GLEN-Bench: A Graph-Language based Benchmark for Nutritional Health

Jiatan Huang[1], Zheyuan Zhang[2], Tianyi Ma[2], Mingchen Li[3], Yaning Zheng[1],
Yanfang Ye[2], Chuxu Zhang[1,†]
[1]University of Connecticut, USA [2]University of Notre Dame, USA
[3]University of Massachusetts Amherst, USA
{jiatan.huang, chuxu.zhang}@uconn.edu, {zzhang42, tma2, yye7}@nd.edu
[†]Corresponding author

Abstract

Nutritional interventions are important for managing chronic health conditions, but current computational methods provide limited support for personalized dietary guidance. We identify three key gaps: (1) dietary pattern studies often ignore real-world constraints such as socioeconomic status, comorbidities, and limited food access; (2) recommendation systems rarely explain why a particular food helps a given patient; and (3) no unified benchmark evaluates methods across the connected tasks needed for nutritional interventions. We introduce GLEN-Bench, the **first comprehensive** **g**raph-**l**anguage based b**e**nchmark for **n**utritional health assessment. We combine NHANES health records, FNDDS food composition data, and USDA food-access metrics to build a knowledge graph that links demographics, health conditions, dietary behaviors, poverty-related constraints, and nutrient needs. We test the benchmark using opioid use disorder, where models must detect subtle nutritional differences across disease stages. GLEN-Bench includes three linked tasks: risk detection identifies at-risk individuals from dietary and socioeconomic patterns; recommendation suggests personalized foods that meet clinical needs within resource constraints; and question answering provides graph-grounded, natural-language explanations to facilitate comprehension. We evaluate these graph-language approaches, including graph neural networks, large language models, and hybrid architectures, to establish solid baselines and identify practical design choices. Our analysis identifies clear dietary patterns linked to health risks, providing insights that can guide practical interventions. We release the benchmark with standardized protocols and evaluation tools at https://github.com/J-Huang01/GLEN-Benchmark.

Keywords

Benchmark; Nutritional Health; Personalization; Graph Neural Networks; Large Language Models; Retrieval-Augmented Generation

1 Introduction

Assessing nutritional health matters for both precision medicine and population health because diet is a modifiable factor that affects disease risk, symptom burden, and recovery [1, 3]. For common conditions such as diabetes, cardiovascular disease, obesity, and kidney disease, diet works together with comorbidities, medications, and social determinants of health (SDoH) to influence patient outcomes [42, 54]. While nutrition-aware assessment is universally acknowledged, existing computational efforts remain fragmented and rarely integrated into end-to-end decision workflows. Population-scale dietary and clinical resources, complemented by SDoH signals (e.g., access and affordability), offer the ingredients for nutrition-aware modeling [50]. Using these signals, recent benchmarks and datasets have focused on specific tasks, including food recommendation [79], diet logging and meal planning [38], risk prediction based on structured records [78], and nutrition-related question answering [10, 76], which typically support comprehensive assessment in a single environment.

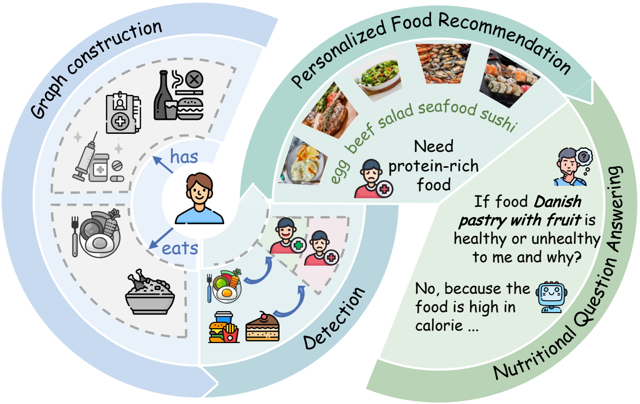


Figure 1: An Overview of GLEN-Bench. Our framework brings scattered nutrition and health signals together into one pipeline for nutritional health assessment.

However, turning these advances into practical, nutrition-aware interventions requires an integrated view that current resources still lack. In practical healthcare settings, effectively implementing dietary interventions involves: (i) identifying individuals who are likely to benefit from dietary interventions; (ii) recommending foods that are clinically appropriate and feasible within individual constraints (e.g., affordability and accessibility); and (iii) offering lucid explanations to build trust and improve adherence. Yet these resources remain fragmented across modalities and task scopes, limiting end-to-end evaluation under realistic clinical and socioeconomic constraints. Despite progress on individual tasks, current benchmarks do not support end-to-end evaluation of nutrition-aware interventions under realistic clinical and socioeconomic constraints. First, nutritional decision-making inherently involves multidimensional reasoning over dietary behaviors, health conditions, socioeconomic factors, and food access barriers; yet prior work often models these dimensions in isolation (e.g., nutrient targets without feasibility constraints). Second, tasks such as food recommendation and health risk detection are usually studied in isolation, even though they depend on one another in real-world settings. Third, the field lacks a standard evaluation 'yardstick'. Fragmented data sources and inconsistent metrics prevent direct comparisons across methods, making it difficult to identify which approaches work in real-world interventions.

To bridge these gaps, we introduce GLEN-Bench, the first comprehensive **g**raph-**l**anguage based b**e**nchmark for **n**utritional health assessment, supporting both structured graph reasoning and natural-language decision support. We build a disease-agnostic heterogeneous graph by integrating NHANES [7], FNDDS [43], and USDA access signals [58]. This structure captures essential dependencies between users, foods, and socioeconomic barriers. The schema is highly scalable; it extends to conditions like diabetes or obesity by simply adding condition-specific requirements. Furthermore, GLEN-Bench centers on an interdependent task suite that reflects end-to-end clinical workflows. Finally, we benchmark over 20 computational approaches, spanning classical ML, graph neural networks, large language models, and hybrid architectures, to establish a systematic baseline under a unified protocol.

We set up three tasks in our benchmark testing: (1) Risk Detection identifies at-risk populations through dietary, social, and economic patterns. This not only supports population screening but also reveals how the model learns signals related to nutrition and health. (2) Personalized Recommendation suggests foods that meet clinical guidelines and individual needs while remaining affordable and accessible. (3) Interpretable Question Answering assesses food-user compatibility via nutrient-level attribution and natural language explanations grounded in the knowledge graph, supporting transparent, informed decisions. These interdependent tasks form real clinical workflows, i.e., detection identifies intervention candidates, recommendation formulates tailored suggestions, and explanation ensures uptake. Beyond these core tasks, our design is scalable and naturally supports meal planning, ingredient substitution, dietary adherence prediction, and food-drug interaction analysis, making GLEN-Bench an evolving AI platform for nutritional intelligence. In summary, the main contributions are as follows:

- **First Comprehensive Multi-Task Benchmark for Nutritional Health**. We introduce the first comprehensive multi-task benchmark for nutritional health, which jointly evaluates risk detection, personalized recommendations, and explainable interpretations. GLEN-Bench is designed as a long-term scalable platform applicable to various nutrition-related diseases and supporting a wide range of downstream tasks.
- **A Novel Multi-Dimensional Nutrition–Health Graph**. We built a large-scale nutrition-health graph to capture the complex inner relationships between diet, health, and social factors. This graph includes not only food and disease data, but also socioeconomic factors such as food insecurity, poverty, and access to healthcare. While previous datasets mostly focused on the association between nutrients and diseases, our approach provides a more comprehensive perspective.
- **Systematic Evaluation Revealing Design Principles**. We conducted the first systematic evaluation of over 20 methods, encompassing classic machine learning (ML), graph neural networks (GNNs), large language models (LLMs), and hybrid architectures. Our findings highlight the advantages of graph-based reasoning methods, where approaches combining graph neural networks with large language models achieved the best overall performance and more consistent interpretations under standardized protocols. We also further revealed interpretable dietary and socioeconomic characteristics associated with health risks, providing practical insights for developing effective interventions.

2 Related Work

Nutritional Health Research. Current computational methods for health risk assessment typically use electronic health records (EHRs), wearable devices, or survey data to predict chronic diseases and adverse health outcomes [28, 34, 69]. Some studies also analyze online behavior to identify at-risk populations and lifestyle patterns [39, 45]. However, most of these methods focus on prediction accuracy rather than providing clear explanations that can guide practical interventions. Research in public health has consistently shown connections between dietary patterns and disease risk, recovery outcomes, and socioeconomic inequalities [3, 42, 54, 62]. These findings have driven the development of computational tools that integrate nutrition and behavioral data [1, 48, 78]. Food recommendation has progressed from preference-based ranking toward health-aware personalization [36, 52, 57], evolving from single-metric optimization (e.g., calories or fat) [16, 52] to integrating nutritional standards and health signals [5, 53, 63, 78, 79]. Nevertheless, many systems still overlook structural feasibility constraints (e.g., affordability and limited access) that shape real dietary choices. Knowledge graphs have been used to model food–health relationships for dietary adaptation and ingredient substitution [9, 13, 14, 18], and recent KGQA and graph-RAG advances combine LLMs with GNNs to support structured reasoning [15, 25–27, 29, 67, 77]. Despite these advances, personalized services are often limited by a lack of user-specific medical information and socioeconomic factors [5], and evaluation results remain inconsistent due to the lack of uniform, intervention-aligned benchmarks that reflect the end-to-end nutritional health workflow.

Nutritional Health Benchmarks and Datasets. Existing nutrition and health resources cover a variety of different modalities and task scenarios, including food-centric corpora, nutritional knowledge bases/ontologies, and health-oriented question-answering datasets. Recipe1M+ [41] provides large-scale multimodal recipes for food understanding and recommendation, while Nutrition5k [56] supports vision-based nutritional analysis and provides fine-grained nutritional component annotation. Structured food knowledge is represented by nutrition knowledge graphs (KGs) such as FoodKG [18] and standardized ontologies such as FoodOn [12]. In terms of nutrition-related question answering, RecipeQA [72] focuses on multimodal procedural reasoning, while other resources integrate nutritional ontologies and health indicators [33, 51] to enable structured food-health queries. Meanwhile, general graph reasoning and retrieval-augmented generation benchmarks advance multi-hop querying and structured reasoning [4, 32, 35], but remain largely generic or disconnected from clinical intervention contexts [20, 37]. Recent domain-specific datasets begin connecting diet and health contexts [76, 78, 79], yet they typically evaluate isolated tasks. Unlike prior benchmarks, we evaluate an end-to-end

