NutriRAG: Unleashing the Power of Large Language Models for Food Identification and Classification through Retrieval Methods

Huixue Zhou[1], Lisa S.Chow[1], Lisa Harnack[1], Satchidananda Panda[2], Emily N.C. Manoogian[2], Minchen Li[1], Yongkang Xiao[1], Rui Zhang[1]

[1]University of Minnesota, Minneapolis, MN, USA; [2]Salk Institute for Biological Studies, San Diego, CA, USA

Introduction

Nutrition is vital for preventing and managing chronic diseases, yet accurately tracking dietary intake in real-world settings remains challenging.

Key Challenges:

1. Unstructured Text: Food entries from apps are ambiguous and unstandardized, making them difficult for traditional ML models that rely on structured data.
2. Limited Databases: Existing resources focus on packaged foods, lacking coverage of homemade or culturally diverse meals.
3. Research Gap: Prior studies emphasize packaged food and social media data, with little work on personalized, free-text dietary analysis using NLP.

We propose **NutriRAG**, a retrieval-augmented LLM framework designed to identify and categorize diverse food descriptions for personalized nutritional analysis.

✓ analyzing real-world dietary logs from nutrition tracking apps

✓ leveraging retrieval-based few-shot learning to overcome limited labeled data.

Framework

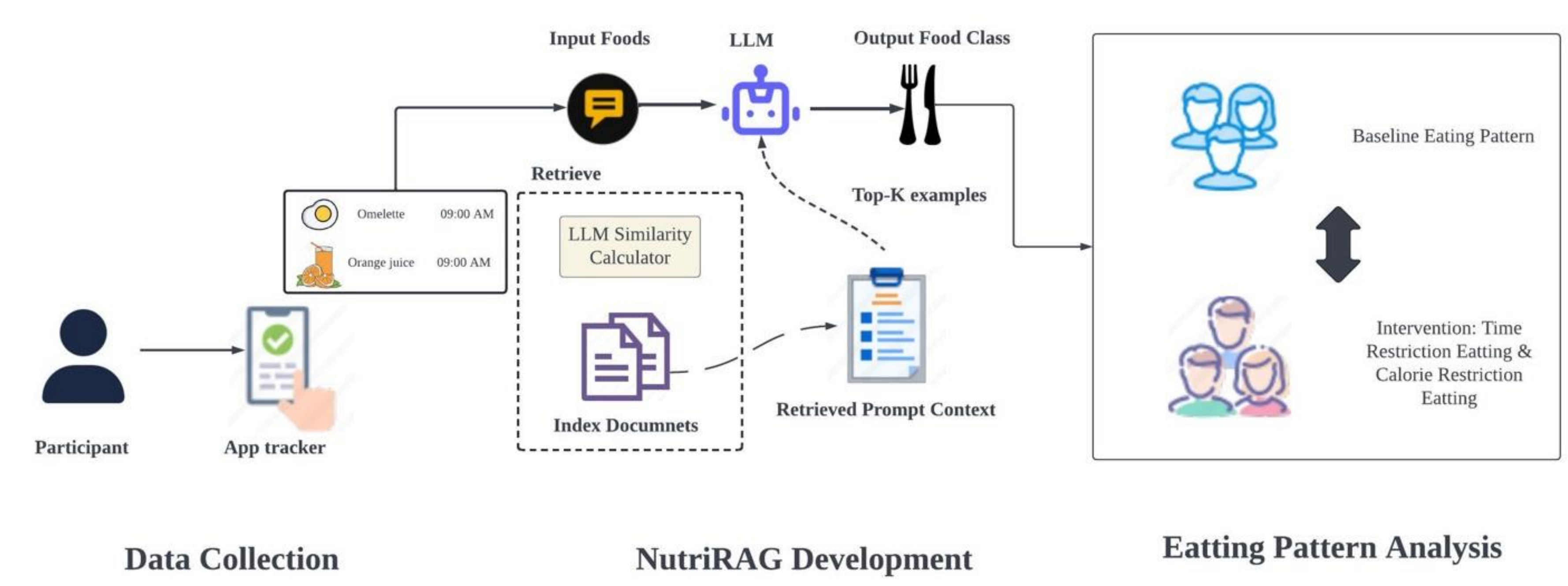
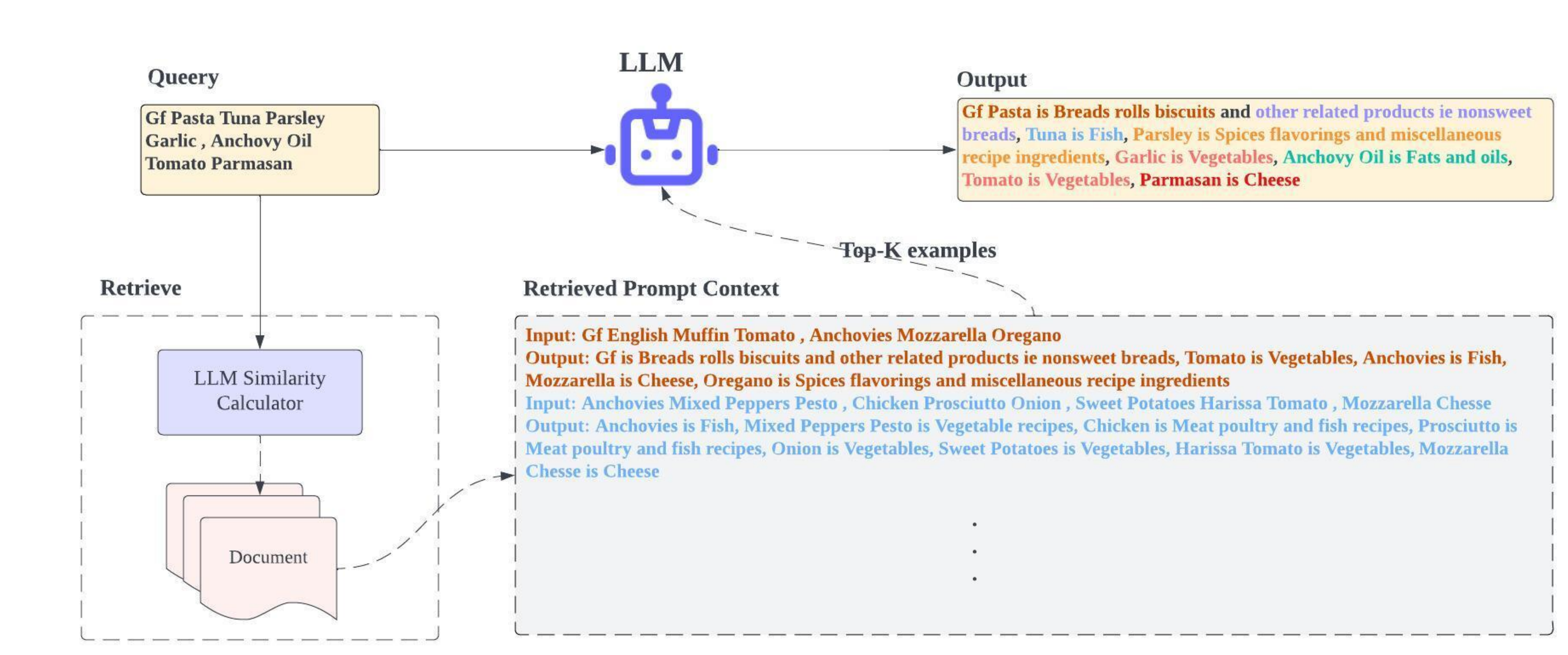


Figure 1. Overview of the study design. A Data Collection: Free-text food entries from the myCircadianClock app are processed, with de-identified user inputs serving as the primary data source. B NutriRAG Development: The pipeline integrates a retrieval-augmented LLM to classify foods. C Application: Analyzed baseline vs. intervention dietary patterns using the RCT dataset (NCT04259632).

NutriRAG



Goal: Automatically identify food categories in large, mostly unlabeled dietary datasets.

Data Source: Free-text food entries collected from the myCircadianClock mobile application (parent RCT dataset).

