BordIRlines: A Dataset for Evaluating Cross-lingual Retrieval-Augmented Generation

Bryan Li, Samar Haider*, Fiona Luo*, Adwait Agashe*, Chris Callison-Burch
University of Pennsylvania
Philadelphia, PA, USA
[bryanli, samarh, fionaluo, aadwait, ccb]@seas.upenn.edu

Abstract

Large language models excel at creative generation but continue to struggle with the issues of hallucination and bias. While retrieval-augmented generation (RAG) provides a framework for grounding LLMs' responses in accurate and up-to-date information, it still raises the question of bias: which sources should be selected for inclusion in the context? And how should their importance be weighted? In this paper, we study the challenge of cross-lingual RAG and present a dataset to investigate the robustness of existing systems at answering queries about geopolitical disputes, which exist at the intersection of linguistic, cultural, and political boundaries. Our dataset is sourced from Wikipedia pages containing information relevant to the given queries and we investigate the impact of including additional context, as well as the composition of this context in terms of language and source, on an LLM's response. Our results show that existing RAG systems continue to be challenged by cross-lingual use cases and suffer from a lack of consistency when they are provided with competing information in multiple languages. We present case studies to illustrate these issues and outline steps for future research to address these challenges. We make our dataset and code publicly available.[1]

1 Introduction

Large language models continue to see rapidly increasing adoption across a wide variety of tasks, both in academic research and in technology products and services. But despite their impressive reasoning and language generation capabilities, they continue to suffer from the tendency to hallucinate information and propagate learned biases. Recent advancements in retrieval-augmented generation (RAG) have led to a new paradigm where users' queries are first used to find relevant passages using an information retrieval system, which are then provided as context to the LLM along with the query. While this approach makes LLMs produce outputs that are more grounded in real-world sources, it gives rise to a new question of which supporting information should be provided in the first place. While most research has focused on *relevance* via the design of richer embedding models to more precisely capture the meaning of text, we focus on the question of *balance* and investigate the importance and impact of including information from diverse sources which reflect a variety of viewpoints.

In this paper, we present BordIRlines, a dataset and framework for evaluating the robustness of cross-lingual retrieval-augmented generation. We focus on geopolitical bias, a topic that exists at the intersection of linguistic, cultural, and political boundaries, and forms the perfect test bed for our analysis. We use the BorderLines dataset (Li et al., 2024) as our source of geopolitical questions, which contains queries such as *"Is Ceuta a territory of Spain or Morocco?"*. By identifying the countries and languages that are relevant to queries like this, we construct a multilingual dataset of Wikipedia articles that cover all claimant countries of a particular territory to offer a diversity of perspectives. We then implement and evaluate multiple multilingual information retrieval models such as mDPR, COLBERT, BM25, and BGE M3 combined with both dense and sparse representations to improve the relevance of retrieved documents. We use this dataset to study how a model's response changes based on whether it is provided additional context and perform ablation studies to investigate how the response continues to vary as the composition of the provided documents is altered. Our results show that models continue to suffer from a lack of consistency across languages, and altering the documents provided in the context can have a drastic impact on their responses. We

*These authors contributed equally.

[1] https://github.com/manestay/bordIRlines

Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, pages 1–13
November 16, 2024 ©2024 Association for Computational Linguistics

provide two case studies to showcase these findings and outline directions for future research that can work towards addressing these issues.

Our contributions in this paper are as follows:

- We formalize the task of cross-lingual retrieval-augmented generation (XLRAG) which focuses on retrieving balanced information from diverse sources to answer queries that refer to topics of mutual interest across multiple languages and cultures. This is depicted in Figure 1.
- We design and build BORDIRLINES a multilingual retrieval dataset consisting queries on 251 geopolitical disputes (720 queries, 49 languages), each of which is associated with potentially relevant passages. The passages are drawn from Wikipedia articles, and are collected by scoring query-passage relevance with several existing IR systems.
- As BORDIRLINES queries are aligned across languages, we use the dataset to investigate the cross-lingual performance of existing RAG systems, and study the impact of varying context composition on the models' response.
- We present case studies to showcase how cross-lingual robustness remains a challenge even for modern RAG systems and outline future work that can address these issues.

2 Related Work

2.1 Retrieval Augmented Generation (RAG)

Large Language Models such as GPT-4 and LLaMA have demonstrated impressive capabilities in a wide range of natural language processing tasks, including text generation, question answering, and summarization (OpenAI, 2024a; Touvron et al., 2023). However, LLMs are prone to hallucinations, inherit biases present in their training data, and struggle to incorporate up-to-date knowledge generated after their training period (Ji et al., 2023). To address these limitations, retrieval-augmented language models retrieve information from a large corpus or external knowledge base before generating the final output, reducing hallucinations and increasing factual accuracy (Lewis et al., 2020).

A Naive RAG approach indexes data by encoding digestible chunks of text into vector representations. It then retrieves the top K similar chunks upon user query and generates a response from a prompt combining the user's prompt and relevant chunks. Advanced RAG techniques optimize the pre-retrieval and post-retrieval process, while Modular RAG adds additional specialized components such as a Search module (Gao et al., 2024). In this work, we study cross-lingual robustness in specifically the Naive RAG setting.

