

Lost in Translation? Exploring the Effectiveness of a Multilingual RAG-Based Travel Assistant Chatbot

Jade Oakes

jadeoakes@brandeis.edu

May 6, 2025

Abstract

This project presents a multilingual RAG-based chatbot for language learning and travel assistance in English and Spanish. It combines prompt-based generation with optional retrieval from (1) multilingual parallel sentence pairs and (2) travel-related passages obtained from Wikivoyage. I evaluated 240 responses across three modes and two decoding settings, annotating each for correctness, fluency, relevance, helpfulness, conciseness, language accuracy, and completeness. While outputs were consistently fluent and concise, many lacked correctness or relevance, and over half were cut off. Semantic similarity analysis revealed weak alignment between responses and retrieved content, suggesting limited grounding. These results highlight challenges in building grounded multilingual systems and suggest that better prompts, output control, and use of context could help improve performance. A Streamlit interface was developed to enable interactive testing and refinement of the system.

1 Introduction

Multilingual chatbots have become popular among language learners and travelers to help navigate different situations. However, many rely solely on generative models without incorporating external knowledge, resulting in generic, inauthentic, or factually inaccurate responses. This project presents a multilingual travel assistant chatbot that combines Retrieval-Augmented Generation (RAG) with prompt engineering to provide grounded, context-aware answers in English and Spanish.

2 Related Work

The following papers highlight the potential and limitations of chatbots in language learning. (Huang et al., 2022) conducted a review of chatbots for language learning. They discussed advantages such as personalization, timeliness, and ease of use, but also discussed challenges like limited chatbot

intelligence and lack of authentic, diverse language use. (Lewis et al., 2020) introduced RAG as an effective method for combining pre-trained language models with external knowledge retrieval, significantly reducing hallucinations and improving factual correctness. With RAG, the knowledge base can be updated without having to retrain the model. (Matar and Mohammad, 2024) demonstrated the benefits of RAG in chatbots through the AWAIT project, showing how RAG integration improves chatbot reliability and reduces misinformation, particularly for providing academic writing help.

3 Data

3.1 Multilingual Sentence Pairs

To support language learning and general grounding, I used parallel sentence pairs from three multilingual datasets: Tatoeba (Tatoeba Project, n.d.), WikiMatrix (Schwenk and Douze, 2020), and OpenSubtitles (Lison and Tiedemann, 2024). Due to the large size of the original corpora, I limited each source to approximately 300,000 samples. Table 1 summarizes the resulting statistics.

Source	Samples	Tokens
Tatoeba	204,102	3,016,625
WikiMatrix	271,275	12,579,975
OpenSubtitles	347,256	5,658,223
Combined	822,633	21,254,823

Table 1: Sentence pair statistics after filtering and sampling.

3.2 Travel Knowledge

For the travel assistant mode, I scraped over 22,000 English and Spanish pages from Wikivoyage (Wikimedia Foundation, n.d.). These were preprocessed, split into manageable passages of around 300 tokens, and stored in a combined JSONL file. Initially, I used NLTK to split by sentence, but I up-

dated the approach to also split on newlines to better handle irregular formatting and long text spans.

Table 2 summarizes the resulting dataset.

Dataset	Chunks	Tokens	Source Pages
Wikivoyage (raw)	22,874	30,995,867	21,880
Travel Info (chunked)	166,919	29,793,905	21,880

Table 2: Wikivoyage-based data before and after chunking.

3.3 Implementation Notes

For generation, I initially tested LLaMA 3.2 3B and 3.1 8B (Touvron et al., 2024), but switched to Mistral 7B Instruct v0.3 (AI, 2024) due to its stronger performance and faster inference. I used Hugging Face’s Transformers pipeline (Wolf et al., 2020) with small batch sizes (batch size = 2) to reduce memory usage and improve stability. Experiments were run on a university-managed GPU cluster using the Slurm workload manager.

4 Methodology

The chatbot system is designed to handle both travel-related and language learning queries in English and Spanish. It follows a RAG framework that combines semantic search over external knowledge sources with large language model generation. The pipeline consists of three main stages: context retrieval, prompt construction, and response generation.

4.1 Retrieval Component

For each user query, the system encodes the text using the SentenceTransformer model all-MiniLM-L12-v2 (Wang et al., 2020). It then retrieves the most relevant entries by comparing the query embedding against a pre-built FAISS index (Johnson et al., 2017), enabling efficient top- k semantic search.

