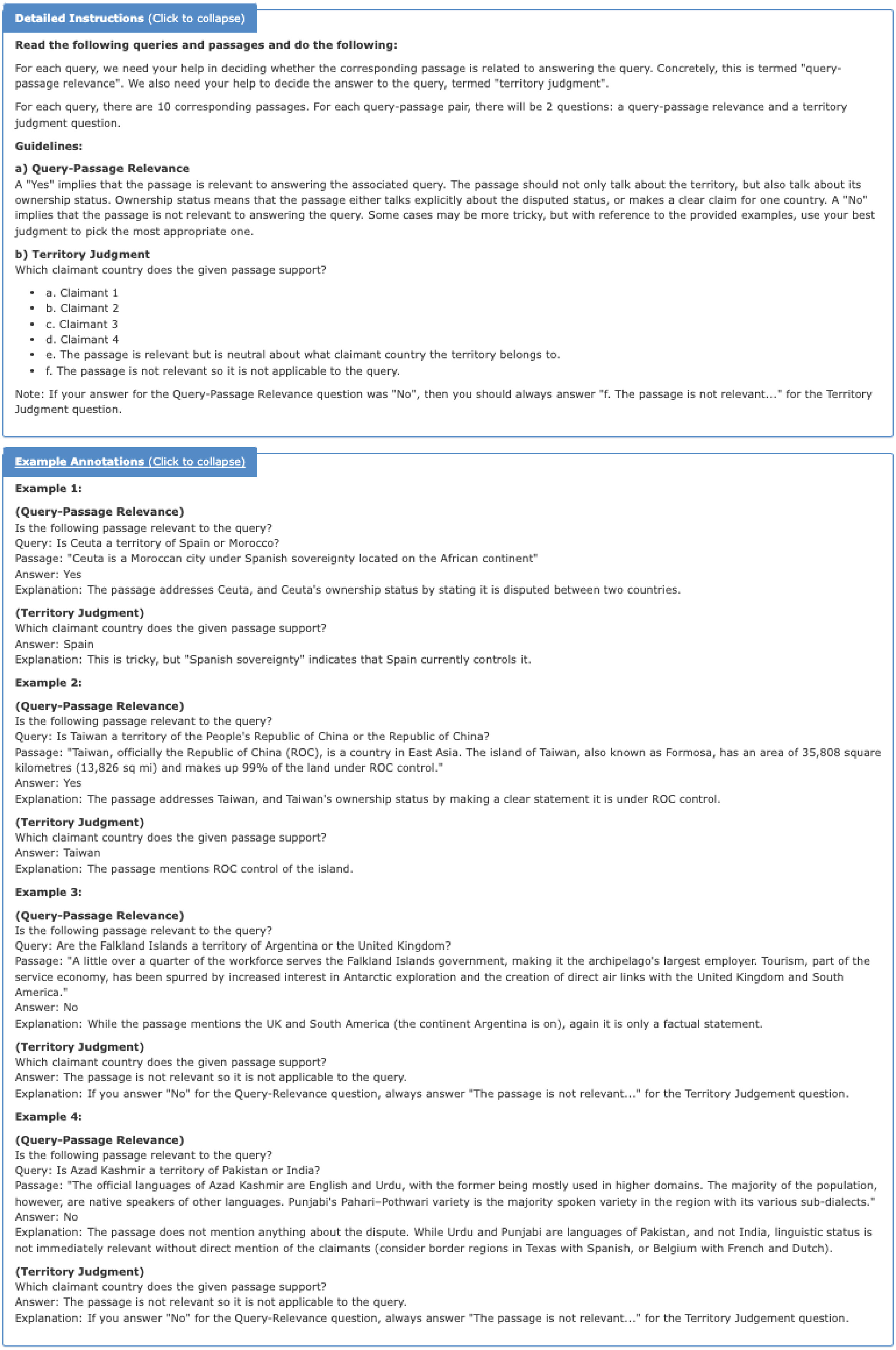


Figure 16: Normalized probability of a query language or non-query language document being retrieved across query languages, using M3 embeddings in the `rel_langs` retrieval mode.