Four-Phase Workflow

1. Query Formulation: Construct nutritional inquiries from participant text (e.g., "Gf Pasta Tuna Parsley Garlic Anchovy Oil Tomato Parmesan").
2. Retrieval and Prompt Context: Retrieve top-k most similar labeled examples using cosine similarity of LLM embeddings to provide in-context demonstrations.
3. LLM Processing and Classification: The LLM uses retrieved examples and task-specific instructions to classify food items into 51 predefined categories.
4. Structured Output Generation: Convert model outputs into structured nutritional categories for downstream dietary pattern analysis.

Key Takeaways

✓NutriRAG effectively classifies unstructured, real-world dietary text using retrieval-augmented LLMs.

✓Retrieval augmentation substantially boosts accuracy across all models, with GPT-4 reaching 82.2% micro-F1.

✓Open-source models like Mixtral 8×7B also benefit, narrowing the gap with proprietary LLMs.

✓Structured, context-aware retrieval improves robustness without requiring model fine-tuning.

✓Eating pattern analysis confirms NutriRAG's utility in detecting changes in dietary behaviors across interventions. recommendation systems.

Results

Table 1. Performance comparison of transformer-based and retrieval-augmented models in food classification (metrics in %).

Model	Micro precision	Micro Recall	Micro F1
Baseline Models			
Bert	56.36	62.74	59.38
BLUEBert	49.84	57.26	53.29
PubmedBert	56.82	59.71	58.22
No examples			
Llama-2-70b	25.58	45.69	32.80
Mixtral 8*7b	59.50	37.60	46.08
GPT3.5	64.54	66.43	65.47
GPT4	75.97	71.53	73.84
Random examples			
Llama-2-70b	37.70	71.72	49.42
Mixtral 8*7b	60.87	75.58	67.43
GPT3.5	64.54	66.43	65.47
GPT4	75.97	71.53	73.84
(UMN) Retrieval-Augmented			
Llama-2-70b	68.81	58.75	63.38

