NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning

Zheyuan Zhang[1*], Yiyang Li[1*], Nhi Ha Lan Le[4*], Zehong Wang[1], Tianyi Ma[1], Vincent Galassi[1], Keerthiram Murugesan [3], Nuno Moniz[1], Werner Geyer[3], Nitesh V Chawla[1], Chuxu Zhang[2], Yanfang Ye[1†]

[1]University of Notre Dame, [2]University of Connecticut, [3]IBM Research, [4]Brandeis University

[*]Equal Contribution [†]Corresponding Author

{zzhang42,yli62,zwang43,tma2,vgalassi,nmoniz2,nchawla,yye7}@nd.edu,

nhihlle@brandeis.edu, keerthiram.murugesa@ibm.com, werner.geyer@us.ibm.com, chuxu.zhang@uconn.edu

Abstract

Diet plays a critical role in human health, yet tailoring dietary reasoning to individual health conditions remains a major challenge. Nutrition Question Answering (QA) has emerged as a popular method for addressing this problem. However, current research faces two critical limitations. On the one hand, the absence of datasets involving user-specific medical information severely limits *personalization*. This challenge is further compounded by the wide variability in individual health needs. On the other hand, while large language models (LLMs), a popular solution for this task, demonstrate strong reasoning abilities, they struggle with the *domain-specific* complexities of personalized healthy dietary reasoning, and existing benchmarks fail to capture these challenges. To address these gaps, we introduce the **N**utritional **G**raph **Q**uestion **A**nswering (**NGQA**) benchmark, the first graph question answering dataset designed for *personalized nutritional health reasoning*. NGQA leverages data from the National Health and Nutrition Examination Survey (NHANES) and the Food and Nutrient Database for Dietary Studies (FNDDS) to evaluate whether a food is healthy for a specific user, supported by explanations of the key contributing nutrients. The benchmark incorporates three question complexity settings and evaluates reasoning across three downstream tasks. Extensive experiments with LLM backbones and baseline models demonstrate that the NGQA benchmark effectively challenges existing models. In sum, NGQA addresses a critical real-world problem while advancing GraphQA research with a novel domain-specific benchmark. Our codebase and dataset are available here.

1 Introduction

Diet is a cornerstone of human health, playing a pivotal role in both maintaining well-being and preventing disease. Despite the well-documented benefits of balanced nutrition, unhealthy eating habits remain alarmingly prevalent in modern society (WHO, 2021). In the United States alone, approximately 42.4% of adults are classified as obese (CDC, 2020a), and in 2017, poor dietary habits contributed to over 11 million deaths and a substantial number of disability-adjusted life-years (DALYs), often linked to factors such as excessive sodium intake (Afshin et al., 2019; WHO, 2023). These statistics underscore an urgent need to promote healthier eating habits on a societal scale. However, nutritional health requires complex domain knowledge, and there is no one-size-fits-all solution for healthy diets, as the nutritional needs of individuals can vary widely based on their health conditions. For example, a diet suitable for someone with a high body mass index (BMI) may differ drastically from that of an individual with a low BMI. Likewise, while individuals recovering from opioid misuse may benefit from a high-protein diet, such dietary choices can be harmful to those managing chronic kidney disease (Mahboub et al., 2021).

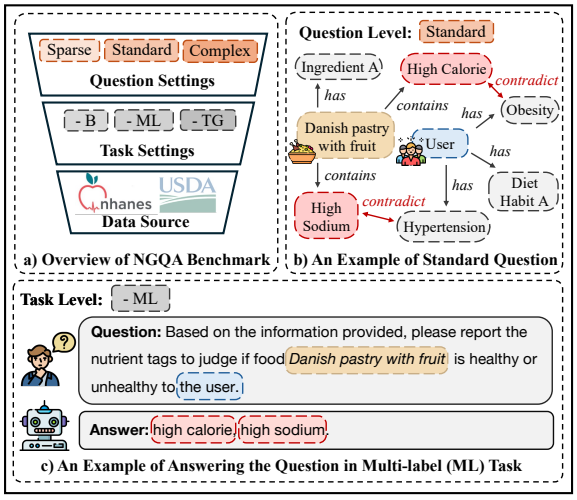

Figure 1: An Overview of NGQA Benchmark (a) along with a data showcase: (b) an example of the knowledge graph used for a standard level question and (c) the question and the answer of that question under the multi-label classification task (-ML) settings.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5934–5966

Why this benchmark matters: Numerous efforts have sought to address the challenges in personalized nutritional health, with Nutrition Question Answering (QA) emerging as a popular task (Min et al., 2022; Bondevik et al., 2024). Recent advancements in large language models (LLMs) have demonstrated significant potential in this domain, offering sophisticated reasoning capabilities to analyze and interpret nutritional information (Mavromatis and Karypis, 2024). However, these efforts remain constrained by two major limitations. First, to the best of our knowledge, no existing benchmark truly personalizes answers based on users' specific health conditions, primarily due to the inaccessibility of individual medical data (Bölz et al., 2023). This lack of user-specific datasets has severely hindered the development of effective solutions. Second, while LLMs exhibit impressive reasoning capabilities in general domains, the medical and nutritional intricacies of this task impose severe limitations on their effectiveness (Mialon et al., 2023). Current benchmarks fail to capture the domain-specific complexities of personalized health-aware dietary reasoning, making it difficult to evaluate, let alone improve, these models in meaningful ways.

To address these critical gaps and advance the understanding of healthy diet personalization, we propose the **N**utritional **G**raph **Q**uestion **A**nswering (**NGQA**) benchmark. This is *the first benchmark in the personalized nutritional health domain* to evaluate whether a specific food is healthy for a user, supported by detailed reasoning of the key contributing nutrients. By recognizing the intricate interplay between a user's medical conditions, dietary behaviors, and the nutrition of foods, we frame this task as a knowledge graph question answering problem. Specifically, using data from the National Health and Nutrition Examination Survey (NHANES) and the Food and Nutrient Database for Dietary Studies (FNDDS), we construct the NGQA benchmark and categorize questions into three complexity settings: sparse, standard, and complex. Each question type is further evaluated through three downstream tasks, binary classification (-B), multi-label classification (-ML), and text generation (-TG), to explore distinct reasoning aspects (Figure-1 (a)). We conduct extensive experiments using various LLM backbones and baseline models to ensure the benchmark is both appropriately challenging and meaningful for advancing the field. Our contributions are summarized as follows:

- **Novel Benchmark for Personalized Nutrition.** We present NGQA, the first benchmark to incorporate users' medical information in a nutritional question answering task, addressing a significant research gap in the domain of personalized healthy diet research.
- **Advancing the GraphQA Ecosystem.** NGQA introduces a domain-specific benchmark and extends GraphQA benchmarks beyond datasets like *WebQSP* and *ExplaGraphs* in the general domain. This addition broadens the scope of GraphQA research, enabling a more comprehensive evaluation of GraphQA models' capabilities beyond general reasoning tasks.
- **Comprehensive Resource and Evaluation.** Through extensive experiments, NGQA provides a challenging benchmark, a complete codebase supporting the full pipeline from data preprocessing to model evaluation, and an extensibility for integrating new models. This comprehensive resource helps advance research in both personalized nutritional health and the broader GraphQA field.

2 Related Work

Question Answering in Nutritional Health Domain. Question answering has become an essential tool in the nutritional and health domain, offering a flexible framework for applications such as food recommendation (Min et al., 2022; Bondevik et al., 2024). Knowledge graphs (KGs) have been widely used to model relationships between foods, ingredients, and health, supporting tasks like ingredient substitution and adaptive dietary recommendations (Haussmann et al., 2019; Chen et al., 2021; Fatemi et al., 2023a; Xu et al., 2024). Recent approaches incorporate health metrics into QA systems, focusing on recipe recommendations and nutritional ontologies (Li et al., 2023; Seneviratne et al., 2021). However, existing methods lack true personalization, as highlighted by (Bölz et al., 2023), due to the absence of user-specific medical data. Our work fills this gap by introducing the first GraphQA benchmark for personalized nutritional health, enabling models to provide tailored nutritional reasoning and explanations.

Graph Retrieval Augmented Generation. Knowledge Graph Question Answering (KGQA) has progressed from early semantic parsing and

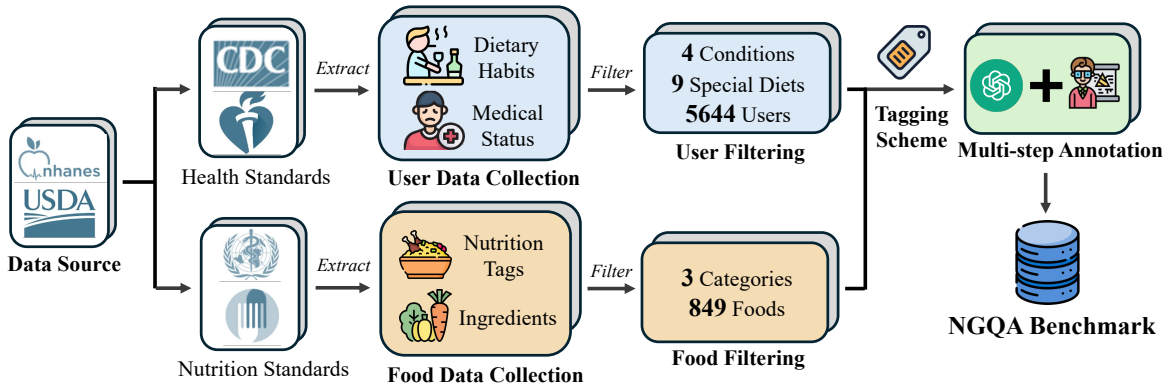

Figure 2: The NGQA benchmark construction process. Each stage shown in the figure is detailed in Section 3.For example, "User Data Collection" block, is introduced in Section 3.1 under the paragraph titled User Data Collection.

retrieval-based methods to advanced techniques leveraging large language models (LLMs) and graph neural networks (GNNs) for reasoning and retrieval (Jiang et al., 2023; Kim et al., 2023; Gao et al., 2024). Building on this progress, Graph-Retrieval Augmented Generation (Graph-RAG) has emerged as a widely studied method, offering more precise, context- and structure-aware reasoning compared to traditional text-based RAG methods (Lewis et al., 2020; Lazaridou et al., 2022; Guo et al., 2024; Wen et al., 2023). Despite the development of various LLM-powered models, benchmarks for the Graph-RAG task remain scarce and lack standardization. Early benchmarks focus primarily on general graph tasks such as shortest paths and node degree (Fatemi et al., 2023b; Wang et al., 2024a), while (He et al., 2024) introduces a GraphQA benchmark for complex reasoning using general-purpose datasets. Building on their framework, we develop the first domain-specific benchmark in the nutritional health domain, bridging the gap between general GraphQA research and personalized health-aware reasoning. More detailed literature is available in Appendix-A.

3 NGQA Benchmark

3.1 Data Collection

Data Source. Using data from the National Health and Nutrition Examination Survey (NHANES) and the Food and Nutrient Database for Dietary Studies (FNDDS), we construct the first GraphQA benchmark designed to address personalized healthy nutrition intake questions. This benchmark integrates detailed user health profiles, dietary behaviors, and comprehensive food nutritional information, enabling a fine-grained analysis of how individual health conditions interact with food nutrition. By representing these relationships through graph structures, the benchmark supports answering complex nutritional questions while capturing the intricate interplay between users' medical conditions and dietary choices. The following sections provide a detailed discussion of these datasets and their integration into our benchmark.