Detailed Instructions (Click to collapse)

Read the following queries and passages and do the following:

For each query, we need your help in deciding whether the corresponding passage is related to answering the query. Concretely, this is termed "query-passage relevance". We also need your help to decide the answer to the query, termed "territory judgment".

For each query, there are 10 corresponding passages. For each query-passage pair, there will be 2 questions: a query-passage relevance and a territory judgment question.

Guidelines:

a) Query-Passage Relevance
A "Yes" implies that the passage is relevant to answering the associated query. The passage should not only talk about the territory, but also talk about its ownership status. Ownership status means that the passage either talks explicitly about the disputed status, or makes a clear claim for one country. A "No" implies that the passage is not relevant to answering the query. Some cases may be more tricky, but with reference to the provided examples, use your best judgment to pick the most appropriate one.

b) Territory Judgment
Which claimant country does the given passage support?

- a. Claimant 1
- b. Claimant 2
- c. Claimant 3
- d. Claimant 4
- e. The passage is relevant but is neutral about what claimant country the territory belongs to.
- f. The passage is not relevant so it is not applicable to the query.

Note: If your answer for the Query-Passage Relevance question was "No", then you should always answer "f. The passage is not relevant..." for the Territory Judgment question.

Example Annotations (Click to collapse)

Example 1:

(Query-Passage Relevance)
Is the following passage relevant to the query?
Query: Is Ceuta a territory of Spain or Morocco?
Passage: "Ceuta is a Moroccan city under Spanish sovereignty located on the African continent"
Answer: Yes
Explanation: The passage addresses Ceuta, and Ceuta's ownership status by stating it is disputed between two countries.

(Territory Judgment)
Which claimant country does the given passage support?
Answer: Spain
Explanation: This is tricky, but "Spanish sovereignty" indicates that Spain currently controls it.

Example 2:

(Query-Passage Relevance)
Is the following passage relevant to the query?
Query: Is Taiwan a territory of the People's Republic of China or the Republic of China?
Passage: "Taiwan, officially the Republic of China (ROC), is a country in East Asia. The island of Taiwan, also known as Formosa, has an area of 35,808 square kilometres (13,826 sq mi) and makes up 99% of the land under ROC control."
Answer: Yes
Explanation: The passage addresses Taiwan, and Taiwan's ownership status by making a clear statement it is under ROC control.

(Territory Judgment)
Which claimant country does the given passage support?
Answer: Taiwan
Explanation: The passage mentions ROC control of the island.

Example 3:

(Query-Passage Relevance)
Is the following passage relevant to the query?
Query: Are the Falkland Islands a territory of Argentina or the United Kingdom?
Passage: "A little over a quarter of the workforce serves the Falkland Islands government, making it the archipelago's largest employer. Tourism, part of the service economy, has been spurred by increased interest in Antarctic exploration and the creation of direct air links with the United Kingdom and South America."
Answer: No
Explanation: While the passage mentions the UK and South America (the continent Argentina is on), again it is only a factual statement.

