

# Improving LLM Metalinguistic Knowledge Incorporating FundéuRAE's Recommendations through RAG

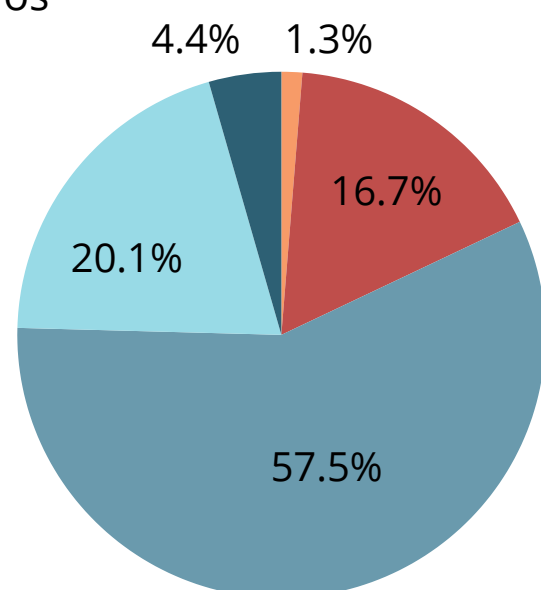
## INTRODUCTION

Large Language Models (LLMs) show remarkable abilities, but they still face challenges such as hallucinations and outdated knowledge. Retrieval-Augmented Generation (RAG) offers a promising solution by integrating information from external databases to enhance their responses (Gao et al., 2023). One of the key areas where these models can be improved is in their metalinguistic knowledge. Therefore, the objective of this project is to improve LLM metalinguistics knowledge incorporating FundéuRAE's recommendations through this technique.

## DATASET

- Created through web scraping.
- It contains **FundéuRAE's recommendations**, an institution dedicated to promoting the proper use of Spanish.
- Total recommendations: 5,065.
- The entries are divided into categories:

