Personalized Causal Graph Reasoning for LLMs: A Case Study on Dietary Recommendations

Anonymous ACL submission

Abstract

Large Language Models (LLMs) effectively leverage common sense knowledge for general reasoning, yet they struggle with personalized reasoning when tasked with interpreting multifactor personal data. This limitation restricts their applicability in domains that require context-aware decision-making tailored to individuals. This paper introduces Personalized Causal Graph Reasoning as an agentic framework that enhances LLM reasoning by incorporating personal causal graphs derived from data of individuals. These graphs provide a foundation that guides the LLM's reasoning process. We evaluate it on a case study on nutrient-oriented dietary recommendations, which requires personal reasoning due to the implicit unique dietary effects. We propose a counterfactual evaluation to estimate the efficiency of LLM-recommended foods for glucose management. Results demonstrate that the proposed method efficiently provides personalized dietary recommendations to reduce average glucose iAUC across three time windows, which outperforms the previous approach. LLM-as-a-judge evaluation results indicate that our proposed method enhances personalization in the reasoning process.

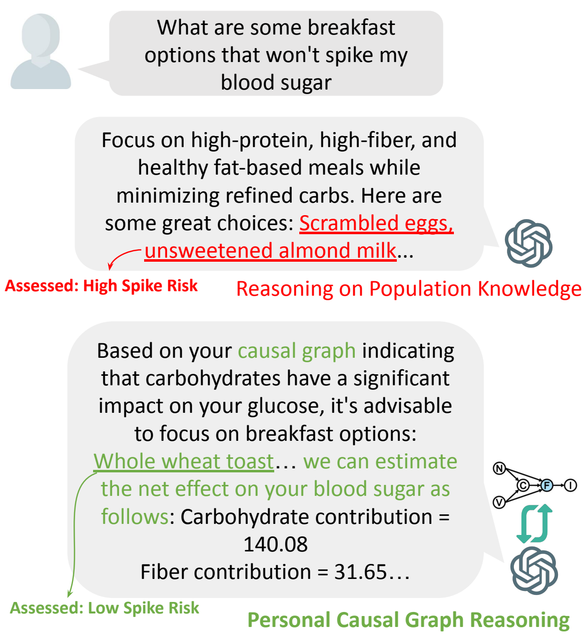

Figure 1: Comparison between a standalone LLM and the proposed Personalized Causal Graph Reasoning for dietary recommendations. The standalone LLM relies on generic reasoning and may provide risky advice, while our method utilizes a personal causal graph to assess individual metabolic responses for more precise recommendations.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generic reasoning by leveraging inherent knowledge to generalize across diverse domains. However, they struggle to incorporate complex, multifactor personal data, which is a critical requirement for real-world decision-making tasks (Chen et al., 2024; Halevy and Dwivedi-Yu, 2023). In domains where context-aware reasoning is essential, such as healthcare, LLMs fail to go beyond broad, population-level knowledge and instead produce generic responses that overlook individual-specific dependencies (Tanneru et al., 2024; Yu et al., 2024; Subramanian et al., 2024). This limitation reduces their practicality when required to align with a user's unique characteristics and needs.

This limitation arises from LLMs' reasoning process that relies solely on population-level knowledge, which impairs their ability to model relationships between personal factors (Hu et al., 2024; Yang et al., 2024a). For comparison, human decision-making is inherently contextual in its understanding of how personal factors interact (Weiner, 2004). For instance, in nutrient-based health interventions, the effectiveness of dietary changes depends on a combination of an individual's metabolic history, underlying conditions, specific nutrient deficiencies, and general nutritional principles. As demonstrated in Figure 1, standalone LLM often fails in such settings because they do not have a structured mechanism to reason over personal causal understanding of dietary effects from data (Yang et al., 2024b).

To address this challenge, we introduce Personalized Causal Graph Reasoning that enhances LLMs' personalized reasoning within an agentic framework. The framework constructs a personalized causal graph for each user based on longitudinal health data for capturing the unique user characteristics. LLMs then reason over this structured representation by dynamically exploring causal graphs and retrieving relevant external knowledge to generate personalized recommendations. Unlike conventional LLMs that process user queries under a generic reasoning paradigm, our approach provides a structured foundation that allows LLMs to perform personalized inference over explicit causal relationships.