User Data Collection. The NHANES dataset forms the foundation of our work for collecting user data. We extract medical information, dietary habits, and food intake records to construct the graph. Specifically, NHANES provides laboratory reports detailing body metrics like Body Mass Index (BMI) and blood pressure, along with biochemical markers such as blood urea nitrogen. It also includes questionnaire responses on prescription drug usage, adherence to special diets, and overall health status. Additionally, NHANES records users' food intake history and dietary behaviors, such as the frequency of adding salt at the table. Our study incorporates 54 distinct dietary habits, with detailed data processing methods provided in Appendix-B. This comprehensive dataset serves as the backbone of our graph, capturing user health conditions and dietary patterns with granular detail.

Food Data Collection. Nutritional information for food items is sourced from FNDDS. FNDDS connects NHANES food codes to detailed nutritional data cataloged in the What We Eat in America (WWEIA) database. Using FNDDS, we associate each food item in NHANES with its full nutritional composition. Additionally, FNDDS links food items to ingredient information and classifies them into broader food categories. For example, a food item like "apple" is linked to its nutrient values (e.g., sugars, vitamins) and assigned to the category "fruits." These associations enrich the graph by providing node-level data for food, ingredients, and

categories.

Tagging Scheme. To evaluate whether a food is specifically healthy for a user based on their personal health conditions, we propose a tagging scheme that assigns nutrition-related tags to both users and foods. This systematic framework aligns food nutritional properties with user health needs, enabling robust assessments of food suitability.

For food tagging, we build upon established guidelines and introduce newly applied standards. Prior works have utilized recommendations from the World Health Organization (WHO) and the Food Standards Agency (FSA) (Wang et al., 2021), while we extend this by incorporating the more detailed EU Nutrition & Health Claims Regulation (Commission, 2006) and the Codex Alimentarius Commission (CAC) (Alimentarius, 1985, 1997). These standards define precise thresholds for nutrient claims. For instance, the EU regulation permits labeling a food as "low sodium" only if it contains no more than 0.12 g of sodium per 100 g (Commission, 2006). Foods meeting such criteria are tagged with corresponding labels like "low_sodium" or "high_protein", reflecting their nutritional properties.

On the user side, health tags are derived from the NHANES dataset, which includes laboratory results and self-reported health information. For example, users with high blood pressure, as defined by American Heart Association (AHA) thresholds or similar guidelines, are tagged with "hypertension," indicating that a low-sodium diet would be beneficial (Grillo et al., 2019; Smyth et al., 2014).

By linking health and food tags, our scheme effectively represents personalized dietary needs and captures the interplay between medical conditions and nutritional requirements. The detailed standards and additional tags for other nutrients and health conditions are described in Appendix-B. By integrating this methodology into our graph-based benchmark, we provide a framework for advancing personalized dietary reasoning and evaluating models in this domain.

3.2 Data Annotation

Real-world data is inherently messy and incomplete, and the datasets we use are no exception. Spanning from 2003 to 2020, NHANES provides data for approximately 100,000 users and over 2 million food records. While this dataset offers an invaluable resource for studying nutrition and health, it includes inconsistencies, ambiguities, and irrelevant entries. To establish a scientifically robust and meaningful benchmark, precise data annotation is essential. This involves not only cleaning and filtering the data but also carefully defining and validating annotations to accurately capture real-world relationships between health conditions, dietary behaviors, and food options. Our annotation process refines both user and food datasets to ensure relevance, accuracy, and applicability to real-life scenarios.

User Filtering. Annotating user data requires careful consideration of the complex interactions between nutrition and health. For instance, elevated blood urea nitrogen (BUN) levels may indicate kidney dysfunction, warranting a low-protein diet, but could also result from insufficient water intake. To maintain scientific rigor and practical relevance, we focus on annotating four prevalent health statuses—obesity, hypertension, opioid misuse, and diabetes—that are directly influenced by dietary interventions. Additionally, we annotate nine special diets reported by users, reflecting health-related dietary practices. Further details on the definitions and implications of these health statuses and diets are provided in the Appendix-B. To ensure consistency and relevance, we exclude users under 18, focusing solely on adult dietary patterns.

Food Filtering. For food annotation, we identify practical entries in the FNDDS database that align with real-world dietary reasoning. While FNDDS supports comprehensive nutritional analysis, it includes many entries unsuitable for practical use, such as raw ingredients or standalone additives. To address this, we restrict our focus to the "mixed dishes" category, as it represents combined recipes closest to real-life diets. Additionally, we include other relevant categories, such as bakery products and desserts (definitions of FNDDS categories are available in the Appendix-I). Finally, we apply a keyword-based deduplication method to remove highly similar entries.

Multi-step Annotation. Using the previously defined standards and tagging schemes, our annotation process systematically establishes "match" or "contradict" relationships between user health conditions and food nutritional profiles. For example, the tag "high_calorie" contradicts the condition "obesity", while "low_sodium" matches with "hypertension". To ensure accuracy and reliability, we adopt a multi-step annotation process. After initial filtering and tagging, large language models (LLMs) perform an initial sanity check to iden-

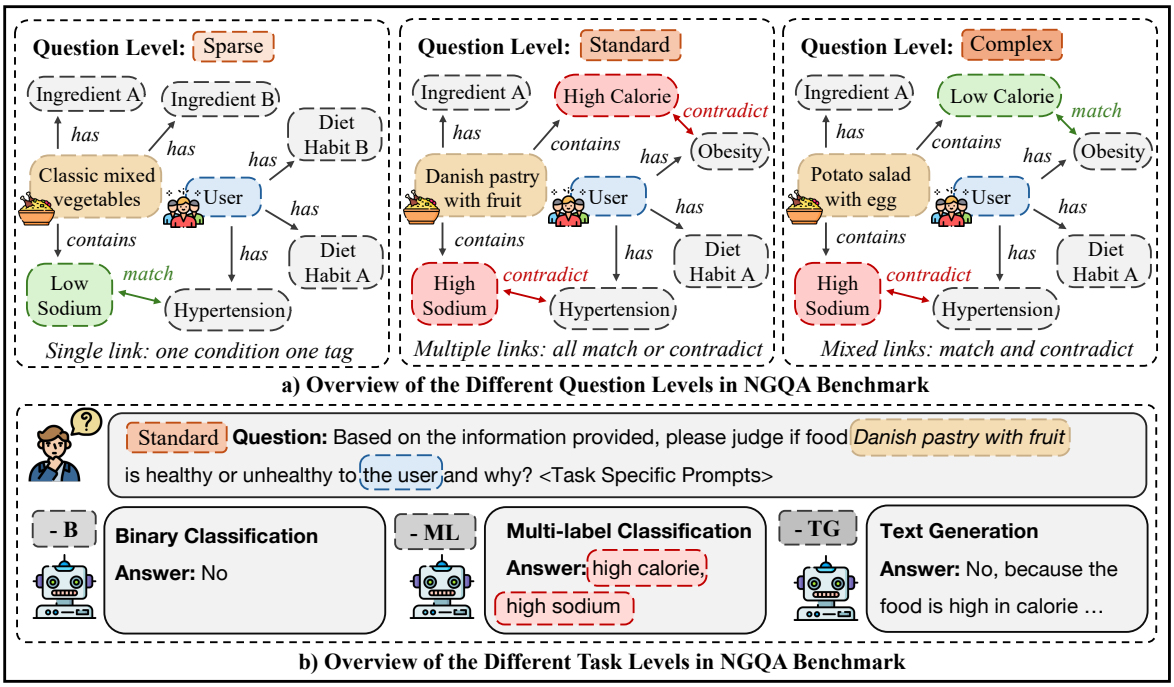

Figure 3: The illustration of different question levels and task levels.

tify inconsistencies or anomalies in the annotations. Subsequently, three human annotators with domain expertise review and cross-validate the results to eliminate remaining inaccuracies. By combining automated checks with human validation, our rigorous annotation strategy captures the real-life complexities of personalized nutrition while maintaining high standards of quality and reliability.

4 Task Definition and Evaluation

4.1 Question Setting

With the annotated data in place, we designed three distinct types of questions, i.e., *sparse*, *standard*, and *complex*, to capture varying levels of difficulty and emulate real-world scenarios in personalized nutrition reasoning. This stratification ensures that our benchmark accommodates a wide range of research and application needs, spanning from controlled, idealized setups to challenging, real-life cases, as illustrated in Figure-3 (a).

Sparse questions address scenarios with minimal available information. In this setting, each food has only one nutrition tag linked to a single user health condition. This setup reflects real-world cases where labels are scarce or data is incomplete, challenging models to reason effectively with limited information. Although sparse questions may appear simple to human observers, the unique link between the user and the food significantly increases the difficulty of subgraph retrieval, making models vulnerable to interference from irrelevant nodes.

Standard questions represent the balanced and idealized scenarios in our benchmark. In this category, foods are linked to multiple nutrition tags, which either match or contradict several user health conditions. This configuration reflects controlled cases where the relationship between dietary choices and health outcomes is clear-cut, enabling a focused evaluation of model performance. Standard questions serve as a foundation for benchmarking in structured and well-defined environments.

Complex questions are designed to replicate the intricacies of real-life nutritional decision-making. Foods in this category may simultaneously have tags that both match with and contradict a user's health conditions. For instance, a food may be low in sodium (beneficial for hypertension) but also high in sugar (problematic for diabetes). These scenarios require models to navigate conflicting information, prioritize user health needs, and perform nuanced trade-off reasoning. This category closely mirrors the ambiguous and multifaceted challenges of real-world dietary decisions.

The benchmark's statistical breakdown is presented in Table-1. To further evaluate the complexity and informativeness of the questions, we introduce the Signal-to-Noise Ratio (SNR). SNR measures the ratio of nodes or tags relevant to the answer (*signal*) against the total nodes or tags in the graph (*noise*). As shown in Table-2, sparse questions exhibit the lowest SNR, reflecting the limited

Question Level	# Records	Avg. # Nodes	Avg. # Edges
Sparse	8,490	25.8	24.9
Standard	3,622	28.2	29.0
Complex	1,690	30.9	34.0

Table 1: Statistics of the Benchmark by Question Level.

Question Level	Avg. Node SNR	Avg. Tag SNR
Sparse	16.4	19.3
Standard	24.7	49.4
Complex	31.6	76.3

Table 2: Signal-to-Noise Ratio (SNR) by Question Level.

resources available for these tasks. Conversely, complex questions, despite containing conflicting information, achieve the highest SNR, underscoring the rich contextual information necessary for accurate reasoning. More statistics of the benchmark are available in Appendix-E.

4.2 Task Setting

To enhance the generality and versatility of our benchmark, we design three distinct downstream task types, each centered on the same domain question but requiring different forms of output, as illustrated in Figure-3 (b). This diversity ensures the benchmark accommodates a wide range of methodologies and research focuses while fostering innovation in addressing personalized nutrition challenges. The tasks are defined as follows:

Binary Classification (-B): This task requires a simple "yes" or "no" response, indicating whether a specific food is suitable for a user based on their health profile. It emphasizes straightforward decision-making, reflecting applications like automated diet advisories or recommendation systems.

Multi-Label Classification (-ML): In this task, models must retrieve the nutritional tags associated with a food and determine which match with or contradict the user's health conditions. By demanding richer output, this task evaluates the model's ability to leverage graph information and identify nuanced relationships.

Text Generation (-TG): The output is a natural language explanation of why a food is healthy or unhealthy for a user. This task assesses a model's capability for interpretable and user-friendly reasoning, which is crucial for real-world applications such as personalized dietary assistant chatbots.

4.3 Evaluation Metrics

To evaluate model performance, we adopt task-specific metrics tailored to each type. For classification tasks, we use standard metrics like accuracy, recall, precision, and F1 score for comprehensive performance assessment. Multi-label classification tasks extend these metrics to their weighted versions, accounting for the distribution of multiple labels across samples. Text generation tasks are evaluated with widely used metrics such as ROUGE, BLEU, and BERT scores, which collectively assess relevance and semantic similarity to reference texts. The definition of ground truths is available in Appendix-B. This multifaceted design supports diverse model architectures and evaluation strategies, providing a robust foundation for advancing personalized nutrition research. By bridging the gap between controlled research environments and the complexities of real-world applications, our benchmark fosters innovation and opens new avenues for addressing healthy dietary reasoning.

