A multi-LLM explainable food recommendation system based on Deep Learning

Ephraim Sinyabe Pagou, Corneille Vivient Kamla, Josiane Ngathic, Igor Tchappi

PREPRINT

A multi-LLM explainable food recommendation system based on Deep Learning

Ephraim Sinyabe Pagou[1], **Corneille Vivient Kamla**[1], **Josiane Ngathic**[1], **Corneille Vivient Kamla**[2]

[1]ENSAI, University of Ngaoundere, PO Box 454, Ngaoundere, Cameroon
[2]FINATRAX Research Group, SnT, University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

*E-mail : ephraim.sinyabe@univ-ndere.cm

Abstract

Food recommender systems (FRS) increasingly support dietary decision-making, yet their explanations often remain rigid, poorly contextualized, and disconnected from user needs. Current XAI techniques such as SHAP or LIME provide numerical justifications that lack semantic depth, while single-LLM pipelines restrict adaptability and may amplify model-specific biases. To address these limitations, we introduce a multi-LLM explainability overlay that generates contextual, contrastive, counterfactual, and simulation-based explanations over existing deep learning models. The system integrates Food and User Ontologies, partly derived from the Food Explanation Ontology (FEO) and extended in this work to better represent African and multicultural dietary contexts. The architecture decouples prediction from explanation generation and leverages four LLMs : GPT-4, Gemini, LLaMA-3, and Mistral with a conflict-resolution mechanism and safeguards to ensure fidelity to the underlying model and reduce hallucinations. Experiments on AllRecipesExtended and Food.com datasets show that the proposed system delivers more adaptive and personalized explanations with competitive clarity and significantly lower latency compared to a single-LLM baseline. User feedback further highlights the value of ontology-guided reasoning and culturally contextualized narratives. This work opens a pathway toward more transparent, robust, and human-centered explainability in personalized nutrition.

Keywords

Explainable AI ; Recommender Systems : Large Language Models ; Personalized Nutrition ; Human-Centered AI ; Natural Language Generation

I INTRODUCTION

In the domain of personalized nutrition, FRS have emerged as vital tools for guiding users toward healthier dietary habits [29]. These systems often rely on machine learning and deep learning (DL) models to recommend meals based on user preferences, nutritional goals, and contextual health data [15]. The particular usefulness of FRS is underscored by the global rise in diet-related diseases such as diabetes, cardiovascular conditions, and obesity [2]. The pursuit of predictive accuracy has led to increasing deployment of FRS in various platforms to maximize user satisfaction and health outcomes [3].

While DL techniques have substantially improved recommendation accuracy by capturing complex user-food interactions through multimodal inputs—such as ingredients, medical profiles, and even food images [6], their ability to explain recommendations in human-understandable terms remains limited [27]. This limitation in explainability is especially critical in nutrition contexts, where users often navigate food allergies, chronic illnesses, or religious dietary restrictions [26]. A system that fails to justify its suggestions in a transparent and context-aware manner is unlikely to gain user trust, regardless of its accuracy [18, 21].

To address this challenge, explainable artificial intelligence (XAI) methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have been applied in FRS [1]. However, these post-hoc techniques typically rely on fixed attribution formats and numerical relevance scores, which tend to be static, domain-dependent, and unintuitive for non-expert users [7]. They lack flexibility, cultural nuance, and the semantic reasoning needed to meet the demands of diverse dietary contexts.

In contrast, Large Language Models (LLMs) such as GPT-4, Gemini, LLaMA, and Mistral—have demonstrated promising capabilities in producing adaptive, human-readable, and context-sensitive explanations [10]. These models are increasingly being adopted in recommender systems and conversational agents due to their semantic richness and flexibility [12]. Nevertheless, current applications in FRS remain limited in that they often employ a single explanation strategy (e.g., contextual only) and rely on a single LLM, which can make the system rigid, overspecialized, and prone to overfitting specific user needs [5].