Two types of knowledge are available:

- **Multilingual sentence pairs** from Tatoeba (Tatoeba Project, n.d.), WikiMatrix (Schwenk and Douze, 2020), and OpenSubtitles (Lison and Tiedemann, 2024), to support language-related tasks.
- **Travel data** scraped from Wikivoyage (Wikimedia Foundation, n.d.) in English and Spanish, chunked into ~300-token passages for geographic and cultural grounding.

Depending on the query and system mode, the appropriate index is selected and the top- k relevant chunks are returned to guide generation.

4.2 Prompt Construction

Prompts are built automatically using predefined templates. Each prompt includes a task instruction, an optional context block (containing retrieved information), and the user query. The prompt format is determined by three modes:

- `general` – uses sentence pair retrieval for translation, grammar checking, etc.
- `travel` – uses Wikivoyage passages for travel-related Q&A
- `no_retrieval` – omits retrieved context and relies solely on the LLM

I tested different prompt templates to see how they affected the quality, fluency, and relevance of the responses.

4.3 Response Generation

Responses were generated using the Mistral 7B Instruct v0.3 model via Hugging Face’s Transformers pipeline. I applied two decoding strategies: deterministic and sampled, to explore how generation settings affect response quality. Responses were post-processed to remove unnecessary formatting. Full experimental details are provided in Section 5.

Example: Prompt and response with retrieved context

Query: What is the best time of year to visit Ecuador?

Mode: travel

Prompt: You are a multilingual travel assistant. Provide helpful information about destinations and travel tips.

Retrieved Context (excerpt): Ecuador has two main seasons: dry (June–September) and wet (October–May). The dry season is ideal for visiting the Galápagos, while the Amazon is better in the wet season. Temperatures are relatively consistent year-round.

Model Output: The best time to visit Ecuador is during the dry season, especially from June to September. This is when the Galápagos Islands have stable weather and less rain. The highlands and Amazon region can be visited year-round, depending on your interests.

(This response was marked as fluent and helpful, with strong alignment to the retrieved content.)

5 Experiments

5.1 Experimental Setup

To evaluate the chatbot’s performance, I generated a total of 240 responses across three operational modes: *general*, *travel*, and *no_retrieval*. Each mode used multiple prompt templates (2 for general, 2 for travel, and 4 for no_retrieval), with each prompt paired with 10 queries.

The queries were designed to reflect two key use cases: language learning (e.g., translation, grammar correction, clarification) and travel assistance (e.g., directions, food, activities). I used 10 English and 10 Spanish queries applied across the different prompt-mode combinations. The full list of prompts and queries is included in Appendix B.

Two decoding settings were tested:

- **Deterministic:** `do_sample=False`
- **Balanced Sampling:** `do_sample=True, top_p=0.85, temperature=0.6`

Each unique combination of query, prompt, and generation setting produced one response, resulting in 240 samples. Generations were run in small batches using Hugging Face’s Transformers pipeline and saved for manual evaluation.

5.2 Annotation and Evaluation

Each response was manually annotated across seven dimensions:

- **Correctness:** factual or linguistic accuracy
- **Fluency:** grammaticality and naturalness
- **Relevance:** alignment with the user’s query
- **Helpfulness:** whether the response met the user’s goal
- **Conciseness:** clarity and efficiency without extra detail
- **Expected Language:** whether the response was in the correct language
- **Cut-off:** whether the response was complete or truncated

Most categories were scored on a 1–5 scale, except for *Expected Language* and *Cut-off*, which were binary (0 or 1). Detailed scoring rubrics for each dimension are provided in Appendix A. I completed all annotations manually after all response generations were produced.

5.3 Comparison Dimensions

To better understand what factors influence response quality, I evaluated responses across the following dimensions:

- **Generation setting:** Comparing deterministic and balanced decoding strategies to assess how sampling impacts factual accuracy, relevance, and helpfulness.
- **Chatbot mode:** Analyzing results across three operational modes—general, travel, and no_retrieval—each with different retrieval and prompt configurations.
- **Prompt template:** Examining how mode-specific prompt wording influences language quality and response grounding.

For each setting and mode, I computed human evaluation statistics (mean, standard deviation, and score distributions) across five dimensions: correctness, fluency, relevance, helpfulness, and conciseness. I also analyzed system-level metrics, including cut-off rate and expected language match, across all 240 samples. Bar charts illustrating evaluation scores by mode and generation setting are shown in Figures 1 and 2 respectively.

6 Results

6.1 Overall Evaluation Scores

Across all 240 samples, the model produced highly fluent and concise responses, with average fluency and conciseness scores of 4.89 and 4.45 respectively. These results suggest that the Mistral 7B Instruct v0.3 model consistently generates well-formed and readable responses. However, the model’s performance was lower in terms of correctness (2.62), relevance (2.83), and helpfulness (2.53), indicating issues with factual grounding and task alignment.