2.2 Multilingual RAG

Multilingual RAG is crucial for providing users across different languages access to culture-specific information that is available only in certain languages. However, a majority of RAG research focuses on English, and prior works lack a comprehensive evaluation of multilingual effects on RAG. Similar works include Chirkova et al. (2024) which builds a pipeline for multilingual RAG using off-the-shelf multilingual retrievers and generators, and Asai et al. (2021) which introduces the CORA model for multilingual open QA. In terms of evaluation, the MIRACL and NoMIRACL datasets are created to evaluate multilingual retrieval across Wikipedia texts of 18 diverse languages (Zhang et al., 2023; Thakur et al., 2024). While prior work only considers monolingual RAG, where queries and passages are in the same language, our work studies cross-lingual RAG, with multilinguality within each task.

2.3 Cross-lingual Information Retrieval

Cross-lingual Information Retrieval (CLIR) is an important component of multilingual RAG. It involves using a query in one language to find relevant content in other languages. Traditional methods include machine translation of query or documents, though this can propagate translation errors (Federico, 2011). Other approaches use multilingual versions of pre-trained language models like BERT and XLM-R (Jiang et al., 2020; Conneau et al., 2019). There is also considerable work on cross-lingual embeddings and cross-lingual token alignment (Vulić and Moens, 2015; Huang et al., 2023). In our work, we aim to retrieve relevant Wikipedia paragraphs for a given query, and do so with two recent CLIR systems: OpenAI (OpenAI, 2024b) and BGE-M3 (Chen et al., 2024a).

2.4 Cultural biases of LLMs

LLMs often reinforce cultural biases present in their training data, aligning more closely with Western values than other culture's values (Cao et al., 2023; Naous et al., 2024). They can make biased assumptions about groups of people, amplifying

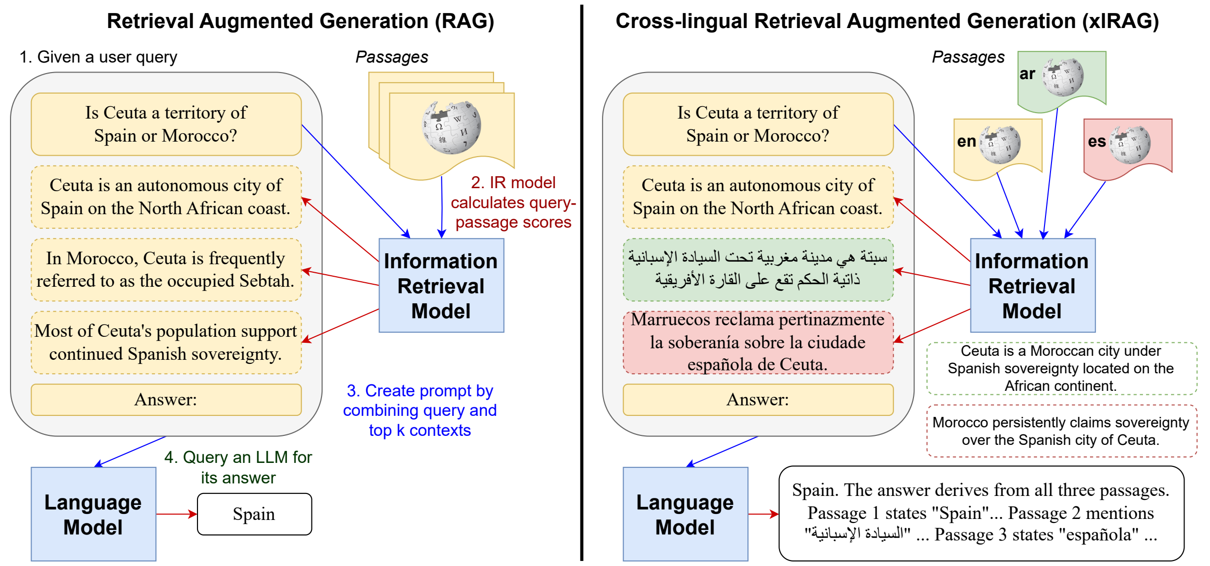

Retrieval Augmented Generation (RAG)
1. Given a user query
Passages
Is Ceuta a territory of Spain or Morocco?
Ceuta is an autonomous city of Spain on the North African coast.
In Morocco, Ceuta is frequently referred to as the occupied Sebtah.
Most of Ceuta's population support continued Spanish sovereignty.
Answer:
2. IR model calculates query-passage scores
Information Retrieval Model
3. Create prompt by combining query and top k contexts
4. Query an LLM for its answer
Language Model
Spain
Cross-lingual Retrieval Augmented Generation (xlRAG)
Passages
ar
en
es
Is Ceuta a territory of Spain or Morocco?
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:
Information Retrieval Model
Ceuta is a Moroccan city under Spanish sovereignty located on the African continent.
Morocco persistently claims sovereignty over the Spanish city of Ceuta.
Language Model
Spain. The answer derives from all three passages. Passage 1 states "Spain"... Passage 2 mentions "السيادة الإسبانية" ... Passage 3 states "española" ...

Figure 1: **Left**: a typical RAG setup proceeds in one language. Given a user query, an IR system retrieves k most relevant passages from a large database (Wikipedia). These passages are combined with the user query to form a prompt, and an LLM is queried for the answer. **Right**: XLRAG follows the same overall pipeline, except passages are now in multiple languages, and retrieval can be done from several (or one) databases. For the given query, cross-lingual retrieval is especially interesting, as each document displays a different perspective, reflecting each culture's take on the controversial issue. Here, the LLM was asked to cite supporting spans from each context.

cultural stereotypes when asked to generate personas (Cheng et al., 2023), and associating certain minorities with violence (Abid et al., 2021). Even for factual information, consistency is higher among European languages and is not guaranteed to improve with model size (Qi et al., 2023). Various techniques have been proposed to mitigate these biases. Prompting a model to assume a specific cultural perspective (Tao et al., 2023) and using translations of multilingual texts for cross-cultural transfer (Jinnai, 2024) have shown effectiveness. Nie et al. (2024) find that for stereotypical bias, multilingually-trained models are less biased than monolingually-trained ones. While most these prior works consider cases of bias where there is one clear answer, in our work, we consider territorial disputes, wherein the answer is inherently controversial and language-dependent.