To evaluate the effectiveness of the proposed framework, we conduct a case study on dietary recommendations. Using a dataset comprising continuous glucose monitoring data, food intake logs, and physical activity records, we construct personal causal graphs that capture the relationship between nutrient intake and glucose regulation. The LLM utilizes these graphs to simulate dietary interventions and recommend foods that are expected to improve glucose stability. We propose counterfactual evaluation to assess whether the model's recommended foods would have led to actual health improvements (Mahoney and Barrenechea, 2019). Additionally, we employ LLM-as-a-judge to assess whether our method improves the reasoning process by making it more personalized.

This paper's contribution is two-fold:

- We introduce Personalized Causal Graph Reasoning that enables LLMs to perform personalized reasoning by incorporating causal graphs derived from personal data.
- We evaluate the proposed framework through a case study on personalized dietary recommendations. To assess its effectiveness, we introduce a counterfactual evaluation method that estimates the potential glucose impact of LLM-generated food recommendations.

2 Related Works

2.1 LLMs and Reasoning

Several techniques were proposed to elevate LLMs' general reasoning tasks. Chain-of-Thought (CoT) as a classic prompting method enhances problem-solving by enabling the generation of intermediate reasoning steps (Wei et al., 2022). Building upon CoT, approaches such as Tree of Thoughts (ToT) and Graph of Thoughts (GoT) have been proposed to further refine LLM reasoning in a more structured manner (Yao et al., 2024; Besta et al., 2024). ToT allows models to explore multiple reasoning paths, while GoT models information as an arbitrary graph, to combine various reasoning paths into cohesive outcomes. By incorporating structured reasoning techniques, LLMs have demonstrated promising performance in decision-making tasks that require dietary knowledge(Azimi et al., 2025).

Beyond prompting techniques, studies have explored iterative reasoning refinements. These include generating multiple reasoning paths and selecting the most consistent one, applying stepwise verification, and integrating feedback mechanisms to improve logical consistency (Havrilla et al., 2024; Li et al., 2022; Nathani et al., 2023). Additionally, Gao et al. propose meta-reasoning, where LLMs dynamically select and apply different reasoning strategies based on the problem context (Gao et al., 2024). The Reasoning on Graphs (RoG) synergizes LLMs with knowledge graphs to enable faithful and interpretable reasoning (Luo et al., 2023). RoG employs a planning-retrieval-reasoning framework, where relation paths grounded in knowledge graphs are generated as faithful plans. These plans are then used to retrieve valid reasoning paths from the graphs.

Their reliance on large-scale, population-level data limits the applicability in contexts requiring precise, personalized reasoning. This limitation arises because they primarily operate on unstructured text prompts and lack mechanisms to incorporate structured representations of personal information. Consequently, they struggle to model the intricate interplay of personal factors necessary for tailored decision-making. However, approaches like RoG offer promising directions to overcome these challenges by using graphs as a bridge to connect personal data to LLM reasoning.

2.2 Nutrition-Oriented Recommendations

Nutrition recommendation systems aim to provide dietary advice tailored to individual health needs. Traditional systems often use collaborative filtering, leveraging user interactions and preferences to generate suggestions (de Hoogh et al., 2023; Abhari et al., 2019; Nijman et al., 2007). However, they fail to capture the complex causal relationships between dietary factors and health outcomes and struggle to adapt dynamically to changes in an individual's health status (Luo et al., 2024; Verma et al., 2018).

We focus on recent advancements that have explored the performance of LLMs on personalized dietary recommendations (Xue et al., 2024; Anjanamma et al., 2024; Yang et al., 2024b). For instance, ChatDiet combines personal and population models to generate tailored food suggestions (Yang et al., 2024b). It employs Retrieval-Augmented Generation (RAG) to retrieve triplets from a pre-constructed causal graph, then structures them into prompts that guide the LLM in recommendation generation. While this approach enhances personalization, it relies on a fixed pattern of retrieving specific triplets to inform the LLM's responses. Despite its promise, further improvements are needed to enable more structured, adaptive reasoning in LLM-based nutrition systems.

3 Personalized Causal Graph Reasoning for Dietary Recommendations

This section introduces the Personalized Causal Graph Reasoning framework. The objective of the proposed framework is to enable an LLM agent to reason over a personal causal graph, which encodes the individual's dietary-health interactions. Figure 2 illustrates the workflow of this reasoning process on dietary recommendation. Unlike conventional LLM-based recommendation approaches that rely purely on text-based correlations, our method focuses on guiding the LLM's reasoning by leveraging the structured causal dependencies between nutrients, biomarkers and health outcomes.