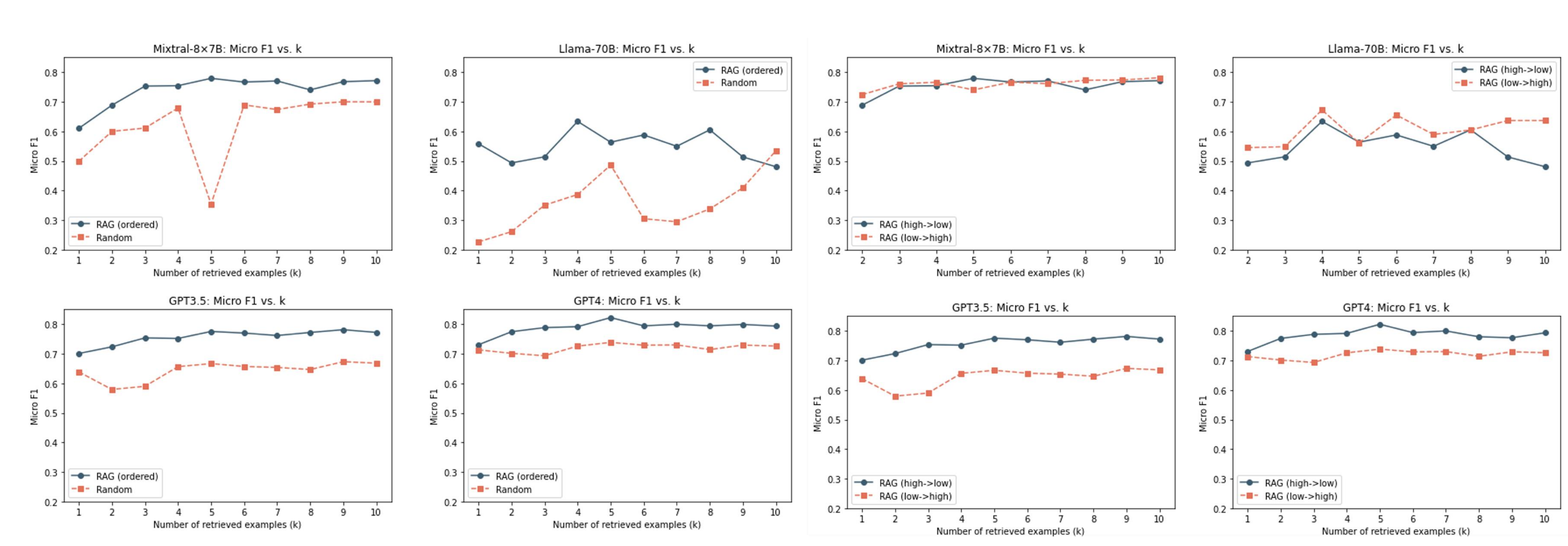


Figure 3. Performance across different retrieval set sizes (k = 1–10), comparing NutriRAG with ordered vs. random retrieval and different example orderings (high-to-low vs. low-to-high similarity).

NutriRAG:

- LLMs outperform baselines: GPT-4 achieved 73.8% micro-F1 without retrieval, surpassing transformer baselines (BERT 59.4%, PubMedBERT 58.2%).
- Retrieval improves accuracy: With RAG, GPT-4 reached 82.2%, and Mixtral 8×7B improved from 46.1% → 77.9%, showing large gains across all models.
- Retrieval robustness: Ordered retrieval (k = 4–5) yielded stable, strong results; open-source models were more sensitive to example ordering.

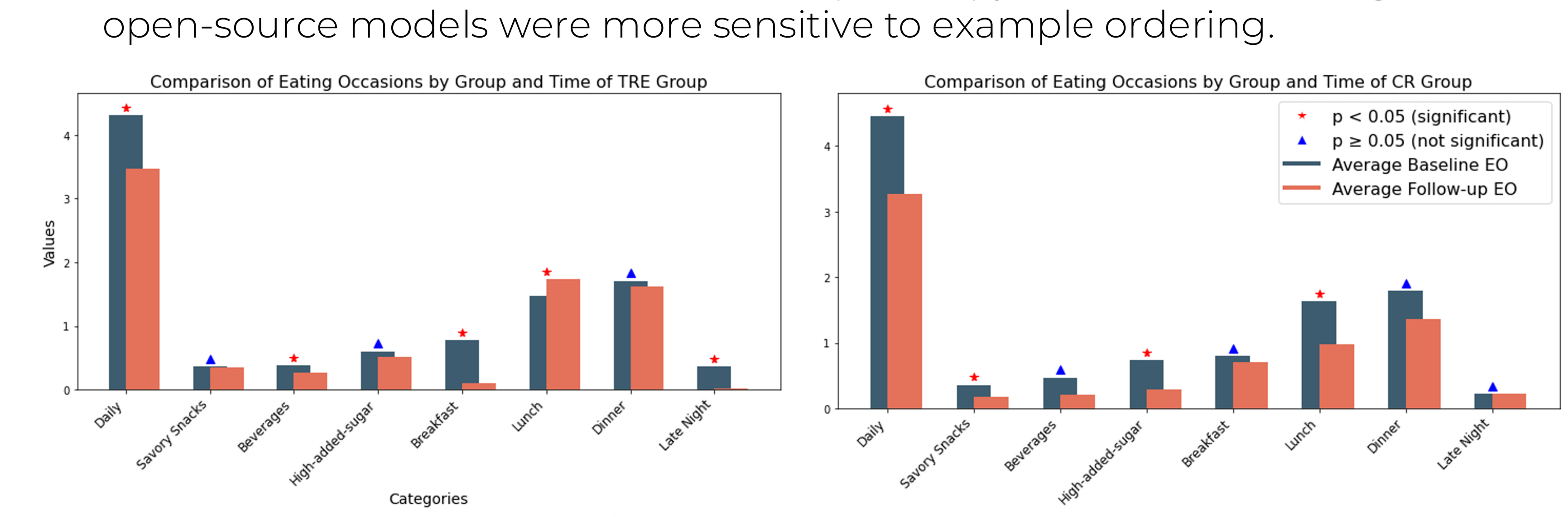


Figure 6. Paired T-test analysis of baseline vs. end-intervention eating occasion (EO).

Eating Pattern Results:

➢ The CR group reduced intake of savory snacks and high–added-sugar foods, while meal timing remained stable.

➢ The TRE group showed fewer beverages and significantly fewer breakfast and late-night eating occasions, reflecting a shortened eating window.