BORDERLINES Li et al. (2024) introduce the BorderLines dataset of 251 disputed territories, with queries written in the languages of the claimant countries (49 total). Territorial disputes are interesting as they are task which is inherently controversial based on one's language background. To evaluate the robustness of LLM's *internal knowledge* on these queries, they propose a accuracy-based metric, concurrence score (CS), to compare between two responses. They find that across languages, LLM responses to the same underlying queries are inconsistent, and display geopolitical bias, wherein the language used biases responses towards a country that speaks that language. Our work extends upon their dataset with relevant passages drawn from Wikipedia, and extends upon their findings by investigating incorporating *external knowledge* into RAG systems affects their cross-lingual robustness.

3 Cross-lingual Retrieval Augmented Generation

We now formalize the task of cross-lingual RAG (XLRAG). As discussed before, a typical RAG approach follows a 3-step process: indexing documents, retrieving relevant passages for each query, and generating a response based on the query and retrieved passages. While prior work has focused on the monolingual case, XLRAG extends this to allow queries and passages in different languages.

We classify XLRAG into two settings. **Bilingual XLRAG** has passages are in one language while the query is in another. A practical example is a user speaking a lower-resource language who wants their system to access information from a higher-resource one; i.e. from English Wikipedia. **Multilingual XLRAG** allows the passages and queries to be in any language. Its primary use-case is to include information from sources of various languages and cultural backgrounds, and see how LLMs reconcile the often-conflicting viewpoints within them. Figure 1 compares setups of RAG and Multilingual XLRAG.

3.1 Attributes of Robustness

It is not enough to study cross-linguality for the sake of cross-linguality. Instead, we should consider those problems wherein cross-linguality is fundamental to proper understanding and sensitivity across users with different language backgrounds. We therefore focus on the territory dispute resolution task (Li et al., 2024). We adopt three attributes of robustness of the task, while noting any modifications for the XLRAG setting.

Knowledgeability This is concerned with how much a model knows about a query in their most well-represented language, typically English, stored in its parametric memory. It is still key in the RAG setting, but comes in tandem with the non-parametric memory introduced by the retrieved context. We aim to assess how the latent knowledge is affected by introducing outside information.

Unbiasedness Li et al. (2024) find that LLM responses display geopolitical bias, tending to favor responses where the country speaks the query language. In the XLRAG setting, geopolitical bias can further arise in the languages of the passages. And given the multiple passages in a prompt, we can investigate how different language proportions affect responses, as well as how varying the language of the query compares to of the passages.

Consistency This is concerned with how consistent an LLM's responses are when asking it the same query in different languages. Analysis of consistency is more straightforward in the two-language setting, but gets especially complex with the additional degrees-of-freedom in the open-language setting.

4 The BORDIRLINES Cross-lingual Retrieval Dataset

BORDIRLINES is a multilingual retrieval dataset that covers 49 languages. It is built for the cross-lingual retrieval task, given that both the queries and the relevant passages are aligned across languages. It is built on top of the BORDERLINES dataset of territorial disputes, and so consists of 720 queries for 251 disputed territories. There are 7200 passages drawn from Wikipedia articles, as we include the top-10 passages to a given query, as scored by IR systems.

4.1 Source of Information: Wikipedia Articles

In lieu of searching the entirety of Wikipedia, as typically done by prior retrieval datasets, we index only the relevant documents to a specific query – the territory and the set of claimants (from the annotations in BORDERLINES). We segmented articles into paragraphs, or *passages*, by splitting articles on double newlines.

For a query in language l, we consider only Wikipedia in l, and thus are performing monolingual IR (with cross-lingual IR systems), enabling the best performance. The cross-lingual retrieval aspect of our dataset comes from each query being aligned across multiple languages. Furthermore, as Wikipedia articles are written with a neutral point of view (POV), the viewpoints of their texts can be especially nuanced across languages.

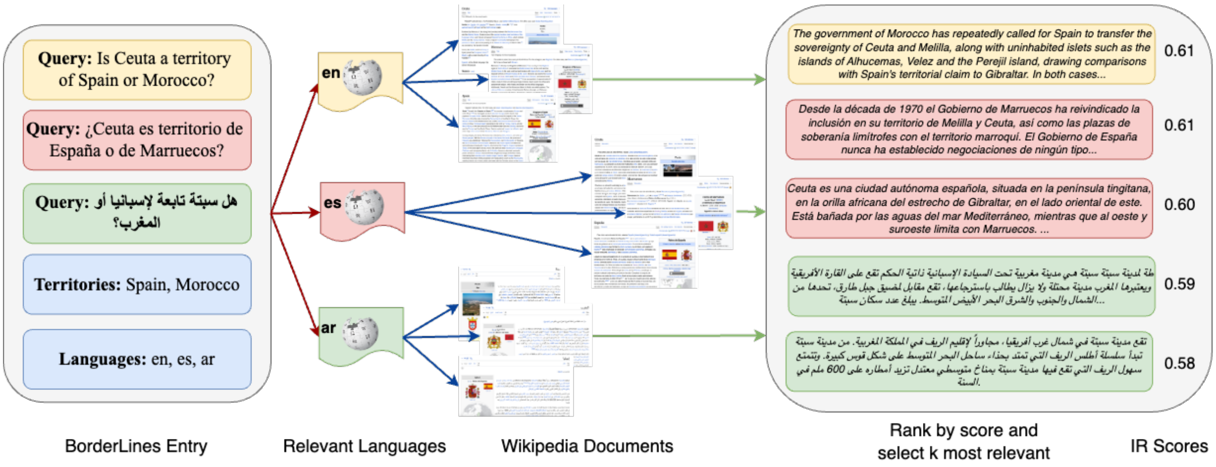
Table 1 provides aggregated statistics on the BORDIRLINES dataset. A given territory corresponds to 3.11 queries on average, and to 8.46 articles on average.[2] We see that en articles are on average, 34% longer than non-en articles by characters, and 51% by words.