5 Experiments

5.1 Experiment Settings

In this section, we conduct extensive experiments to evaluate existing Graph-RAG models' reasoning capability on the proposed benchmark. For baseline models, we select the five most classical baselines: KAPING (Baek et al., 2023), CoT-Zero (Kojima et al., 2022), CoT-BAG (Wang et al., 2024a), ToG (Sun et al., 2024), and a naive plain Graph-RAG pipeline (implementation details in Appendix-C). For the main experiments, we choose GPT-4o-mini and Llama-3.1-70B-instruct as the LLM backbone, we also conduct additional experiments on a series of GPT-3.5 in Appendix-D. Note that we didn't select the most advanced LLM backbones or the most sophisticated fine-tuned baselines because we argue our contributions focus primarily on the proposed benchmark with the novel tasks for this specific domain, and the experiment results along with the hallucination analyses have demonstrated our tasks are properly designed where the classic baselines can be adequately challenged while maintaining efficiency. In the following sections, we go through the experiment results for each task.

5.2 Binary Classification Task

Table-3,4 (a) presents the performance of baseline models on the binary classification task, which evaluates the models' ability to provide a decisive "yes"

Question Level	Method	a) Binary Classification (-B)				b) Multi-label Classification (-ML)				c) Text Generation (-TG)				
		Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Sparse	Plain	0.597	0.163	**1.000**	0.281	0.180	0.994	0.211	0.344	0.539	0.478	0.539	0.284	0.937
	KAPING	0.535	0.054	0.725	0.101	0.175	0.992	0.208	0.339	0.523	0.460	0.523	0.267	0.935
	CoT-Zero	0.660	0.295	0.998	0.456	0.203	0.996	0.244	0.384	0.546	0.484	0.546	0.289	0.939
	CoT-BAG	0.604	0.177	**1.000**	0.301	0.213	**1.000**	0.252	0.394	0.548	0.489	0.548	0.293	0.938
	ToG	**0.773**	**0.538**	0.982	**0.695**	**0.244**	0.913	**0.299**	**0.433**	**0.625**	**0.571**	**0.625**	**0.361**	**0.946**
Standard	Plain	0.576	0.199	**1.000**	0.332	0.491	0.998	0.490	0.653	0.722	0.632	0.694	0.484	0.962
	KAPING	0.502	0.064	0.931	0.119	0.459	0.996	0.462	0.627	0.709	0.624	0.676	0.462	0.960
	CoT-Zero	0.657	0.351	**1.000**	0.519	0.539	0.997	0.545	0.696	0.733	0.644	0.705	0.494	0.963
	CoT-BAG	0.590	0.225	**1.000**	0.367	0.560	**1.000**	0.561	0.709	0.733	0.646	0.703	0.495	0.963
	ToG	**0.863**	**0.741**	0.999	**0.851**	**0.619**	0.884	**0.679**	**0.746**	**0.818**	**0.763**	**0.782**	**0.611**	**0.972**
Complex	Plain	0.660	0.064	0.975	0.119	0.718	0.972	0.737	0.836	0.736	0.651	0.700	0.495	0.960
	KAPING	0.657	0.057	0.972	0.108	0.688	**0.976**	0.713	0.809	0.739	0.663	0.702	0.484	0.960
	CoT-Zero	0.663	0.072	0.978	0.134	0.745	0.974	0.768	0.856	0.748	0.660	0.710	0.505	0.962
	CoT-BAG	0.663	0.070	**1.000**	0.131	**0.755**	0.963	0.780	**0.859**	0.747	0.662	0.708	0.505	0.961
	ToG	**0.747**	**0.396**	0.810	**0.532**	0.615	0.699	**0.812**	0.730	**0.773**	**0.692**	**0.737**	**0.531**	**0.964**

Table 3: Experimental results based on five baseline methods on the three tasks with the three question levels using the GPT-4o-mini. The best performance of each group is bolded.

Question Level	Method	a) Binary Classification (-B)				b) Multi-label Classification (-ML)				c) Text Generation (-TG)				
		Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Sparse	Plain	0.616	0.241	0.862	0.377	0.219	0.896	0.237	0.367	0.565	0.500	0.564	0.309	0.938
	KAPING	0.533	0.073	0.627	0.131	0.195	0.889	0.219	0.347	0.537	0.468	0.537	0.276	0.935
	CoT-Zero	0.605	0.289	0.726	0.413	0.363	0.764	0.427	0.526	0.559	0.502	0.559	0.342	0.887
	CoT-BAG	0.606	0.288	0.731	0.413	**0.420**	0.743	**0.472**	**0.559**	0.548	0.489	0.547	0.333	0.885
	ToG	**0.848**	**0.696**	**0.984**	**0.815**	0.323	**0.956**	0.317	0.467	**0.722**	**0.679**	**0.722**	**0.500**	**0.958**
Standard	Plain	0.590	0.258	0.887	0.400	0.565	0.922	0.567	0.693	0.775	0.707	0.734	0.551	0.966
	KAPING	0.481	0.048	0.622	0.089	0.483	0.895	0.506	0.639	0.720	0.637	0.683	0.475	0.959
	CoT-Zero	0.658	0.353	**1.000**	0.522	0.537	0.996	0.543	0.695	0.733	0.645	0.706	0.494	0.951
	CoT-BAG	0.587	0.220	**1.000**	0.360	0.559	**0.998**	0.560	0.708	0.548	0.489	0.547	0.333	0.885
	ToG	**0.865**	**0.744**	**1.000**	**0.853**	**0.824**	0.924	**0.844**	**0.875**	**0.887**	**0.829**	**0.823**	**0.696**	**0.978**
Complex	Plain	0.625	0.042	0.356	0.0758	0.679	0.868	0.770	0.811	0.761	0.681	0.714	0.510	0.960
	KAPING	0.630	0.047	0.414	0.0849	0.655	0.850	0.752	0.792	0.745	0.664	0.703	0.491	0.959
	CoT-Zero	0.664	0.075	0.979	0.139	0.747	**0.973**	0.769	0.856	0.747	0.660	0.711	0.505	0.948
	CoT-BAG	0.662	0.0685	**1.000**	0.128	**0.753**	0.963	0.778	**0.858**	0.747	0.662	0.708	0.505	0.947
	ToG	**0.722**	**0.294**	0.830	**0.434**	0.687	0.716	**0.895**	0.785	**0.818**	**0.742**	**0.765**	**0.598**	**0.969**

Table 4: Experimental results based on five baseline methods on the three tasks with the three question levels using the Llama-3.1-70B-instruct. The best performance of each group is bolded.

or "no" response based on summarized reasoning. The results reveal a notable conservatism in model behavior, as evidenced by the low recall scores. This likely stems from the sensitive nature of medical questions, where LLMs try to avoid offering simple "yes" answers without explanations unless their confidence is exceptionally high. Despite this challenge, the experiments yield two important insights into how external domain knowledge can support LLMs in this scenario. First, increasing the number of links in the graph (e.g., from Sparse to Standard questions) consistently improves recall across all baselines. This indicates that richer external knowledge provides LLMs with greater context and reassurance, enabling them to produce more confident positive answers. Second, ToG significantly outperforms other baselines, showing performance gains unique to this task. We attribute this improvement to ToG's effective pruning mechanism, which removes irrelevant nodes and increases the SNR. By reducing noise and focusing on relevant information, ToG enhances LLMs' ability to make confident and accurate decisions.

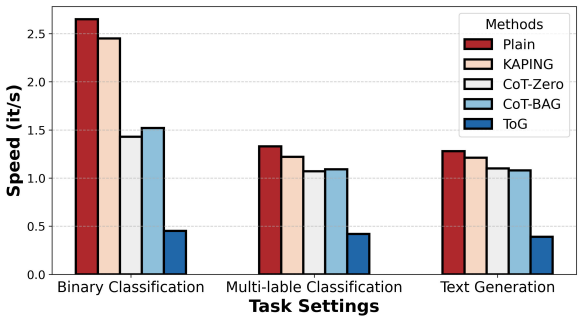

Figure 4: Efficiency analysis of the five baseline methods across three tasks.

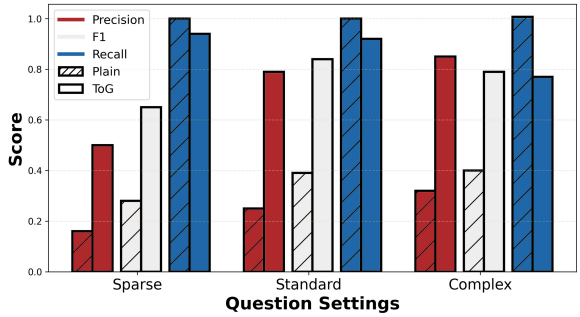

Figure 5: Retrieval quality of ToG vs. Plain across three types of questions on recall, precision and F1.

5.3 Multi-label and Text Generation Task

Table-3,4 (b) and (c) present the performance of baseline models on the multi-label classification (ML) and text generation (TG) tasks. The ML task evaluates models' ability to retrieve nutrition tags associated with foods and user health conditions, while the TG task tests their capacity to generate natural language explanations, offering a more comprehensive and realistic evaluation. The results reveal similar patterns across tasks: while baselines are competent at identifying nutrition tags from the graph, the primary challenge lies in correctly identifying the relevant tags based on user health conditions, as indicated by the overall high recall scores in the ML task.

Both tasks are most challenging on sparse question sets due to their low-resource nature. Conversely, models achieve the best performance on complex question sets, which may appear counterintuitive. However, as shown in Table-2, complex questions have a higher Signal-to-Noise Ratio (SNR), providing models with a clearer signal that offsets their logical complexity. Additionally, the ToG model performs similarly on the standard and complex question sets due to its pruning process, which increases SNR by removing irrelevant nodes. While effective, this process can also discard valuable information, leading to lower performance on complex questions. This trade-off contrasts with ToG's success in binary classification task and highlights the comprehensiveness of our benchmark, which challenges models across diverse scenarios to uncover their strengths and weaknesses.

5.4 Efficiency and Retrieval Quality

Beyond model performance, efficiency is a critical consideration in Graph-RAG systems. To evaluate this, we conduct an efficiency analysis of baseline models on our benchmark, as shown in Figure-4. As can be seen, the binary classification task exhibits the fastest runtime, as it requires the shortest output. In contrast, the multi-label classification and text generation tasks involve longer outputs, leading to slower performance. Due to ToG's reliance on multiple LLM calls during the retrieval process, its runtime is significantly slower compared to other methods. Additionally, the quality of subgraph retrieval plays a crucial role in downstream reasoning. To assess this, we perform a retrieval quality analysis using ToG as a case study, comparing it against a plain Graph-RAG pipeline,

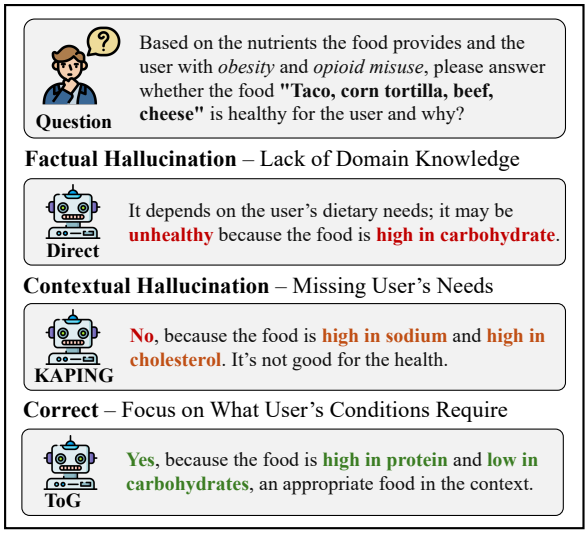

Figure 6: A case study of error analysis.