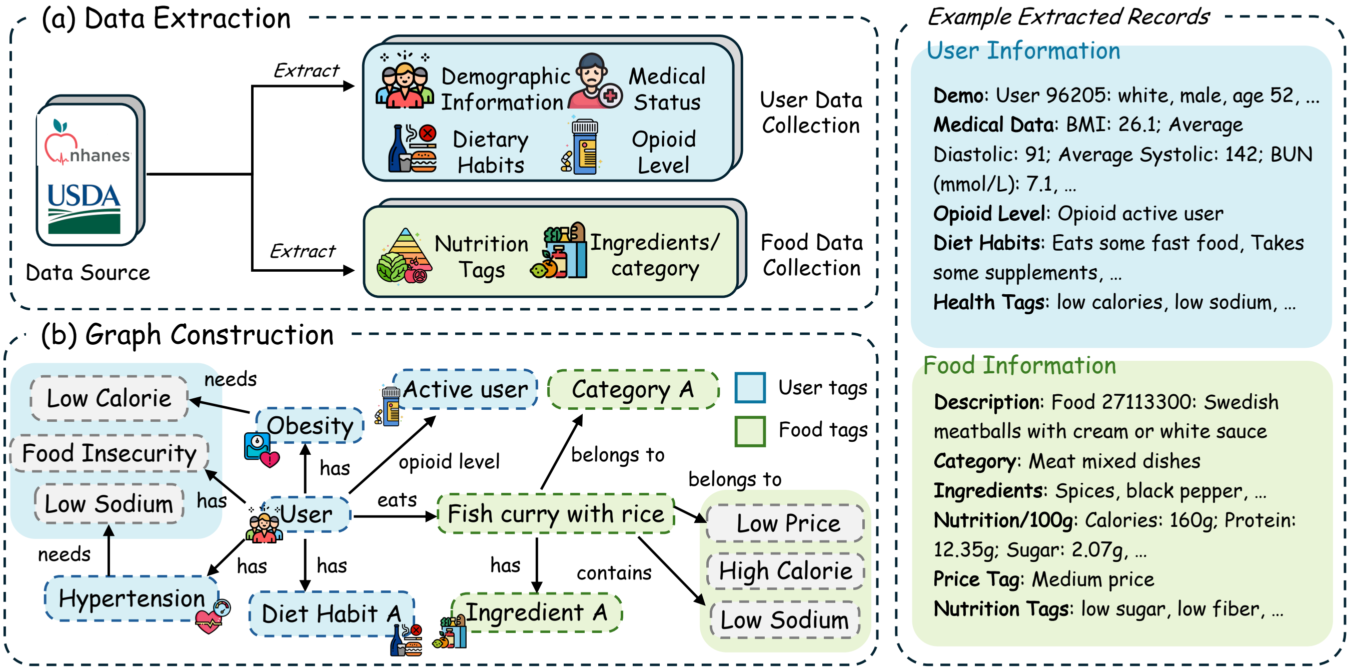


Figure 2: The GLEN-bench construction process. (a) Data extraction from NHANES and USDA sources to collect multi-dimensional user and food information. (b) Graph construction of a heterogeneous knowledge graph modeling the relationships between clinical health status, dietary habits, socioeconomic barriers (e.g., food insecurity), and food nutritional profiles.

workflow with explicit feasibility constraints (e.g., affordability and access). For completeness, we include additional related work in the appendix E.

3 GLEN-Bench

3.1 Data Sources

GLEN-Bench integrates three population-level datasets to capture nutrition, health, and socioeconomic factors. From NHANES, we extract demographics, clinical data, medical conditions, medications, dietary behaviors, and 24-hour food intake records, along with socioeconomic indicators like income and food security status. Using FNDDS/WWEIA, we map foods to nutrient profiles and standardized categories, enabling nutrient-based food representations and threshold-defined nutrition tags. Finally, we augment foods with affordability tiers using USDA Purchase-to-Plate price estimates, allowing models to reason about economic feasibility alongside clinical suitability (detailed source descriptions are provided in Appendix A.1).

3.2 Graph Construction

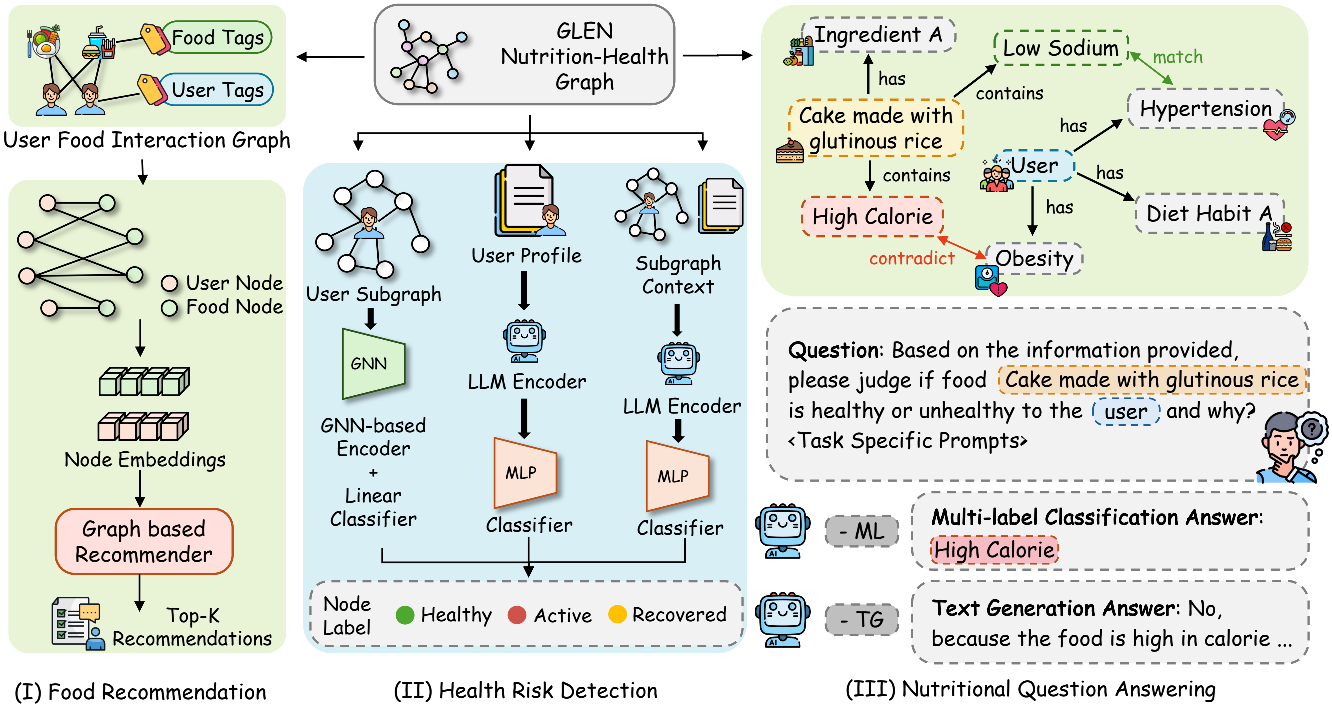
We utilized NHANES data from 2003 to 2020 to construct the GLEN Nutritional-Health Graph, a heterogeneous graph that captures the interactions between users, foods, and their clinical, behavioral, and socioeconomic backgrounds.

Definition 3.1 GLEN Nutrition-Health Graph. We model the graph as a directed heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathcal{T}$ and $\mathcal{R}$ denote node and relation types. Each node $v \in \mathcal{V}$ and edge $e \in \mathcal{E}$ is assigned a type through mappings $\tau : \mathcal{V} \to \mathcal{T}$ and $\varphi : \mathcal{E} \to \mathcal{R}$. Node types include *user*, *food*, *ingredient*, *category*, *dietary_habit*, *health_condition*, *nutrition_tag*, *financial_tag*, and *price_tag*, while edges are used to describe consumption, ingredient composition, health and dietary habit context, and nutrition/price constraints.

User Nodes. Let $\mathcal{U} = \{v \in \mathcal{V} \mid \tau(v) = \text{user}\}$ and $\mathcal{F} = \{v \in \mathcal{V} \mid \tau(v) = \text{food}\}$. Each user node $u \in \mathcal{U}$ is associated with a feature vector $\mathbf{x}_u \in \mathbb{R}^{d_u}$ that concatenates demographic attributes (e.g., age, sex, race) with clinical and laboratory measurements (e.g., BMI, blood pressure, metabolic biomarkers) from NHANES. Diagnosed conditions, self-reported dietary habits, and financial hardship are modeled as explicit neighbors rather than entries in $\mathbf{x}_u$: users connect to *health_condition*, *dietary_habit*, and *financial_tag* nodes (e.g., poverty status and food insecurity). Together, these attributes and links collectively describe the medical and social environment that influences dietary behavior and vulnerability.

Food Nodes. Each food node $f \in \mathcal{F}$ has a feature vector $\mathbf{x}_f \in \mathbb{R}^{d_f}$ derived from FNDDS and USDA Purchase-to-Plate. From FNDDS and WWEIA we obtain nutrient profiles summarizing macro- and micronutrients, while *ingredient* and *category* nodes capture composition and hierarchical food groups; foods connect to these nodes via *contain* and *belong_to* edges. From Purchase-to-Plate we obtain price per 100 grams and discretize it into *price_tag* nodes (e.g., *low_price*/*medium_price*/*high_price*). Thus, each food is grounded by nutrient content and affordability, with additional context provided by ingredient and category neighbors.

Health Conditions and Nutrition Tags. Health-related dietary requirements are encoded by *health_condition* and *nutrition_tag* nodes. Health conditions are derived from NHANES clinical and questionnaire records and linked to users through edges. Nutrition tags represent guideline-based threshold constraints (e.g.,



(I) Food Recommendation (II) Health Risk Detection (III) Nutritional Question Answering

Figure 3: Overview of the GLEN-Bench framework. The system builds on a unified nutrition-health Graph to perform three related tasks: (I) personalized food recommendation, (II) health risk detection to identify intervention needs, and (III) nutritional question answering that provides explainable, graph-based explanations.

low_sodium, *high_protein*); foods satisfying the thresholds connect via *food_has_tag* edges. We further associate conditions with relevant tags (e.g., hypertension with low sodium) to support downstream suitability assessment. Financial hardship is represented by *financial_tag* nodes (poverty and food insecurity) and considered jointly with *price_tag* nodes when modeling economic access.

Edges and Relations. Edges in $\mathcal{E}$ capture observed behavior and constraint structure: users link to consumed foods, conditions, habits, and financial tags, while foods link to ingredients, categories, nutrition tags, and price tags. Together, these relations encode what people eat, how foods are composed and priced, and how clinical status and socioeconomic constraints shape nutritionally appropriate and affordable options. The resulting graph integrates dietary behavior, health status, financial hardship, and food composition in a single heterogeneous structure. Let $\mathbf{X}^{\mathcal{U}} \in \mathbb{R}^{|\mathcal{U}| \times d_u}$ and $\mathbf{X}^{\mathcal{F}} \in \mathbb{R}^{|\mathcal{F}| \times d_f}$ denote the feature matrices for user and food nodes, and let $\mathbf{A}$ denote the adjacency structure induced by $\mathcal{E}$. Together, $(\mathcal{G}, \mathbf{X}^{\mathcal{U}}, \mathbf{X}^{\mathcal{F}})$ provide the backbone for GLEN-Bench detection, recommendation, and question answering tasks across diverse health conditions. Details on feature processing and dataset statistics are provided in Appendix A.

4 Nutritional Health Applications

This section instantiates GLEN-Bench's three core evaluation tasks and reports a standardized benchmarking study of diverse model families. For each task, we describe the task definition, evaluation metrics, baseline models and setup, and present the corresponding results and analyses.

