CENTRO DE INVESTIGACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES

Healthy Cooking with Large Language Models, Supervised Fine-Tuning, and Retrieval Augmented Generation

Andrea Morales-Garzón[1], Oscar A. Rocha, Sara Benel Ramirez, Gabriel Tuco Casquino[2], Alberto Medina[3]

[1]Department of Computer Science and Artificial Intelligence, University of Granada
[2]Universidad Católica de Santa María, Perú, [3]ETSII-UPM, Universidad Politécnica de Madrid

Introduction

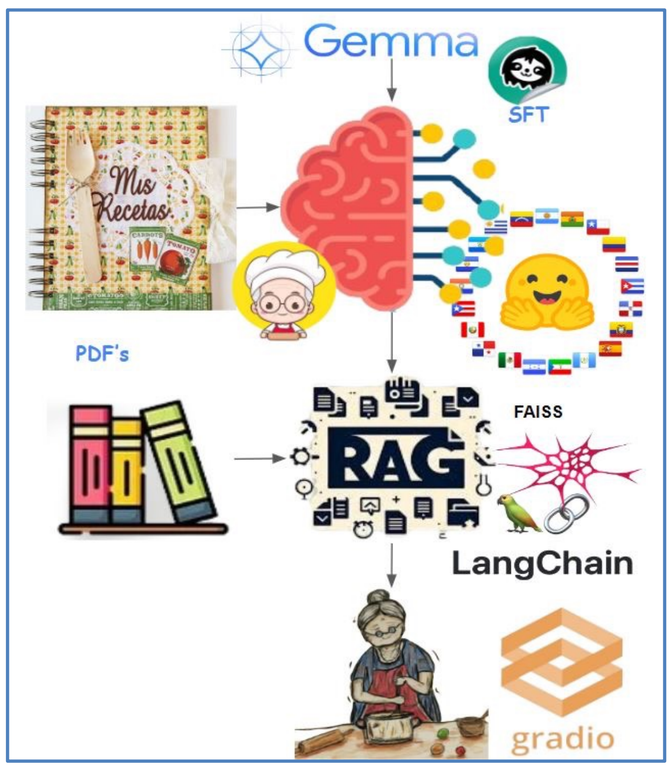
- **According to the WHO**, adopting healthy eating habits helps prevent malnutrition and diseases such as cancer, and diabetes.
- There is a **shortage of Spanish resources** to address nutrition computational problems.
- **LLMs**: **monotonous** in similar scenarios, and potentially **dangerous when dealing with diseases or allergies**.

Our contributions

- We present **RecetasDeLaAbuel@**, a corpus of **20,447 recipes** from Spanish-speaking countries.
- To the best of our knowledge, this is the **largest dataset of recipes in Spanish** available for free use.
- We fine-tuned 4bit Gemma2B (Supervised Fine-Tuning and Retrieval Augmented Generation methodologies) **to provide nutritional and cooking advice.**

ComeBien app: We tested its quality through an application that allows for nutritional and culinary context and food queries.

Methodology



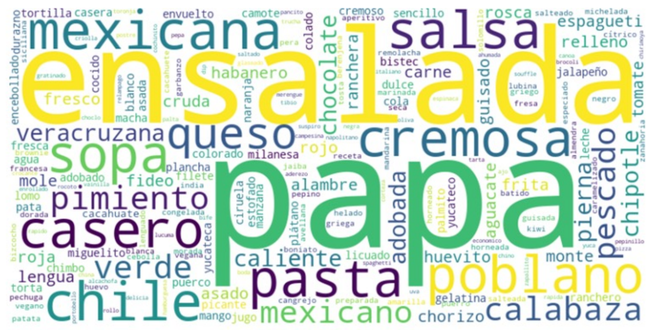
Recipe corpus

1. Web scrapping
- We scraped several web pages on Hispanic/Latin American and international cuisine written in Spanish.
- We tried to complete the recipe information by analysing the keywords of the recipe and quering Mixtral-8x7B-Instruct-v0.1