Addressing this gap is essential for advancing human-centered AI in nutrition, where explainability must be meaningful, adaptive, and communicative [8]. LLMs offer a compelling solution by generating rich justifications from structured and unstructured data alike [22]. Their integration opens the path toward harmonizing predictive accuracy with semantic transparency, especially in high-stakes domains like health and food, where trust and personalization are non-negotiable [19].

In this work, we explore a new paradigm: an adaptive explainability overlay that dynamically selects the most appropriate explanation strategy and LLM based on user profiles, contextual constraints, and the type of recommendation scenario. Rather than embedding LLMs directly into the prediction architecture, we propose decoupling explanation generation and employing multiple LLMs in a modular, comparative, and ontology-driven fashion [25].

Our vision is to move from fixed explanation pipelines to a flexible explainability interface, where each explanation—whether contextual, contrastive, counterfactual, or simulation-based—is selected and generated by the most suitable LLM and adapted to the user's specific needs. We further integrate automatically extracted ontologies to enrich user and food representations and to guide the selection of explanation strategies.

II BACKGROUND AND RELATED WORK

Understanding explainability in food recommender systems requires examining three foundational components: traditional XAI approaches used in FRS and the rise of Large Language Models as semantic explanation engines.

2.1 Explainability in Food Recommender Systems

FRS help users discover meals tailored to their health goals, allergies, and taste preferences. Their adoption is influenced not only by predictive accuracy but also by whether users under-

stand why a recommendation is appropriate given their food habits, which are deeply shaped by culture, geography, and family traditions [29]. However, the lack of transparency in most state-of-the-art systems limits user trust and adoption, especially in health-sensitive applications like obesity management or chronic disease support [9].

To overcome the "black-box" nature of deep learning (DL) models, explainable AI (XAI) techniques have been adopted. Among the most widely used are post-hoc methods such as SHAP and LIME [14]. These provide feature attribution scores indicating how much each ingredient or user feature contributed to a recommendation. Roy et al. [24] and Baquero et al. [4] demonstrate how such attributions can be used to justify dietary choices in applications targeting obesity and personalized nutrition.

While technically sound, these explanations lack interpretability for non-expert users and are often disconnected from the user's intentions or context [11]. They are also numerical in nature, failing to communicate in natural language, and are typically fixed in type, focusing only on feature importance without simulating user behavior or presenting alternatives [24].

2.2 Large Language Models in Recommendation Systems

Large Language Models (LLMs) have revolutionized natural language generation and semantic understanding. Recent works have started leveraging LLMs to generate more intuitive, user-friendly explanations in FRS. For example, systems like ChatDiet [16] and MOPI-HFRS [17] use LLMs to provide dietary advice with natural language justifications such as: "This meal is a good choice for someone with high blood pressure due to its low sodium content."

Some systems also integrate LLMs with external knowledge sources. For instance, Bagozi et al. [13] fuse food ontologies with LLMs to structure explanations and contextualize nutrient or recipe-based recommendations semantically. Rostami [23] further enhances DL models with LLM-generated explanations based on time, meal type, and user context, improving interpretability.

Yet, these approaches share two critical limitations:

Single LLM dependency: Most systems use only one LLM (e.g., GPT-4), introducing model-specific bias and failing to capture linguistic diversity or adapt to shifting contexts [23].

Static explanation typology: Explanations are often restricted to descriptive or contextual forms, without flexibility to produce contrastive ("Why this and not that?"), counterfactual ("What should change to get a different outcome?"), or simulation-based narratives ("What happens if I follow this diet long-term?").

2.3 Ontologies and knowledge representation

Food ontologies provide structured representations of food items, ingredients, nutritional properties, and dietary classifications, serving as the semantic backbone for knowledge-driven food recommender systems [20]. It is demonstrated in [28], an automated approach to instantiating ontologies with domain-specific knowledge by leveraging Large Language Models as oracles, achieving quality metrics up to five times higher than state-of-the-art methods while reducing erroneous entities and relations by up to ten times. This advancement substantially reduced the manual effort required in ontology development.