We define an individual's personal causal graph $G_i = (V_i, E_i, W_i)$, where V_i represents the set of nodes (dietary factors, biomarkers, metabolic conditions), E_i denotes the directed causal edges between variables, and W_i encodes the strength of causal relationships. Given the graph, the LLM agent performs a structured reasoning process that consists of five key stages: goal identification, graph traversal, external knowledge retrieval, verification, and structured response generation.

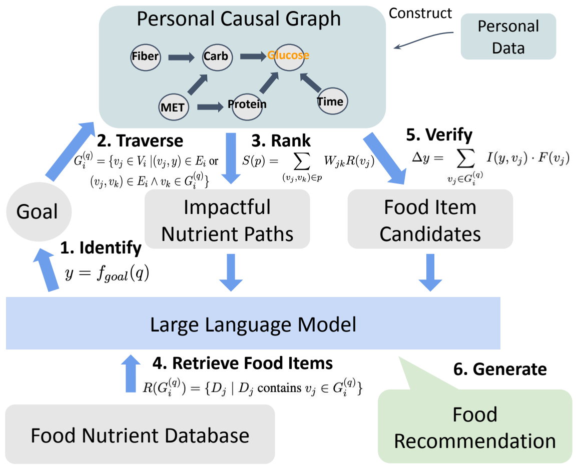

Figure 2: Demonstration of the workflow of the proposed Personalized Causal Graph Reasoning framework on dietary recommendation.

3.1 Goal Identification

When a user submits a query q, the LLM first identifies the primary objective and map it to a corresponding node y in the personal causal graph. Formally, given the query q, the LLM applies a mapping function f_{goal} to determine the target variable:

$$y = f_{goal}(q), \quad y \in V_i \tag{1}$$

For instance, if the user asks, "How can I prevent glucose spikes?", the target y would correspond to the glucose incremental area under the curve in the personal causal graph.

3.2 Personal Causal Graph Traversal and Paths Ranking

Once the target variable is identified, the LLM agent traverses the personal causal graph to identify relevant dietary factors. The objective is to find upstream nodes (nutrient intake variables) that causally influence y. The model retrieves the subgraph $G_i^{(q)}$ consisting of all relevant causal paths leading to y:

$$G_i^{(q)} = \{v_j \in V_i \,|(v_j, y) \in E_i \text{ or } (v_j, v_k) \in E_i \wedge v_k \in G_i^{(q)}\} \tag{2}$$

The traversal process prioritizes paths based on causal effect strength. Each retrieved path $p = \{v_1, v_2, ..., y\}$ is assigned a causal relevance score

$S(p)$ computed as:

$$S(p) = \sum_{(v_j, v_k) \in p} W_{jk} R(v_j) \tag{3}$$

where W_{jk} represents the causal strength between v_j and v_k, and $R(v_j)$ captures the individual's historical consumption of v_j. Paths with higher causal scores are given greater weight in generating recommendations. For instance, if an individual has consistently consumed high-glycemic carbohydrates, those pathways might be ranked lower in favor of fiber-rich interventions.

3.3 External Knowledge Retrieval

The personal causal graph identifies key dietary factors, but it does not specify which foods to recommend. To bridge this gap, the LLM agent queries a food database from Yang et al. (Yang et al., 2024b) to retrieve relevant nutritional information:

$$R(G_i^{(q)}) = \{D_j \mid D_j \text{ contains } v_j \in G_i^{(q)}\} \tag{4}$$

where $R(G_i^{(q)})$ represents the set of foods with a high concentration of impactful nutrients, and D_j denotes an individual food item. The retrieved food items are ranked based on their concentration of the identified nutrient.