Table 3 summarizes these evaluation trends. The disparity between language quality (fluency and conciseness) versus content quality (correctness, relevance, and helpfulness) motivated a deeper

analysis across chatbot modes and decoding strategies to better understand these patterns.

Eval Dimension	Mean	Std. Deviation
Correctness	2.62	1.86
Fluency	4.89	0.53
Relevance	2.83	1.76
Helpfulness	2.53	1.82
Conciseness	4.45	0.82

Table 3: Overall human evaluation scores across all 240 samples.

6.2 Evaluation by Chatbot Mode

To better understand how the chatbot’s behavior affected response quality, I analyzed evaluation scores across the three different modes: general, travel, and no_retrieval. Each mode varies in both prompt design and retrieval configuration. travel mode uses a domain-specific prompt along with retrieved context passages, general mode relies on a broader, more open-ended prompt with retrieval, and no_retrieval mode disables retrieval entirely.

As shown in Table 4, travel mode consistently achieved the highest mean scores across all five evaluation dimensions, particularly in correctness and helpfulness. In contrast, no_retrieval mode struggled with correctness and task alignment, though fluency remained high. Figure 1 further illustrates these trends.

Dimension	General	Travel	No Retrieval
Correctness	2.70	3.35	2.41
Fluency	4.85	5.00	4.88
Relevance	3.05	3.48	2.62
Helpfulness	2.55	3.32	2.33
Conciseness	4.47	4.72	4.38

Table 4: Average human evaluation scores by chatbot mode. Travel mode consistently outperformed other modes across all dimensions.

6.3 Evaluation by Generation Setting

In addition to prompt and retrieval variation, I compared two decoding strategies: balanced sampling and deterministic generation. These settings affect output diversity, which may influence factual accuracy, wordiness, and user satisfaction.

As shown in Table 5, balanced decoding achieved slightly higher scores than deterministic decoding across all evaluation dimensions, particularly in correctness and helpfulness. Fluency

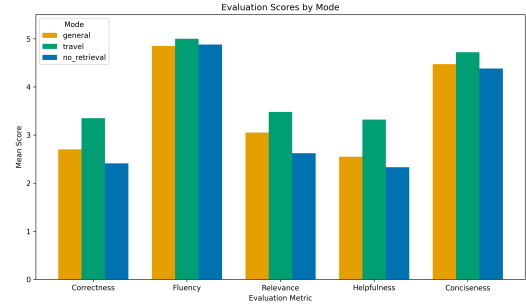


Figure 1: Bar chart comparing average human evaluation scores by chatbot mode.

remained high under both settings. These patterns are also illustrated in Figure 2.

Dimension	Balanced	Deterministic
Correctness	2.69	2.54
Fluency	4.92	4.87
Relevance	2.88	2.79
Helpfulness	2.63	2.42
Conciseness	4.46	4.44

Table 5: Average human evaluation scores by generation setting. Balanced decoding led to slightly better scores across all dimensions.

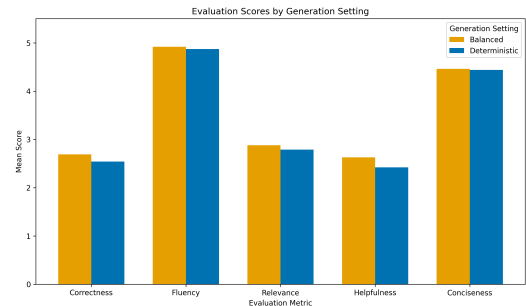


Figure 2: Bar chart comparing evaluation scores between balanced and deterministic generation settings.

6.4 System-Level Binary Metrics

To evaluate response control and language consistency, I tracked two binary metrics across all 240 samples: **expected language match** and **cut-off rate**. Overall, 83.33% of responses were in the expected language, while 51.25% were cut off before completion. These truncations likely contributed to lower scores in correctness and helpfulness, particularly in no_retrieval mode.

Table 6 summarizes these global trends. Additional breakdowns by mode and generation setting are visualized in Figures 3 and 4.

Metric	Percent Yes (1)	Percent No (0)
Expected Language	83.33%	16.67%
Cut-Off	51.25%	48.75%

Table 6: Percentage of responses in the expected language and percentage that were cut off.

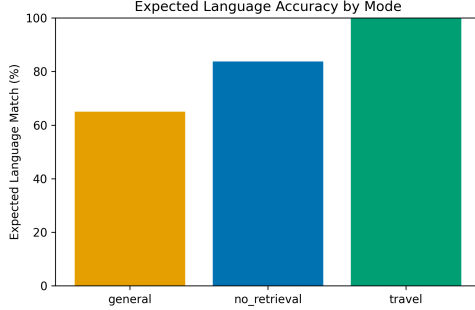


Figure 3: Expected language match rates across chatbot modes.