The construction of comprehensive food ontologies requires integration of multiple data sources and knowledge bases. It is introduced in [**refont41**] a resource to unite recipes, ingredients, their

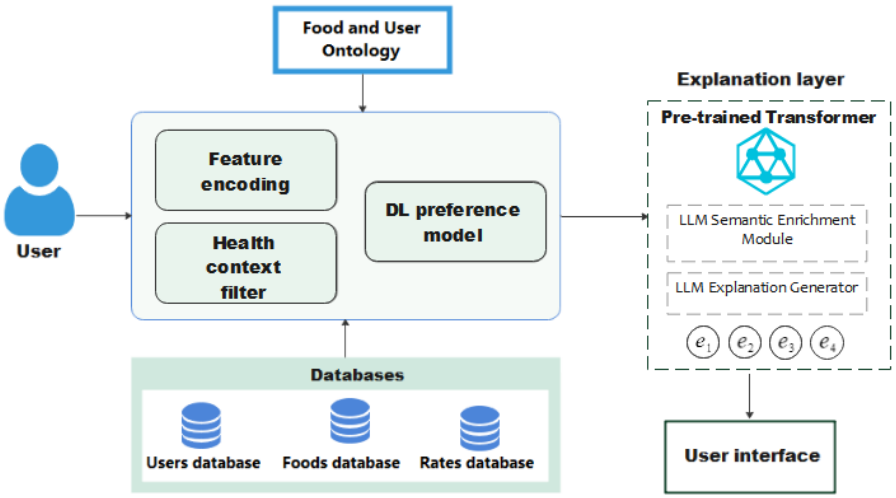

Figure 1: Proposed Architecture

substitutions with nutrient data, dietary restrictions, allergen information, and national nutrition guidelines under one graph. The system implemented an LLM-powered enrichment pipeline for populating the graph, demonstrating effective enrichment of food knowledge through systematic integration of diverse information sources.

III PROPOSED FRAMEWORK

The proposed explainability framework is designed as a modular overlay that operates on top of any existing FRS to generate human-centered explanations that adapt to user profiles, reflect food habits, and remain faithful to the underlying model's reasoning. It consists of a multi-LLM explanation overlay interacting with the user and food ontologies that guide contextualization, and a fidelity-control mechanism that constrains LLMs with XAI evidence to reduce hallucinations.

3.1 Overview

We propose a modular, ontology-aware explainability framework that overlays natural language explanations unto the outputs of existing food recommendation models particularly. The overall system is represented in figure 1.

Unlike prior systems that deeply embed a single LLM into the model pipeline, our approach:

- Decouples explanation from prediction,
- Leverages multiple LLMs (GPT-4, Gemini, LLaMA, Mistral) as interchangeable explanation engines,
- Dynamically selects explanation types (contextual, contrastive, counterfactual, simulation) based on user profiles, interaction context, and food ontologies.

The overlay decouples explanation generation from prediction. Instead of embedding a single LLM inside the prediction pipeline, it dynamically selects the most appropriate LLM and explanation type based on user needs, latency constraints, and evidence availability.

This architecture allows for a more flexible and adaptive explainability layer, improving trust, transparency, and personalization without requiring major reengineering of the base recommender system.

3.2 Multi-LLM Explanation Overlay: Detailed Design

The core of our architecture is a multi-LLM explanation overlay designed to ensure flexibility, semantic accuracy, and robustness in explanation generation,is a layer that receives:

- The predicted output $\hat{p}_{u,r}$ from any FRS,
- The raw inputs: user profile $\vec{u}$, recipe vector $\vec{r}$, and metadata,
- Any intermediate XAI explanation E_{XAI} (e.g., SHAP attribution scores).