4.1 Opioid Misuse Detection

We apply GLEN-Bench to opioid use disorder (OUD), a challenging scenario where models must identify subtle nutritional and behavioral differences between active use, recovery, and healthy individuals. The opioid crisis has claimed over 500,000 U.S. lives since 1999 [22, 23], yet nutrition-informed support remains understudied despite evidence of poor diet quality and socioeconomic hardship among individuals with OUD [8, 11, 21, 44, 60, 68]. Accordingly, we include OUD risk detection to evaluate screening under realistic clinical and socioeconomic constraints.

Task Definition. The Health Risk Detection task aims to classify each user node in the GLEN Nutrition-Health Graph into health-status categories based on dietary and socioeconomic signatures. For our OUD instantiation, this is formulated as a three-class classification problem: *active opioid user*, *opioid-recovered user*, or *normal user*. Unlike binary settings, we formulate a three-class problem and leverage multi-relational graph context. We obtain labels from NHANES opioid prescription and self-report records (Appendix A). To classify individuals, models must integrate information about their diet, health conditions, and socioeconomic factors. Distinguishing active users from those in recovery is particularly hard, as both groups share similar risk factors and differ only in minor dietary and behavioral patterns. For all predictive models in this task, we remove User–Opioid_Level edges and never expose opioid_level nodes or features during training and inference, to avoid any label leakage.

Evaluation Metrics. In our dataset, the three classes are highly imbalanced, with at-risk users being relatively rare. Therefore, we

Table 1: Opioid Misuse Detection over GLEN Nutrition-Health Graph with two train-validation-test splits (60/20/20 and 70/15/15). LLM methods use LLaMA-3.1-8B, Qwen3-8B, and DeepSeek-R1-Distill-Qwen-7B.

Methods	**Train 60% – Valid 20% – Test 20%**				**Train 70% – Valid 15% – Test 15%**			
	F1-macro	AUC	GMean	Accuracy	F1-macro	AUC	GMean	Accuracy
MLP	11.42±1.09	56.08±0.36	49.17±0.08	17.38±0.32	12.75±1.03	56.95±0.22	52.58±0.11	18.85±0.32
GCN[30]	12.52±0.94	57.01±0.31	52.39±0.13	18.37±2.88	13.49±1.07	58.41±0.36	53.46±0.09	20.07±0.31
GraphSAGE[17]	16.96±1.05	55.63±0.15	51.51±0.12	28.57±0.28	18.75±0.54	59.62±0.14	56.39±0.23	32.41±1.02
GAT[59]	17.84±0.89	56.70±0.25	54.15±0.19	33.04±2.59	19.60±0.97	61.52±0.28	57.19±0.22	33.82±0.29
RGCN[49]	19.22±0.51	58.85±0.34	54.31±0.24	33.13±0.77	20.56±0.13	56.91±0.04	50.86±0.03	37.49±0.22
HGT[24]	21.63±0.74	57.42±0.35	50.80±0.39	39.79±0.23	20.80±0.89	56.48±0.32	52.41±0.27	37.70±0.27
HAN [65]	25.85±0.06	70.58±0.09	**65.08±0.12**	47.43±0.19	24.25±0.09	71.23±0.05	**65.67±0.10**	43.40±0.24
LLaMA-3.1	22.94±0.07	67.99±0.19	60.33±0.19	42.24±0.20	21.61±0.04	68.21±0.15	59.27±0.12	38.97±0.12
Qwen3	23.82±0.07	67.90±0.17	60.89±0.18	44.73±0.29	22.91±0.08	69.76±0.07	60.73±0.06	42.84±0.27
DeepSeek-R1	24.86±0.03	64.31±0.17	56.29±0.16	48.51±0.12	24.11±0.11	69.05±0.11	60.42±0.06	45.74±0.37
LLaMA-3.1 + Graph	27.99±0.14	70.28±0.33	61.61±0.23	54.73±0.31	29.28±0.10	69.74±0.29	60.29±0.38	58.63±0.56
Qwen3 + Graph	28.55±0.34	70.57±0.37	61.34±0.24	55.91±0.76	29.55±0.43	71.31±0.42	63.39±0.32	59.15±1.04
DeepSeek-R1 + Graph	**31.45±0.15**	**70.74±0.37**	60.43±0.20	**67.20±0.49**	**30.22±0.21**	**71.97±0.33**	63.29±0.28	**60.82±0.66**

use metrics that evaluate performance across all classes rather than favoring the majority class. We report four metrics: macro-averaged F1 (F1-macro), one-vs-rest AUC averaged across classes, geometric mean of per-class recalls (GMean), and overall accuracy. F1-macro gives equal weight to each class to assess balanced performance. GMean is particularly sensitive to how well the model identifies minority classes and reflects the balance between detecting at-risk users and avoiding bias toward the majority class. Accuracy and AUC offer additional perspectives on overall correctness and ranking quality.

Baseline Models and Setup. We evaluate representative meth-ods using two data split configurations that simulate different train-ing scenarios: **60/20/20** and **70/15/15** for train/validation/test sets. For each split, we randomly divide users and run experiments with multiple random seeds, then report the mean and standard devia-tion of results. We organize baseline methods into three categories (Table 1). (i) Non-graph baseline: a multi-layer perceptron (MLP) operating on tabular user features, which quantifies the benefit of explicitly modeling interactions. (ii) Graph baselines: homoge-neous GNNs (GCN [30], GraphSAGE [17], GAT [59]) trained on a simplified homogeneous graph that ignores node/edge types (or concatenates type features), enabling us to isolate the gain from con-nectivity alone; and heterogeneous GNNs (RGCN [49], HGT [24], HAN [65]) that preserve multi-relational structure via type-specific parameters. (iii) LLM baselines: large language models (LLaMA-3.1, Qwen3, DeepSeek-R1) applied to a serialized representation of each user's profile information (without graph neighborhood). We fur-ther evaluate *LLM + Graph* hybrids (e.g., DeepSeek-R1 + Graph) by injecting graph-derived neighborhood context into prompts as structured evidence. Table 1 reports F1-macro, AUC, GMean, and Accuracy for all baselines under both splits.

Results Table 1 shows that graph structure is essential for identifying at-risk users whose patterns of misuse are often latent or "hidden". Compared to non-graph MLP baseline, message passing models such as GCN, GraphSAGE, and GAT significantly improve macro-F1 and GMean under both splits. This shows that node fea-tures alone are not enough. The graph structure captures complex relationships and risk patterns that cannot be detected from indi-vidual data points. Heterogeneous modeling further amplifies these gains. HAN attains the best GMean among classical GNNs in the 70/15/15 split, indicating that separating relation types (e.g., user-to-health vs. user-to-habit) reduces feature interference and improves sensitivity toward minority classes. LLMs perform well on average, but they struggle with class balance in highly imbalanced datasets. While their AUC values are competitive, their macro F1 and GMean scores lag behind the best graph models. This suggests that LLMs are susceptible to semantic bias from majority classes, and seman-tic reasoning alone cannot capture the fine-grained relational cues needed for minority class detection. The hybrid approach addresses this gap. DeepSeek-R1-Distill-Qwen-7B model with graph achieves the best results across metrics in both splits. The improvement over text-only LLMs suggests that graph structure acts as a guide, help-ing the LLM focus on relevant neighborhood patterns and make more stable predictions for rare users. The results across the two dataset splits show a consistent trend. Adjusting the split ratio from 60/20/20 to 70/15/15 decreased the number of training samples, but the relative ranking of the models remained stable. Although the accuracy of DeepSeek-R1 + Graph adjusts from 67.20±0.49 to 60.82±0.66, its continued dominance confirms that the model cap-tures invariant structural properties rather than relying on specific overfitting. While pure LLMs improve primarily on accuracy and AUC with flatter macro F1 scores, the consistent gains of hybrid approaches across all metrics suggest that graph structure helps extract deeper value from additional training data.

These results suggest a clear strategy: use graphs to organize clinical information and let LLMs analyze the resulting patterns. Graphs capture relationships that are hard to describe in text, while LLMs provide flexible reasoning across different user profiles. This approach translates graph structure into better predictions while keeping strong performance for identifying at-risk individuals.

4.2 Personalized Food Recommendation

Personalized food recommendation is central to nutrition-aware intervention because it translates risk screening into actionable daily

Table 2: Personalized Food Recommendation over GLEN Nutrition-Health Graph. Results are reported as mean ± std%. The best performance is bolded and runner-ups are underlined.

Methods	GNN	Graph CF	Contrastive	Specialized	Recall@20	NDCG@20	H-Score@20	PA@20	AvgTags@20
GCN[30]	✓	–	–	–	9.65±0.27	7.68±0.16	26.45±1.17	14.41±0.54	6.64±0.39
GraphSAGE[17]	✓	–	–	–	9.75±0.80	6.74±0.32	30.27±1.17	15.14±0.33	6.46±0.19
GAT[59]	✓	–	–	–	10.17±0.30	8.58±0.17	32.24±0.30	16.39±0.38	6.90±0.32
NGCF[64]	–	✓	–	–	12.93±0.05	7.98±0.07	33.98±1.11	17.62±0.14	7.09±0.40
LightGCN[19]	–	✓	–	–	11.99±0.41	7.67±0.28	33.86±0.31	17.35±0.33	6.81±0.20
SimGCL[70]	–	✓	✓	–	12.79±0.39	8.06±0.24	34.11±0.34	16.85±0.26	6.74±0.15
SGL[74]	–	✓	✓	–	11.02±0.15	6.80±0.09	33.34±0.09	15.95±0.11	6.73±0.67
LightGCL[6]	–	✓	✓	–	11.48±0.24	7.60±0.18	33.07±0.21	16.17±0.20	6.71±0.14
RecipeRec[57]	–	–	–	✓	12.72±0.28	9.17±0.75	39.64±0.20	26.55±0.63	6.78±0.93
HFRS-DA[14]	–	–	–	✓	12.78±0.27	9.20±0.75	34.21±0.17	**27.78±0.75**	**7.30±0.72**
MOPI-HFRS[79]	–	–	–	✓	**13.25±0.34**	**9.97±0.61**	**38.18±0.74**	26.84±1.07	7.21±0.73

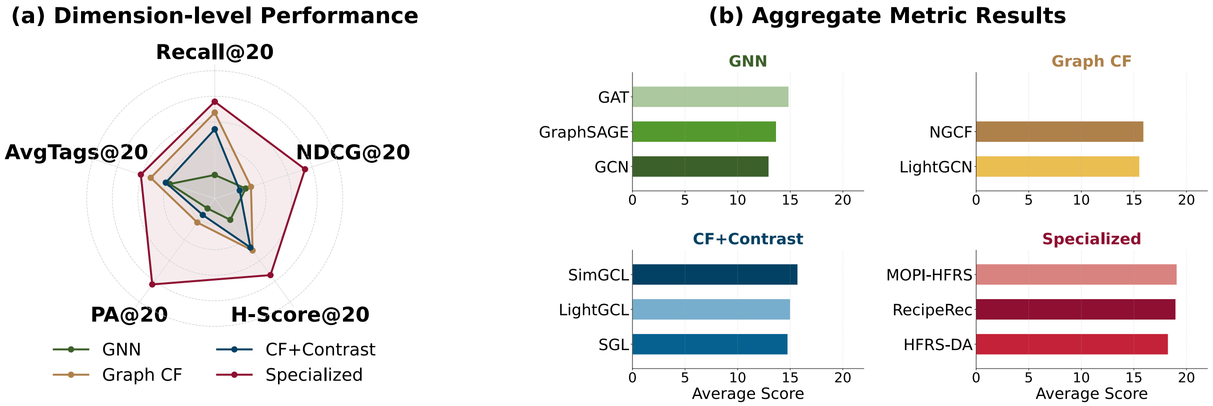


Figure 4: Personalized Food Recommendation performance across four model families (GNNs, collaborative filtering methods, contrastive-enhanced CF, and domain-specialized models) on GLEN-Bench.

choices. In practice, recommendations must be not only preference-aligned but also clinically appropriate under condition-specific constraints (e.g., sodium restriction for hypertension) and feasible un-der socioeconomic limitations such as affordability and food access. These interconnected needs constitute a multi-objective scenario that requires balancing relevance, health compliance, and feasibil-ity, rather than simply optimizing user interaction. Therefore, we treat personalized food recommendation as a core task to evaluate whether the model can leverage the GLEN Nutrition-Health Graph to generate clinically and economically valuable recommendations.

