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

Correspondence: amoralesg@decsai.ugr.es

Abstract

In response to the growing demand of the global society to adopt healthy habits, this paper presents the design, development and validation of a Spanish culinary recipe dataset focused on healthy nutrition that includes representative dishes from the gastronomy of Spanish-speaking countries. We also evaluate the dataset using Gemma 2B, Supervised Fine-Tuning 4bit and Retrieval Augmented Generation methodologies, showing its use to solve concrete problems related to nutrition.

1 Introduction

Food is essential in human development, and maintaining proper eating habits prevents diet-related risk factors and diseases. According to the World Health Organization, adopting healthy eating habits helps prevent malnutrition and diseases such as cancer, diabetes and obesity. Previous works propose scrapping a corpus of recipes from specialized cooking websites and their following homogenization (Li et al., 2022; Majumder et al., 2019; Marın et al., 2021; Salvador et al., 2017; Yagcioglu et al., 2018), or extending existing resources, tackling data curation and dataset extension (Bień et al., 2020). Despite their relevance, there is a notable shortage of Spanish resources to address nutrition-related computational problems. While large language models (LLMs) are a commonly used tool for providing culinary knowledge, their limitations, such as being error-prone, monotonous in similar scenarios, and potentially dangerous when dealing with diseases or allergies, underscore the need for a more accurate and reliable tool (Niszczota and Rybicka, 2023). We aim to alleviate these drawbacks: the need for a large recipe corpus in Spanish and the inefficient use of LLMs to improve the model's responsiveness. We present a corpus of 20,447 recipes from Spanish-speaking countries to provide open resources to address societal questions and make them accessible to the Spanish-speaking community. This dataset, called *RececetasDeLaAbuel@*[1], includes recipes belonging to the gastronomy of different Spanish-speaking countries obtained from multiple recipe websites while taking into account possible geographical and linguistic biases. To our knowledge, this is the largest dataset of recipes of Spanish-speaking origin available for free use. In addition, we tested its quality through an application that allows for nutritional and culinary context and food queries. For this, we use Gemma 2B and the latest Supervised Fine-Tuning (SFT) 4bit + Retrieval Augmented Generation (RAG) methodologies (Lewis et al., 2020).

2 Methodology

Our methodology is structured as follows: (1) Design for corpus creation, (2) Collection of recipes by web scraping, (3) Data curation and unification, (4) Corpus generation with instructions, (5) SFT training on lightweight 4-bit LLMs, (6) RAG with FAISS/LangChain and (7) Development and deployment of a demo in Gradio. Fig. 1 shows a diagram of the applied methodology.

3 Recipe corpus

Web scraping. We scraped several web pages on Hispanic/Latin American and international cuisine written in Spanish to create the dataset. We collected a total of 20,447 recipes using the Newspaper3k, Scrapy and BeautifulSoup libraries from the repository Frorozcoloa/ChatCocina[2], an open-source project for recipe extraction. In the extraction, attributes such as name, ingredients, preparation steps, duration, category, context or description, rating and votes, diners and difficulty of each recipe were obtained. In some recipes,

[1]https://huggingface.co/datasets/somosnlp/RecetasDeLaAbuela

[2]https://github.com/Frorozcoloa/ChatCocina

Figure 1: Methodology (LLMs SFT 4bit +RAG)

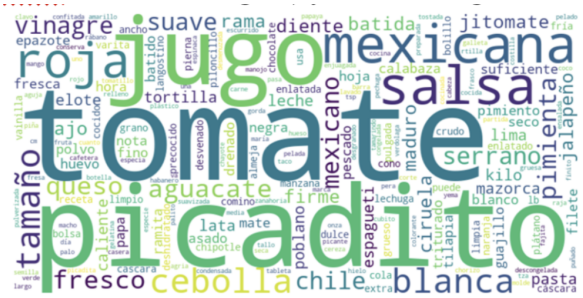

Figure 2: Most used ingredients in Mexico

Figure 3: Most used terms in recipe titles

we calculated the country of origin by analyzing the keywords of the recipe (name of the country, gentilism and keywords of ancient cultures such as Aztec, Inca, gaucho, Guarani, Mapuche or Bolivarian). In adverse cases, we solved it by querying using the Together.AI API [3] with `Mixtral-8x7B-Instruct-v0.1`. It is essential to highlight that some recipes lack values in specific attributes in the final corpus due to differences in structure or data in their origins.

Data preprocessing and homogenization. We performed data curation on specific fields of the dataset. For "*Ingredients*" and "*Steps*", we have separated quantities from metric units (e.g., 1ml → 1 ml). For "*Name*" and "*Country*", we have implemented a filter for duplicate recipes with the same name and same country. We have homogenized the corpus fields to facilitate the acquisition of statistics and visualization. We have used the HH:MM format for "*Duration*". Finally, for "*Country*", we have implemented the ISO format with the country codes.

Statistics and biases. We have obtained statistics through WordCloud visualizations, making use of TF-IDF to get three types of results:

(1) *Most used ingredients by country*. For example, Fig. 2 shows a Wordcloud with the most used ingredients in Mexican recipes.