3.4 Verification by Simulating Dietary Effects

After selecting a food item and its corresponding nutrient composition, the LLM agent simulates hypothetical dietary interventions using the personal causal graph. Given the ranked causal paths influencing y the agent estimates the expected change in y under different dietary adjustments. The intervention effect of modifying nutrient intake v_j is computed as:

$$\Delta y = \sum_{v_j \in G_i^{(q)}} I(y, v_j) \cdot F(v_j) \tag{5}$$

where $I(y, v_j)$ represents the aggregated causal influence of v_j on y, and $F(v_j)$ models the individual response function:

$$F(v_j) = \beta_j \cdot \Delta v_j + \epsilon \tag{6}$$

where β_j is the personalized response coefficient estimated from historical glucose responses, Δv_j is the proposed change in (e.g., increasing fiber intake by 15g), ϵ accounts for errors in predictions with expectation $E[\epsilon] = 0$. Through this process, the LLM agent predicts the potential benefit of dietary adjustments before finalizing a recommendation. To ensure that the proposed intervention is causally valid, the LLM conducts a counterfactual reasoning step using the calculated effect to assess whether alternative dietary modifications would yield more effective outcomes. If the predicted impact of the initial recommendation aligns with the user's goal y, the agent reiterates through the reasoning process to select alternative recommendations with valid causal justifications.

3.5 Response Generation

To generate the final recommendation, we construct a structured prompt that integrates the retrieved causal graph information, food-nutrient associations, and supporting evidence. The prompt explicitly states the target health outcome y, presents the causal pathways derived from $G_i^{(q)}$ in natural language, includes ranked dietary factors based on their relevance, and appends the retrieved food-nutrient content data. The LLM is prompted to first explain the causal reasoning before presenting the recommendation to keep responses personalized, interpretable, and grounded in causal inference rather than relying on generic correlations.

4 Case Study on Dietary Recommendations

4.1 Dataset and Pre-Processing

We utilize a publicly available dataset comprising 49 participants aged 18 to 69 years with a BMI range of 21–46 kg/m², collected between 2021 and 2024 (Gutierrez-Osuna et al., 2025). The cohort includes 15 individuals without diabetes (HbA1c $< 5.7\%$), 16 with prediabetes ($5.7\% \leq$ HbA1c $\leq 6.4\%$), and 14 with type 2 diabetes (HbA1c $> 6.4\%$). The dataset spans approximately ten days per participant. We select data from 34 participants who have complete MET recordings. For each participant, we select the data of continuous glucose monitoring (CGM) and fitness tracker readings recorded at 1-minute intervals, along with detailed meal records, including total caloric intake and macronutrient composition (carbohydrates, protein, fat, fiber) for each meal (breakfast, lunch, and dinner), and daily average MET.

In this case study, we define the objective as the **incremental area under the curve (iAUC) of postprandial glucose levels**. The iAUC quantifies the body's glycemic response to dietary intake and captures both the magnitude and duration of postprandial glucose excursions (Zeevi et al., 2015; Reynolds et al., 2020; Floch et al., 1990).

We compute the iAUC over three distinct intervals: 30 minutes, 1 hour, and 2 hours following food intake. The 30-minute interval captures the initial glucose rise, which reflects absorption kinetics and early insulin dynamics. The 2-hour interval represents the full postprandial phase and characterizes prolonged glycemic effects and glucose regulation efficiency. The 1-hour interval serves as an intermediate measure to distinguish between transient fluctuations and sustained metabolic responses. For baseline glucose estimation, conventional approaches define the baseline as the fasting glucose level measured immediately before food intake. This definition may not fully account for individual variability in glycemic patterns. To obtain a more representative baseline, we use the average glucose level over the 24 hour period preceding the meal (Chkroun et al., 2023).

4.2 Personal Causal Graph Construction

To enable personalized causal graph-based reasoning, we construct a personal causal graph using dietary intake, glucose monitoring, and MET data for each user. The construction process consists of two key steps: inferring the causal structure using a causal discovery method and estimating causal effects.

4.2.1 Inferring Causal Structure

The first step in constructing the causal graph is to determine the structure of causal relationships between dietary factors, metabolic biomarkers, and external modulators. We apply the Peter-Clark (PC) algorithm (Spirtes et al., 2001) to infer a directed acyclic graph that represents the direct causal dependencies between these variables. We use the first half of each user's data for the causal graph construction.

The PC algorithm first detects conditional independence relationships to eliminate non-causal edges, ensuring that only direct dependencies are retained. It then orients the edges by leveraging causal constraints, ensuring that dietary intake variables precede metabolic changes in a physiologically meaningful manner. Finally, it adjusts for confounders such as physical activity and baseline glucose levels to prevent spurious associations. The output of this step is a causal graph $G_i = (V_i, E_i)$, where nodes V_i represent dietary intake variables, metabolic biomarkers, and external modulators, while edges E_i encode directed causal relationships between these variables. We employ the Causal Discovery Tool library to conduct PC algorithm (Kalainathan et al., 2020). At this stage, the edges indicate causal influence but do not yet quantify the strength of these effects so there are no weights for edges.