Task Definition. This task is formulated as a top-K food recommendation problem where, for each user in GLEN Nutrition-Health Graph, the model must rank candidate foods and return a list of K items that are nutritionally appropriate and economically feasible, while still matching the user's preferences. Unlike conventional collaborative filtering that relies primarily on user-item interac-tion patterns, our formulation leverages the rich heterogeneous structure of GLEN Nutrition-Health Graph. User nodes connect not only to consumed foods but also to health condition nodes encoding clinical diagnoses (e.g., hypertension, diabetes), nutri-tion tag nodes specifying dietary requirements (e.g., low_sodium, high_fiber), and socioeconomic indicator nodes capturing financial constraints (poverty status, food insecurity). Food nodes link to ingredient and category nodes describing compositional structure, nutrition tags reflecting nutrient profiles, and price tags denot-ing affordability tiers. A successful recommender must therefore generate foods that align with user preferences while respecting health-derived nutritonal constraints (e.g., restricting sodium for hypertensive users) and accommodating economic realities im-posed by poverty and limited food access.

Evaluation Metrics. We fix the recommendation list length to $K = 20$ and evaluate ranking quality using two standard metrics, Recall@20 and NDCG@20 (Normalized Discounted Cumulative Gain). To assess whether recommendations are consistent with user-specific dietary needs, we use the health-aware H-Score metric from prior work on personalized food recommendation: H-Score@20 measures, for each user, the proportion of recommended foods that share at least one required health or nutrition tag with the user, and then averages this value across users. To capture economic accessibility, we introduce PA@20 (Poverty Awareness@20). For each user, PA@20 computes the proportion of recommended foods whose price tags are compatible with that user's financial tags (for example low-price foods for users tagged with low income or food insecurity) and averages this proportion across users. Higher PA@20 indicates that the model adapts recommendations to the user's budget constraints. Finally, we report AvgTags@20, the average number of distinct relevant nutrition or health tags covered by the top-20 list, which reflects the informational richness and tag-level diversity of the recommended menu.

Baseline Models and Setup. Baselines for this task (Table 2) fall into three categories. (i) Classical graph neural network baselines: GCN [30], GraphSAGE [17], and GAT [59], applied to the user-food interaction graph extracted from GLEN Nutrition-Health Graph. (ii) SOTA graph-based recommendation baselines: SGL [74], LightGCN [19], SimGCL [70], LightGCL [6], and NGCF [64], which are widely used in general recommendation benchmarks and optimized for ranking metrics. (iii) SOTA food recommendation baselines: RecipeRec [57], HFRS-DA [14], and MOPI-HFRS [79], which explicitly incorporate nutrition or health constraints and support multi-objective recommendation. All methods are trained and evaluated on the same user-food splits. We repeat experiments across multiple random seeds and report mean and standard deviation.

Results. Table 2 and Figure 4 link model behavior to architectural traits. Figure 4 summarizes these effects at the family level, with panel (a) reporting dimension-wise performance and panel (b) aggregating metric averages. Baseline GNNs (GCN/GraphSAGE/GAT) leverage local neighborhood aggregation on the user–food graph. GAT achieves 10.17±0.30 Recall@20 and 8.58±0.17 NDCG@20, indicating that neighborhood structure supports preference aligned ranking. However, without explicitly modeling health and price, improvements on H-Score@20 and PA@20 are limited. This indicates that interaction patterns alone cannot guarantee clinically and economically viable recommendations. Collaborative filtering variants further strengthen collaborative signals. NGCF improves ranking quality (e.g., 12.93±0.05 Recall@20 and 7.98±0.07 NDCG@20), while LightGCN simplifies propagation and SimGCL/SGL introduce contrastive regularization. These design choices yield steady gains on Recall and NDCG, yet H-Score and PA improve only marginally, reflecting the preference-centered objective. Constraint-aware models exhibit a different pattern. HFRS-DA encodes typed relations among users, foods, health conditions, nutrition tags, and prices, and applies relation-specific attention, achieving 34.21±0.17 H-Score@20 and 27.78±0.75 PA@20 while maintaining 12.78±0.27 Recall@20. This indicates that explicitly modeling relational constraints can improve compliance without sacrificing ranking quality. MOPI-HFRS further couples relevance and constraint objectives via prompt-driven multi-objective optimization. It reaches 13.25±0.34 Recall@20, 9.97±0.61 NDCG@20, 38.18±0.74 H-Score@20, and 26.84±1.07 PA@20, with 7.21±0.73 AvgTags@20, indicating broader coverage of nutrition- and health-related attributes.

Overall, these results highlight that improving ranking quality does not necessarily translate into better constraint satisfaction. Methods that explicitly encode multi-relational health and socioeconomic signals, and optimize for constraint-aware objectives, achieve more clinically appropriate and economically feasible recommendations while preserving competitive Recall and NDCG.

4.3 Nutritional Question Answering

Nutritional question answering helps build trust in dietary recommendations. Patients and providers need clear explanations before changing diets, particularly in clinical contexts. Even when a model can rank foods well, real-world adoption depends on whether it can explain why a food is suitable (or unsuitable) given the user's health conditions, dietary requirements, and socioeconomic constraints. This task therefore complements risk detection and recommendation by evaluating evidence-grounded reasoning and interpretability. By requiring both tag-level attribution and language-based explanations grounded in the GLEN Nutrition-Health Graph, nutritional QA assesses whether models can produce faithful, personalized rationales rather than generic nutrition statements, supporting transparency, user trust, and informed intervention decisions.

Task Definition. Nutritional Question Answering (QA) tests whether models can reason over the GLEN Nutrition-Health Graph and generate accurate, interpretable answers to personalized nutrition questions. Each question presents a user, a candidate food, and relevant graph context, asking whether the food is suitable and which nutrients or tags support the decision. We consider two coupled views: multi-label classification to predict the supporting nutrition tags, and text generation to produce a natural-language justification.

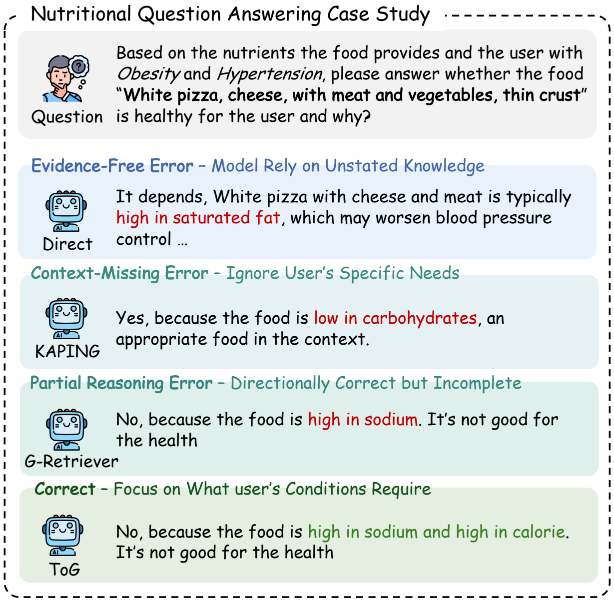
Evaluation Metrics. We evaluate QA along two corresponding output views. For multi-label classification (ML), we report Accuracy, Recall, Precision, F1, and AUC to measure how well the model recovers the ground-truth supporting nutrition tags, balancing coverage (Recall) against spurious predictions (Precision/F1). For text generation (TG), we use ROUGE-1/2/L, BLEU, and BERTScore to compare generated explanations with references in terms of overlap and semantic similarity, reflecting fluency and content fidelity.

Baseline Models and Setup. We benchmark a range of large language models and graph enhanced prompting strategies. For each backbone model (LLaMA3 and GPT-4), we first test a Plain setting that directly prompts the model with the question, without requiring explicit reasoning steps. We then include several reasoning oriented prompting methods such as CoT-Zero [31] and CoT-BAG [61], which encourage chain of thought reasoning, and ToT [73] and GoT [4], which organize intermediate thoughts in a tree or graph structure. To examine how graph knowledge affects performance, we test methods that combine retrieval from the GLEN Nutrition-Health Graph with LLM reasoning. We evaluate four Graph RAG frameworks: KAPING [2] and ToG [55] retrieve and simplify subgraphs before generating answers, G-retriever [20] focuses on retrieving high-quality graph information for reasoning, and KAR [71] incorporates retrieved knowledge triples and refines answers iteratively. All methods share the same question templates and are evaluated on both ML and TG metrics, which allows us to compare pure prompting against graph aware pipelines and to quantify how much structured nutritional knowledge improves personalized nutritional QA.