(2) *Geographic biases*. The geographic representation of the dataset is closely linked to the nationality of those involved in the project. There is a higher proportion of recipes from Spain, and to a lesser extent from Mexico and Peru. Fig. 4 shows the heat map of the number of recipes by country, where the higher the number, the darker the color. We have observed very few recipes from countries such as Panama or Paraguay since the original web pages do not focus on these countries.

(3) *Most present terms in recipe titles*. Some terms that refer to an origin may not necessarily be closely related to the recipe's origin. Fig. 3 shows the WordCloud generated from the recipe names.

Creation of the instruction corpus. We have created the instruction corpus in two phases. First, we curated the initial corpus. We removed records with no data (nulls, blanks and line breaks) in the "Ingredients" and "Steps" columns and then concatenated the information from all columns into one. Secondly, we created a corpus of synthetic instructions using the LLM Genstruct in the distilabel environment. We used this LLM to generate question-answer pairs from the previously aggregated information. The main advantage of Genstruct is that it can massively simulate questions from a human user in a very varied and natural way instead of always asking for ingredients or preparation steps of different recipes, e.g., "Tell me a vegetarian dish" or "Recommend me a healthy chicken-based dish".

[3] http://together.ai

Figure 4: Distribution of recipes per country

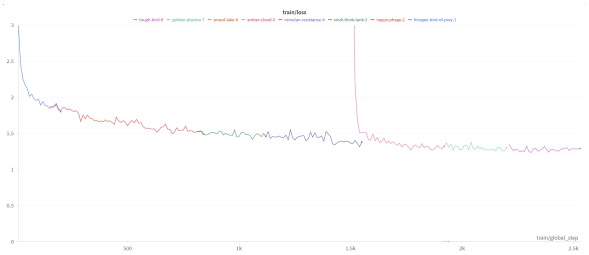

Figure 5: Model loss visualization

4 Fine-tuning of LLMs y RAG

SFT with Gemma. The SFT training has been performed using the `unsloth/gemma-2b-bnb-4bit` model since 4-bit quantization offers a superior effort/quality ratio (2.4x faster and 58% less VRAM vs. the Gemma 7B LLM). Mistral 7B (slower) and TinyLlama 1.1B (slightly faster but less accurate) have also been tested. A UTF-8 subset of 2.5k recipes of 3 attributes forming the question/answer pair (Recipe name vs ingredients and preparation steps) is extracted from the RecetasDeLaAbuel@ corpus. The SFT training lasts approximately two hours (equivalent to 8 epochs and 1550 steps) on HuggingFace Nvidia T4 medium (8 vCPU, 30 GV RAM, 16GB VRAM), and we obtained the RecetasDeLaAbuela5k model[4].

SFT Hiperparameters. We trained the model using CUDA 12.1 Pytorch 2.2.2 with maximum input tokens 4096, LoRA r=16, α=16, no dropout/bias/rslora/LoftQ, two processes, two batches, four gradient accumulation steps, five warmup steps, 2e-4 learning rate, adamw_8bit optimizer, 0.01 weight drop rate and linear scheduler. A total of 19M parameters are trained for Gemma 2B. The maximum peak SFT/total memory reaches 12/14.4Gb (83%/98%).

Training results. We saved the SFT training every 10 steps in Wandb. Fig. 5 shows a plot with the model loss, which is decreasing (1.8 at 200 steps and 1.3 at 1500 steps) with ±0.05 shallow ripple. We used BERTScore (Zhang et al., 2019) to evaluate the resemblance of the original recipes to those generated by the model, obtaining precision: 0.67, recall: 0.71 and f1: 0.69.

Environmental impact. Experiments were performed using HuggingFace (AWS) in the sa-east-1 region, which has a carbon efficiency of 0.2 kg CO2 eq/kWh. A cumulative of 50 hours of computation was performed on HW type T4 (TDP of 70W). Total estimated emissions are 0.7 kg eq. CO2, obtained through the ML CO2 Impact website (Lacoste et al., 2019).

5 Results

The trained models can be tested in the Gradio demos RecetasDeLaAbuel@[5] and ComeBien[6]. The user can enter their question about a recipe and add either a nutritional context (e.g.,"You are an AI expert on cooking and nutrition") or a culinary context extracted from cookbook PDFs using RAG and FAISS LangChain.

6 Conclusions and future work

The RecetasDelaAbuel@ corpus has been successfully generated, trained and validated. With it, we have contributed to creating an open-source strategy to compile the most extensive corpus of recipes from Spanish-speaking countries. In future work, we will extend the experimentation with Mistral and TinyLlama and use MoE and RAG of traditional recipe books.

Acknowledgments

This work was developed as part of the SomosNLP Spanish Hackathon 2024.

References

Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. Recipenlg: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28.

[4] https://huggingface.co/somosnlp/RecetasDeLaAbuela5k_gemma-2b-bnb-4bit

[5] https://huggingface.co/spaces/somosnlp/RecetasDeLaAbuela_Demo

[6] https://huggingface.co/spaces/somosnlp/ComeBien_Demo

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ming Li, Lin Li, Qing Xie, Jingling Yuan, and Xiaohui Tao. 2022. Mealrec: a meal recommendation dataset. *arXiv preprint arXiv:2205.12133*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. *arXiv preprint arXiv:1909.00105*.

Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Paweł Niszczota and Iga Rybicka. 2023. The credibility of dietary advice formulated by chatgpt: robodiets for people with food allergies. *Nutrition*, 112:112076.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.