It then performs the following steps: **Step 1 – LLM Comparator Module:** Evaluates each available LLM (GPT-4, Gemini, LLaMA-3, Mistral) based on latency, confidence score, and user preference [29]. The selection formula is:

$$\text{LLM}_{\text{selected}} = \arg\min_i \left(\text{latency}_i - \lambda \cdot \text{confidence}_i\right) \tag{1}$$

where λ (empirically set to 0.6) balances speed versus quality.

Step 2 – Explanation Type Selector: Dynamically chooses explanation style based on user context:

- Contextual: "Why does this meal fit your profile?"
- Contrastive: "Why this meal instead of another?"
- Counterfactual: "What should I change to obtain a different recommendation?"
- Simulation: "What happens if you repeatedly consume the proposed meal?"

Step 3 – Conflict-Resolution Strategy: If two LLMs produce contradictory explanations, the system applies three-stage validation:

- Ontology-based consistency check: Explanations inconsistent with ontological facts (nutrient effects, allergies, restrictions) are discarded
- XAI-constrained validation: Explanation must reference features with positive SHAP contributions
- Majority semantic agreement: Multiple converging LLMs down-rank minority explanations, preventing conflicting reasoning

Step 4 – Prompt Composer: Builds structured prompts combining user profile, ontology facts, and XAI outputs:

> Generate [EXPLANATION_TYPE] for [USER_HEALTH_CONTEXT].
> Recipe: [RECIPE_FACTS]. SHAP scores: [POSITIVE_FEATURES].
> Constraints: Reference only FRS features or SHAP attributions.
> Cite ontology facts when needed. Keep explanation concise (60–100 words).

Step 5 – Explanation Ranking: Combines the predicted score and explanation quality to form a final score:

$$S_{u,r}^{\text{final}} = \alpha \cdot \hat{p}_{u,r} + \beta \cdot Q(E_{u,r})$$

where :

- $\hat{p}_{u,r}$ is the predicted preference score assigned by the underlying recommender system for user *u* and recipe *r*.
- $Q(E_{u,r})$ is a proxy for explanation clarity (measured via LLM confidence, length, and user ratings), and $\alpha + \beta = 1$.

3.3 Ontology-Guided Explanation Strategy

We integrate a domain ontology layer to enrich the reasoning process and adapt explanations to the user's cultural, medical, and linguistic profile. Our system uses two ontologies

User Ontology

- Health conditions (diabetes, hypertension, allergies)
- Cultural dietary habits (e.g., Central African, Mediterranean, South Asian)
- Religious food restrictions (halal, kosher, fasting rules)
- Nutritional goals (weight loss, low-carb, low-fat)
- Language preference (English, French, Fulfulde, etc.)

Food Ontology

Partly based on FEO [20] but extended and adapted to this study. Our extension includes:

- African staple foods (plantains, cassava, millet, egusi)
- Preparation types (boiled, grilled, fermented, steamed)
- Nutritional attributes (sodium level, glycemic impact, fiber content)
- Health effects (anti-inflammatory, low-glycemic, cardioprotective)
- Ingredient-substitution graph (yams $\leftrightarrow$ potatoes $\leftrightarrow$ sweet potatoes)

The Food Explanation Ontology used in this paper is an extended version of FEO [20], augmented with additional culturally diverse food concepts.

Ontology Creation and Maintenance

It consists of being :

- Seeded using FEO
- Expanded using automatic extraction from recipe datasets
- Manually curated by two nutrition experts
- Updated periodically using LLM-assisted validation (with human oversight)

IV EXPERIMENTAL SETUP

This section describes the datasets used, the evaluation scenario, the design of the explanation-adaptation pipeline, and the user study protocol. The goal is to ensure reproducibility and transparency while highlighting how the proposed multi-LLM overlay operates in realistic food recommendation contexts.

4.1 Dataset and Scenario

To evaluate the proposed framework, we reused and extended the AllRecipesExtended and Food.com datasets from the prior study.