Backbone	Factual	Contextual
gpt-3.5-turbo	0.308	0.692
gpt-4o-mini	0.155	0.845
llama-3.1-70b-instruct	0.040	0.960

Table 5: Backbone-wise Analysis

as illustrated in Figure-5. As shown, the retrieval scores of ToG align with its performance in the main experiments, confirming our assumption that fluctuations in ToG's performance are rooted in its pruning process during the subgraph retrieval phase.

5.5 Error Analysis

In this section, we analyze the types of hallucinations observed in our experiments both quantitatively and qualitatively. We conduct a detailed analysis of hallucination errors observed in the text generation task, categorizing them into three types: *retrieval errors*, where the retrieved subgraph lacks sufficient information; *contextual hallucinations*, where the model is misled by irrelevant but retrieved content; and *factual hallucinations*, where the model produces incorrect or unsupported claims despite having access to relevant evidence. Our focus is on the relative prevalence of contextual versus factual hallucinations, as retrieval errors are more attributable to retriever design and fall outside the scope of reasoning-focused evaluation. Importantly, we report relative ratios rather than absolute rates, since overall performance trends are already captured in the main experiments.

To investigate reasoning trends, we analyze

Method	Question Setting	Factual	Contextual
CoT_Zero	Sparse	0.180	0.820
CoT_Zero	Standard	0.201	0.799
CoT_Zero	Complex	0.143	0.857
ToG	Sparse	0.246	0.754
ToG	Standard	0.114	0.886
ToG	Complex	0.089	0.911
Plain	Sparse	0.191	0.809
Plain	Standard	0.196	0.804
Plain	Complex	0.147	0.853

Table 6: Method-wise Analysis

Question Setting	Factual	Contextual
Sparse	0.206	0.794
Standard	0.170	0.830
Complex	0.126	0.874

Table 7: Question Setting-wise Analysis

model outputs across all three question settings using three representative methods—Plain, CoT-Zero, and ToG—on three backbone models. We examine hallucination behaviors from three dimensions: model backbone, method design, and question type.

From the backbone-wise perspective (Table-5), GPT-4o-mini consistently exhibits fewer factual hallucinations than GPT-3.5-turbo, suggesting a stronger capacity for grounding its responses in retrieved knowledge. LLaMA-3.1 shows the lowest factual hallucination rate overall, but given its lower task performance, this may indicate over-reliance on the retrieved content without sufficient reasoning flexibility.

Method-wise (Table-6), ToG demonstrates increased factual hallucinations under sparse question settings. This is expected, as its strict pruning can limit necessary context. However, under standard and complex settings, ToG performs better than other methods, reflecting its ability to reason more accurately when supported by sufficient and relevant subgraphs.

And across question settings (Table-7), sparse questions consistently yield the highest factual hallucination rates, while complex questions yield the lowest. This aligns with our earlier discussion on the signal-to-noise ratio (Section 4.1), where sparse settings pose greater challenges in both retrieval and downstream reasoning.

Qualitatively, we observe that hallucinations often emerge when the model fails to appropriately contextualize retrieved evidence or misinterprets domain-specific signals. To provide an intuitive example, as shown in Figure-6, in evaluating whether the food item "Taco, corn tortilla, beef, cheese" suits an obese user recovering from opioid misuse, models may overlook relevant dietary features (e.g., low carbohydrate content in corn tortillas) or improperly emphasize general attributes like sodium content. These errors reflect both factual and contextual hallucinations. They underscore the importance of domain knowledge and graph relevance in guiding model reasoning. In this regard, ToG's focused subgraph retrieval appears to reduce such errors compared to baseline methods, which often include irrelevant or noisy information.

In summary, our findings highlight how hallucination patterns vary based on model architecture, retrieval strategy, and input conditions. These results support the value of our benchmark in revealing reasoning failures and guiding improvements in retrieval-augmented generation systems for complex, health-related domains. Additional case studies are provided in Appendix H.

6 Conclusion

In this work, we introduce the Nutritional Graph Question Answering (NGQA) benchmark, the first dataset designed to address the critical challenges of personalized nutritional health reasoning. By leveraging user-specific medical data and framing the problem as a knowledge graph question answering task, NGQA bridges the gap between general-purpose benchmarks and domain-specific applications. Our benchmark not only advances the scope of GraphQA research by incorporating complex, real-world nutritional scenarios but also provides a comprehensive resource for evaluating and improving models in this domain. We believe NGQA lays the foundation for future research in personalized diet and health-aware reasoning, fostering innovation in both nutritional health and GraphQA.

7 Acknowledgments

This work was partially supported by the NSF under grants IIS-2321504, IIS-2334193, IIS-2340346, IIS-2217239, CNS-2426514, and CMMI-2146076, ND-IBM Tech Ethics Lab Program and Notre Dame Strategic Framework Research Grant (2025). Any expressed opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the sponsors.

Limitation

In this section, we discuss the limitations of this work and outline directions for future research. First, the benchmark includes a limited number of health conditions, though more are available. For example, osteoporosis suggests a high-calcium diet, a renal diet indicates low protein intake, and high low-density lipoprotein (LDL) levels may call for a low-cholesterol diet. As noted in the paper, we prioritized conditions most prevalent in the United States and most relevant to dietary interventions, but expanding to include additional conditions could enhance coverage and utility. Second, while we focus on the interplay between dietary behaviors and medical conditions, other factors, such as food insecurity, remain unexplored. NHANES offers extensive socioeconomic data, presenting opportunities to extend the benchmark to account for broader determinants of dietary decision-making. Third, for simplicity, complex questions are reduced to binary classification by counting "match" and "contradict" tags. However, real-life dietary decisions require nuanced trade-offs and reasoning that go beyond this approach. More sophisticated evaluation methods could better reflect practical scenarios. Lastly, the benchmark could benefit from additional tasks. For example, the existing graphs support questions like, "What alternative foods could meet a user's dietary preferences and medical needs?" Incorporating such tasks would broaden the benchmark's scope and encourage further innovation. Despite these limitations, this work establishes a robust baseline as a pioneering effort in personalized nutrition reasoning. We defer these challenges to future work, envisioning the benchmark as a foundation for ongoing advancements in this critical domain.

Ethics and Privacy Statement

Safeguarding privacy and adhering to ethical principles are paramount when working with sensitive health-related data. The National Health and Nutrition Examination Survey (NHANES) serves as a benchmark in this regard, strictly complying with confidentiality protocols mandated by public legislation. These robust privacy measures enable us to achieve our research goals while remaining fully aligned with the survey's established guidelines. Notably, the NHANES dataset is anonymized, with personally identifiable information (PII)—such as social security numbers and physical addresses—removed. Despite the absence of PII, the dataset retains its utility for detailed analyses, allowing us to investigate the relationship between users' medical data and health-aware food recommendations as presented in this study. Additionally, in practical applications, the generated recommendations and interpretations are treated as personal medical records, ensuring sustained privacy protection. By adhering to these principles, our research maintains the highest levels of ethical responsibility and data privacy.

References

Ashkan Afshin, Patrick J Sur, Kairsten A Fay, Leslie Cornaby, Giannina Ferrara, Jason S Salama, and Christopher J L Murray. 2019. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*.

FAO/WHO Codex Alimentarius. 1985. Guidelines on nutrition labelling. Accessed: 2024-07-12.

FAO/WHO Codex Alimentarius. 1997. Guidelines for use of nutrition and health claims. Accessed: 2024-07-12.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *ACL*.

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 *RecSys*.

Jon Nicolas Bondevik, Kwabena Ebo Bennin, Önder Babur, and Carsten Ersch. 2024. A systematic review on food recommender systems. *Expert Systems with Applications*.

Yuxuan Cao, Jiarong Xu, Carl Yang, Jiaan Wang, Yunchao Zhang, Chunping Wang, Lei Chen, and Yang Yang. 2023. When to pre-train graph neural networks? from data generation perspective! In *KDD*.

CDC. 2020a. Adult obesity facts.

CDC. 2020b. *Americans Share Hopeful Stories of Recovery From Opioid Use Disorder*. https://www.cdc.gov/rxawareness/pdf/articles/TA-T3D2-English_MatteArticle_Release_508.pdf.

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 *WSDM*.

European Commission. 2006. Eu nutrition & health claims regulation legislation (ec) 1924/2006. Accessed: 2024-07-12.

Carrie Dennett. 2021. Diet's role in opioid recovery. *Today's Dietitian*.

Yujie Fan, Mingxuan Ju, Chuxu Zhang, and Yanfang Ye. 2022. Heterogeneous temporal graph neural network. In *SDM*.

Bahare Fatemi, Quentin Duval, Rohit Girdhar, Michal Drozdzal, and Adriana Romero-Soriano. 2023a. Learning to substitute ingredients in recipes. *arXiv*.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2023b. Talk like a graph: Encoding graphs for large language models. *arXiv*.

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*.

Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware food recommender system. In *RecSys*.

Andrea Grillo, Lucia Salvi, Paolo Coruzzi, Paolo Salvi, and Gianfranco Parati. 2019. Sodium intake and hypertension. *Nutrients*.

Ja K Gu, Penelope Allison, Alexis Grimes Trotter, Luenda E Charles, Claudia C Ma, Matthew Groenewold, Michael E Andrew, and Sara E Luckhaupt. 2022. Prevalence of self-reported prescription opioid use and illicit drug use among us adults: Nhanes 2005–2016. *Journal of occupational and environmental medicine*.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph. *Complex & Intelligent Systems*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *NeurIPS*.

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 *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv*.

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. In *EMNLP*.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. In *EMNLP*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeuralIPS*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot

prompting for open-domain question answering. *arXiv*.

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. *NeuralIPS*.

Diya Li, Mohammed J Zaki, and Ching-hua Chen. 2023. Health-guided recipe recommendation over knowledge graphs. *Journal of Web Semantics*.

Zheyuan Liu, Xiaoxin He, Yijun Tian, and Nitesh V Chawla. 2024. Can we soft prompt llms for graph learning tasks? In *WWW*, pages 481–484.

Tianyi Ma, Yiyue Qian, Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. 2025a. Llm-empowered class imbalanced graph prompt learning for online drug trafficking detection. *arXiv*.

Tianyi Ma, Yiyue Qian, Chuxu Zhang, and Yanfang Ye. 2023. Hypergraph contrastive learning for drug trafficking community detection. In *ICDM*.

Tianyi Ma, Yiyue Qian, Shinan Zhang, Chuxu Zhang, and Yanfang Ye. 2025b. Adaptive expansion for hypergraph learning. *arXiv preprint arXiv:2502.15564*.

Nadine Mahboub, Rana Rizk, Mirey Karavetian, and Nanne de Vries. 2021. Nutritional status and eating habits of people who use drugs and/or are undergoing treatment for recovery: a narrative review. *Nutrition reviews*.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv*.

Weiqing Min, Chunlin Liu, Leyi Xu, and Shuqiang Jiang. 2022. Applications of knowledge graphs for food science and industry. *Patterns*.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, et al. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv*.

NIDA. 2024. *Opioids*. https://www.drugabuse.gov/drug-topics/opioids.

Yiyue Qian, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2024. Dual-level hypergraph contrastive learning with adaptive temperature enhancement. In *WWW*.

Yiyue Qian, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2025. Adaptive graph enhancement for imbalanced multi-relation graph learning. In *CIKM*.

Khary K Rigg and Gladys E Ibañez. 2010. Motivations for non-medical prescription drug use: A mixed methods analysis. *Journal of Substance Abuse Treatment*.