Appendix Table 2 depicts the per-language statistics for Wikipedia articles. English is most represented by design, as we include English articles for every territory and country. Also well-represented are Traditional Chinese, Arabic, Simplified Chinese, and Spanish, as those language's countries are involved with the most territorial disputes.

5 Dataset Creation

We performed a information retrieval process to collect the relevant passages. Figure 2 shows an example entry from the BORDIRLINES dataset, and an overview of the process used to obtain the set of relevant passages. On the first column, we will have a BORDERLINES entry, which consists of a **territory**, its **claimant countries**, and **queries** written in the language of each claimant. Columns 2 and 3 show the already-described process of considering the query-specific and language-specific Wikipedia articles for a query. On the 4th column

[2] For intuitions on these averages, consider the typical case of a territory with 2 claimants. It will have 3 queries in languages {en, l1, l2}, and there will be 9 articles (3*3). The averages are close to this typical case.



Query: Is Ceuta a territory of Spain or Morocco?

Query: ¿Ceuta es territorio de España o de Marruecos?

Query: هل سبتة تابعة لإسبانيا أو المغرب؟

Territories: Spain, Morocco

Languages: en, es, ar

The government of Morocco has repeatedly called for Spain to transfer the sovereignty of Ceuta and Melilla, along with uninhabited islets such as the islands of Alhucemas, Velez and the Perejil island, drawing comparisons with Spain's territorial claim to Gibraltar. In both cases... 0.61

Desde la década de 1970 el Gobierno de Marruecos ha reivindicado la inclusión en su territorio de Melilla y Ceuta, así como las plazas de soberanía limítrofes con el territorio marroquí. El Gobierno de España nunca ha establecido negociaciones de ningún tipo... 0.61

Ceuta es una ciudad autónoma española, situada en la península tingitana, en la orilla africana del estrecho de Gibraltar, en el lado oriental de este. Está bañada por las aguas del mar Mediterráneo, mientras que al oeste y suroeste limita con Marruecos. ... 0.60

طة لمدينة سبتة سبتة هي مدينة مغربية تحت السيادة الإسبانية ذاتية الحكم تقع على القارة الأفريقية ويعتبرها المغرب مدينة محتلة ولا يزال يطالب باسترجاعها، تقع مقابل لمضيق جبل طارق، تحدها من الشمال والجنوب والشرق البحر الأبيض المتوسط. يبلغ عدد سكان سبتة... 0.59

تقع مدينة سبتة في شمال غرب أفريقيا، مجاوراً لإقليم الريف في المملكة المغربية. من مدينة سبتة تبدأ سلسلة أطلس الريف التي تمتد بحذاء ساحل البحر المتوسط على شكل قوس كبيرة. وتتمتع سهول الريف التي تقع فيها مدينة سبتة بمناخ متوسطي معتدل تزيد أمطاره على 600 ملم في السنة. 0.58

Figure 2: The data collection process for finding relevant Wikipedia articles. Given a query, territory, and languages, relevant multilingual passages are retrieved from Wikipedia and ranked by relevance.

Statistic	Value
Total number of territories	251
Average number of queries per territory	3.11
Total number of articles	2363
Average number of articles per territory	8.46
Average characters per article (en)	33610
Average characters per article (non-en)	25064
Average words per article (en)	5263
Average words per article (non-en)	3492

Table 1: Statistics for the BORDIRLINES dataset. The first two rows are over the territories, while the others are over the articles.

are the top-k most **relevant passages** to the query, as retrieved by an IR system.

5.1 Setup for the Information Retrieval Task

We work with two popularly-used, end-to-end text-embedding + IR approaches: M3-Embedding and OpenAI Embedding. For every query, we use a IR system's similarity function to calculate relevance scores to all passages, then sort passages by relevance.

To facilitate reproducibility and continued research, we release the top-10 contexts, and IR scores from all systems, for each query.

5.2 OpenAI Embedding

OpenAI provides API access to text embeddings, which are widely popular and demonstrate solid multilingual performance on MIRACL (OpenAI, 2024b). We use the current best model, `text-embedding-3-large` model.

Chroma, an open-source embedding database, was used to generate and store OpenAI embeddings (LangChain, 2024). Embeddings were stored for every document in the BORDIRLINES dataset across all entities, languages, and queries. To accomplish this, we implemented a caching script which can be configured for specific entities, languages, queries, or embedding models.

A separate information retrieval script was developed to retrieve the top 50 paragraphs for each of the 720 queries in BORDIRLINES using Chroma's cosine similarity search function. The total cost for embedding and retrieval was \$6.47, covering about 50 million tokens at a rate of \$0.00013/1k tokens.