4.2.2 Estimating Causal Effects

Once the causal structure is identified, we estimate the causal effect strengths of the edges in the graph, quantifying how changes in dietary intake influence metabolic outcomes. We employ Structural Causal Models (SCMs), where each variable is expressed as a function of its direct causes and an independent noise term (Elwert, 2013). To assign causal effect strengths, we assume a linear SCM that models the impact of each dietary factor on a metabolic outcome as a weighted relationship. We then apply regression-based inference to estimate the magnitude of these effects with the first half of each user's data, and use the resulting values as the edge weights E_i in the personalized causal graph G_i.

4.3 Counterfactual Evaluation

In order to qualitatively determine whether the recommended food intake truly contributes to achieving the user's goal, we propose to simulate the counterfactual outcome using a ground truth causal graph and validate the recommendation against it. To obtain a reliable reference graph, we construct a causal graph using the full personal dataset to serve as the ground truth for validation. Since this graph is inferred from more data, it provides a more robust representation of the individual's nutrient-glucose interactions to approximate whether the model's recommendations hold under real-world conditions.

For each recommendation, we conduct a counterfactual simulation based on the ground truth graph to estimate its expected impact on the user's target y, which is the glucose $iAUC$ in this case study. Given a user query, the LLM selects a food item through the estimated graph in section 4.2.1 and 4.2.2. To eliminate scale bias, the recommended food portion is standardized to 500 kcal. We then introduce this food into the ground truth causal graph and use the estimated causal effects to compute the predicted change in glucose $iAUC$. This follows the inference:

$$i\hat{AUC} = \mathbb{E}[y \mid do(v_j)] \tag{7}$$

where y refers to the user's goal, v_j denotes the nutrient content, and X_c refers to the relevel confounding variables such as MET. How the food consumption would have affected the user's goal is estimated by conditioning on X_c. Given the counterfactual estimate $i\hat{AUC}$, we compare it with the expected glucose response under the user's historical dietary pattern, denoted by $iA\bar{U}C_i$, which corresponds to the iAUC when the user consumes a meal with an average nutrient composition based on their past intake. We report the **Mean Glucose Reduction (MGR)** of the food recommendation, which is computed as :

$$MGR = \frac{\Sigma_{i=1}^{N} iA\bar{U}C_i - iA\hat{U}C_i}{N} \tag{8}$$

where N is the number of food recommendations evaluated. Note that it is not an absolute value. A positive MGR indicates that, on average, the LLM-recommended foods lead to lower glucose responses compared to the user's typical dietary choices.

4.4 Experiment Settings

To generate personalized food recommendations, we employ GPT-4o as the LLM agent. The LLM is instructed to follow the process outlined in Figure 2. The prompt combines these components: instruction, the user's query, the retrieved causal paths in a structured format, and the retrieved food-nutrient data. As we retrieve the most influential causal paths, these paths are then summarized into a natural language description. For instance, if the graph indicates that carbohydrate intake strongly increases postprandial glucose levels, while fiber consumption reduces glucose spikes, the extracted causal summary would be formatted as: *"Carbohydrates have a strong positive causal effect on glucose levels (ranked 1). Fiber has a moderate negative effect, reducing glucose spikes (ranked 2)."* Additionally, LLM is instructed to first analyze the causal relationships, use the retrieved nutrient information to generate a food recommendation, and then verify the food recommendation.

For testing, we query the agent five times per participant, requesting food recommendations for glucose management across three time windows (30 minutes, 1 hour, and 2 hours), using the baseline glucose levels from the past 2 hours. To ensure diversity in recommendations, we impose a constraint preventing the agent from suggesting any food items that were previously recommended in earlier queries for the same participant.

5 Results

We compare the proposed method against several baseline models. The baselines include RAG approaches, such as ChatDiet (Yang et al., 2024b) and vanilla RAG models augmented with general dietary guidelines, leveraging either Chain-of-Thought (CoT) prompting or Tree-of-Thought (ToT) reasoning. We also include non-RAG baselines, where a vanilla LLM is tested with and without CoT or ToT prompting. The performance of each method is assessed based on MGR and its standard deviation, as reported in Table 1.