Results Table 3 shows a clear pattern across backbones. Plain prompting reasons over the input without structured evidence, so it tends to over predict supporting tags and generates rationales that contain generic phrases. Adding chain of thought variants such as CoT-Zero and CoT-BAG, or search style controllers such as ToT and GoT, organizes intermediate steps but still operates without external nutritional structure. The result is small and inconsistent gains in F1 and AUC and only modest movement in ROUGE and BERTScore. Reasoning tokens help with step ordering yet do not

Table 3: Nutritional Question Answering over GLEN Nutrition-Health Graph

Model	Method	Multi-label Classification (-ML)					Text Generation (-TG)				
		Accuracy	Recall	Precision	F1	AUC	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
LLaMA3	Plain	0.3274	0.9880	0.3248	0.4863	0.7336	0.6088	0.5531	0.5997	0.3626	0.9436
	CoT-Zero[31]	0.3282	**0.9896**	0.3262	0.4882	0.7349	0.6118	0.5566	0.6027	0.3649	0.9435
	CoT-BAG[61]	0.3496	0.9335	0.3856	0.5292	0.7701	0.6743	0.5971	0.6569	0.4213	0.9516
	ToT[73]	0.3153	0.7376	0.3832	0.4787	0.7076	0.6408	0.5678	0.6232	0.3844	0.9436
	GoT[4]	0.3117	0.9566	0.3135	0.4709	0.7043	0.6122	0.5487	0.6045	0.3547	0.9422
	KAPING[2]	0.3177	0.9876	0.3191	0.4803	0.7163	0.5931	0.5286	0.5840	0.3704	0.9416
	KAR[71]	0.3275	0.9216	0.3648	0.5102	0.7625	0.6250	0.5589	0.6141	0.3677	0.9444
	ToG[55]	**0.3640**	0.7556	0.4082	0.5118	0.7359	**0.7359**	**0.6697**	**0.7180**	**0.5086**	**0.9591**
	G-retriever[20]	0.3558	0.7811	**0.5054**	**0.5723**	**0.7948**	0.6493	0.5763	0.6330	0.3907	0.9462
GPT4	Plain	0.3089	**0.9938**	0.3096	0.4701	0.7168	0.6080	0.5242	0.5944	0.3489	0.9405
	CoT-Zero[31]	0.3114	0.9912	0.3124	0.4727	0.7216	0.6227	0.5425	0.6090	0.3627	0.9430
	CoT-BAG[61]	**0.3213**	0.9880	0.3250	0.4856	0.7383	0.5307	0.4172	0.5168	0.2563	0.9194
	ToT[73]	0.2568	0.8158	0.3171	0.4474	0.6915	0.6651	0.5904	0.6514	0.4135	**0.9457**
	GoT[4]	0.3088	0.9619	0.3129	0.4705	0.7076	0.6261	0.5476	0.6123	0.3674	0.9390
	KAPING[2]	0.3038	0.9903	0.3065	0.4658	0.7149	0.5906	0.5024	0.5783	0.3318	0.9371
	KAR[71]	0.3117	0.9357	0.3481	**0.4951**	**0.7512**	0.6120	0.5227	0.5982	0.3499	0.9392
	ToG[55]	0.2894	0.5780	**0.3934**	0.4220	0.6663	0.6002	0.4818	0.5853	0.3356	0.9347
	G-retriever[20]	0.3210	0.7882	0.3808	0.4923	0.7371	**0.6830**	**0.6087**	**0.6686**	**0.4551**	0.9455


Nutritional Question Answering Case Study

Question: Based on the nutrients the food provides and the user with *Obesity* and *Hypertension*, please answer whether the food **"White pizza, cheese, with meat and vegetables, thin crust"** is healthy for the user and why?

Evidence-Free Error – Model Rely on Unstated Knowledge

Direct: It depends, White pizza with cheese and meat is typically high in saturated fat, which may worsen blood pressure control ...

Context-Missing Error – Ignore User's Specific Needs

KAPING: Yes, because the food is low in carbohydrates, an appropriate food in the context.

Partial Reasoning Error – Directionally Correct but Incomplete

G-Retriever: No, because the food is high in sodium. It's not good for the health

Correct – Focus on What user's Conditions Require

ToG: No, because the food is high in sodium and high in calorie. It's not good for the health

Figure 5: A case study of error analysis.

supply the fine grained tag level facts the questions require. Graph-based retrieval methods perform much better because they focus on finding high-quality evidence. KAPING and ToG first retrieve and filter relevant subgraphs, then use them for generation. This improves AUC and ROUGE compared to direct prompting because the model now has access to specific nutrient and health information. G-retriever goes further by emphasizing retrieval accuracy. With LLaMA3, it achieves the best multi-label scores in its group, with the highest F1 and AUC, and also improves ROUGE and BERTScore. This shows that better subgraphs lead to fewer incorrect tags and more accurate explanations. KAR uses a different approach by incorporating retrieved knowledge triples during answer generation. With GPT-4, this method achieves the best F1 score among GPT-4 variants, while G-retriever produces the best ROUGE-1, ROUGE-L, and BERTScore. This demonstrates that strong retrieval improves explanation quality even with powerful language models.

Two general patterns emerge from these results. First, retrieval quality determines multi-label performance: accurate graph information leads to better AUC and F1 scores and fewer incorrect tags. Second, model size primarily affects text generation: GPT-4 produces higher ROUGE and BERTScore than LLaMA3 with the same retrieval method, though the relative ranking of retrieval approaches remains stable across both models. These findings suggest choosing the retrieval method before selecting the language model. Use methods like KAR when tag accuracy matters most, and use high-quality retrievers like G-retriever when clear, well-supported explanations are the priority.

5 Conclusion

GLEN-Bench advances nutritional health research by combining diet, clinical data, and socioeconomic factors in a single benchmark. Our evaluation shows that integrating graph-based reasoning with language models improves performance on risk detection, personalized recommendation, and interpretable explanations. This approach better reflects real-world intervention workflows than treating these tasks separately. Results further highlight the importance of explicit relational structure and constraint-aware modeling for learning meaningful representations and generating clinically appropriate, economically feasible recommendations. Beyond opioid use disorder, GLEN-Bench generalizes to diverse nutrition-sensitive conditions (e.g., diabetes) and supports broader tasks such as meal planning, adherence prediction, and food–drug interaction analysis. By standardizing data, tasks, and evaluation protocols, GLEN-Bench enables systematic, reproducible progress toward fair and accessible nutrition-aware AI.

References

[1] Tagne Poupi Theodore Armand, Kintoh Allen Nfor, Jung-In Kim, and Hee-Cheol Kim. 2024. Applications of artificial intelligence, machine learning, and deep

learning in nutrition: a systematic review. *Nutrients* 16, 7 (2024), 1073.

[2] Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136* (2023).

[3] Hygerta Berisha, Reham Hattab, Laura Comi, Claudia Giglione, Silvia Migliac-cio, and Paolo Magni. 2025. Nutrition and lifestyle interventions in managing dyslipidemia and cardiometabolic risk. *Nutrients* 17, 5 (2025), 776.

[4] 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*, Vol. 38. 17682–17690.

[5] Felix Bölz, Diana Nurbakova, Sylvie Calabretto, Armin Gerl, Lionel Brunie, and Harald Kosch. 2023. HUMMUS: a linked, healthiness-aware, user-centered and argument-enabling recipe data set for recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1–11.

[6] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple yet effective graph contrastive learning for recommendation. *arXiv preprint arXiv:2302.08191* (2023).

[7] CDC. 2024. *National Health and Nutrition Examination Survey*. https://wwwn.d.gov/nchs/nhanes/default.aspx

[8] Melody N Chavez and Khary K Rigg. 2020. Nutritional implications of opioid use disorder: A guide for drug treatment providers. *Psychology of Addictive Behaviors* 34, 6 (2020), 699.

[9] Yu Chen, Ananya Subburathinam, Ching-Hua Chen, and Mohammed J Zaki. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proceedings of the 14th ACM international conference on web search and data mining*. 544–552.

[10] Bruce Coburn, Jiangpeng He, Megan Rollo, Satvinder S Dhaliwal, Deborah Kerr, and Fengqing Zhu. [n. d.]. Evaluating Large Multimodal Models for Nutrition Analysis: A New Benchmark Enriched with Contextual Metadata. In *IEEE-EMBS International Conference on Biomedical and Health Informatics 2025*.

[11] PM Cunningham. 2016. The use of sobriety nutritional therapy in the treatment of opioid addiction. *J Addict Res Ther* 7, 3 (2016), Article–3.

[12] Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food* 2, 1 (2018), 23.

[13] Bahare Fatemi, Quentin Duval, Rohit Girdhar, Michal Drozdzal, and Adriana Romero-Soriano. 2023. Learning to substitute ingredients in recipes. *arXiv preprint arXiv:2302.07960* (2023).

[14] Saman Forouzandeh, Mehrdad Rostami, Kamal Berahmand, and Razieh Sheikhpour. 2024. Health-aware food recommendation system with dual attention in heterogeneous graphs. *Computers in Biology and Medicine* 169 (2024), 107882.

[15] Yifu Gao, Linbo Qiao, Zhigang Kan, Zhihua Wen, Yongquan He, and Dongsheng Li. 2024. Two-stage generative question answering on temporal knowledge graph using large language models. *arXiv preprint arXiv:2402.16568* (2024).

[16] Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware food recommender system. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 333–334.

[17] William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive Representation Learning on Large Graphs. arXiv:1706.02216 [cs.SI] https://arxiv.org/abs/1706.02216

[18] Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L McGuinness, and Mohammed J Zaki. 2019. FoodKG: a semantics-driven knowledge graph for food recommendation. In *International Semantic Web Conference*. Springer, 146–162.

[19] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[20] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented gen-eration for textual graph understanding and question answering. *Advances in Neural Information Processing Systems* 37 (2024), 132876–132907.

[21] Colleen Heflin and Xiaohan Sun. 2022. Food insecurity and the opioid crisis. *The ANNALS of the American Academy of Political and Social Science* 703, 1 (2022), 262–284.

[22] Andrew A Herring, Allison D Rosen, Elizabeth A Samuels, Chunqing Lin, Melissa Speener, John Kaleekal, Steven J Shoptaw, Aimee K Moulin, Arianna Campbell, Erik Anderson, et al. 2024. Emergency department access to buprenorphine for opioid use disorder. *JAMA network open* 7, 1 (2024), e2353771–e2353771.

[23] Kim A Hoffman, Javier Ponce Terashima, and Dennis McCarty. 2019. Opioid use disorder and treatment: challenges and opportunities. *BMC health services research* 19, 1 (2019), 884.

[24] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. arXiv:2003.01332 [cs.LG] https://arxiv.org/abs/2003.01332

[25] Jiatan Huang, Mingchen Li, Zonghai Yao, Dawei Li, Yuxin Zhang, Zhichao Yang, Yongkang Xiao, Feiyun Ouyang, Xiaohan Li, Shuo Han, and Hong Yu. 2026. RiTeK: A Dataset for Large Language Models Complex Reasoning over Textual Knowledge Graphs in Medicine. arXiv:2410.13987 [cs.CL] https://arxiv.org/abs/2410.13987

[26] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645* (2023).

[27] Mingxuan Ju, Wenhao Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. 2022. Grape: Knowledge graph enhanced passage reader for open-domain question answering. *arXiv preprint arXiv:2210.02933* (2022).

[28] Dimitrios-Ioannis Kasartzian and Thomas Tsiampalis. 2025. Transforming cardiovascular risk prediction: a review of machine learning and artificial intelligence innovations. *Life* 15, 1 (2025), 94.

[29] Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *arXiv preprint arXiv:2310.11220* (2023).

[30] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs.LG] https://arxiv.org/abs/ 1609.02907

[31] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[32] 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 (2020), 9459–9474.

[33] Diya Li, Mohammed J Zaki, and Ching-hua Chen. 2023. Health-guided recipe recommendation over knowledge graphs. *Journal of Web Semantics* 75 (2023), 100743.

[34] Mingchen Li, Jiatan Huang, Jeremy Yeung, Anne Blaes, Steven Johnson, Hongfang Liu, Hua Xu, and Rui Zhang. 2024. Cancerllm: A large language model in cancer domain. *arXiv preprint arXiv:2406.10459* (2024).