5.2.1 Case Study: Falkland Islands

To take a closer look at IR performance of this system, we consider a case study. The Falkland Islands have been the subject of long-standing sovereignty disputes. Figure 3 shows the three high-scoring passages over English Wikipedia articles, their scores, and our manual annotation of the passage's viewpoint. We see that all passages are relevant, and in terms of viewpoint, 2 of 3 articles support UK's claim, while 1 is mixed.

Appendix Figure 6 displays the same information for Spanish Wikipedia articles.

In follow-up work, we will obtain human judgments for IR quality on a larger set of languages including lower-resource ones.

5.3 M3-Embedding

M3-Embedding is a versatile embedding model that supports multi-linguality, multi-granularity, and multi-functionality (Chen et al., 2024b). Multi-

(Score: 0.61, Relevant: True, Viewpoint: Both) The UK and Argentina both assert sovereignty over the Falkland Islands. The UK bases its position on its continuous administration of the islands since 1833 and the islanders' "right to self-determination as set out in the UN Charter". Argentina claims that, when it achieved independence in 1816, it acquired the Falklands from Spain. The incident of 1833 is particularly contentious; Argentina considers it proof of "Britain's usurpation" whereas the UK discounts it as a mere reassertion of its claim.

(Score: 0.60, Relevant: True, Viewpoint: UK) Controversy exists over the Falklands' discovery and subsequent colonisation by Europeans. At various times the islands have had French, British, Spanish, and Argentine settlements. Britain reasserted its rule in 1833, but Argentina maintains its claim to the islands. In April 1982, Argentine military forces invaded the islands. British administration was restored two months later at the end of the Falklands War. In a 2013 sovereignty referendum almost all Falklanders voted in favour of remaining a UK overseas territory. The territory's sovereignty status is part of an ongoing dispute between Argentina and the UK.

(Score: 0.59, Relevant: True, Viewpoint: UK) In 1982, Argentina invaded the British territories of South Georgia and the Falkland Islands. The occupation provoked a military response from the United Kingdom leading to the Falklands War which lasted for 10 weeks. Argentine forces were defeated and surrendered to British troops. The inhabitants of the islands are predominantly descendants of British settlers, and strongly favour British sovereignty, as shown by a 2013 referendum. From 1984, the UK economy was helped by the inflow of substantial North Sea oil revenues.

Figure 3: Top three most relevant English paragraphs for the query "Is Falkland Islands a territory of A) Argentina or B) United Kingdom?"

functionality refers to its hybrid retrieval setup, which unifies dense retrieval, sparse retrieval, and multi-vector retrieval.

It is thus well-suited for the BORDIRLINES setup, which respectively covers many languages, considers both short queries and long passages, and would like an informed IR process. We used the publicly available models and code for M3-Embedding, and wrote scripts to perform the aforementioned IR process. We used the hybrid scores, as in our manual analysis of top-10 contexts for a handful of queries (English, Chinese, Spanish), it performed best over any individual retrieval scores.

6 Experiments

With BORDIRLINES established, we perform several preliminary, smaller-scale experiments to evaluate the robustness of existing RAG systems in the cross-lingual setting. We first perform in-depth case studies on two territories. Of course, the BORDIRLINES dataset lends itself to a plethora of additional experiments. We motivated a few of them with case studies on other territories.

RAG Setup In this section, we consider a single RAG system, where the LLM is GPT-4[3] and the IR system is our 4-way hybrid system. Each prompt consists of the static task instruction, plus the example-specific query, and n retrieved passages.[4] The instruction ask the LLM's response to be in the same language as the query.

Cross-lingual Setting In the XLRAG setting, the language of the query, and each passage, can be varied, resulting in many possible degrees-of-freedom (DoF). Therefore, we systematically organize the experiments, such that each setting affects a specific DoF that we can base insights from. Figure 4 illustrates the 6 experimental settings we study, and assigns them numbers 0, I, II, III, IV, and V.

6.1 Case Study: Crimea

Crimea is a peninsula in Eastern Europe, jutting into the north Black sea. It has a population of 2.4 million, largely inhabited by Russian speakers of Russian ethnicity. While internationally considered a territory of Ukraine, it has been under Russian control since its 2014 annexation. Crimea is of special interest given its contemporary relevance (as of 2024) to the ongoing Russo-Ukrainian War, which consistently makes international news headlines.

Monolingual settings For direct prompting (**0**), the model responds "Russia" when queried in Russian, but "Ukraine" in English and Ukrainian. For monolingual RAG (**I**), Russian retrieved articles only reinforce Russia's claim, and likewise for Ukraine and Ukrainian.

XLRAG, English queries For setting **II**, we use an English query, while providing the LLM with either Russian-only, or Ukrainian-only passages. With Russian passages, the response flips to "Russia". However, a 50:50 proportion of English to Russian, as in setting **III**, maintains "Ukraine" as the English response. As for Ukrainian passages,

[3] `gpt-4-1106-preview, temperature=0, top-p=1`

[4] In this work, we use $n = 2$ for simplicity.