Andrew Rosenblum, Lisa A Marsch, Herman Joseph, and Russell K Portenoy. 2008. Opioids and the treatment of chronic pain: controversies, current status, and future directions. *Experimental and Clinical Psychopharmacology*.

Oshani Seneviratne, Jonathan Harris, Ching-Hua Chen, and Deborah L McGuinness. 2021. Personal health knowledge graph for clinically relevant diet recommendations. *arXiv*.

Sola S Shirai, Oshani Seneviratne, Minor E Gordon, Ching-Hua Chen, and Deborah L McGuinness. 2021. Identifying ingredient substitutions using a knowledge graph of food. *Frontiers in Artificial Intelligence*.

Andrew Smyth, Martin J O'Donnell, Salim Yusuf, Catherine M Clase, Koon K Teo, Michelle Canavan, Donal N Reddan, and Johannes FE Mann. 2014. Sodium intake and renal outcomes: a systematic review. *American journal of hypertension*.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *EMNLP-IJCNLP*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *ICLR*.

Lauren J. Tanz, Amanda T. Dinwiddie, Christine L. Mattson, Julie O'Donnell, and Nicole L. Davis. 2022. Drug overdose deaths among persons aged 10–19 years - united states, july 2019-december 2021. *Morbidity and Mortality Weekly Report*.

Dhaval Taunk, Lakshya Khanna, Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. 2023. Grapeqa: Graph augmentation and pruning to enhance question-answering. In *WWW*.

Y Tian, C Zhang, Z Guo, C Huang, R Metoyer, and N Chawla. 2022a. Reciperec: A heterogeneous graph learning model for recipe recommendation. In *IJCAI*.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Yihong Ma, Ronald Metoyer, and Nitesh V Chawla. 2022b. Recipe2vec: Multi-modal recipe representation learning with graph neural networks. *arXiv*.

Yijun Tian, Chuxu Zhang, Ronald Metoyer, and Nitesh V Chawla. 2021. Recipe representation learning with networks. In *CIKM*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv*.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2024a. Can language models solve graph problems in natural language? *NeuralIPS*.

Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. 2021. Market2dish: health-aware food recommendation. *TOMM*.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv*.

Zehong Wang, Sidney Liu, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2025a. Can llms convert graphs to text-attributed graphs? In *NAACL*.

Zehong Wang, Zheyuan Zhang, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. 2024b. Gft: Graph foundation model with transferable tree vocabulary. *NeruIPS*.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. 2025b. Learning cross-task generalities across graphs via task-trees. *ICML*.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. 2025c. Neural graph pattern machine. *ICML*.

Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. 2024c. Subgraph pooling: tackling negative transfer on graphs. In *IJCAI*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv*.

WHO. 2021. Healthy diet.

WHO. 2023. Obesity info page of the world health organization.

Yuanbo Xu, Tian Li, Yongjian Yang, Weitong Chen, and Lin Yue. 2024. An adaptive category-aware recommender based on dual knowledge graphs. *Information Processing & Management*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *NAACL*.

Wenbin Yue, Zidong Wang, Jieyu Zhang, and Xiaohui Liu. 2021. An overview of recommendation techniques and their applications in healthcare. *IEEE/CAA Journal of Automatica Sinica*.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *KDD*.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *ACL*.

Lingzi Zhang, Yinan Zhang, Xin Zhou, and Zhiqi Shen. 2024a. Greenrec: A large-scale dataset for green food recommendation. In *WWW*.

Zheyuan Zhang, Zehong Wang, Shifu Hou, Evan Hall, Landon Bachman, Jasmine White, Vincent Galassi, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. 2024b. Diet-odin: A novel framework for opioid misuse detection with interpretable dietary patterns. In *KDD*.

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. 2024c. Mopi-hfrs: A multi-objective personalized health-aware food recommendation system with llm-enhanced interpretation. *arXiv*.

A Additional Related Work

A.1 Prior Works in Nutrition Personalization

With growing awareness of the importance of dietary health, various studies have sought to incorporate health metrics into applications such as food recommendation systems (Tian et al., 2022a,b, 2021). These approaches can be grouped into three primary categories. First, some research emphasizes single indicators like calorie or fat content, as highlighted in works by Ge et al. (Ge et al., 2015) and Shirai et al. (Shirai et al., 2021), though such metrics often fail to represent the multifaceted nature of a balanced diet. Second, simulated health data has been utilized, as demonstrated by Wang et al. (Wang et al., 2021), but these methods often diverge from real-world data distributions. Finally, recent studies have applied global health guidelines to develop composite health scores, such as (Bölz et al., 2023; Zhang et al., 2024a). However, foods deemed healthy by general standards can still negatively affect certain individuals (Yue et al., 2021; Zhang et al., 2024c), highlighting the absence of a universal solution. The primary challenge remains the scarcity of accurate user health data, a gap our benchmark uniquely addresses.

A.2 Knowledge Graph Question Answering

Knowledge Graph Question Answering (KGQA) has undergone significant advancements, evolving from early approaches such as semantic parsing and retrieval-based methods. Initial models translated natural language queries into structured formats like SPARQL for execution on knowledge graphs (Sun et al., 2019; Zhang et al., 2022). Many of these methods employed pre-trained models like BERT for query encoding and used frameworks such as GNNs or LSTMs for retrieving entities and subgraphs (Yasunaga et al., 2021; Taunk et al., 2023).

More recent progress integrates large language models (LLMs) to improve both retrieval efficiency and reasoning ability. Approaches like Jiang et al. (Jiang et al., 2023) and Wang et al. (Wang et al., 2023) utilize LLMs to transform queries into formats such as SQL or SPARQL, enhancing retrieval accuracy. Others, such as Kim et al. (Kim et al., 2023) and Gao et al. (Gao et al., 2024), focus on reasoning over retrieved subgraphs or triples, tackling multi-hop reasoning tasks in KGQA. However, most benchmarks in this field are designed for general-purpose datasets and fail to address domain-specific complexities, such as the challenges unique to nutritional health reasoning.

A.3 Graph-Retrieval Augmented Generation

Graph-Retrieval Augmented Generation (Graph-RAG) extends the Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020; Ni et al., 2025) by enriching large language models with structured knowledge retrieval. While traditional RAG retrieves unstructured text, Graph-RAG leverages GNNs to retrieve structured subgraphs encoded as triples, improving reasoning precision and minimizing redundancy (Guo et al., 2024; Wen et al., 2023; Lazaridou et al., 2022; Liu et al., 2024).

Existing Graph-RAG benchmarks primarily evaluate basic graph reasoning tasks, such as shortest paths, node degree, and edge existence (Fatemi et al., 2023b; Wang et al., 2024a, 2025a). Although these benchmarks provide insights into foundational reasoning, they lack domain specificity. Recent work by He et al. (He et al., 2024) introduced benchmarks targeting advanced reasoning in general graph contexts, but domain-specific benchmarks for applications such as nutrition remain underdeveloped. By adapting the principles of Graph-RAG, our work introduces the first benchmark designed to tackle personalized health-aware reasoning, addressing this critical gap in the literature.

A.4 Graph Neural Networks.

GNNs are a class of learning models specifically designed to operate on graph-structured data and have demonstrated substantial success across a variety of domains (Kipf and Welling, 2016; Veličković et al., 2017; Hamilton et al., 2017; Zhang et al., 2019; Fan et al., 2022; Ma et al., 2025a; Qian et al., 2025). They have been effectively applied in social network analysis, recommendation systems, biological interaction networks, and molecular property prediction, among others, by leveraging the relational inductive biases inherent in graph structures. A key strength of GNNs lies in their ability to generalize across graph instances of varying sizes and topologies, enabling their deployment in diverse and dynamic real-world settings (Ma et al., 2023; Qian et al., 2024; Ma et al., 2025b). Moreover, recent efforts have explored the transferability of GNNs across tasks and domains (Wang et al., 2024c; Cao et al., 2023), including pretraining strategies and task-agnostic embeddings, which

draw inspiration from the success of transfer learning in language and vision domains. However, challenges such as oversmoothing, limited expressiveness, and poor scalability remain active research areas. Looking ahead, the field is shifting toward the development of graph foundation models: large-scale, pretrained GNNs designed to capture generalizable structural and semantic patterns across graph corpora (Wang et al., 2024b, 2025c,b). These models aim to provide reusable, adaptable representations and serve as backbones for a wide range of downstream tasks, mirroring the transformative impact of foundation models in NLP and vision.

B Benchmark Details

B.1 Data Source Description

NHANES. National Health and Nutrition Examination Survey (NHANES) is a publicly available dataset collected by the U.S. Centers for Disease Control and Prevention (CDC) to assess the health and nutritional status of the U.S. population through interviews, physical examinations, and laboratory tests. Data is released every two years and encompasses five main categories: Demographics, Dietary Data, Examination Data, Laboratory Data, and Questionnaire Data. These comprehensive datasets provide a wealth of information on health indicators, dietary behaviors, and medical conditions.

FNDDS and WWEIA. The Food and Nutrient Database for Dietary Studies (FNDDS) is a comprehensive resource developed by the U.S. Department of Agriculture (USDA) to facilitate dietary intake analysis by providing detailed nutritional information for foods and beverages consumed in the United States. It serves as the backbone for analyzing dietary recall data collected through the What We Eat in America (WWEIA) program, which is a component of NHANES. WWEIA captures dietary intake data through 24-hour dietary recall interviews, linking reported food and beverage items to their corresponding nutrient profiles in FNDDS. Together, FNDDS and WWEIA enable researchers to study dietary patterns, nutrient intake, and their relationship to health outcomes, making them critical tools for advancing nutrition research and public health policy.

B.2 Dietary Habit Processing Details

Dietary habit data was sourced from various NHANES tables, including the Diet Behavior and

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 8: Nutrient Reference Values (NRV) and thresholds (per 100g of food) used based on the nutritional standards.

Health Indicator	High Threshold	Low Threshold
BMI	30	18.5
Waist Circumference (cm)	102 (88)	-
Blood Pressure (mmHg)	140	90
Osteoporosis	-	-
Blood Urea Nitrogen (mmol/L)	7.1	-
Low-Density Lipoprotein (mmol/L)	3.3	-
Red Blood Cell (million cells/uL)	-	4
Glucose (mmol/L)	7	-
Glycohemoglobin (%)	6.5	-
Hemoglobin (g/dL)	-	13.2 (11.6)

Table 9: Health Indicators with Corresponding High and Low Thresholds. Parentheses indicate sex-specific: male (female) thresholds where applicable.

Consumer Behavior datasets, which capture user-reported behaviors and preferences related to food choices, preparation methods, and consumption patterns. Traditional processing approaches proved insufficient for the complexity and diversity of these features. To address this, a thorough manual review was conducted by a team of four researchers. Key features indicative of dietary habits, such as awareness of healthy eating practices or frequency of consuming processed or frozen foods, were identified and categorized. Users were then grouped into high and low habit categories based on their responses, with the top 10% and bottom 10% assigned corresponding habit tags. For instance, users reporting the highest milk consumption were tagged with "drink lots of milk," while those with minimal consumption were labeled as "drink little or no milk." This process generated 54 distinct dietary habit tags, which were incorporated as nodes in the graph. These habit nodes provide critical insights into user behaviors, enabling a nuanced understanding of the relationship between dietary patterns and health outcomes.

B.3 Full Mappings of Nutrition Tags

In this section, we discuss the overall mapping relationship between health indicators and nutrition. In total, we involve nutrition tags for 16 different nutrients focusing on various health aspects, including 7 for macro-nutrients (calories, carbohydrates, protein, saturated fat, cholesterol, sugar, and dietary fiber) and 9 for micro-nutrients (sodium, potassium, phosphorus, iron, calcium, folic acid, and vitamin C, D, and B12). A detailed table of thresholds can be seen in Table-8. As discussed in the paper, these thresholds are derived from existing standards and legislation, from World Health Organization (WHO), Food Standards Agency (FSA)m EU Nutrition & Health Claims Regulation (Commission, 2006) and the Codex Alimentarius Commission (CAC) (Alimentarius, 1985, 1997). An even more detailed standards are listed in Appendix-I. Following the similar practice, we also extract the thresholds for health conditions, as shown in Table-9, Since we have the thresholds for both nutrition and health, we demonstrate the full mapping relationship can be seen in Table-10. Note that the special diet data can be retrieved from NHANES data, which directly indicates a user needs certain nutrients.