[35] Mingchen Li and Shihao Ji. 2022. Semantic Structure based Query Graph Predic-tion for Question Answering over Knowledge Graph. arXiv:2204.10194 [cs.CL] https://arxiv.org/abs/2204.10194

[36] Ming Li, Lin Li, Xiaohui Tao, and Ning Zhong. 2023. User-meal interaction learning for meal recommendation: A reproducibility study. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 104–113.

[37] Mingchen Li, Chen Ling, Rui Zhang, and Liang Zhao. 2024. A Condensed Tran-sition Graph Framework for Zero-shot Link Prediction with Large Language Models. arXiv:2402.10779 [cs.CL] https://arxiv.org/abs/2402.10779

[38] Xinyi Li, Annabelle Yin, Ha Young Choi, Virginia Chan, Margaret Allman-Farinelli, and Juliana Chen. 2024. Evaluating the quality and comparative validity of manual food logging and artificial intelligence-enabled food image recognition in apps for nutrition care. *Nutrients* 16, 15 (2024), 2573.

[39] Yu Liu and Boyuan Wang. 2025. Advanced applications in chronic disease monitoring using IoT mobile sensing device data, machine learning algorithms and frame theory: a systematic review. *Frontiers in Public Health* 13 (2025), 1510456.

[40] Zheyuan Liu, Chunhui Zhang, Yijun Tian, Erchi Zhang, Chao Huang, Yanfang Ye, and Chuxu Zhang. 2023. Fair graph representation learning via diverse mixture-of-experts. In *Proceedings of the ACM web conference 2023*. 28–38.

[41] 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 (2021), 187–203.

[42] Javier Maroto-Rodriguez, Rosario Ortolá, Veronica Cabanas-Sanchez, David Martinez-Gomez, Fernando Rodriguez-Artalejo, and Mercedes Sotos-Prieto. 2025. Diet quality patterns and chronic kidney disease incidence: a UK Biobank cobbler study. *The American Journal of Clinical Nutrition* 121, 2 (2025), 445–453.

[43] Janice B Montville, Jaspreet KC Ahuja, Carrie L Martin, Kaushalya Y Heendeniya, Grace Omolewa-Tomobi, Lois C Steinfeldt, Jaswinder Anand, Meghan E Adler, Randy P LaComb, and Alanna Moshfegh. 2013. USDA food and nutrient database for dietary studies (FNDDS), 5.0. *Procedia Food Science* 2 (2013), 99–112.

[44] MK Nagarajan and D Goodman. 2020. Not just substance use: the critical gap in nutritional interventions for pregnant women with opioid use disorders. *Public Health* 180 (2020), 114–116.

[45] Joel Paul. 2025. Real-time predictive health monitoring using AI-driven wearable sensors: Enhancing early detection and personalized interventions in chronic disease management. (2025).

[46] Yiyue Qian, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2025. Adaptive Graph Enhancement for Imbalanced Multi-relation Graph Learning. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* (Hannover, Germany) *(WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 717–725. doi:10.1145/3701551.3703553

[47] Yiyue Qian, Chunhui Zhang, Yiming Zhang, Qianlong Wen, Yanfang Ye, and Chuxu Zhang. 2022. Co-modality graph contrastive learning for imbalanced node classification. *Advances in Neural Information Processing Systems* 35 (2022), 15862–15874.

[48] Alessia Salinari, Michele Machì, Yasmany Armas Diaz, Danila Cianciosi, Zexiu Qi, Bei Yang, Maria Soledad Ferreiro Cotorruelo, Santos Gracia Villar, Luis Alonso Dzul Lopez, Maurizio Battino, et al. 2023. The application of digital technologies and artificial intelligence in healthcare: an overview on nutrition assessment. *Diseases* 11, 3 (2023), 97.

[49] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling Relational Data with Graph Convolutional Networks. arXiv:1703.06103 [stat.ML] https://arxiv.org/abs/1703.06103

[50] Hilary K Seligman, Ronli Levi, Victoria O Adebiyi, Alisha Coleman-Jensen, Joanne F Guthrie, and Edward A Frongillo. 2023. Assessing and monitoring nutrition security to promote healthy dietary intake and outcomes in the United States. *Annual Review of Nutrition* 43, 1 (2023), 409–429.

[51] Oshani Seneviratne, Jonathan Harris, Ching-Hua Chen, and Deborah L McGuinness. 2021. Personal health knowledge graph for clinically relevant diet recommendations. *arXiv preprint arXiv:2110.10131* (2021).

[52] Jialiang Shi, Takahiro Komamizu, Keisuke Doman, Haruya Kyutoku, and Ichiro Ide. 2023. Recipemeta: Metapath-enhanced recipe recommendation on heterogeneous recipe network. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia.* 1–7.

[53] Kaiwen Shi, Zheyuan Zhang, Zhengqing Yuan, Keerthiram Murugesan, Vincent Galass, Chuxu Zhang, and Yanfang Ye. 2025. NG-Router: Graph-Supervised Multi-Agent Collaboration for Nutrition Question Answering. *arXiv preprint arXiv:2510.09854* (2025).

[54] Daniel Simancas-Racines, Giuseppe Annunziata, Ludovica Verde, Federica Fascì-Spurio, Claudia Reytor-González, Giovanna Muscogiuri, Evelyn Frias-Toral, and Luigi Barrea. 2025. Nutritional strategies for battling Obesity-Linked liver disease: the role of medical nutritional therapy in metabolic Dysfunction-Associated steatotic liver disease (MASLD) management. *Current Obesity Reports* 14, 1 (2025), 7.

[55] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697* (2023).

[56] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. 2021. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 8903–8911.

[57] Yijun Tian, Chuxu Zhang, Zhichun Guo, Chao Huang, Ronald Metoyer, and Nitesh V Chawla. 2022. RecipeRec: A heterogeneous graph learning model for recipe recommendation. *arXiv preprint arXiv:2205.14005* (2022).

[58] United States Department of Agriculture. 2024. Purchase-to-Plate Crosswalk. https://www.ers.usda.gov/data-products/purchase-to-plate-crosswalk/. Accessed: 2025-02-XX.

[59] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. arXiv:1710.10903 [stat.ML] https://arxiv.org/abs/1710.10903

[60] Freya Waddington, Mark Naunton, Jackson Thomas, Greg Kyle, Brendon Wheatley, and Victor Oguoma. 2023. Examination of the nutritional intake of patients undergoing opioid replacement therapy: A systematic review. *Nutrition & Dietetics* 80, 1 (2023), 55–64.

[61] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems* 36 (2023), 30840–30861.

[62] Peilu Wang, Xiang Gao, Walter C Willett, and Edward L Giovannucci. 2024. Socioeconomic status, diet, and behavioral factors and cardiometabolic diseases and mortality. *JAMA Network Open* 7, 12 (2024), e2451837–e2451837.

[63] Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. 2021. Market2Dish: Health-aware food recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 1 (2021), 1–19.

[64] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval.* 165–174.

[65] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P. Yu, and Yanfang Ye. 2021. Heterogeneous Graph Attention Network. arXiv:1903.07293 [cs.SI] https://arxiv.org/abs/1903.07293

[66] Zehong Wang, Zheyuan Zhang, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. 2024. Gft: Graph foundation model with transferable tree vocabulary. *Advances in Neural Information Processing Systems* 37 (2024), 107403–107443.

[67] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 10370–10388.

[68] David A Wiss. 2019. A biopsychosocial overview of the opioid crisis: considering nutrition and gastrointestinal health. *Frontiers in Public Health* 7 (2019), 193.

[69] Jionglin Wu, Jason Roy, and Walter F Stewart. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care* 48, 6 (2010), S106–S113.

[70] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 726–735.

[71] Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A Rossi, Haoliang Wang, and Julian McAuley. 2025. Knowledge-aware query expansion with large language models for textual and relational retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* 4275–4286.

[72] 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* (2018).

[73] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[74] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval.* 1294–1303.

[75] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining.* 793–803.

[76] Zheyuan Zhang, Yiyang Li, Nhi Ha Lan Le, Zehong Wang, Tianyi Ma, Vincent Galassi, Keerthiram Murugesan, Nuno Moniz, Werner Geyer, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. 2024. NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning. arXiv:2412.15547 [cs.CL] https://arxiv.org/abs/2412.15547

[77] Zheyuan Zhang, Kaiwen Shi, Zhengqing Yuan, Zehong Wang, Tianyi Ma, Keerthiram Murugesan, Vincent Galassi, Chuxu Zhang, and Yanfang Ye. 2025. AgentRouter: A Knowledge-Graph-Guided LLM Router for Collaborative Multi-Agent Question Answering. *arXiv preprint arXiv:2510.05445* (2025).

[78] Zheyuan Zhang, Zehong Wang, Shifu Hou, Evan Hall, Landon Bachman, Jasmine White, Vincent Galassi, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. 2024. Diet-odin: A novel framework for opioid misuse detection with interpretable dietary patterns. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 6312–6323.

[79] Zheyuan Zhang, Zehong Wang, Tianyi Ma, Varun Sameer Taneja, Sofia Nelson, Nhi Ha Lan Le, Keerthiram Murugesan, Mingxuan Ju, Nitesh V Chawla, Chuxu Zhang, et al. 2025. Mopi-hfrs: A multi-objective personalized health-aware food recommendation system with llm-enhanced interpretation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1.* 2860–2871.

[80] Jianan Zhao, Qianlong Wen, Mingxuan Ju, Chuxu Zhang, and Yanfang Ye. 2023. Self-supervised graph structure refinement for graph neural networks. In *Proceedings of the sixteenth ACM international conference on web search and data mining.* 159–167.

A GRAPH CONSTRUCTION DETAILS

A.1 Data Sources and Preprocessing

The GLEN Nutrition-Health Graph is constructed by integrating three population-scale resources to jointly model *clinical status*, *dietary intake*, and *socioeconomic affordability*. Together, these sources provide a unified foundation for studying how health conditions, dietary behaviors, food composition, and real-world access constraints interact at scale.

NHANES (2003–2020). We use NHANES cycles from 2003 to 2020 and extract participant-level information from: (i) *Demographics* (e.g., age, sex, race/ethnicity), (ii) *Examination* and *Laboratory* sections (e.g., BMI, blood pressure, metabolic biomarkers), (iii) *Questionnaire* modules for diagnosed medical conditions (e.g., obesity, hypertension, depression) and behavioral factors (e.g., salt use, supplement intake), and (iv) *Dietary* modules for 24-hour dietary recalls and food consumption logs. We treat each participant as a **user**

node and consolidate all available attributes into a unified user profile via ID-based joins across modules. In addition, we derive socioeconomic attributes including annual household income and food insecurity indicators, which enable annotating users with poverty status and food insecurity risk—introducing real-world feasibility constraints that are rarely modeled in nutrition-health datasets. When multiple records exist across modules, we keep a single unified user record; missing values are preserved as blank (or imputed with simple statistics depending on the downstream model setting).

FNDDS and WWEIA. Each NHANES consumed food item is recorded with a food code. We map NHANES food codes to the Food and Nutrient Database for Dietary Studies (FNDDS) to obtain comprehensive nutrient profiles, and to What We Eat In America (WWEIA) to obtain standardized hierarchical food categories. This mapping allows us to create **food** nodes equipped with nutrient vectors, and to organize foods through **category** nodes and structured **ingredient** composition. Based on standardized nutrient thresholds, we further derive interpretable **nutrition_tag** labels to support downstream tasks that require explicit nutritional justifications.