- Fonética
- Gramática
- Léxico
- Ortografía
- Otros

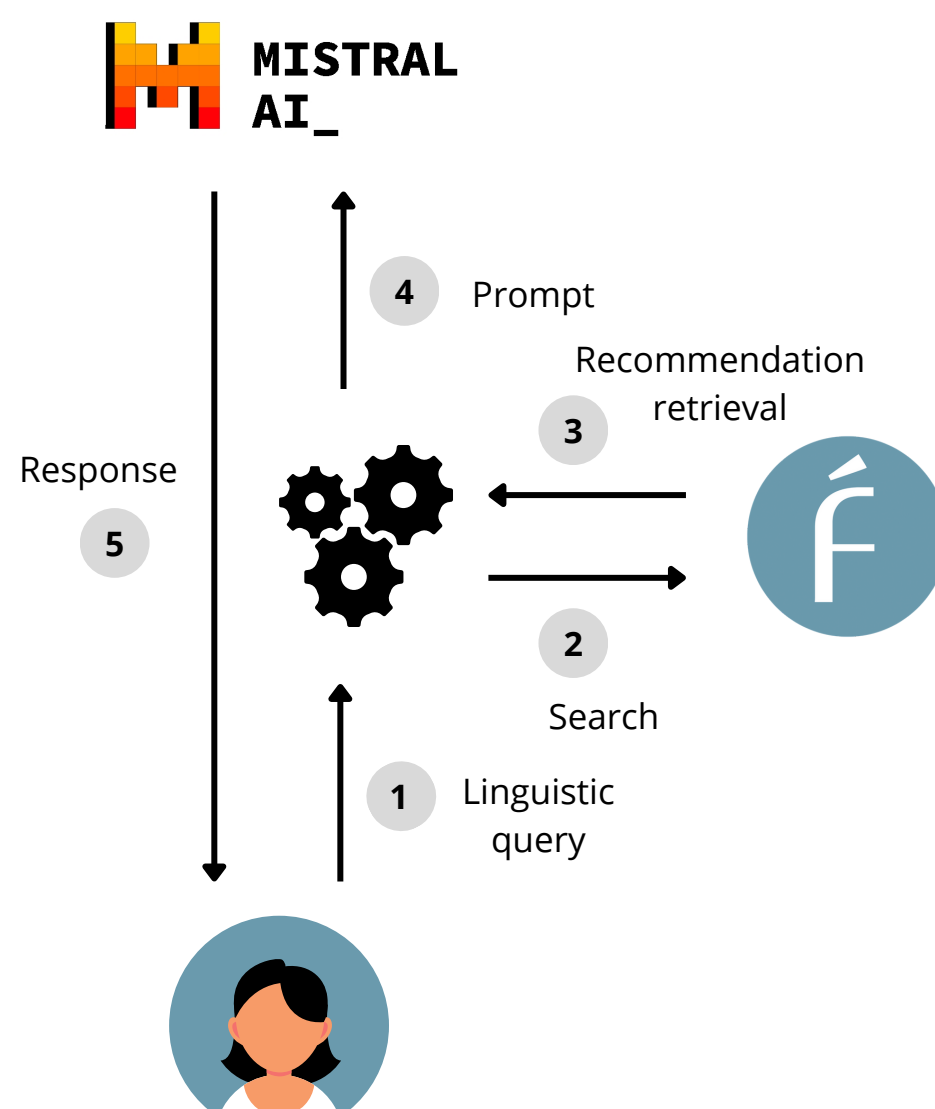


## METHODOLOGY

- Create and implement the dataset.
- LLM Model: **Mistral-7B-Instruct-v0.1** by Mistral AI. It showed the best production in Spanish.
- Information retrieval system to further perform RAG.
- Select 10 linguistic queries from different categories as prompts.
- Qualitative analysis of the answers before and after RAG.

## RAG

**Retrieval-Augmented Generation (RAG)** is a method that improves the accuracy and reliability of generative AI models by incorporating factual information from external sources.



## LINGUISTIC QUERIES

¿Se puede considerar 'gepetear' un neologismo?

¿Cuál es la forma adecuada para abreviar 'Reino Unido'?

¿Cuáles son los equivalentes de 'password' en español?

¿La palabra 'solo' se escribe con tilde cuando es adverbio?

¿Cómo se escribe 'reduflación' o 'reduflación'?

¿Cómo se escriben los títulos de las obras de creación?

¿Cómo se acentúa 'adecua' o 'adecúa'?

¿Se puede escribir coma entre sujeto y verbo de una oración?

¿Cuál es la diferencia entre 'sino' y 'si no'?

¿Cuál es la forma recomendada para referirse a los habitantes de Ucrania?

Check all the responses here!



## QUALITATIVE ANALYSIS

- Evolution of the accuracy of responses.
- Renewed content.
- Reduction of irrelevant information.
- Improvement of examples quality, obtained in many occasions directly from FundéuRAE.
- Persistent issues and misconceptions, such as the introduction of non-existent terms.

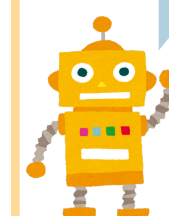
¡Hola! ¿En qué puedo ayudarte?

¿Se puede considerar 'gepetear' un neologismo?

¡Sí! Gepetear es un neologismo válido en ciertas circunstancias, como cuando se utiliza para referirse al uso de la aplicación ChatGPT en redes sociales o entretenimiento.

¿Cuál es la forma adecuada para abreviar 'Reino Unido'?

La forma adecuada para abreviar 'Reino Unido' es R.U. y no UK, que es la sigla del nombre inglés del país, United Kingdom.



## CONCLUSIONS

The implementation of RAG significantly improves the responses to the linguistic queries, making them more concise and appropriate, while also providing new metalinguistic information to the model. However, hallucinations are still present, and the output can also be highly variable and unpredictable, as we have obtained diverse responses using the same prompts. Finally, creating a chatbot with this technology would be very interesting to improve the user's information retrieval in Fundéu's web.

## REFERENCES

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv*. <https://arxiv.org/abs/2312.10997>

FundéuRAE. (2024, May 14). <https://www.fundeu.es/dudas/>



amaiamurillo/FundeuRAE