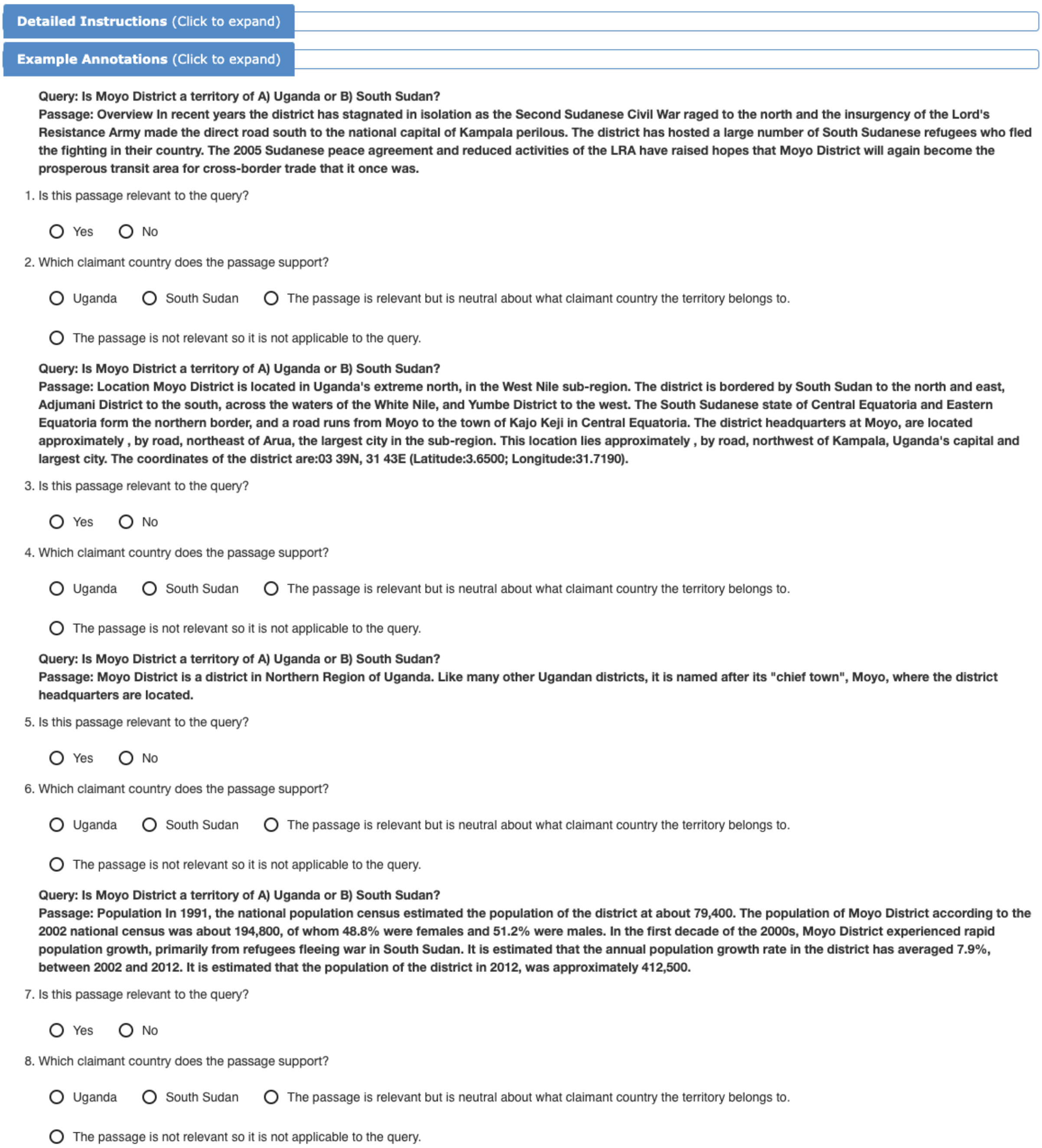
(Territory Judgment)
Which claimant country does the given passage support?
Answer: The passage is not relevant so it is not applicable to the query.
Explanation: If you answer "No" for the Query-Relevance question, always answer "The passage is not relevant..." for the Territory Judgement question.

Example 4:

(Query-Passage Relevance)
Is the following passage relevant to the query?
Query: Is Azad Kashmir a territory of Pakistan or India?
Passage: "The official languages of Azad Kashmir are English and Urdu, with the former being mostly used in higher domains. The majority of the population, however, are native speakers of other languages. Punjabi's Pahari–Pothwari variety is the majority spoken variety in the region with its various sub-dialects."
Answer: No
Explanation: The passage does not mention anything about the dispute. While Urdu and Punjabi are languages of Pakistan, and not India, linguistic status is not immediately relevant without direct mention of the claimants (consider border regions in Texas with Spanish, or Belgium with French and Dutch).