USDA Purchase-to-Plate. To incorporate real-world affordability constraints, we augment each food with price estimates from the USDA Purchase-to-Plate pipeline. Prices are normalized to *cost per 100 grams* and discretized into a small set of affordability tiers, represented as **price_tag** nodes (Low/Medium/High). These tiers approximate the relative economic burden of selecting specific foods and enable models to jointly reason about nutritional suitability and accessibility, which distinguishes GLEN-Bench from prior nutrition-oriented graph resources that typically ignore economic feasibility.

A.2 Graph Schema and Statistics

We model GLEN Nutrition-Health Graph as a heterogeneous graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R})$, where each node $v \in \mathcal{V}$ is assigned a node type $\tau(v) \in \mathcal{T}$ and each edge $e \in \mathcal{E}$ is assigned a relation type $\phi(e) \in \mathcal{R}$. The graph includes user, food, ingredient, category, dietary habit, health condition, nutrition tag, price tag, poverty condition, and opioid-level nodes, with edges capturing observed consumption, food composition/category structure, nutrition and price tagging, and user-specific clinical/socioeconomic associations. We report the detailed node/relation statistics in Table 4.

Item	Count
Nodes	# User = 104,244, # Food = 9,640, # Ingredient = 36,591, # Category = 36,718, # Habit = 48, # Health Condition = 19, # Nutrition Tag = 18, # Poverty Condition = 1, # Price Tag = 3, # Opioid Level = 3
Relations	# User–Food = 1,803,215, # User–Habit = 652,277, # User–Health Condition = 145,650, # User–Poverty Condition = 35,033, # User–Opioid Level = 98,448, # Food–Ingredient = 31,510, # Food–Category = 8,388, # Food–Nutrition Tag = 52,582, # Food–Price Tag = 7,683, # Health Condition–Nutrition Tag = 23, # Poverty Condition–Price Tag = 1

Table 4: The statistics of GLEN Nutrition-Health Graph.

A.3 Node Features and Representations

We associate each node with a feature vector, with different sources depending on the node type.

User Nodes. Each user node u is associated with a feature vector $x_u \in \mathbb{R}^{d_u}$, constructed by concatenating demographic attributes and clinical/laboratory measurements extracted from NHANES (e.g., BMI, blood pressure, metabolic biomarkers). Diagnosed conditions, dietary habits, poverty-related constraints, and opioid status are modeled as explicit neighbors rather than being fully absorbed into x_u.

Food Nodes. Each food node f is associated with a continuous nutrient vector $x_f \in \mathbb{R}^{d_f}$, derived from FNDDS nutrient profiles. These nutrient values are also used to derive threshold-based nutrition tags (Section 6).

Categorical/semantic nodes. For non-numeric node types such as ingredient, category, habit, health_condition, nutrition_tag, price_tag, and poverty_condition, we represent their semantics using pre-trained BERT embeddings over their textual names or descriptions, producing fixed-length vectors as node representations. This provides a unified embedding space for symbolic concepts that do not have natural numeric features.

A.4 Dietary Habit Extraction

Dietary habit information is compiled from NHANES diet- and behavior-related questionnaires. Due to the categorical diversity of these responses, we perform a structured feature-to-habit mapping: a team of four reviewers identifies questionnaire indicators that describe habitual dietary patterns (e.g., awareness toward healthy eating, frequency of consuming certain food types). For each habit-related feature, we apply a **top/bottom quantile strategy**: participants in the top 10% and bottom 10% (by frequency or ordinal response) are assigned corresponding habit tags. This process yields **48** distinct habit concepts, each modeled as a **habit** node, connected to users via User–Habit edges.

A.5 Health Risk Label Generation (Opioid Status)

We construct a three-level opioid status label for evaluation and create **opioid_level** nodes (3 levels). Users are connected to their corresponding level via User-Opioid_Level edges. Following NHANES drug-use and prescription-related records, we operationalize:

- **Active Opioid Users:** users with evidence of heroin use within the past year, or continuous prescription opioid use for over 90 days.
- **Opioid-Recovered Users:** users with a history of opioid use but without the criteria of active misuse in the current period (derived from NHANES history/current-use signals).
- **Normal Users:** users with no record indicating opioid misuse or long-term prescription history.

Note that not all users have opioid-related records; therefore, we only create User-Opioid_Level edges when the relevant modules are available (98,448 labeled users in the graph statistics).

A.6 Tagging Scheme

GLEN includes two coupled tagging mechanisms: (i) **food-level nutrition tags** derived from nutrient thresholds, and (ii) **health-driven dietary requirement tags** that connect clinical indicators to recommended nutrition constraints.

Nutrition Tags from Nutrient Thresholds (Food → Tag) We define a fixed set of **nutrition_tag** nodes and assign them to foods using nutrient thresholding. Specifically, for each food f and nutrient dimension k, we compare the nutrient value against predefined low/high thresholds to determine whether f satisfies a tag. The thresholds used in GLEN are summarized in Table 6, together with recommended reference values (NRV) when applicable.

Tag assignment. Given food nutrient vector x_f, a tag is added if the corresponding threshold condition holds. For example:

- **Low Sodium**: $\text{sodium}(f) \leq 120$ mg;
- **High Protein**: $\text{protein}(f) \geq 15$ g;
- **Low Sugar**: $\text{sugar}(f) \leq 5$ g;
- **High Fiber**: $\text{fiber}(f) \geq 6$ g.

We then create Food-NutritionTag edges for all satisfied tags (52,582 edges in the final graph).

Health Indicators to Dietary Requirement Tags (Condition → Tag) In addition to food-level tags, we encode clinical dietary requirements by connecting health_condition nodes to recommended nutrition_tag nodes. This captures domain-guided associations such as "hypertension → low sodium" and enables downstream tasks (e.g., personalized recommendation and QA) to reason over interpretable constraint paths.

We summarize the condition-to-tag mapping used in GLEN in Table 5. Each row defines a set of recommended tags for a health indicator. We instantiate these as Health_Condition-Nutrition_Tag edges (23 edges in total).

Price Tags and Affordability Constraints (Food → PriceTag) We discretize normalized price (cost per 100g) into three tiers and represent them with price_tag nodes: low_price, medium_price and high_price.

Poverty Condition and Economic Access (User → Poverty-Condition) We represent extreme socioeconomic constraint with a poverty_condition node, and connect users meeting the criterion through User-PovertyCondition edges. We further anchor affordability constraints by linking PovertyCondition–PriceTag, enabling models to jointly reason about nutritional suitability and economic feasibility.

Health Indicator	Associated Tags
Obesity	Low Calorie; High Fiber
Diabetes	Low Sugar; Low Carb; High Fiber
Anemia	High Iron; High Vitamin C; High Folate Acid; High Vitamin B_{12}
Chronic kidney disease (CKD)	Low Protein; Low Sodium; Low Phosphorus; Low Potassium
Dyslipidemia	Low Saturated Fat; Low Cholesterol; High Fiber
Hyperuricemia	Low Purine
Sleep disorder	High Vitamin D
Depression	High Folate Acid; High Vitamin D
Liver disease	Low Sodium
Weight loss/Low calorie diet	Low Calorie
Low fat/Low cholesterol diet	Low Cholesterol; Low Saturated Fat; High Fiber
Low salt/Low sodium diet	Low Sodium
Sugar free/Low sugar diet	Low Sugar
Diabetic diet	Low Sugar; Low Carb; High Fiber
Weight gain/Muscle building diet	High Calorie; High Protein
Low carbohydrate diet	Low Carb
High protein diet	High Protein
Renal/Kidney diet	Low Protein; Low Sodium; Low Phosphorus; Low Potassium

Table 5: Health indicators and associated nutrition tags in GLEN Nutrition-Health Graph.

Nutrients	Low Threshold	High Threshold	NRV
Calories (kcal)	40	225	2000
Carbohydrates (g)	55	75	-
Protein (g)	10	15	50
Saturated Fat (g)	1.5	5	20
Cholesterol (mg)	20	40	300
Sugar (g)	5	22.5	-
Dietary Fiber (g)	3	6	-
Sodium (mg)	120	200	2000
Potassium (mg)	0	525	3500
Phosphorus (mg)	0	105	700
Iron (mg)	0	3.3	22
Calcium (mg)	0	150	1000
Folic Acid (μg)	0	60	400
Vitamin C (mg)	0	15	100
Vitamin D (μg)	0	2.25	15
Vitamin B12 (μg)	0	0.36	2.4

Table 6: Nutrient thresholds and recommended values for nutrition tags.

B EXPERIMENTAL SETTINGS

B.1 Environmental Settings

All experiments are conducted on a Linux server equipped with two NVIDIA A100 GPUs (40GB each). We implement all methods in Python 3.10.18 with PyTorch 2.4.0 and PyTorch Geometric 2.7.0. We use standard scientific computing libraries (NumPy, SciPy, scikit-learn) for preprocessing and evaluation, and HuggingFace Transformers (v4.57.1) for LLM-based baselines. When applicable, we enable mixed-precision training/inference to improve throughput on A100 GPUs.

B.2 Data Splits and Reproducibility

Train/valid/test splits. For opioid misuse detection (Section 4.1), we evaluate under two split configurations: **60/20/20** and **70/15/15** for train/validation/test, split at the **user** level to prevent leakage. For personalized food recommendation (Section 4.2), we use a single **60/20/20** split under the same user-level partitioning protocol.

Repeated runs. For each split configuration, we evaluate each method using a fixed set of random seeds and report the mean and standard deviation across runs. For training-based methods, we select the checkpoint with the best validation performance and report test results using that checkpoint.

B.3 Hyperparameters

We adopt unified training settings across methods whenever applicable to ensure fair comparison, and select model checkpoints based on validation performance.

Opioid misuse detection. All GNN-based models are optimized with Adam for 500 epochs using hidden dimension 256 and dropout 0.6. We report mean±std over 10 random seeds. We apply dropout in message passing layers and choose learning rate and weight decay from a small preset range on the validation set.

Personalized food recommendation. We formulate recommendation as implicit-feedback ranking and train all recommenders using the BPR objective with negative sampling. We fix the recommendation list length to K=20. For a balanced comparison across methods, we use a unified configuration with hidden dimension 128 and Adam optimizer (learning rate 1e-3, L2 regularization 1e-6) for up to 500 epochs, and apply a learning-rate decay step every 200 epochs. All methods are trained and evaluated on the same 60/20/20 user-level split (performed once with a fixed split seed), and we select the best validation checkpoint for final test reporting. If a baseline is unstable under the shared configuration, we apply minimal adjustments within the same hyperparameter family to ensure convergence, while keeping the protocol and evaluation unchanged.

Nutritional question answering (multi-label + explanation). For QA baselines (prompting, reasoning-oriented prompting, and graph-grounded pipelines), we evaluate both multi-label tag prediction and explanation generation following Section 4.3.2. To reduce variance, we keep decoding settings fixed within each LLM backbone and do not tune decoding hyperparameters on the test set.