Example: Cut-off response

Query: ¿Cómo puedo responder a “How’s the weather today?” en inglés?

Mode: general

Prompt: You are a language tutor, translator, and conversation partner. Help the user naturally. Answer clearly.

Retrieved Context (excerpt): How’s the weather today? → ¿Cómo está el clima hoy?

Model Output: You can respond to “How’s the weather?” in English by saying “What’s the weather like

(This response was marked as fluent but cut off, and did not provide a relevant or complete answer.)

6.5 Semantic Grounding Analysis

To assess how closely the generated responses aligned with retrieved content, I computed cosine similarity between each output and its most similar retrieved passage. The average maximum similarity was relatively low at 0.17, with values ranging from -0.15 to 0.85. As shown in Figure 5, most responses had low or near-zero similarity, suggesting that the model often relied on background knowledge or prompt framing rather than directly incorporating retrieved context.

7 Conclusion

This project presented a multilingual RAG-based chatbot that supports both language learning and travel-related queries in English and Spanish. By combining FAISS-based semantic retrieval with

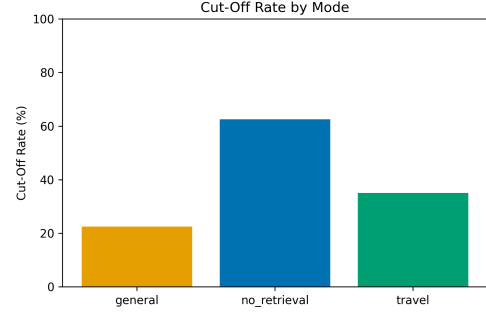


Figure 4: Cut-off rates across chatbot modes.

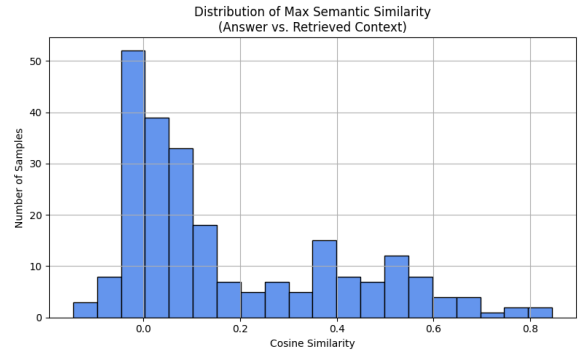


Figure 5: Distribution of max cosine similarity between each generated response and the most similar retrieved context chunk.

prompt-conditioned generation using Mistral 7B Instruct v0.3, the system produced fluent and concise responses across a range of tasks. However, evaluation results revealed trade-offs between fluency and factual correctness, as well as frequent cut-off issues and limited reliance on retrieved context.

To support interactive exploration and qualitative testing, I developed a simple web-based demo using *Streamlit* ([Streamlit Inc., n.d.](https://streamlit.io/)). The app allows users to input queries, choose a mode, and view the generated response alongside retrieved passages. This interface offers a practical entry point for future user studies and iterative refinement.

8 Limitations and Future Work

One key limitation of the current system is the frequent generation of incomplete or cut-off responses, likely due to prompt length and generation constraints. While fluency was consistently strong, correctness and helpfulness varied widely, particularly in the absence of tightly grounded context.

Another limitation lies in the use of cosine similarity as a proxy for semantic grounding. While useful, it does not fully capture whether the re-

trieved information meaningfully influenced the generation. Alternative grounding metrics or methods for tracking how context influences responses could offer deeper insights.

The system currently supports only two languages and fixed modes. Future work could explore extending support to additional languages, dynamic task prompting, and fine-tuning the language model on in-domain or instructional data. Incorporating user feedback from the Streamlit demo could also help improve prompt design and overall performance in real-world use.