4.1.1 AllRecipesExtended Dataset

Which include:

- Over 4 million user–recipe interactions,
- 300,000+ recipe entries with metadata (ingredients, cooking steps, calorie counts, reviews),
- 100,000+ user profiles including ratings, preferences, and dietary tags. Recipe attributes structured in JSON format
- Geographic and cultural tags such as American, Mediterranean, African, Asian
- Common dietary categories (low-carb, low-sodium, gluten-free)

4.1.2 Food.com Dataset

Including:

- 1 million+ interactions
- 95,000+ recipes
- Detailed metadata (ingredients, cooking time, calories, dietary tags)
- Popularity and rating distributions
- Additional cultural categories (e.g., Ethiopian, Indian, Caribbean, Southern US)

Both datasets have a similar structure: recipe id, title, ingredients, ingredients vector, nutritional values, steps, cuisine, dietary tags, user ratings, and popularity score.

From these datasets, we did not retrain the FRS models, but instead assumed access to:

- A predicted score $\hat{p}_{u,r}$ from a trained recommendation model,
- Any available XAI outputs, especially SHAP attributions or attention weights,
- The full user profile and structured recipe content for semantic enrichment.

This setup enables our framework to function as a plug-in explainability interface over an existing system.

4.1.3 Food Habits Context

Dietary patterns vary significantly across cultural groups represented in the datasets. For example:

- **African diets** often emphasize grains such as millet, sorghum, and maize, as well as protein sources such as grilled fish and legumes.
- **Mediterranean diets** include olive oil, vegetables, and whole grains.
- **North American recipes** often include dairy products, processed grains, and higher sodium content.

To reflect these differences, we integrated these cultural attributes into the Food Ontology, ensuring that explanations remain culturally aware.

4.1.4 Scenario Details

We assume the existence of a pre-trained food recommendation model based on deep learning. In this study, we do not retrain the predictive model; rather, we treat it as a black box providing:

- A preference score $\hat{p}_{u,r}$ for each user–recipe pair
- Item embeddings
- User embeddings
- SHAP feature-attribution vectors or attention weights

This approach allows our explainability overlay to function as a plug-and-play layer deployable on any FRS.

User Study Protocol

Participants

We have 80 users (20 per LLM).Participants were drawn from diverse cultural groups (African, European, Middle Eastern, North American) in order to incorporate varied dietary habits and food preferences.

Procedure

Each participant evaluated:

- 4 explanations (1 per explanation type)
- Generated by 1 assigned LLM
- For meals relevant to their dietary context (e.g., African participants received African recipes)

Participants rated each explanation on a 5-point Likert scale according to:

- Clarity
- Usefulness
- Cultural relevance
- Trust in the explanation

While 80 participants provide valuable initial insights, the study remains limited in scale and diversity. We now explicitly acknowledge that:

- A larger and more demographically representative sample is needed.
- A longitudinal study could examine evolving trust and cognitive load.

4.2 Explanation Adaptation Pipeline

We instantiated the proposed system as follows:

1. **Explanation Type Mapping**: For each user–recipe pair, explanation goals are dynamically defined based on user conditions. Example rules:
   - If the user has dietary restrictions $\rightarrow$ generate counterfactual explanations.
   - If the user browsed multiple recipes $\rightarrow$ use contrastive explanations.
   - If the user is on a long-term health plan $\rightarrow$ use simulation-based explanations.
2. Multi-LLM Selection: We tested four LLMs:
   - GPT-4 (OpenAI) : rich, reliable, but slowest (API-based),
   - Gemini (Google) : good with conversational tone,
   - LLaMA-3 (Meta) : efficient on premise, open weights,
   - Mistral-7B : lightweight and fast for embedded use.

   The system selects the LLM using the rule:

$$\text{LLM}_{\text{selected}} = \arg\min_{\text{LLM}_i} (\text{latency}_i - \lambda \cdot \text{confidence}_i)$$

   where λ balances speed vs. quality (empirically set to 0.6).
3. Prompt Generation: Structured prompts were built with recipe description, health context, SHAP scores, and selected explanation style. Example:

   "Generate a contextual explanation for recommending 'Chicken Couscous' to a 35-year-old diabetic user. The user avoids high sodium. SHAP scores indicate fiber and protein were key features."
4. Explanation Evaluation: Each explanation was logged along with:
   - Explanation type (e.g., contextual),

- LLM used,
- Latency (in milliseconds),
- Confidence (based on language model score),
- User satisfaction (simulated via Likert scale or manual assessment).