As shown in the results, our approach outperforms the baselines over longer time horizons (1 hour and 2 hours), achieving significantly higher MGR ($p < 0.05$) with a lower standard deviation. Among the baselines, ChatDiet, a retrieval-augmented model, performs competitively in the short-term window but remains less effective in longer time frames compared to our method. The effect of dietary intake over short durations is inherently variable, making it difficult to determine a significant performance difference. However, over extended time windows, where the physiological impact of food consumption can be estimated with greater confidence, the superior performance of our approach more reliably demonstrates the added value of personalized causal reasoning over static retrieval-based systems.

Models that rely solely on general dietary guidelines or prompting techniques such as CoT and ToT exhibit highly unstable performance, with some configurations even leading to an increase in glucose levels. This instability arises because these models lack access to personalized context, making it impossible to capture an individual's unique metabolic patterns. These findings reinforce the necessity of explicit causal modeling for effective personalized nutrition recommendations. Overall, our results highlight the crucial role of personalized causal graph reasoning, particularly in dietary interventions. Our framework enables the model to generate more effective, stable, and context-aware dietary recommendations tailored to individual metabolic responses.

	30 mins MGR	1hr MGR	2hr MGR
Proposed	19.84 (31.00)	**158.21 (61.73)**	**411.56 (77.21)**
ChatDiet(Yang et al., 2024b)	**33.92 (36.01)**	120.45 (88.64)	307.12 (123.84)
LLM + General Diet Guidelines + CoT	16.38 (57.28)	-45.72 (252.71)	-79.61 (217.99)
LLM + General Diet Guidelines + ToT	-18.70 (78.42)	62.19 (229.45)	13.88 (179.41)
LLM + CoT	-10.59 (65.12)	-49.23 (208.57)	-64.11 (254.30)
LLM + ToT	8.77 (81.64)	-6.43 (173.90)	63.40 (251.85)
Sole LLM	21.40 (51.93)	44.83 (226.57)	-149.89 (308.46)

Table 1: MGR and standard deviation for baseline models and the proposed Personalized Causal Graph Reasoning framework

5.1 Ablation Study

We conduct an ablation study by progressively removing key components and evaluating their impact. Specifically, we examine the effect of removing the verification step, disabling the path ranking mechanism, and completely excluding the personal causal graph, thereby testing the model's performance when relying solely on the LLM. The results are summarized in Table 2.

The full model achieves the highest glucose reduction, particularly in the more stable 1-hour and 2-hour time windows. Removing the verification step results in only a slight decline in performance, indicating that while it is not the primary driver of improvement, it helps refine recommendations in certain corner cases. In contrast, disabling path ranking leads to a substantial increase in variance, as it plays a core role in prioritizing the most influential nutrients, which is essential for stabilizing glucose impact predictions. Removing the personal causal graph entirely prevents the agent from performing personalized reasoning, rendering the model ineffective at generating meaningful dietary recommendations.

5.2 Evaluating Reasoning Personalization with LLM-as-a-Judge

To assess the personalization level of our Personalized Causal Graph Reasoning framework, we employ LLaMA-3 70B (Dubey et al., 2024) as an LLM-as-a-judge (Zheng et al., 2023) to compare its reasoning process against the previous method. The evaluation follows a blind comparison setup, where the judge is presented with two outputs in a random order without knowing their source. The judge is instructed to select the response that demonstrates a higher degree of personalization of the reasoning process, considering factors such as whether the reasoning incorporates the user's unique metabolic patterns, past dietary responses, and personalized causal dependencies; whether the response adapts to the specific health context of the user rather than relying on generic dietary principles; and whether the explanation leverages structured causal insights instead of relying on general nutritional heuristics.

Each comparison is conducted across multiple test cases, and the LLM-as-a-judge selects the more personalized reasoning in each instance. The final win rate reflects the percentage of cases where our model was preferred over ChatDiet. The results, presented in Table 3, show that our Personalized Causal Graph Reasoning framework achieves a dominant win rate of 98.43%.

6 Limitations

Our current framework constrains LLM reasoning to a single, well-defined objective. While this ensures a focused decision-making process, real-world dietary planning often involves multiple, uncertain health goals, such as cardiovascular health, weight management, and micronutrient balance. The model does not yet support multi-objective reasoning, limiting its applicability to users with diverse and evolving dietary needs.