C FINE-GRAINED CORRELATION BETWEEN DIETARY PATTERNS AND OPIOID MISUSE

To contextualize the opioid misuse detection task (Section 4.1), we analyze fine-grained differences in medical status and dietary/lifestyle habits across opioid, recovered, and normal user groups in NHANES. Table 7 summarizes the number of users exhibiting each indicator in the three groups and reports significance using Chi-square tests.

As shown in Table 7, we observe consistent and statistically significant distribution shifts across both medical conditions (e.g., sleep disorder, depression, obesity, hypertension, diabetes, CKD, dyslipidemia, hyperuricemia, liver disease) and dietary habits (e.g., diet quality, frozen/ready-to-eat food consumption, eating outside the home, supplement intake, salt-related behaviors), together with tobacco- and alcohol-related indicators. These results indicate that opioid misuse is associated with systematic behavioral and clinical signatures beyond any single variable. This motivates Task 1 to evaluate whether computational models can screen at-risk users by jointly leveraging multi-view signals encoded in the GLEN Nutrition-Health Graph.

Table 7: Statistical analysis of medical status and dietary habits across opioid, recovered, and normal user groups.

	Category	# Opioid Users	# Recov. Users	# Normal Users	p-value
NHANES Data		3104	437	94907	-
Medical Status	Sleep Disorder	989	99	7607	< 0.001
	Depression	1337	167	10733	< 0.001
	Obesity	1155	138	17554	< 0.001
	Hypertension	667	80	10038	< 0.001
	Diabetes	313	23	3681	< 0.001
	Anemia	749	69	20606	< 0.001
	CKD	589	48	5286	< 0.001
	Dyslipidemia	1328	188	26867	< 0.001
	Hyperuricemia	692	80	10570	< 0.001
	Liver disease	480	122	13267	< 0.001
Dietary Habits	Drinks little or no milk	591	77	11609	< 0.001
	Adds lots of salt at table	507	96	9331	< 0.001
	Eats lots of frozen food	296	49	7022	< 0.001
	Eats often outside the home	266	65	9318	< 0.001
	Eats many ready to eat meals	366	57	9271	< 0.001
	Takes few or no supplements	1217	222	52799	< 0.001
	Uses lots of salt in preparation	968	185	33868	< 0.001
	Claims to have a poor diet	1090	149	15975	< 0.001
	Light cigarette smoker	529	65	11443	< 0.001
	Heavy cigarette smoker	172	57	3193	< 0.001
	Uses tobacco rarely	90	21	1765	< 0.001
	Uses tobacco often	407	65	2771	< 0.001
	Drinks Alcohol more than average	923	200	7799	< 0.001

Note: p-values are calculated using Chi-square tests to indicate statistical significance across the three groups.

D PROMPT DESIGN

For nutritional question answering (Section 4.3), we use a unified prompt template with method-specific wrappers. Each instance provides (i) a user and a candidate food, (ii) a compact graph context serialized as a node list and an edge list, and (iii) an explicit output format constraint to ensure parseable predictions.

We control output strictness through difficulty-conditioned notes. For medium difficulty, the model outputs a comma-separated list of nutrition tags (e.g., high_carb, low_sodium) chosen from a fixed option set. For hard difficulty, the model outputs a binary decision (Yes/No) followed by a short rationale that explicitly references the selected tags. Across prompting baselines, we keep the core instruction fixed and only vary the wrapper that introduces the auxiliary evidence (e.g., plain prompting, chain-of-thought style instruction, or graph-grounded retrieval outputs such as KAPING/ToG/GoT/G-Retriever/KAR).

While the core template remains identical, different baselines introduce the auxiliary evidence with lightweight wrapper instructions that mirror their intended reasoning or retrieval mechanism. **Plain** uses only the shared instruction and the serialized graph context without additional reasoning cues. **CoT-Zero** appends a step-by-step instruction, encouraging the model to explicitly (i) extract the food's nutrition tags from the graph, (ii) compare them against the user's health conditions and required tags, and (iii) form a final judgement. **CoT-BaG** follows the same core objective but emphasizes that the input evidence is a directed graph description, prompting the model to interpret nodes and edges as structured facts.

For retrieval-augmented graph baselines, the wrapper primarily clarifies the provenance and intended use of the provided evidence.

Prompt 1: Nutritional QA Prompt Template

System Message: You are a nutrition-aware QA agent. Use **only** the provided graph evidence. Follow the output format strictly.

Difficulty Note:

Medium: output nutrition tags only, as a comma-separated list.

Hard: output `Yes/No` + a short evidence-grounded rationale that explicitly mentions the tags.

Graph Evidence (Serialized):

Node List: `[(id, {type, attr}), ...]`

Edge List: `[(src, rel, dst), ...]`

Question:

Decision: Is the candidate food healthy for the user? (`Yes/No`)

Attribution: Which nutrition tags support the decision?

Explanation (Hard only): Provide a short rationale grounded in the given evidence.

Output Format:

Medium: `low_carb, low_sodium, low_protein, ...`

Hard: `Yes, because the food is low_sodium, low_carb, ...`

Figure 6: Prompt template for nutritional question answering. The prompt combines a unified system instruction, difficulty-conditioned output constraints, and a serialized graph evidence block to encourage evidence-grounded decisions and explanations.

KAPING and **ToG** are framed as graph-grounded pipelines that supply a relevant subgraph/context block to be used as the sole evidence for answering. **ToT** and **GoT** additionally provide intermediate reasoning artifacts produced by tree/graph-structured search and aggregation, and the wrapper instructs the model to use these artifacts as supporting evidence rather than generating unsupported facts. **G-Retriever** specifies that the context is a compact subgraph selected by Prize-Collecting Steiner Tree optimization to balance relevance and brevity, encouraging faithful use of the retrieved neighborhood. Finally, **KAR** introduces knowledge-aware retrieved information (e.g., filtered triples and refined context) and instructs the model to ground both tag attribution and explanation in these retrieved facts.

Across all wrappers, we keep the allowed nutrition-tag vocabulary fixed and enforce the same output-format constraints, so that performance differences reflect the quality of evidence selection and reasoning rather than prompt leakage or formatting freedom.

E ADDITIONAL RELATED WORK

Graph Representation Learning Graph neural networks have become essential for learning from structured relational data. Heterogeneous graph neural networks [75] extend standard GNNs to handle multiple node and edge types, enabling richer modeling of complex systems with diverse entities and relationships. This multi-relational capability is particularly valuable when reasoning requires understanding interactions across different types of connections. Real-world graphs often contain noise, missing edges, and structural errors. Graph structure refinement methods [80] address these challenges by learning to adaptively adjust topology during training, improving robustness to data quality issues. Many graph learning tasks also face severe class imbalance, where target nodes or patterns are rare. Contrastive learning approaches [46, 47] have shown effectiveness in such settings through self-supervised objectives that learn representations without requiring extensive labels, enabling pre-training on large unlabeled graphs before fine-tuning on specific tasks. Fairness in graph learning has gained attention as models may amplify biases present in training data, leading to disparate performance across subgroups [40]. Recent work on graph foundation models [66] aims to build transferable representations that generalize across different graphs and domains, reducing the need for task-specific training data and enabling efficient adaptation to new applications.

F ETHICS AND PRIVACY STATEMENT

GLEN-Bench is constructed from publicly available, population-level resources and is designed with privacy and ethical considerations as first-class requirements. Our primary human data source, the National Health and Nutrition Examination Survey (NHANES), is released under strict confidentiality safeguards and public-data policies. The NHANES records we use are de-identified and do not contain personally identifiable information (PII) such as social security numbers, names, phone numbers, or physical addresses. We additionally follow a data minimization principle: our benchmark focuses on variables necessary for nutrition-health assessment (e.g., demographic attributes, clinical measurements, dietary recalls, and questionnaire-derived behaviors) and does not attempt to re-identify individuals or link records to external identity sources.

GLEN-Bench integrates NHANES with food composition and category information (FNDDS/WWEIA) and aggregated price/access signals (USDA Purchase-to-Plate) to study clinically appropriate and feasible nutrition interventions. These auxiliary resources do not introduce individual-level identifiers; they provide food-level nutrition and affordability context that supports constraint-aware modeling. We report results in aggregate and do not release any information that could enable tracing predictions back to specific survey participants.

We recognize that nutrition and substance-use–related variables can be sensitive. As a benchmark, GLEN-Bench is intended for research on equitable decision support rather than automated clinical decision-making. Any real-world deployment should include additional safeguards, including access control, secure storage, audit logging, and human oversight, and should treat recommendations and explanations as protected health information within clinical workflows. By operating within established survey policies and emphasizing de-identification, minimization, and responsible use, we aim to uphold strong ethical integrity and privacy protection throughout dataset construction, evaluation, and dissemination.

G LIMITATIONS AND DISCUSSION

GLEN-Bench provides a unified setting for end-to-end, constraint-aware nutritional health assessment, but several limitations remain. First, the benchmark is built on observational survey data (NHANES) and derived resources, which introduces confounding and measurement noise (e.g., imperfect 24-hour recalls and questionnaire responses). As a result, strong performance should be interpreted as capturing reliable associations rather than causal effects, and models may exploit spurious shortcuts. A promising direction is to incorporate evaluations that stress-test robustness to confounding and missingness, such as sensitivity analyses, subgroup robustness checks, and counterfactual or intervention-style variants (e.g., whether a model's recommendation changes appropriately when specific constraints or nutrient targets are perturbed).

Second, the feasibility signals in GLEN-Bench are necessarily simplified. Price tags and access-related constraints are estimated at an aggregate level and may not reflect local availability, seasonal price variation, or individual purchasing context. Similarly, practical dietary decision-making often depends on additional constraints not fully represented here—cultural preferences, allergies, cooking skills, time budget, household composition, and condition- or medication-specific restrictions. Future versions of the benchmark could enrich these dimensions by adding finer-grained SDoH variables, dynamic affordability/availability indicators, and structured preference profiles, enabling more realistic evaluation of what it means for recommendations to be simultaneously clinically appropriate and actionable.

Third, while our task suite captures a realistic workflow (screening → recommending → explaining), the current metrics do not fully measure clinical utility, long-term adherence, or safety beyond tag-level compliance. In particular, LLM-based methods can generate fluent rationales that are not faithfully grounded in graph evidence. An important future direction is to strengthen faithfulness and safety evaluation: auditing whether explanations cite retrieved evidence, checking consistency between predicted tags and generated rationales, calibrating model confidence (including abstention on uncertain cases), and incorporating human-in-the-loop evaluation protocols that reflect clinical and public-health decision processes. More broadly, GLEN-Bench opens opportunities for additional tasks such as nutrient-aware substitution, meal planning under budget and nutrient constraints, adherence prediction, and food–drug interaction analysis, which would further align benchmarking with real-world nutrition interventions.

Overall, we view GLEN-Bench as an evolving platform. The current release establishes a reproducible, multi-task benchmark with explicit socioeconomic constraints, while future work can deepen causal rigor, broaden population coverage, enrich constraint modeling, and improve faithfulness and safety assessment for nutrition-aware decision support.