Aprendiendo a cocinar de manera saludable con Large Language Models, Supervised Fine Tuning y Retrieval Augmented Generation

Anonymous ACL submission

Abstract

En respuesta a la creciente demanda de la sociedad global en adoptar hábitos saludables, este artículo presenta el diseño, desarrollo y validación de un dataset de recetas culinarias en español enfocado a nutrición saludable que incluye platos representativos de la gastronomía de países hispanohablantes. Además, evaluamos el dataset utilizando Gemma 2B, junto con las metodologías Supervised Fine-Tuning 4bit y Retrieval Augmented Generation, mostrando su uso para resolver problemas concretos relacionados con la nutrición.

1 Introducción

La alimentación constituye un pilar fundamental en el desarrollo humano, y mantener unos hábitos alimenticios adecuados es esencial para prevenir factores de riesgo asociados con la dieta y enfermedades directamente vinculadas con la alimentación. Según la Organización Mundial de la Salud, adoptar hábitos alimentarios saludables ayuda a prevenir la desnutrición y enfermedades como el cáncer, la diabetes y la obesidad, entre otras. Trabajos previos proponen escrapear un corpus de recetas de páginas web especializadas, junto con su posterior homogeneización (Li et al., 2022; Majumder et al., 2019; Marın et al., 2021; Salvador et al., 2017; Yagcioglu et al., 2018), o extender recursos existentes en términos de curado de datos y extensión del dataset (Bień et al., 2020). A pesar de su relevancia, hay una escasez notable de recursos en español que permitan abordar problemas computacionales relacionados con la nutrición. Los modelos grandes de lenguaje (LLMs) son una herramienta útil para proporcionar conocimiento culinario. Sin embargo, su uso es propenso a errores, se comportan de forma monótona en escenarios similares, y pueden ser peligrosos cuando tratamos con enfermedades o alergias (Niszczota and Rybicka, 2023). Nuestro objetivo es paliar estos inconvenientes: la falta de corpus de recetas en español, y su aprovechamiento ineficiente con LLMs para mejorar la respuesta del modelo. Presentamos un corpus de 20,447 recetas de países hispanohablantes con el propósito de proporcionar recursos abiertos para abordar interrogantes de la sociedad y hacerlos accesibles a la comunidad hispanohablante. Este dataset, llamado *RecetasDeLaAbuel@*[1], incluye recetas pertenecientes a la gastronomía de distintos países hispanohablantes obtenida de múltiples páginas web de recetas, a la vez que se tiene en cuenta posibles sesgos geográficos y lingüísticos. Hasta donde sabemos, este es el mayor dataset de recetas de origen hispanohablante disponible para su libre uso. Además, probamos su calidad a través de una aplicación que permite incluir contexto nutricional y culinario y realizar consultas sobre alimentación. Para ello, usamos Gemma 2B, junto con las últimas metodologías Supervised Fine-Tuning (SFT) 4bit + Retrieval Augmented Generation (RAG) (Lewis et al., 2020).

2 Metodología

La metodología se estructura siguiendo los siguientes pasos: (1) Diseño para creación del corpus, (2) Recopilación recetas mediante web scraping, (3) Curado y unificación de los datos, (4) Generación del corpus con instrucciones, (5) Entrenamiento SFT sobre LLMs ligeros de 4 bits, (6) RAG con FAISS/LangChain y (7) Desarrollo y despliegue de una demo en Gradio. La Fig. 1 muestra un diagrama de la metodología aplicada.