The method also lacks an early stopping mechanism in personal causal graph traversal. As the graph grows in complexity, the LLM agent does not have a mechanism to determine when sufficient causal evidence has been gathered, potentially leading to redundant or inefficient reasoning. This is sufficient for the specific case study, but a more adaptive traversal strategy is needed to dynamically assess when to terminate search paths based on confidence in the retrieved causal relationships.

Regarding the case study on LLM dietary rec-

	30 mins MGR	1hr MGR	2hr MGR
Full	19.84 (31.00)	**158.21 (61.73)**	**411.56 (77.21)**
Remove Verification step	19.16 (32.46)	1163.98 (67.51)	402.74 (86.54)
Remove Path Ranking	**23.88 (38.52)**	952.34 (77.96)	367.02 (92.03)
Remove Personal Graph (Sole LLM)	21.40 (51.93)	44.83 (226.57)	-149.89 (308.46)

Table 2: Ablation study results on removing key components.

	Win Rate
Proposed	98.43%
ChatDiet(Yang et al., 2024b)	1.57%

Table 3: LLM-as-a-Judge Results on Reasoning Personalization

ommendations, the limited amount of nutrient intake and glucose response data presents another challenge. Inferring causal relationships requires a sufficient number of observations and interventions, but the available dataset is relatively small, leading to the uncertainties in causal estimation as we can see in the high standard deviation of the results. The reliance on short-term observational data may not fully capture the complex, long-term metabolic effects of dietary interventions. Incorporating larger datasets, or self-reported dietary logs could improve the reliability of causal inference.

The causal graph construction does not explicitly model all potential confounders. While glucose regulation is influenced by macronutrient intake and physical activity, other physiological factors such as gut microbiome composition, hormonal fluctuations, and sleep patterns play critical roles. The current framework does not account for these influences, which may affect the accuracy of its dietary recommendations. Expanding the causal model to incorporate a broader range of physiological variables would provide a more complete understanding of individual dietary responses.

Finally, the evaluation relies on counterfactual simulation rather than in vivo validation. While causal inference techniques estimate the potential impact of dietary changes, real-world outcomes are influenced by adherence variability, behavioral responses, and external lifestyle factors. Without real-world validation, there is a risk that LLM-generated recommendations may not translate into actual health improvements or could lead to unintended dietary imbalances if misinterpreted or applied inconsistently. Conducting controlled trials to measure the actual impact of LLM-recommended dietary interventions would be necessary to validate the model's real-world effectiveness and ensure its safety and reliability.

7 Conclusion

We presented Personalized Causal Graph Reasoning to address the need for personalized LLM reasoning in real-world scenarios. A case study was conducted by integrating the proposed framework into personalized dietary recommendations. A counterfactual evaluation method was employed to assess performance without requiring human experts. The results showed that the proposed approach improved glucose management compared to retrieval-augmented and prompt-based baselines. LLM-as-a-judge results indicated that the proposed method provided more personalized reasoning than existing approaches.

Overall, we have demonstrated the importance of personalized LLM reasoning and the effectiveness of personalized causal graph reasoning in a domain where complex personal data plays a critical role-dietary recommendation. A deeper analysis is needed for developing more refined personalized reasoning mechanisms to handle multi-objective decision-making and large-scale personal graph reasoning. The dietary recommendation study could be extended to incorporate additional confounders and include real-world trials to evaluate its practical effectiveness, which we leave for future works

References

Shahabeddin Abhari, Reza Safdari, L Azadbakht, K Bagheri Lankarani, Sh R Niakan Kalhori, B Honarvar, Kh Abhari, SM Ayyoubzadeh, Z Karbasi, Somayyeh Zakerabasali, et al. 2019. A systematic review of nutrition recommendation systems: with focus on technical aspects. *Journal of biomedical physics & engineering*, 9(6):591.

C Anjanamma, G Sirisha, B Sravani, K Shilpa, CV Lakshmi Narayana, and V Vivekanandhan. 2024. Personalized food nutrient recommendations for kids using ai and behavior analysis. In *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, pages 1805–1810. IEEE.

Iman Azimi, Mohan Qi, Li Wang, Amir M Rahmani, and Youlin Li. 2025. Evaluation of llms accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval. *Scientific Reports*, 15(1):1506.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Célina Chkroun, Inez Trouwborst, Anna Cherta-Murillo, Lauren Owen, Christian Darimont, and Andreas Rytz. 2023. Defining a continuous glucose baseline to assess the impact of nutritional interventions. *Frontiers in Nutrition*, 10:1203899.