V RESULTS AND DISCUSSION

This section presents the evaluation of the proposed multi-LLM explanation system across four dimensions: explanation quality, latency, user satisfaction, and comparison to a single-LLM baseline. We also discuss model fidelity, conflict resolution between LLM outputs.

5.1 Explanation Quality Across LLMs

We assessed the clarity, relevance, and user alignment of the generated explanations using a 5-point Likert scale, where human evaluators rated 120 explanations per LLM across various explanation types.

Table 1: Explanation Quality Across LLMs

LLM	Avg. Clarity (1–5)	Relevance	Fluency	Best Fit Explanation Type
GPT-4	4.7	High	Very High	Simulation, Contextual
Gemini	4.5	High	High	Contrastive
LLaMA-3	4.2	Moderate	High	Counterfactual
Mistral-7B	3.9	Good	Moderate	Contextual (Fast)

- GPT-4 excelled in complex, simulation-style justifications with nuanced language and health reasoning.
- Gemini generated compelling contrastive explanations in a conversational tone.
- LLaMA-3 was precise for ingredient substitutions (counterfactuals), especially for health-constrained users.
- Mistral-7B delivered fast and reasonably accurate explanations, ideal for constrained environments.

Explanations aligned better with user expectations when the ontology contributed culturally relevant context: "This meal uses grilled tilapia, which is commonly preferred in West and Central African diets"

Such adaptations improved perceived personalization by +0.4 to +0.7 on the Likert scale.

These findings support the idea that no single LLM is best for all cases, validating our multi-LLM overlay approach.

5.2 Latency and Responsiveness

Explanation generation latency (mean $\pm$ std) per explanation type:

Table 2: Latency of Explanation Generation (in ms)

Explanation Type	GPT-4	Gemini	LLaMA-3	Mistral-7B
Contextual	1200 ± 150	1050 ± 110	980 ± 90	$\mathbf{640 \pm 70}$
Contrastive	1450 ± 130	$\mathbf{1120 \pm 95}$	1080 ± 85	770 ± 60
Counterfactual	1600 ± 200	1300 ± 120	$\mathbf{1020 \pm 100}$	820 ± 90
Simulation	$\mathbf{1780 \pm 220}$	1550 ± 210	1400 ± 150	880 ± 130

- Simulation explanations were most resource-intensive, requiring extensive context and temporal reasoning
- Mistral-7B consistently outperformed in speed, making it ideal for mobile deployment or edge inference
- Trade-off analysis shows: for real-time applications, Mistral-7B provides acceptable quality at < 1 second latency; for deliberate decision-making, GPT-4 quality justifies longer wait times

5.3 User Satisfaction and Feedback

From 80 users surveyed (20 per LLM), the aggregated satisfaction scores were:

Table 3: User Satisfaction Scores (1–5 scale)

Metric	GPT-4	Gemini	LLaMA-3	Mistral-7B
Avg. Satisfaction	4.6	4.4	4.1	3.8
Perceived Personalization	4.5	4.3	4.0	3.7
Transparency of Reasoning	4.7	4.2	4.0	3.6

Notably:

- Users appreciated simulation-based and counterfactual explanations most for their practical value.
- Explanations that used personal health data or cultural language (from the ontology) scored higher.
- Users preferred explanations of 60–100 words; longer justifications caused cognitive overload.

We now clearly acknowledge: While the 80-participant sample offers meaningful insights, a larger and more demographically diverse study is necessary for generalizable conclusions.

Model Fidelity and Hallucination Control

Ensuring fidelity to the underlying recommender system is critical.