(Territory Judgment)
Which claimant country does the given passage support?
Answer: The passage is not relevant so it is not applicable to the query.
Explanation: If you answer "No" for the Query-Relevance question, always answer "The passage is not relevant..." for the Territory Judgement question.

Figure 17: Screenshot of the instructions for the human annotation task in Section H.1. The detailed instructions define *query-passage relevance* and *judgment*, and the example annotations list four annotated cases.



Detailed Instructions (Click to expand)
Example Annotations (Click to expand)

Query: Is Moyo District a territory of A) Uganda or B) South Sudan?
Passage: Overview In recent years the district has stagnated in isolation as the Second Sudanese Civil War raged to the north and the insurgency of the Lord's Resistance Army made the direct road south to the national capital of Kampala perilous. The district has hosted a large number of South Sudanese refugees who fled the fighting in their country. The 2005 Sudanese peace agreement and reduced activities of the LRA have raised hopes that Moyo District will again become the prosperous transit area for cross-border trade that it once was.

1. Is this passage relevant to the query?

○ Yes ○ No

2. Which claimant country does the passage support?

○ Uganda ○ South Sudan ○ The passage is relevant but is neutral about what claimant country the territory belongs to.

○ The passage is not relevant so it is not applicable to the query.

Query: Is Moyo District a territory of A) Uganda or B) South Sudan?
Passage: Location Moyo District is located in Uganda's extreme north, in the West Nile sub-region. The district is bordered by South Sudan to the north and east, Adjumani District to the south, across the waters of the White Nile, and Yumbe District to the west. The South Sudanese state of Central Equatoria and Eastern Equatoria form the northern border, and a road runs from Moyo to the town of Kajo Keji in Central Equatoria. The district headquarters at Moyo, are located approximately , by road, northeast of Arua, the largest city in the sub-region. This location lies approximately , by road, northwest of Kampala, Uganda's capital and largest city. The coordinates of the district are:03 39N, 31 43E (Latitude:3.6500; Longitude:31.7190).

3. Is this passage relevant to the query?

○ Yes ○ No

4. Which claimant country does the passage support?

○ Uganda ○ South Sudan ○ The passage is relevant but is neutral about what claimant country the territory belongs to.

○ The passage is not relevant so it is not applicable to the query.

Query: Is Moyo District a territory of A) Uganda or B) South Sudan?
Passage: Moyo District is a district in Northern Region of Uganda. Like many other Ugandan districts, it is named after its "chief town", Moyo, where the district headquarters are located.

5. Is this passage relevant to the query?

○ Yes ○ No

6. Which claimant country does the passage support?

○ Uganda ○ South Sudan ○ The passage is relevant but is neutral about what claimant country the territory belongs to.

○ The passage is not relevant so it is not applicable to the query.

Query: Is Moyo District a territory of A) Uganda or B) South Sudan?
Passage: Population In 1991, the national population census estimated the population of the district at about 79,400. The population of Moyo District according to the 2002 national census was about 194,800, of whom 48.8% were females and 51.2% were males. In the first decade of the 2000s, Moyo District experienced rapid population growth, primarily from refugees fleeing war in South Sudan. It is estimated that the annual population growth rate in the district has averaged 7.9%, between 2002 and 2012. It is estimated that the population of the district in 2012, was approximately 412,500.

7. Is this passage relevant to the query?

○ Yes ○ No

8. Which claimant country does the passage support?

○ Uganda ○ South Sudan ○ The passage is relevant but is neutral about what claimant country the territory belongs to.

○ The passage is not relevant so it is not applicable to the query.

Figure 18: A screenshot of an example HIT presented to a human annotator on MTurk for the Moyo District territory. Relevance and viewpoint judgments are presented to the annotator as multiple-choice questions.