However, as we emphasize in the paper, the interactions between nutrition and health are complex and multi-facet. To maintain scientific rigor and practical relevance, we focus on annotating four prevalent health statues, of which diet has been proved to be beneficial for intervention. Their mapping to nutrition tags can be seen in Table-11. The definition of these major health statues are discussed in the next section.

B.4 The Definition of Health Conditions

In the paper, we focus on annotating the four prevalent health statuses—obesity, hypertension, opioid misuse, and diabetes—that are directly influenced by dietary interventions. Among them, WHO and American Heart Association (AHA) provide clear and well-known definitions for obesity and hypertension. We mark a user obesity if the BMI is 30 or greater, and we mark a user hypertension if the average of 4 test of systolic pressure is 140 mm Hg or higher or diastolic pressure is 90 mm Hg or higher. This is classified as stage-2 hypertension and require medical control. For Diabetes, NHANES provides specific questionnaire for diabetic users, and we also mark a user diabetic if the user's Glucose (mmol/L) level is over 7.0 AND Glycohemoglobin (%) is over 6.5.

Opioid misuse, on the other hand, is a tricky health condition to be defined. However we argue this health condition is of vital importance, as the opioid crisis has been one of the most critical society concerns in the United States. Opioids are a category of drugs that include the illegal substance heroin, synthetic opioids such as fentanyl, and prescription painkillers like oxycodone (NIDA, 2024). While primarily used for pain management, opioids can induce euphoria, making them prone to misuse (Dennett, 2021; Rigg and Ibañez, 2010; Rosenblum et al., 2008). For instance, in 2019, 10.1 million Americans reported opioid misuse, and in 2021, there were an estimated 108,000 drug overdose deaths in the United States, 90% of which were linked to opioids (CDC, 2020b; Tanz et al., 2022). In this work, we follow prior work (Zhang et al., 2024b) to define misuse by the following criteria: (1) records of illicit opioid drug use, like heroin, within a year, or (2) records of prescription opioid medication use for over 90 days, which is a threshold commonly employed in the medical domain (Gu et al., 2022).

NHANES dataset provides illicit drug usage data, and we can track down the opioid prescription medicine usage data using the Multum Lexicon Therapeutic Classification Scheme, a 3-level nested category system that assigns a therapeutic classification to each drug and each ingredient of the drug. Category codes used to identify prescription opioid use were: Level 1: 57 = central nervous system agents; Level 2: 58 = Analgesics; Level 3: 60 = narcotic analgesics, or 191 = narcotic analgesics combinations (Detail in Appendix-I).

B.5 Definitions of Ground Truth

In this section, we outline how ground truths are determined for each task. For the multi-label classification task, the process is straightforward. As discussed earlier, nutrition tags are created and linked to users' health conditions based on predefined standards. The ground truths for this task are simply the lists of nutrition tags relevant to each user's health profile.

For the binary classification task, we use the relationship between the user's condition and the food's nutrition tags. A "Yes" label is assigned if the relationship is a "match," and "No" is assigned if the relationship is a "contradict." In the case of complex question settings, where multiple

Nutrient Category	Tag Name	Source Health Indicators
Macro-nutrients	High Calories	Low BMI; Low waist circumference; Weight gain/Muscle building diet
	Low Calories	High BMI; High waist circumference; Weight loss diet
	Low Carb	Low carbohydrate diet; High BMI; High waist circumference
	High Protein	Opioid misuse; Weight gain/Muscle building diet; High protein diet
	Low Protein	High blood urea nitrogen; Renal/Kidney diet
	Low Saturated Fat	High low-density lipoprotein; Low fat/Low cholesterol diet
	Low Sugar	Opioid misuse; Diabetic Diet; Low sugar Diet
	Low Cholesterol	High low-density lipoprotein; Low fat/Low cholesterol diet
	High Fiber	High low-density lipoprotein; Opioid misuse; Diabetic Diet
Micro-nutrients	Low Sodium	High blood pressure; Renal/Kidney diet; Low salt diet
	High Potassium	High blood pressure
	Low Phosphorus	Renal/Kidney diet
	High Iron	Low red blood cell/Low hemoglobin
	High Calcium	Osteoporosis/brittle bones
	High Folic Acid	Low red blood cell count
	High Vitamin C	Low red blood cell/Low hemoglobin; Osteoporosis/brittle bones
	High Vitamin D	Osteoporosis/brittle bones
	High Vitamin B12	Low red blood cell count

Table 10: Nutrient Categories, Tag Names, and Associated Source Health Indicators. Nutrient categories are organized to consolidate related tags and their respective health indicators for clarity.

Health Indicator	Associated Tags
Obesity	Low Calorie
Opioid Misuse	High Protein; Low Sugar; Low Sodium
Hypertension	Low Sodium
Diabetes	Low Sugar; Low Carb
Weight Loss/Low Calorie Diet	Low Calorie
Low Fat/Low Cholesterol Diet	Low Cholesterol; Low Saturated Fat
Low Salt/Low Sodium Diet	Low Sodium
Sugar-Free/Low Sugar Diet	Low Sugar
Diabetic Diet	Low Sugar; Low Carb
Weight Gain/Muscle Building Diet	High Calorie; High Protein
Low Carbohydrate Diet	Low Carb
High Protein Diet	High Protein
Renal/Kidney Diet	Low Protein

Table 11: Health Indicators and Their Associated Nutritional Tags. Each indicator is linked to relevant tags reflecting dietary requirements.

"match" and "contradict" links exist, we calculate the count of each. A question is marked as "Yes" if the number of "match" links exceeds the number of "contradict" links.

For the text generation task, we generate reference texts using a combined approach. First, the overall healthiness of the food is determined using the binary classification result ("Yes" or "No"). This is followed by a natural language explanation that lists the relevant nutrition tags. For example, a reference text might read: "Yes, because the food is low in calories and high in protein." This method ensures that the reference text provides a clear and natural explanation for the decision.

C Implementation Details

In this section, we discuss the implementation details of the baseline models. Specially how we set the hyper-parameters and how we make adaption to our task. All codes all provided in the codebase mentioned in the abstract.

Plain refers to a naive GraphRAG pipeline. Unlike approaches that directly input natural language text or tabular data, we transform the user and food information from the knowledge graph structure into multiple triples, each consisting of an entity, a relationship, and another entity, then concatenate them before feeding into the LLMs.

KAPING answers questions based on a subgraph composed of the entities mentioned in the query and their neighboring nodes. Following the methodology described in the original paper, we first extract the entities present in the query—specifically the user and food—from the provided knowledge graph. Then, we include their respective neighboring nodes to construct a subgraph via retrieval. This subgraph is subsequently transformed into triples and concatenated before feeding into the LLMs. Note that in the original implementation, the authors also used top-k filtering to prune the retrieval results. However, since we don't have any other entities in the question, this pruning based on embedding similarities with the question doesn't generate any reasonable results. We skip this step in our implementation.

CoT-Zero is a two-stage prompting stategy. In the first stage, "Let's think step by step" is appended after the question to guide the model towards producing a reasoning path. In the second

Question Level	Method	a) Binary Classification (-B)				b) Multi-label Classification (-ML)				c) Text Generation (-TG)				
		Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Sparse	Plain	0.536	0.0384	0.957	0.0739	0.197	0.810	0.272	0.377	**0.457**	**0.381**	**0.456**	**0.214**	**0.920**
	KAPING	0.537	0.0399	**0.959**	0.0766	0.196	0.812	0.271	0.377	0.457	0.380	0.455	0.214	0.920
	CoT-Zero	0.532	0.0301	0.954	0.0583	0.254	0.827	**0.393**	0.466	0.435	0.358	0.433	0.199	0.873
	CoT-BAG	0.589	0.298	0.661	0.411	**0.270**	**0.872**	0.352	**0.469**	0.450	0.377	0.449	0.212	0.878
	ToG	**0.634**	**0.403**	0.711	**0.514**	0.210	0.705	0.249	0.356	0.448	0.344	0.443	0.194	0.907
Standard	Plain	0.527	0.105	**1.000**	0.191	0.460	0.821	0.539	0.622	0.626	0.518	0.607	0.361	0.938
	KAPING	0.525	0.101	**1.000**	0.183	0.461	0.821	0.540	0.623	0.627	0.519	**0.608**	**0.362**	**0.939**
	CoT-Zero	0.492	0.0391	**1.000**	0.0753	0.528	0.843	0.622	0.688	0.585	**0.571**	0.475	0.321	0.912
	CoT-BAG	0.595	0.310	0.805	0.448	**0.565**	**0.858**	**0.622**	**0.707**	0.615	0.513	0.597	0.350	0.918
	ToG	**0.839**	**0.763**	0.918	**0.833**	0.515	0.761	0.577	0.638	**0.630**	0.506	0.598	0.353	0.928
Complex	Plain	0.663	0.0799	0.891	0.147	0.599	**0.792**	0.751	0.748	0.664	**0.573**	0.643	**0.395**	**0.940**
	KAPING	0.665	0.0865	0.883	0.158	0.600	0.788	0.752	0.746	**0.664**	0.571	**0.645**	0.393	0.940
	CoT-Zero	0.647	0.0277	**0.944**	0.0539	**0.635**	0.783	**0.807**	**0.776**	0.630	0.534	0.615	0.357	0.918
	CoT-BAG	0.656	0.219	0.565	0.315	0.630	0.769	0.800	0.771	0.651	0.562	0.632	0.383	0.922
	ToG	**0.771**	**0.773**	0.657	**0.710**	0.522	0.616	0.753	0.641	0.630	0.511	0.598	0.350	0.927

Table 12: Experimental results based on five baseline methods on the three tasks with the three question levels using the GPT-3.5-turbo. The best performance of each group is bolded.

stage, the reasoning path is fed to the model to extract the final answer. However, our initial experiments showed that we can combine these two steps, by having both "Let's think step by step" and final output requirements in one prompt, while still achieving the same performance. This allows us to save computational and API resources, avoiding potential inconsistencies and information loss that arise when feeding the reasoning output into a second step. This is because with the one-step approach, the model can make a final decision based on both the original graph, and its own reasoning path, whereas in the second-step approach, the original graph is not available to the model.

CoT-BAG is designed to improve the graph reasoning capabilities of LLMs by first encouraing the model to "build" an implicit graph representation of the problem, and then using chain-of-thought reasoning to solve it. For this approach, a single prompt is sufficient to guide the model through both the graph construction and reasoning, by combining both "Let's construct a graph from the given nodes and edges" and "Let's think step by step to arrive at the final answer". Adapting CoT-BaG to our benchmark requires creating a textual description of the graph triples, in the following format: "The graph contains an edge between node [source] and node [target] with attribute [relationship], an edge between..." to include in the input prompt, alongside the question, and output requirements.