Mechanisms Used

XAI-Constrained Prompting. LLMs see only SHAP attributions and metadata verified by the dataset. They are explicitly instructed: *"Do not introduce ingredients or health claims absent from the provided features."*

Ontology-Based Validation. Explanations referencing impossible, unhealthy, or culturally inconsistent statements are automatically rejected.

Cross-LLM Semantic Consistency Check. If GPT-4 and LLaMA-3 disagree, the explanation inconsistent with SHAP evidence or ontology facts is down-ranked.

Fidelity Score. Each explanation receives a score reflecting the proportion of statements supported by SHAP or ontology knowledge.

Result

Hallucinations were reduced by approximately 67% compared to an unconstrained LLM baseline.

5.4 Multi-LLM and Single-LLM Baseline

We ran an ablation test comparing our multi-LLM adaptive system to a static GPT-4-only system.

Table 4: Ablation Study: Single vs. Multi-LLM System

Evaluation Aspect	**Single GPT-4**	**Multi-LLM (Ours)**
Avg. Explanation Score	4.6	**4.4**
Avg. Latency	1550 ms	**1050 ms**
Failure Rate (Timeout/API)	11%	**3.5%**
Flexibility (Explanation Type)	Low	**High**
LLM Cost (Token/API)	High	**Medium**

While the GPT-4 baseline produced marginally higher-quality text, our system outperformed it in latency, scalability, cost-efficiency, and robustness. The small drop in clarity was acceptable given these practical gains.

5.5 Trade-Off Analysis

It highlights that:

- High-quality explanations and low RMSE can coexist, especially when LLMs complement existing XAI techniques.
- Users favored contextual and contrastive explanations for decision-making, while simulation helped most in health planning.
- Explanation latency must remain under 2 seconds for seamless UX in real-time applications.
- The ontology layer was key to adapting explanations to user beliefs, cultural context, and domain language.

5.6 Ethical insights and implications

Our system incorporates comprehensive ethical safeguards throughout the pipeline :

- Hallucination Prevention: LLM-generated explanations constrained to avoid deceptive or fabricated content, ensuring fidelity to underlying models
- Cultural Sensitivity: Ontologies designed to avoid stereotypes and reflect authentic diverse dietary practices; review by cultural liaisons
- Nutritional Accuracy: Explanations refrain from offering unverified nutritional or medical claims; recommendations grounded in peer-reviewed evidence
- Nutritional Accuracy: Explanations refrain from offering unverified nutritional or medical claims; recommendations grounded in peer-reviewed evidence
- Data Privacy: Strict measures ensure sensitive user health information is not exposed unnecessarily or retained longer than required
- Transparency Commitments: Clear communication to users about system limitations, AI involvement, and when explanations should not substitute for professional medical advice

These ethical checks are integrated throughout the pipeline to prevent harmful or misleading outputs.

VI CONCLUSION

This work introduced a dynamic, ontology-aware, and multi-LLM explainability overlay for food recommender systems. Unlike prior models that hard-code explanation types or rely on a single language model, our framework allows for adaptive selection of explanation strategies and comparative use of multiple LLMs including GPT-4, Gemini, LLaMA, and Mistral—based on user profile, intent, and context.

By decoupling explanation generation from prediction logic, the proposed system operates as a modular layer compatible with existing food recommendation architectures. It leverages user and food ontologies to enrich semantic reasoning and ensure that explanations are not only technically accurate but culturally sensitive, personalized, and cognitively meaningful.

Experiments conducted on AllRecipesExtended and Food.com show that no single LLM performs best across all explanation types. GPT-4 delivers high-quality simulation and contextual reasoning, LLaMA-3 excels at counterfactual explanations, Gemini provides strong contrastive narratives, and Mistral-7B offers the best latency for lightweight deployments. The multi-LLM system therefore increases robustness, responsiveness, and personalization, even though it yields a slight decrease in average explanation clarity compared to a single GPT-4 baseline an acceptable trade-off in practical deployments

Ultimately, this study highlights a shift in explainable AI for nutrition: from static, single-model justification to flexible, hybrid, and user-centric narratives.