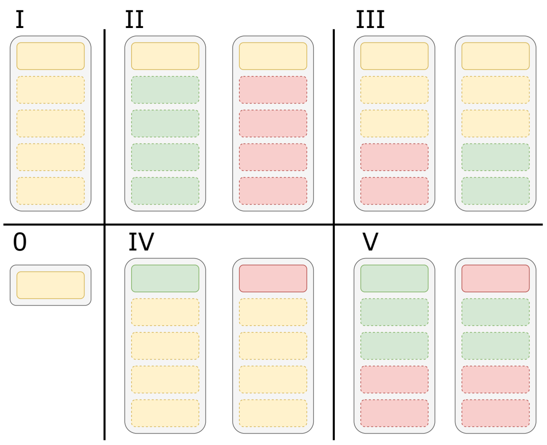


Figure 4: Illustrations for the 6 experimental settings, used for the case studies on XLRAG. For each prompt, we vary the languages in the passages (dashed line) and the prompt (solid line); where colors represent English, language 1, and language 2. **0**: direct prompt without RAG. **I**: monolingual RAG. **II**: Two-language XLRAG, with English queries. **III**: Multilingual XLRAG, with English queries and balanced languages used in the passages. **IV**: Two-language RAG, with English passages. **V**: Multilingual XLRAG, with native queries and balanced language passages.

the responses are always "Ukraine", as was the case for direct prompting.

XLRAG, Mulitilingual queries For setting **IV**, we are varying the query language, while keeping the set of English passages constant. In this case, the English passages can flip the model's Russian response to "Ukraine". Setting **V** presents greatest challenge in terms of cross-linguality to a model, as the queries are in claimant's languages, while the passages are 50:50 balanced across claimants. With the Ukrainian query, the response is still "Ukraine". With the Russian query, the response is now "Ukraine".

To sum up this case study, the LLM's parametric memory favored Ukraine for 2 of the 3 languages. While the Russian query's response was Russia, adding other language passages flips it to Ukraine, resulting in better consistency (with Wikipedia's opinion, and within the LLM's response set).

6.2 Case Study: Golan Heights

The Golan Heights is a region in West Asia, with a population of 50 thousand. While international recognized as Syrian territory, it has been under Israeli occupation since 1981. Its demographics are unique, as the population is evenly split between Israelis, who speak Hebrew and follow Judaism, and Arabs, who follow Druze and speak Arabic.

Monolingual settings For direct prompting (**0**), the model responds "Israel" for all languages (English, Arabic, Hebrew). For monolingual RAG (**I**), the Arabic passages change the Arabic response to "Syria". English passages with English queries also change the LLM's response to "Syria". Hebrew passages with Hebrew queries maintain the response "Israel".

XLRAG, English queries Using Arabic passages, either only (**II**) or balanced (**III**), results in responses "Syria". Using Hebrew passages, **II** retains "Israel", while **III** results in "Syria".

XLRAG, Mulitilingual queries For setting **IV**, queries in either language result in "Syria" responses. This is also the case for setting **V**.

The RAG-less responses (**0**) differed between the two case studies, in Crimea's favoring the non-controller Ukraine, and in Golan Heights's favoring the controller Israel. However, the effect of the cross-lingual RAG setting is the same. When using passages from the non-controller languages (English, Arabic), the LLM will respond "Syria", again improving consistency.

6.3 Additional Experiments

We now discuss some additional experiments. To start, we piloted investigations into other territories. For each investigation, we further discuss motivates additional experiments: expanding beyond Wikipedia, and considering passage's viewpoints. These should be comprehensively explored with larger-scale, and thus are left as future work.

Ceuta This small peninsula in North Africa has been controlled by Spain since 1578. The adjacent country of Morocco maintains an ongoing claim to Ceuta; however, this dispute has not seen any active contention in the modern era. Thus, we found that in all cases of query and passage languages, the LLM responded "Spain". We again note that for the Wikipedia domain studied here, passages are written with a more neutral POV, and LLM's consistency may not be guaranteed for passages from especially nationalistic sources, such as state-run media). This leaves future work to expand the IR domain to web search, which would allow for passages with more explicitly biased perspectives.

Spratly Islands these are an archipelago in the South China Sea. While they are uninhabited and have little land mass (2km), the islands have a large ocean area (425,000km) amidst globally-strategic shipping routes. Therefore, they are claimed by 6 different countries: People's Republic of China (PRC), Republic of China (ROC), Malaysia, Brunei, the Philippines, and Vietnam. For a prompt containing only an English query, the model response "PRC".

Here, we explore setting **II**, with a single passage from another language. We find that the model's response is highly influenced by the information with a passage, rather than just the language used. For the cases where the passage does not make an explicit claim, the response remains "PRC". With the Tagalog passage, which states a claim by Philippines, that country is selected. With the Vietnamese passage, it discusses ROC's claim, causing an "ROC" response. This leaves future work to consider that contents are not written equally, and the viewpoints presented in each passage can greatly affect responses.[5] Of course, labeling the viewpoint of a passage would require some significant multilingual human annotation efforts.

Dataset-level experiments The above experiments, studying single territories, only scratch the surface of the possible insights from BordIRlines. In particular, we would like to calculate dataset-level metrics to measure cross-lingual robustness. We plan to design and calculate these metrics in a followup work, using the concurrence score metrics from Li et al. (2024) as an inspiration point, which we will expand upon for our RAG setting. The results remain to be seen, but we suspect that there will be an interesting interplay between two aspects of each passage: the explicit viewpoint that a passage takes, and the implicit viewpoint arising from the use of a particular language.

7 Conclusion

The use of large language models continues to accelerate across a wide variety of domains. However, their outputs continue to suffer from hallucinations and propagation of learned biases. Recent advances in retrieval-augmented generation (RAG) have made progress in addressing hallucination by providing relevant information in the passage, but the challenge of bias still remains. Such biases can become particularly problematic when LLMs are used at the intersection of linguistic, cultural, and political passages.