ToG introduces a strategy that iteratively searches and prunes reasoning paths on a knowledge graph starting from entities mentioned in the query to identify suitable paths. However, the open-source ToG codebase is implemented based on Wikidata and Freebase databases, making it incompatible with private datasets. To evaluate ToG on our benchmark, we reimplemented it following the original methodology. Furthermore, we adapted ToG to better suit the characteristics of our benchmark with the following adjustments: 1). Adjusting the width parameter to 5: ToG's original width parameter is set to 3, which retains three reasoning paths during pruning. However, answering questions in our benchmark sometimes requires more than three reasoning paths. By setting the width parameter to 5, ToG preserves five reasoning paths at each pruning step and generates answers based on these paths. 2). Delaying pruning until the second iteration: In ToG's first iteration, the information gathered is often insufficient to evaluate the importance of each reasoning path. Pruning too early risks discarding paths that may be critical for answering the query. Delaying pruning allows ToG to collect more comprehensive information before making pruning decisions. These modifications ensure that ToG is better aligned with the requirements and complexities of our benchmark, enabling more effective performance evaluation.

D Additional Experiments

To further demonstrate the performance of different LLM backbones on our benchmark, we conducted additional tests using GPT-3.5-Turbo as backbones for various baselines. As shown in Table-12, ToG exhibited a noticeable performance degradation when GPT-3.5-Turbo was used as the backbone, particularly when addressing standard and complex questions. This decline is primarily due to GPT-3.5-Turbo's relatively weaker reasoning abilities, which often lead to the retrieval of suboptimal information. Such information provides minimal support—or even introduces negative impacts—on subsequent answer generation. These two sets of experiments highlight the stringent reasoning requirements imposed by our benchmark on the

Diet Type	Obesity	Hypertension	Opioid Misuse	Diabetes
Weight Loss/Low Calorie Diet	2,253	647	222	267
Low Fat/Low Cholesterol Diet	448	247	76	116
Low Salt/Low Sodium Diet	442	350	86	115
Sugar-Free/Low Sugar Diet	170	89	20	78
Diabetic Diet	692	432	126	647
Weight Gain/Muscle Building Diet	3	20	12	1
Low Carbohydrate Diet	244	69	25	57
High Protein Diet	47	12	9	8
Renal/Kidney Diet	25	24	13	7

Table 13: Adoption of Diet Types Across Health Conditions. Each entry represents the number of users with a specific condition following a corresponding diet type.

Status	# Users
Weight Loss/Low Calorie Diet	4,693
Low Fat/Low Cholesterol Diet	1,196
Low Salt/Low Sodium Diet	1,037
Sugar-Free/Low Sugar Diet	417
Diabetic Diet	1,403
Weight Gain/Muscle Building Diet	274
Low Carbohydrate Diet	489
High Protein Diet	146
Renal/Kidney Diet	59
Obesity	18,271
Hypertension	10,257
Opioid Misuse	2,822
Diabetes	3,837

Table 14: Distribution of Users Across Health Conditions and Special Diets.

tested models.

E Additional Statistics

In addition to the basic statistics provided above, we also provide an in detailed benchmark discussing the user distribution on health conditions and the overlap between the four major conditions and the special diets.

Spanning from 2003 to 2020, the latest available NHANES data includes a total of 95,872 unique users. Table-14 illustrates the distribution of health conditions across this population, highlighting the significant prevalence of obesity (18,271 users) and hypertension (10,257 users). These numbers emphasize the widespread impact of these conditions on public health and underscore the urgent need for dietary interventions. However, the stark contrast between the prevalence of these conditions and the adoption of relevant dietary interventions—such as low-calorie diets (4,693 users) or low-sodium diets (1,037 users)—reveals a significant gap. While conditions like obesity and hypertension demand immediate dietary action, far fewer individuals engage in corresponding interventions. This disparity highlights the critical need for personalized dietary reasoning to encourage healthier eating habits tailored to individual health conditions.

A similar trend emerges in Table-13, which examines the alignment between specific health conditions and diet types. While there is some adoption of relevant dietary actions, such as weight loss diets (2,253 for obesity, 647 for hypertension) and low-sodium diets (442 for obesity, 350 for hypertension), these numbers remain disproportionately low relative to the overall prevalence of these conditions. The gap is even more pronounced for diabetes, where fewer than half of diagnosed individuals (647 users) follow diabetic diets out of 3,837 diagnosed users. Specialized interventions, such as renal/kidney or muscle-building diets, see minimal adoption across all conditions, suggesting a lack of accessibility or awareness for these targeted approaches. These patterns reinforce the need for tailored, actionable dietary recommendations to address the divide between health condition prevalence and effective dietary responses, ensuring broader access to appropriate and impactful interventions.

F Prompt Design

In this section, we will demonstrate our carefully designed prompts for the three task settings and selected baselines. The principle of our prompt design is to let LLMs become familiar with nutri-

Role	Category		Content
System	-		"Act as a nutritionist. Analyze if a given food is healthy to a user following further instructions."
User	Question		"Based on the nutrients the food provides and the user needs, please answer whether the food 'Fish curry with rice' is healthy for the user?"
	Method prompt	Default	"Below is the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information."
		Customized	"…"
	Textualized graph		(Fish curry with rice belongs to Seafood mixed dishes), (Fish curry with rice has fish curry), (Fish curry with rice contains high_sodium)…
	Task prompt	Binary classification	"Your output will strictly be Yes or No with no other words."
		Multi-label classification	"Your output must be strictly formatted as a comma-separated list of nutrients prefixed with "high" or "low", based solely on the provided options: carb, protein, sugar, sodium, cholesterol, saturated_fat, calorie. For example, a valid output would be: high_carb, low_protein, high_sugar. No extra words or deviations are allowed."
		Text generation	"Your output must consist of "Yes" or "No", followed by a list of nutrients addressed with "high" or "low," selected from the following options: carb, protein, sugar, sodium, cholesterol, saturated fat, and calorie. For example, a valid output would be: Yes, because the food is high in carb, low in protein, high in sugar. Ensure the output adheres to this format without any additional words or deviations."

Figure 7: The paradigm of prompt for final output.

Role	Content
System	"Identify the top-<width> reasoning paths that are most likely to lead to the answer for the query. Please respond with the indices of the reasoning paths, starting from 1, and separate them with commas (e.g., 1,2,5). Include nothing else in your response."
User	"The query is <question>, and the reasoning paths are: <reasoning_path_list>. Your selected top-<width> reasoning paths are:"

Figure 8: The prompt used in ToG.

tional domain knowledge while avoiding providing explicit guidance.

When querying LLMs for the final output, the paradigm of our prompt is shown as Figure-7. The system prompt is fixed while the user prompt consists of four flexible parts: *question*, *method prompt*, *textualized graph*, and *task prompt*. The *question* and *task prompt* will be automatically adjusted according to the experiment settings. The *method prompt* can be customized to the methods proposed by the benchmark users, e.g., adding "Let's think step by step." for CoT-Zero and adding "Let's construct a graph from the given nodes and edges" for CoT-BAG. We encourage benchmark users to further explore the potential of method prompts. The *textualized graph* is by default generated by concatenating the triplets in the retrieved knowledge graph. Benchmark users can also customize their own textualization method.

Additionally, the prompt we used to prune the relations and entities when testing ToG is shown in Figure-8.

G Case Study

We present 7 case studies across 3 Tasks (Binary Classification, Multi-label Classification, Text Generation), 3 Question Levels (Sparse, Standard,

Complex) and 5 Baselines (Plain, KAPING, CoT-Zero, CoT-BaG, ToG). This section provides insights into how the prompts are structured across different baselines and the reasoning path behind the LLM's final answer, as detailed in Tables 15-21. The case studies provide critical insights into the strengths and limitations of each baseline, while emphasizing the challenges posed by personalized dietary reasoning, highlighting our benchmark's role in advancing the development of robust, domain-specific AI models for personalized health-aware nutrition reasoning.

H Addtional Error Analysis

Our experiments showed that in the specific task of health-aware nutrition reasoning, LLMs are prone to two main types of errors: contextual hallucination and factual hallucination. To understand these shortcomings, we perform an error analysis focusing on the Text Quality Evaluation task, using 3 methods (KAPING, CoT-Zero, ToG) as a representative setting. We prompt the models to also include the reasonings behind their final answer, which then go through a human review process, revealing 2 types of reasoning failures: Contextual Hallucination and Factual Hallucination. Note that we do not check for KG topology errors, as our KG generation process ensures there are no structural problems in the knowledge base that would affect the model's information retrieval and processing performance. Exemplary demonstrations of these 2 error types are shown in Table-22 and Table-23.

I Standards and Regulation

In this section, we provide the standards and regulations used in this paper and attach their links of original document in footnote. There in general three categories: 1) The FNDDS category code [1] used for filtering food candidates (Figure-9). 2) Nutrition claim regulations from WHO, FSA[2], CAC[3][4], and EU legislation[5]. used for defining nutrition thresholds (Figure-10 and Figure-11) . Note that since there are discrepancies in the regulation. We adopt a stricter measure and make it sure it fits NHANES data. The Vitamins and Minerals high thresholds are calculated from the Daily Nutritional Reference Value (NRV), where CAC defines if a food (per 100g) contains over 15% of NRV, it can claim itself a source of such nutrient. The Codex Alimentarius, or "Food Code" is a collection of standards, guidelines and codes of practice adopted by the Codex Alimentarius Commission. The Commission, also known as CAC, is the central part of the Joint FAO/WHO Food Standards Program and was established by FAO and WHO to protect consumer health and promote fair practices in food trade. 3) The Multum Lexicon Therapeutic Classification Scheme[6], used to define opioid prescription medicines and later mark opioid misuse (Figure-12).

[1] Full documention of FNDDS at here

[2] FSA Guideline

[3] Guidelines on Nutrition Labeling

[4] Guidelines for Use of Nutrition and Health Claims

[5] EU Nutrition & Health Claims Regulation legislation (EC)

[6] Full document of Multum Lexicon Therapeutic Classification Scheme at here

Configurations	Task: Binary Classification Question: Complex Model: GPT-4o-mini Method: Plain
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer whether the food **"Black beans and brown rice"** is healthy for the user? Please answer with yes or no. *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. *[Textualized graph]* (Black beans and brown rice belongs to Bean, pea, legume dishes), (Black beans and brown rice has Black beans, from dried, fat added), (Black beans and brown rice has Rice, brown, cooked, no added fat), (Black beans and brown rice belongs to low_carb), (Black beans and brown rice belongs to low_sugar), (Black beans and brown rice belongs to high_sodium), (Black beans and brown rice belongs to low_protein), (Black beans and brown rice belongs to low_cholesterol), (Black beans and brown rice belongs to low_saturated_fat), (user has Eats lots of fish), (user has Adds little to no salt at table), (user has Eats little to no fast food), (user has Eats few to no ready to eat meals), (user has Takes more supplements), (user has Uses lots of salt in preparation), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats gluten free diet), (user has Eats high protein diet), (user has diabetes), (user has Low carbohydrate diet), (user has High protein diet), (diabetes match low_sugar), (diabetes match low_carb), (Low carbohydrate diet match low_carb), (High protein diet contradict low_protein) *[Note prompt]* Important Note: Your output will strictly be Yes or No with no other words.
True Answer	No
LLM Answer	No
LLM Reasoning	While black beans and brown rice provide some beneficial nutrients, they are low in protein, which contradicts the user's high protein diet requirement. Additionally, the dish is high in sodium, which may not be suitable for someone managing diabetes and following a low carbohydrate and low sugar diet. Therefore, it does not align well with the user's nutritional needs.
Observation	Factual hallucination: Although the model is able to correctly determine that the food is unhealthy to the user because it is low in protein compared to the user's High protein diet requirement, upon inspecting its reasonings, it is found that the model mistakenly deduces that high_sodium food tag is unsuitable for a low carb and low sugar diet, which is not correct.