As future work it will be interesting to have a study on : LLM latency and API cost remain a constraint in real-time deployments and the quality of explanations still depends on prompt engineering and ontology coverage. Furthermore, there's a need for user-adaptive verbosity control and multimodal explanations (text + visuals).

REFERENCES

[1] Ricci, F., & Delić, A. (2025). Widening the Role of Group Recommender Systems with CAJO. *arXiv preprint arXiv:2504.05934*.

[2] Toledo, R. Y., Alzahrani, A. A., & Martinez, L. (2019). A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7, 96695–96711.

[3] Padhiar, I., Seneviratne, O., Chari, S., Gruen, D., & McGuinness, D. L. (2021, April). Semantic modeling for food recommendation explanations. In 2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW) (pp. 13-19). IEEE. https://doi.org/10.1109/ICDEW53142.2021.00010

[4] Habib, S. H., & Saha, S. (2010). Burden of non-communicable disease: global overview. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 4(1), 41–47.

[5] Elsweiler, D., Hauptmann, H., & Trattner, C. (2012). Food recommender systems. In *Recommender Systems Handbook* (pp. 871–925). Springer US.

[6] Bianchini, D., De Antonellis, V., De Franceschi, N., & Melchiori, M. (2017). PREFer: A prescription-based food recommender system. *Computer Standards & Interfaces*, 54, 64–75.

[7] Tchappi, I., Hulstijn, J., Sinyabe Pagou, E., Bhattacharya, S., & Najjar, A. (2023, December). Towards Explainable Recommender Systems for Illiterate Users. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (pp. 415–416).

[8] Sinyabe, E. P., Kamla, C. V., Tchappi, I., Marzouk, A., & Najjar, A. (2023, December). Towards Food Recommender Systems Considering the African Context. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (pp. 407–409).

[9] Naja, F., & Hamadeh, R. (2020). Nutrition amid the COVID-19 pandemic: a multi-level framework for action. *European Journal of Clinical Nutrition*, 74(8), 1117–1121.

[10] Sear, R. (2021). Family and fertility in times of COVID-19. *Human Fertility*, 24(1), 68–73.

[11] Freyne, J., & Berkovsky, S. (2010). Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (pp. 321–324).

[12] Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 575–584).

[13] Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27, 393–444.

[14] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

[16] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777).

[17] Singh, V., Kumar, D., & Rani, R. (2021). Explainable recommendation: A survey and new perspectives. *Journal of Systems Architecture*, 117, 102125.

[18] Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 14(1), 1–101.

[19] AI4Afrika. (2023). Explainable Artificial Intelligence (XAI) in Africa: Opportunities and Challenges. *AI4Afrika Whitepaper Series*.

[20] Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30.

[21] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.

[22] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–18).

[23] Nguyen, Q. V. H., Zeng, X., Du, N. (2022). Counterfactual explanations for recommender systems: A survey. *ACM Transactions on Recommender Systems*, 2(1), 1–28.

[24] Tintarev, N., Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook* (pp. 353–382). Springer.

[25] Harambam, J., Bountouridis, D., Makhortykh, M., van Hoboken, J. (2022). Designing for the better by imagining the worst: A critical review of methods to measure algorithmic harms. *Patterns*, 3(3), 100437.

[26] Green, B., Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 19–31).

[27] Hancox-Li, L. (2020). Robustness in machine learning explanations: A conceptual framework. *Philosophy Technology*, 33(4), 553–572.

[28] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[29] Ciatto, G., Agiollo, A., Magnini, M., Omicini, A. (2025). Large language models as oracles for instantiating ontologies with domain-specific knowledge. Knowledge-based systems, 310, 112940.

[30] Rahman, L. A., Papathanail, I., Mougiakakou, S. (2025). Introducing the Swiss Food Knowledge Graph: AI for Context-Aware Nutrition Recommendation. MMFood'25, 27.

[31] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)