References

- Mistral AI. 2024. [Mistral 7b instruct v0.3](#). Accessed: 2025-05-05.
- Weiqin Huang, Khe Foon Hew, and Luke K. Fryer. 2022. [Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning](#). *Journal of Computer Assisted Learning*, 38(1):237–257.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. [Billion-scale similarity search with gpus](#). *arXiv preprint arXiv:1702.08734*.
- 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 (NeurIPS)*, volume 33.
- Pierre Lison and Jörg Tiedemann. 2024. [Opensubtitles: Parallel subtitles for machine translation](#). Retrieved April 24, 2025.
- Khalid Matar and Yaser Mohammad. 2024. Improving the reliability of educational ai chatbots using retrieval-augmented generation. <https://www.diva-portal.org/smash/record.jsf?pid=diva2:1888614>. Preprint, Digitala Vetenskapliga Arkivet (DiVA).
- Holger Schwenk and Matthijs Douze. 2020. [Wikimatix: Mining parallel sentences from wikipedia](#). Retrieved April 24, 2025.
- Streamlit Inc. n.d. [Streamlit: The fastest way to build and share data apps](#). Retrieved April 24, 2025.
- Tatoeba Project. n.d. [Tatoeba downloads](#). Retrieved April 24, 2025.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Wang, Alex Andonian, Faysal Azhar, Vijay Badrinarayanan, et al. 2024. [Llama 3: Open foundation and instruction models](#). Accessed May 2025.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers](#). In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*.
- Wikimedia Foundation. n.d. [Wikivoyage: The free worldwide travel guide](#). Retrieved April 24, 2025.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#).

A Manual Evaluation Rubric

Category	Scoring Guidelines
Correctness	5: Completely correct — accurate translation or info, no errors. 4: Minor inaccuracy — mostly correct, with small factual or grammatical issues. 3: Mixed — contains both correct and incorrect info, or a partial answer. 2: Mostly incorrect — major factual or linguistic errors, misleading content. 1: Completely incorrect — wrong answer, nonsensical, or contradicts known facts.
Fluency	5: Native-level fluency — perfectly natural, grammatically correct. 4: Minor grammar or awkwardness — but still easy to understand. 3: Understandable, but noticeably awkward or includes grammar issues. 2: Poor grammar or awkward phrasing that interferes with understanding. 1: Unintelligible or very broken language.
Relevance	5: Directly relevant — fully addresses the query without distraction. 4: Mostly relevant — slight off-topic info, but main point is addressed. 3: Partially relevant — includes some on-topic info, but with distractions or tangents. 2: Barely relevant — the connection to the prompt is weak. 1: Completely off-topic or unrelated to the prompt.
Helpfulness	5: Very helpful — fully meets the user’s need (e.g., gives correct answer, usable explanation, or translation). 4: Helpful — minor missing detail or vague part, but still usable. 3: Somewhat helpful — limited usefulness or only partially solves the task. 2: Barely helpful — unclear, missing key info, or requires user to guess. 1: Not helpful — misleading, uninformative, or confusing.
Conciseness	5: Very concise — clear, direct, and efficient with no extra fluff. 4: Mostly concise — slight repetition or minor extra detail, but overall well-paced. 3: Acceptable — some redundancy or wordiness, but not overly distracting. 2: Wordy — includes unnecessary details or repeats ideas, making it less readable. 1: Rambling or verbose — long-winded, repetitive, or padded with filler that hinders clarity.
Expected Language	1: Yes — correct language 0: No — incorrect language
Cut-off	1: Yes — response was cut off 0: No — response was complete

B Prompt Templates and Evaluation Queries

Prompt Templates

- **General (English):** You are a language tutor, translator, and conversation partner. Help the user naturally. Answer clearly.
- **General (Spanish):** Eres un tutor de lenguas y compañero de conversación. Ayuda al usuario de forma natural. Responde con claridad.
- **Travel (English):** You are a multilingual travel assistant. Provide helpful information about destinations and travel tips.
- **Travel (Spanish):** Eres un asistente de viajes multilingüe. Proporciona información útil sobre destinos y consejos de viaje.

English Queries

General

- How do you say 'flower' in Spanish?
- Is this sentence correct: 'Yo fue al mercado'?
- What's a good way to reply to '¿Qué tiempo hace hoy?' in Spanish?
- What does 'mesa' mean in Spanish?
- Translate this into Spanish: She didn't want to see the movie but she went anyway.

Travel

- Where is Cordoba located?
- What are 3 things to do in Seville?
- How can I get from Cuzco to Lima?
- What are typical foods in Oaxaca?
- What is the best time of year to visit Ecuador?

Spanish Queries

General

- ¿Cómo se dice 'libro' en inglés?
- ¿Es correcta esta frase: 'She don't like apples'?
- ¿Cómo puedo responder a 'How's the weather today?' en inglés?
- ¿Qué significa 'fridge' en inglés?
- Traduce esto al inglés: No pudieron venir porque no tenían suficiente dinero.

Travel

- ¿Dónde está Vancouver?
- ¿Qué son 3 cosas que puedo hacer en Chicago?
- ¿Cómo llego de Nueva York a Boston?
- ¿Qué comidas típicas hay en Londres?
- ¿Cuál es la mejor época del año para visitar Estados Unidos?