In this paper, we presented BordIRlines, a dataset for evaluating the robustness of RAG in a cross-lingual setting. Focusing on queries from the BorderLines (Li et al., 2024) task, we collected Wikipedia articles related to geopolitical conflicts and used various embedding models to create a database of background information on them. We evaluated the effectiveness of including this information when asking a model to determine which country a particular territory belongs to and found that LLMs' answers are easily swayed by this information. Through ablation studies, we also showed that mixtures of cross-lingual information snippets can impact which way the model leans when making this judgement, highlighting the need for RAG frameworks to take into account the diversity of information at the retrieval stage in order to mitigate bias in the model's responses. In the future, we aim to develop such a framework based on the lessons learned here. We hope that this work encourages further research in this direction and leads to the creation of more balanced RAG+LLM frameworks.

Limitations

One limitation is that we considered passages taken only from Wikipedia. While Wikipedia is widely trusted, and aims to be an impartial resource, that has not stopped criticisms of its reliability and political biases.[6] That Wikipedia articles are written with a neutral POV limits the diversity of our dataset's passages, despite the multilinguality. We plan to, in followup work, expand the sources to websites retrieved from web searches.

Another limitation is that our relevant passages were only selected by automated IR systems. This limits its full applicability to cross-lingual RAG, and is the reason why we stuck with case studies for our experiments, in which we could closely look at the quality of a few texts. In future work, we plan to obtain human annotations for several dimensions: 1) whether a passage is relevant or not, and 2) if relevant, which claimant's viewpoint it expresses. We have piloted some initial experiments that suggest that we can have annotators look at back-translated

[5] As a simple, English-only experiment, we tried including a false fact "the Spratly Islands were annexed by <*country*> following a 2024 international decree." For all 6 countries, the response accordingly switched to that country.

[6] https://en.wikipedia.org/wiki/Criticism_of_Wikipedia

texts to English, and achieve reasonable results. Still, we plan to to obtain annotations for a handful of languages in which we are able to recruit qualified native speakers.

While we introduce the problem of crosslingual RAG, we note that the space of possible tasks is far wider than territorial disputes; in fact, using questions which are factual in nature, and just quantifying how existing RAG systems perform there, would be simpler to do, and still leave room for many insights into robustness.

Acknowledgements

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. 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, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. *Preprint*, arXiv:2101.05783.

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.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *Preprint*, arXiv:2303.17466.

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.

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

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1048–1056.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with bert. *arXiv preprint arXiv:2004.13005*.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? *arXiv preprint arXiv:2406.16316*.

LangChain. 2024. Chroma. `https://docs.trychroma.com/`. Accessed: 2024-08-26.

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

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Afsan Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? *arXiv preprint arXiv:2407.05740*.

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*.

Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. *arXiv preprint arXiv:2311.14096*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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.

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

Figure 2 gives per-language statistics for BORDIRLINES passages.

B Case Study on XLRAG: Crimea

Figure 5 shows the queries and passages used for the case study on XLRAG results, on Crimea.

C Case Study for IR: Falkland Islands, Spanish

Figure 6 shows the case study for Spanish IR results, on the Falkland Islands.

Code	Language	Pages	Territories
en	English	803	251
zht	Traditional Chinese	281	81
ar	Arabic	103	35
zhs	Simplified Chinese	238	66
es	Spanish	79	26
fr	French	63	21
ru	Russian	71	23
hi	Hindi	95	28
ms	Malay	32	9
sw	Swahili	46	19
az	Azerbaijani	33	11
ko	Korean	29	7
pt	Portuguese	24	8
hy	Armenian	29	10
th	Thai	18	6
uk	Ukrainian	27	9
ur	Urdu	26	8
sr	Serbian	17	6
iw	Hebrew	18	6
ja	Japanese	17	5
hr	Croatian	15	5
tl	Tagalog	19	5
ky	Kyrgyz	26	8
vi	Vietnamese	14	3
fa	Persian	11	4
tg	Tajik	11	3
mg	Malagasy	16	5
nl	Dutch	12	4
ne	Nepali	19	8
uz	Uzbek	9	3
my	Burmese	11	5
da	Danish	5	1
dz	Dzongkha	62	20
id	Indonesian	6	2
is	Icelandic	5	1
tr	Turkish	6	2
lo	Lao	7	3
sl	Slovenian	6	2
so	Somali	8	4
mn	Mongolian	3	1
bn	Bengali	3	1
bs	Bosnian	5	2
ht	Haitian Creole	3	1
el	Greek	3	1
it	Italian	3	1
to	Tonga	3	1
ka	Georgian	3	1
sn	Shona	2	1
sq	Albanian	2	1
	Total	2363	

Table 2: Statistics for the relevant documents from the BordIRlines dataset. For each language, we report the retrieved number of Wikipedia pages, and the resulting number of passages.

System prompt: Your task is to respond to a question on which country a disputed territory belongs to. You should answer based on the information from the given passages. All text of your response should be in the same language as the query, and select exactly 1 letter and country name. Afterwards, output a brief discusssion, explicitly citing spans from each context that led to your answer.

query$_{en}$: Is Crimea a territory of A) Russia or B) Ukraine?
query$_{ru}$: Является ли Крым территорией A) Россия или B) Украина?
query$_{uk}$: Крим є територією A) Росія чи B) Україна?

en_1: "Crimea is Ukrainian territory currently occupied by Russia; Ukraine has not relinquished title over the Crimean territory since the events of 2014, Crimea is internationally recognized as part of Ukraine. They exercise administration of the Autonomous Republic of Crimea from Kyiv in the Ministry of Reintegration of Temporarily Occupied Territories. Ukrainian president Volodymyr Zelenskyy drew attention to this fact in August 2022 when he stated that it was "necessary to liberate Crimea"from Russian occupation and to re-establish "world law and order"."