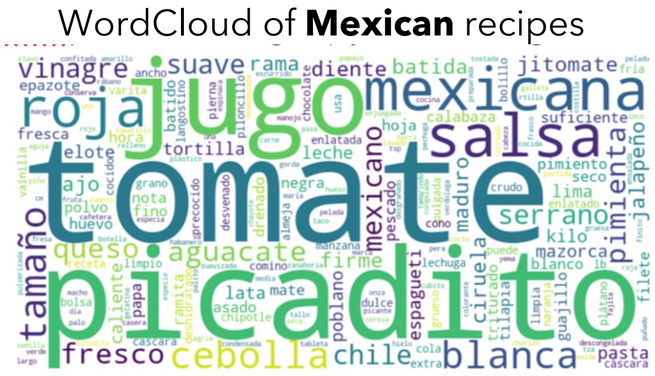
2. Data preprocessing and homogenisation.
- We performed data curation on specific fields.
- We have homogenized the corpus fields to facilitate the acquisition of statis- tics and visualization.

3. Creation of the instruction corpus in two fases:
1. We curated the corpus removing NAs, line breaks and concatenating texts.
2. We created a corpus of synthetic instructions using the LLM Genstruct.

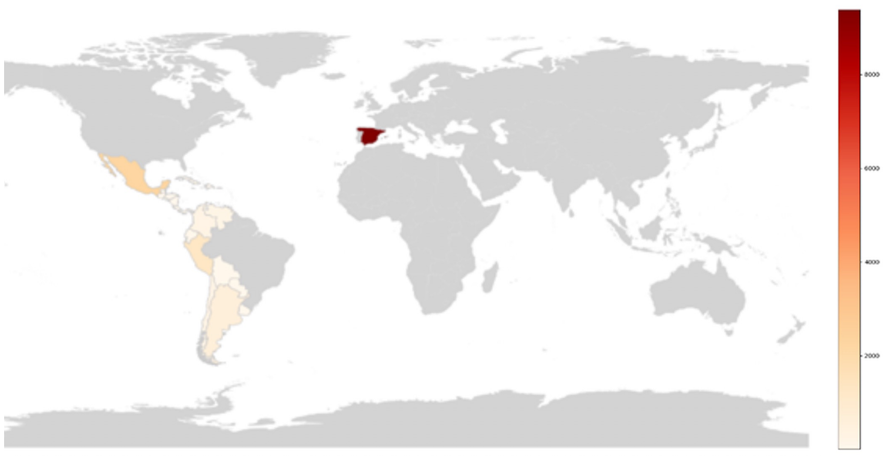
WordCloud of **recipe titles**



WordCloud of **Mexican** recipes



Country representation



Heatmap with represented countries in the corpus recipes.

Supervised Fine Tuning (SFT)

1. SFT with Gemma
- Gemma2B with 4Bit Quantization.
- We extracted a UTF-8 subset of 2.5k recipes of 3 attributes forming the question/answer pair (Recipe name vs ingredients and preparation steps).
- 8 epochs and 1550 steps.

2. SFT Hiperparameters
- Max. input tokens 4096
- LoRA r=16, α=16.
- Two batches
- Four gradient accumulation steps, 5 warmup steps
- 2e-4 learning rate
- Adamw_8bit optimizer, 0.01 weight drop rate and linear Schedule.
- 19M trained parameters

3. Training results
- Model **loss**, which is decreasing (1.8 at 200 steps and 1.3 at 1500 steps) with ±0.05 shallow ripple.
- **BERTScore** to compare the original recipes to those generated by the Gemma model. Precision: 0.67, recall: 0.71 and f1: 0.69.

Conclusions

- **RecetasDelaAbuel@** has been generated, trained and validated.
- **Most extensive corpus of recipes in Spanish.**

Future work

- **Extend experimentation** with other LLMs.
- Use **MoE** and **Retrieval Augmented Generation** (RAG) of traditional recipe books.

Environmental impact

- **Total estimated emissions** are 0.7 kg eq. CO2, obtained through the ML CO2 Impact website

https://huggingface.co/datasets/somosnlp/RecetasDeLaAbuela

This project was developed during the international Spanish NLP Hackathon Somos600M organized by SomosNLP; we thank the organizers and sponsors, especially SomosNLP and HuggingFace, for GPU credits and model endpoints. We also thank to Tomás Vergara Browne for contributing with the English translation of this paper.

UNIVERSIDAD DE GRANADA

Andrea Morales Garzón - amoralesg@decsai.ugr.es – 16/06/2024
Dept. of Computer Science and Artificial Intelligence, University of Granada