3 Corpus de recetas

Web scrapping. Para la creación del dataset se realizó Web Scraping en diferentes páginas web sobre cocina hispano/latinoamericana e internacional redactada en español. En total se recopilaron

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

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

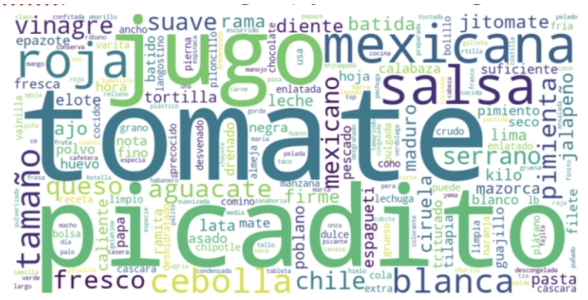

Figure 2: Ingredientes más usados en México

Figure 3: Términos más usados en las recetas

20,447 recetas mediante las librerías Newspaper3k, Scrapy y BeautifulSoup a partir del repositorio Frorozcoloa/ChatCocina[2] (proyecto open-source para extracción de recetas). En la extracción se obtuvieron atributos como nombre, ingredientes, pasos de preparación, duración, categoría, contexto o descripción, valoración y votos, comensales y dificultad de cada receta. En algunas recetas, el país de procedencia se ha calculado mediante análisis de palabras clave de la receta (nombre de países, gentilicios y palabras clave de culturas antiguas como azteca, inca, gaucho, guaraní, mapuche o bolivariano) y en caso negativo se ha resuelto preguntando mediante la API Together.AI[3] al modelo `Mixtral-8x7B-Instruct-v0.1`. Es importante resaltar que algunas recetas carecen de valores en ciertos atributos en el corpus final debido a diferencias en su estructura o datos en sus orígenes.

Preprocesamiento y homogeneización. Se realizó un curado de datos en campos específicos del dataset. Para "*Ingredientes*" y "*Pasos*" se implementó una separación de cantidades con unidades métricas. Ejemplo: 1ml -> 1 ml. Para "*Nombre*" y "*País*"se implementó un filtro de recetas duplicadas que tuvieran mismo nombre y estuvieran asociadas a un mismo país. Hemos realizado una homogeneización de campos del corpus para facilitar las tareas de obtención de estadísticas y visualización. Para "*Duración*" se utilizó el formato HH:MM, y para "*País*" se implementó el formato ISO_A3 con los códigos de los países.

Estadísticas y sesgos. Las estadísticas se obtuvieron mediante visualizaciones de WordCloud, haciendo uso de TF-IDF para obtener tres tipos de resultados:

(1) *Ingredientes más utilizados por país*. Por ejemplo, Fig. 2 muestra un Wordcloud con los ingredientes más utilizados en recetas mexicanas.

(2) *Sesgos geográficos.* Con respecto a la representación por país que tiene el dataset, que está estrechamente vinculado a la nacionalidad de los involucrados en el proyecto. Existe mayor proporción de recetas de España, y en menor medida de México y Perú. La Fig. 4 muestra el mapa de calor de la cantidad de recetas por país, donde a mayor cantidad, más oscuro es el color. Se observan muy pocas recetas de países como Panamá o Paraguay puesto que las páginas web originales no se centran en estos países.

(3) *Términos más presentes a la hora de definir los nombres de las recetas*. Algunos términos que hacen referencia a un origen, no necesariamente deben de estar estrechamente relacionados al origen de la receta. El WordCloud generado a partir de los nombres de las recetas se muestra en la Fig. 3.

Creación del Corpus de Instrucciones. La creación del corpus de instrucciones se ha realizado en 2 fases. Primero, depuramos corpus inicial. Se eliminaron los registros sin datos (nulos, espacios vacíos y saltos de línea) en las columnas de "Ingredientes" y "Pasos" y luego se concatenó la información de todas las columnas en una sola. A