Iris M de Hoogh, Machiel J Reinders, Esmée L Doets, Femke PM Hoevenaars, and Jan L Top. 2023. Design issues in personalized nutrition advice systems. *Journal of Medical Internet Research*, 25:e37667.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Felix Elwert. 2013. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer.

Jean-Pierre Le Floch, Philippe Escuyer, Eric Baudin, Dominique Baudon, and Léon Perlemuter. 1990. Blood glucose area under the curve: methodological aspects. *Diabetes care*, 13(2):172–175.

Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. 2024. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*.

Ricardo Gutierrez-Osuna, David Kerr, Bobak Mortazavi, and Anurag Das. 2025. Cgmacros: a scientific dataset for personalized nutrition and diet monitoring. *Scientific Data (under review)*.

Alon Halevy and Jane Dwivedi-Yu. 2023. Learnings from data integration for augmented language models. *arXiv preprint arXiv:2304.04576*.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. Glore: When, where, and how to improve llm reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*.

Peng Hu, Changjiang Gao, Ruiqi Gao, Jiajun Chen, and Shujian Huang. 2024. Large language models are limited in out-of-context knowledge reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3144–3155.

Diviyan Kalainathan, Olivier Goudet, and Ritik Dutta. 2020. Causal discovery toolbox: Uncovering causal relationships in python. *Journal of Machine Learning Research*, 21(37):1–5.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. Making large language models better reasoners with step-aware verifier. *arXiv preprint arXiv:2206.02336*.

Huishi Luo, Fuzhen Zhuang, Ruobing Xie, Hengshu Zhu, Deqing Wang, Zhulin An, and Yongjun Xu. 2024. A survey on causal inference for recommendation. *The Innovation*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

James Mahoney and Rodrigo Barrenechea. 2019. The logic of counterfactual analysis in case-study explanation. *The British journal of sociology*, 70(1):306–338.

Deepak Nathani, David Wang, Liangming Pan, and William Yang Wang. 2023. Maf: Multi-aspect feedback for improving reasoning in large language models. *arXiv preprint arXiv:2310.12426*.

CAJ Nijman, IM Zijp, A Sierksma, AJC Roodenburg, R Leenen, C Van den Kerkhoff, JA Weststrate, and GW Meijer. 2007. A method to improve the nutritional quality of foods and beverages based on dietary recommendations. *European Journal of Clinical Nutrition*, 61(4):461–471.

Andrew N Reynolds, Jim Mann, Mona Elbalshy, Evelyn Mete, Caleb Robinson, Indrawati Oey, Pat Silcock, Nerida Downes, Tracy Perry, and Lisa Te Morenga. 2020. Wholegrain particle size influences postprandial glycemia in type 2 diabetes: a randomized crossover study comparing four wholegrain breads. *Diabetes care*, 43(2):476–479.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, prediction, and search*. MIT press.

Ajan Subramanian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. 2024. Graph-augmented llms for personalized health insights: A case study in sleep analysis. In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pages 1–4. IEEE.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*.

Meghna Verma, Raquel Hontecillas, Nuria Tubau-Juni, Vida Abedi, and Josep Bassaganya-Riera. 2018. Challenges in personalized nutrition and health.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Saul J Weiner. 2004. Contextualizing medical decisions to individualize care: lessons from the qualitative sciences. *Journal of General Internal Medicine*, 19(3):281–285.

Cheng Xue, Yijie Guo, Ziyi Wang, Mona Shimizu, Jihong Jeung, and Haipeng Mi. 2024. Dishagent: Enhancing dining experiences through llm-based smart dishes. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–4.

Zhongqi Yang, Iman Azimi, Mohammed J Zaki, Manas Gaur, Oshani Seneviratne, Deborah L McGuinness, Sabbir M Rashid, and Amir M Rahmani. 2024a. Transforming personal health ai: Integrating knowledge and causal graphs with large language models. *Under Review*.

Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M Rahmani. 2024b. Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an llm-augmented framework. *Smart Health*, 32:100465.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Haoran Yu, Chang Yu, Zihan Wang, Dongxian Zou, and Hao Qin. 2024. Enhancing healthcare through large language models: A study on medical question answering. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 895–900. IEEE.

David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, et al. 2015. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.