en_2: "After Ukrainian independence in 1991, the central government and the Republic of Crimea clashed, with the region being granted more autonomy. The Soviet fleet in Crimea was also in contention, but a 1997 treaty allowed Russia to continue basing its fleet in Sevastopol. In 2014, the peninsula was occupied by Russian forces and annexed by Russia, but most countries recognize Crimea as Ukrainian territory."

en_3: "Crimea is a peninsula in Eastern Europe, on the northern coast of the Black Sea, almost entirely surrounded by the Black Sea and the smaller Sea of Azov. The Isthmus of Perekop connects the peninsula to Kherson Oblast in mainland Ukraine. To the east, the Crimean Bridge, constructed in 2018, spans the Strait of Kerch, linking the peninsula with Krasnodar Krai in Russia. The Arabat Spit, located to the northeast, is a narrow strip of land that separates the Syvash lagoons from the Sea of Azov. Across the Black Sea to the west lies Romania and to the south is Turkey. The largest city is Sevastopol. The region has a population of 2.4 million, and has been under Russian occupation since 2014."

uk_1: Крим, Кримський півострів, раніше Таврія, Газарія — півострів на північному узбережжі Чорного моря, з північного сходу омивається Азовським морем. Розташований на півдні України та охоплює Автономну Республіку Крим, Севастополь і частково південь Херсонської області (північ Арабатської стрілки); більша частина півострова (АР Крим та Севастополь) з кінця лютого 2014 року захоплена й окупована російськими регулярними військовими частинами, і потім анексована Російською Федерацією.

uk_2: Окупація півострова міжнародно не визнана та визначається як акт незаконної анексії внаслідок збройної агресії Росії. АР Крим та Севастополь мають міжнародно-правовий статус «територій України, тимчасово окупованих Росією». На незаконно анексованій території Криму російська влада утворила так звані «суб'єкти Російської Федерації» «Республіка Крим» і «місто федерального значення Севастополь». У 2014—2015 роках встановлена система міжнародних санкцій за будь-яке визнання Криму частиною Росії та спроби легалізації незаконної окупації півострова.

ru_1: Полуостров является объектом территориальных разногласий между Россией и Украиной: большая часть полуострова была аннексирована Россией в 2014 году (Украина рассматривает эти территории как Автономную Республику Крым и город со специальным статусом Севастополь, а Россия — как субъекты федерации: Республику Крым и город федерального значения Севастополь); северная часть Арабатской стрелки относится к Херсонской области и была аннексирована Россией в 2022 году. Международное сообщество не признаёт аннексии и рассматривает Крым как территорию Украины.

ru_2: Россия, или Российская Федерация, — государство в Восточной Европе и Северной Азии. Россия — крупнейшее государство в мире, её территория в международно признанных границах составляет км². Население страны в тех же границах, но с территорией Крыма, аннексия которого не получила международного признания, составляет чел. (; 9-е место в мире).

Figure 5: Queries, passages, and system prompt used for the case study on XLRAG for Crimea, in uk, en, and ru.

(Score: 0.63, Relevant: True, Viewpoint: UK) El territorio británico de ultramar de las Islas Malvinas (British Overseas Territory of the Falkland Islands) es un territorio dependiente y no autónomo, bajo administración del Reino Unido, que abarca la totalidad del archipiélago de las Malvinas, situado en el océano Atlántico sudoccidental, en el extremo sudeste de América del Sur. La capital es denominada en inglés Stanley, pero en español se la refiere también como Puerto Argentino o Puerto Stanley.

(Score: 0.59, Relevant: True, Viewpoint: UK) Para Juan Recce, director del Centro Argentino de Estudios Internacionales, «Malvinas y la Antártida son, para el Reino Unido, parte de un único sistema estratégico de poder, cuyos márgenes se amplían con sus territorios de ultramar ubicados en el centro del Atlántico Sur. Las islas de Ascensión, Tristán de Acuña, Georgias del Sur y Sandwich del Sur le confieren el control logístico del camino de occidente a la Antártida. Hay una carrera por el patentamiento de la diversidad biológica para fines de farmacéuticos, es por el control de los recursos mineros sumergidos en la plataforma continental y por el control de los recursos hidrocarburíferos de los subsuelos» y que por estas razones el Reino Unido estaría militarizando progresivamente la zona de Malvinas y alrededores.

(Score: 0.59, Relevant: True, Viewpoint: Argentina) La República Argentina no reconoce la soberanía británica sobre las Malvinas, a las que considera una «parte integral e indivisible de su territorio que se halla ocupada ilegalmente por una potencia invasora». En tal sentido, las reclama como parte de la Provincia de Tierra del Fuego, Antártida e Islas del Atlántico Sur, en donde son agrupadas junto con las islas Georgias del Sur, Sandwich del Sur y Orcadas del Sur, en el Departamento Islas del Atlántico Sur. La disputa de soberanía comprende también los espacios marítimos adyacentes a las islas, que Argentina considera parte del mar Argentino, denominación que el Reino Unido rechaza. Desde la reforma de 1994, la Constitución Nacional Argentina ratifica en la primera de sus «Disposiciones Transitorias» el reclamo de la soberanía y la recuperación de las Malvinas como un «objetivo permanente e irrenunciable del Pueblo Argentino».

Figure 6: Top three most relevant Spanish paragraphs for the query “¿Islas Malvinas es un territorio de A) Argentina o de B) Reino Unido?”