[2] https://github.com/Frorozcoloa/ChatCocina
[3] http://together.ai

Figure 4: Cantidad de recetas por país

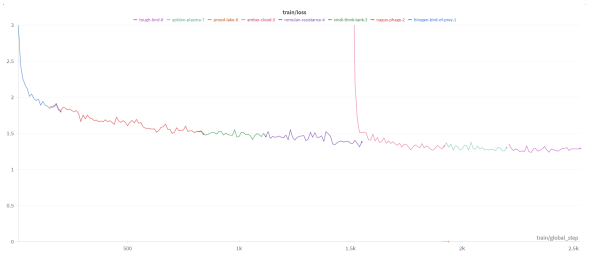

Figure 5: Visualización de la pérdida del modelo

continuación, creamos un corpus de instrucciones sintéticas mediante el LLM Genstruct en el entorno distilabel. Este LLM generó pares de preguntas-respuestas a partir de la información agregada previamente. La ventaja de Genstruct es que es capaz de simular de forma masiva preguntas de un usuario humano de forma muy variada y natural en vez de preguntar siempre por los ingredientes o pasos de preparación de distintas recetas, p. ej., "Dime un plato vegetariano" o "Recomiéndame un plato saludable basado en pollo".

4 Fine-tuning de LLMs y RAG

SFT con Gemma. El entrenamiento SFT se ha realizado tomando como base el modelo gemma-2b-bnb-4bit de sloth puesto que la cuantización de 4 bits ofrece un ratio esfuerzo/calidad superior (2.4x más rápido y 58% menos VRAM frente al LLM Gemma 7b). Se ha probado también Mistral 7B (más lento) y TinyLlama 1.1B (un poco más rápido pero menos preciso). Se extrae del corpus RecetasDeLaAbuel@ un subconjunto UTF-8 de 2.5k recetas de 3 atributos que forman el par pregunta/respuesta (Nombre de la receta vs ingredientes y pasos de preparación). El entrenamiento SFT dura 2h aproximadamente (equivalente a 8 épocas y 1550 pasos) en HuggingFace Nvidia T4 medium (8 vCPU, 30 GV RAM 16GB VRAM) y se obtiene el modelo RecetasDeLaAbuela5k[4].

SFT Hiper-parámetros. El entrenamiento se realiza bajo CUDA 12.1 Pytorch 2.2.2 con un máximo de tokens de entrada 4096, LoRA r=16, =16 sin dropout/bias/rslora/LoftQ, 2 procesos, 2 batches, 4 pasos de acumulación de gradiente, 5 pasos de calentamiento, tasa de aprendizaje 2e-4, optimizador adamw_8bit, tasa de caída de pesos de 0.01 y planificador lineal. Para Gemma 2B se entrenan un total de 19M de parámetros. El máximo pico de memoria SFT/total alcanza 12/14.4Gb (83%/98%).

Resultados del entrenamiento. El entrenamiento SFT se guarda cada 10 pasos en Wandb. La Fig. 5 muestra una gráfica con la pérdida del modelo, la cual es decreciente (1.8 a los 200 pasos y 1.3 a los 1500 pasos) con rizado ±0.05 muy bajo. Utilizamos BERTScore (Zhang et al., 2019) para evaluar el parecido de las recetas originales con las generadas por el modelo, obteniendo precisión: 0.67, recall: 0.71 y f1: 0.69.

Impacto medioambiental. Los experimentos se realizaron utilizando HuggingFace (AWS) en la región sa-east-1, que tiene una eficiencia de carbono de 0.2 kg CO2 eq/kWh. Se realizó un acumulado de 50 horas de cómputo en HW tipo T4 (TDP de 70W). Las emisiones totales estimadas son 0.7 kg eq. CO2., obtenidas a través de la web ML CO2 Impact (Lacoste et al., 2019).

5 Resultados

Los modelos entrenados se pueden probar en las demos Gradio RecetasDeLaAbuel@[5] y ComeBien[6]. El usuario puede introducir su pregunta sobre una receta y añadir ya sea un contexto nutricional ("Eres una IA especialista en cocina y nutrición") o un contexto culinario extraído de PDFs de libros de cocina mediante RAG y FAISS LangChain.

6 Conclusiones y trabajo futuro

El corpus RecetasDelaAbuel@ ha sido generado, entrenado y validado con éxito. Se ha iniciado un trabajo open-source de compilación del corpus más extenso de recetas hispanoamericanas. Como trabajo futuro, ampliaremos la experimentación con pruebas en Mistral, TinyLlama, y utilizando MoE junto con RAG de libros tradicionales de cocina.

[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

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.

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*.