Table 15: Case Study 1

Configurations	Task: Multi-label Classification Question: Standard Model: Llama-3.1-70b-instruct Method: KAPING
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer what nutrient tags are used to determine whether the food **"Nachos, cheese only"** is healthy or unhealthy for the user? *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. *[Textualized graph]* (Nachos, cheese only belongs to Nachos), (Nachos, cheese only has Cheese sauce), (Nachos, cheese only has Snacks, tortilla chips, plain, white corn, salted), (Nachos, cheese only has Fast foods, nachos, with cheese), (Nachos, cheese only belongs to low_carb), (Nachos, cheese only belongs to low_sugar), (Nachos, cheese only belongs to high_sodium), (Nachos, cheese only belongs to high_calorie), (Nachos, cheese only belongs to low_protein), (Nachos, cheese only belongs to low_cholesterol), (user has Drinks little or no milk), (user has Eats little or no shellfish), (user has Adds little to no salt at table), (user has Drinks Alcohol more than average), (user has Eats many ready to eat meals), (user has Drinks lots of water), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats gluten free diet), (user has obesity), (user has High protein diet), (obesity contradict high_calorie), (High protein diet contradict low_protein) *[Note prompt]* Important Note: Your output must be strictly, with no extra words, separated by comma, a list of nutrients with high or low before the nutrients among these options: carb, protein, sugar, sodium, cholesterol, saturated_fat, calorie. For example, the output is: high_carb, low_protein, high_sugar. You should only include the nutrient tags that connect the food with the user.
True Answer	high_calorie, low_protein
LLM Answer	high_calorie, high_sodium, low_protein
LLM Reasoning	It is stated that "Nachos, cheese only belongs to high_calorie" and "high_sodium." Additionally, it is mentioned that "Nachos, cheese only belongs to low_protein." Given the user's condition of obesity and High protein diet requirements, consuming high-calorie food would contradict their health goal of obesity. Furthermore, a high-protein diet requirement contradicts the fact that "Nachos, cheese only belongs to low_protein". User's habit of adding little to no salt at table contradicts the fact that "Nachos, cheese only belong to high_sodium".
Notes	The model includes information about the user's habit of adding little to no salt at the table and concludes that high_sodium is a food tag that is important in determining the food's healthiness, which is irrelevant information in this case because it is supposed to focus on the user's health and diet only - *Factual Hallucination*.

Table 16: Case Study 2

Configurations	Task: Text Generation Question: Complex Model: GPT-4o-mini Method: CoT-Zero
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer whether the food **"Turkey with gravy"** is healthy for the user? Please answer with a short sentence explaining why. *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. Let's think step by step to determine the healthiness of the food, by extracting the nutritional properties of the food from the given graph, then comparing them to the nutrition requirements of the health status, dietary need and habits of the user. A food is unhealthy only if it has certain properties that are unsuitable to the user's health and diet. Do not be too strict with your criteria, only focus on a few main nutritional tags that strongly indicate its healthiness or unhealthiness to the particular diet or health status the user has. Some nutritional tags might not be as important in determining healthiness. *[Textualized graph]* (Turkey with gravy belongs to Poultry mixed dishes), (Turkey with gravy has Turkey, whole, meat only, cooked, roasted), (Turkey with gravy has Salt, table, iodized), (Turkey with gravy has Gravy, chicken, canned or bottled, ready-to-serve), (Turkey with gravy belongs to low_carb), (Turkey with gravy belongs to low_sugar), (Turkey with gravy belongs to high_sodium), (Turkey with gravy belongs to high_protein), (Turkey with gravy belongs to high_cholesterol), (Turkey with gravy belongs to low_saturated_fat), (user has Eats little or no shellfish), (user has Drinks Alcohol less than average), (user has Eats little to no frozen food), (user has Eats few to no meals outside home), (user has Eats few to no ready to eat meals), (user has Takes few or no supplements), (user has Uses little to no salt in preparation), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats weight loss diet), (user has Eats low fat diet), (user has Eats high fiber diet), (user has opioid_misuse), (user has diabetes), (user has Low fat/Low cholesterol diet), (user has Low salt/Low sodium diet), (user has Diabetic diet), (opioid_misuse match high_protein), (opioid_misuse match low_sugar), (opioid_misuse contradict high_sodium), (diabetes match low_sugar), (diabetes match low_carb), (Low fat/Low cholesterol diet contradict high_cholesterol), (Low fat/Low cholesterol diet match low_saturated_fat), (Low salt/Low sodium diet contradict high_sodium), (Diabetic diet match low_sugar), (Diabetic diet match low_carb) *[Note prompt]* Important Note: Your output must be a Yes or No followed by strictly a list of nutrients with high or low as prefix among these options: carb, protein, sugar, sodium, cholesterol, saturated fat, calorie. For example, the output is: Yes, because the food is high in carb, low in protein, high in sugar.
True Answer	Yes, because the food is low in carb, low in sugar, low in saturated_fat, high in protein.
LLM Answer	No, because the food is high in sodium, high in cholesterol.
LLM Reasoning	Turkey with gravy is high in sodium, which contradicts the user's low salt/low sodium diet. It is also high in cholesterol, which contradicts the user's low fat/low cholesterol diet.
Notes	The model fails to recognize there are far more good properties about this food item than bad ones.

Table 17: Case Study 3

Configurations	Task: Multi-label Classification Question: Sparse Model: GPT-3.5-turbo Method: CoT-BaG
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer whether the food **"Sesame chicken"** is healthy for the user? Please answer with a short sentence explaining why. *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. You will be given the textual description of a directed graph. Let's first construct a graph with the given nodes and edges. Then determine the healthiness of the food by traversing the graph and determining the nutritional properties of the food, then compare them to the health status, dietary need and habits of the user. Do not be too strict with your criteria, only focus on a few main nutritional tags that strongly indicate its healthiness or unhealthiness to the particular diet or health status the user has. Some nutritional tags might not be as important in determining healthiness. *[Textualized graph]* Here is the description of the graph: This is the list of edges: an edge between node "Sesame chicken" and "Stir-fry and soy-based sauce mixtures" with attribute "belongs to", an edge between node "Sesame chicken" and "Restaurant, Chinese, sesame chicken" with attribute "has", an edge between node "Sesame chicken" and "low_carb" with attribute "belongs to", an edge between node "Sesame chicken" and "high_sodium" with attribute "belongs to", an edge between node "Sesame chicken" and "high_calorie" with attribute "belongs to", an edge between node "Sesame chicken" and "high_cholesterol" with attribute "belongs to", an edge between node "user" and "Eats little or no shellfish" with attribute "has", an edge between node "user" and "Eats little or no fish" with attribute "has", an edge between node "user" and "Adds little to no salt at table" with attribute "has", an edge between node "user" and "Heavy cigarette smoker" with attribute "has", an edge between node "user" and "Drinks Alcohol more than average" with attribute "has", an edge between node "user" and "Eats lots of frozen food" with attribute "has", an edge between node "user" and "Eats few to no ready to eat meals" with attribute "has", an edge between node "user" and "Takes few or no supplements" with attribute "has", an edge between node "user" and "Drinks lots of water" with attribute "has", an edge between node "user" and "Eats organic food less often" with attribute "has", an edge between node "user" and "Claims to have a poor diet" with attribute "has", an edge between node "user" and "Ate more food than usual" with attribute "has", an edge between node "user" and "obesity" with attribute "has", an edge between node "user" and "hypertension" with attribute "has", an edge between node "obesity" and "high_calorie" with attribute "contradict", an edge between node "hypertension" and "high_sodium" with attribute "contradict". *[Note prompt]* Important Note: Your output must be a Yes or No followed by strictly a list of nutrients with high or low as prefix among these options: carb, protein, sugar, sodium, cholesterol, saturated fat, calorie. For example, the output is: Yes, because the food is high in carb, low in protein, high in sugar.
True Answer	No, because the food is high in sodium, high in calorie.
LLM Answer	No, because the food is high in sodium, high in calorie, high in cholesterol.
LLM Reasoning	Sesame chicken typically contains a lot of sodium due to the soy-based sauce, high calories from frying, and cholesterol from the meat and sauce. These nutritional properties may not align well with the user's dietary needs, especially if they have hypertension or obesity.
Notes	Although the model is able to identify two important food tags that contribute to the food's unhealthiness, upon inspecting its reasonings, it is unclear whether the model deduces this answer from the graph information, or from its general knowledge of the food's ingredients.

Table 18: Case Study 4

Configurations	Task: Text Generation Question: Standard Model: GPT-4o-mini Method: ToG
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer whether the food **"Vegetable rice soup, canned, prepared with water or ready-to-serve"** is healthy for the user? Please answer with a short sentence explaining why. *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. You will be given the textual description of a directed graph. *[Textualized graph] Note: Bolded triples are those that appear in ToG's retrieved subgraph.* (Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to Soups), (Vegetable rice soup, canned, prepared with water or ready-to-serve has Soup, minestrone, canned, chunky, ready-to-serve), **(Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to low_carb)**, (Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to low_sugar), (Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to high_sodium), (Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to low_protein), **(Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to low_cholesterol)**, **(Vegetable rice soup, canned, prepared with water or ready-to-serve belongs to low_saturated_fat)**, (user has Drinks lots of milk), (user has Eats little or no shellfish), (user has Eats little or no fish), (user has Adds little to no salt at table), (user has Eats few to no meals outside home), (user has Eats few to no ready to eat meals), (user has Takes more supplements), (user has Claims to have a good diet), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats weight loss diet), (user has Eats high protein diet), **(user has Low fat/Low cholesterol diet)**, **(user has Low carbohydrate diet)**, **(Low fat/Low cholesterol diet match low_cholesterol)**, **(Low fat/Low cholesterol diet match low_saturated_fat)**, **(Low carbohydrate diet match low_carb)** *[Note prompt]* Important Note: Your output must be a Yes or No followed by strictly a list of nutrients with high or low as prefix among these options: carb, protein, sugar, sodium, cholesterol, saturated fat, calorie. For example, the output is: Yes, because the food is high in carb, low in protein, high in sugar.
True Answer	Yes, because the food is low in carb, low in cholesterol, low in saturated_fat.
LLM Answer	Yes, because the food is low in carb, low in cholesterol, low in saturated fat.
LLM Reasoning	The user has a Low fat/Low cholesterol diet and a Low carbohydrate diet. The given food, Vegetable rice soup, matches the user's dietary requirements as it belongs to low_cholesterol, low_saturated_fat, and low_carb categories. Therefore, the food is suitable for the user's needs.
Observation	ToG's retrieved subgraph contains all the necessary triples, no more no less, to determine that the food is healthy to the user's diet requirements.

Table 19: Case Study 5

Configurations	Task: Text Generation Question: Complex Model: Llama-3.1-70B-instruct Method: ToG
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer whether the food **"Pie, custard, individual size or tart"** is healthy for the user? Please answer with a short sentence explaining why. *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. You will be given the textual description of a directed graph. *[Textualized graph] Note: Bolded triples are those that appear in ToG's retrieved subgraph.* (Pie, custard, individual size or tart belongs to Cakes and pies), (Pie, custard, individual size or tart has Sugars, granulated), (Pie, custard, individual size or tart has Vanilla extract, imitation, no alcohol), (Pie, custard, individual size or tart has Cornstarch), (Pie, custard, individual size or tart has Egg, whole, raw, fresh), (Pie, custard, individual size or tart has Wheat flour, white, all-purpose, enriched, bleached), (Pie, custard, individual size or tart has Shortening, vegetable, household, composite), (Pie, custard, individual size or tart has Salt, table, iodized), (Pie, custard, individual size or tart has Milk, nonfat, fluid, without added vitamin A and vitamin D (fat free or skim)), (Pie, custard, individual size or tart has Beverages, water, tap, municipal), **(Pie, custard, individual size or tart belongs to low_carb)**, (Pie, custard, individual size or tart belongs to high_sodium), **(Pie, custard, individual size or tart belongs to low_protein)**, **(Pie, custard, individual size or tart belongs to high_cholesterol)**, **(Pie, custard, individual size or tart belongs to low_saturated_fat)**, (user has Drinks lots of milk), (user has Uses tobacco often), (user has Drinks Alcohol more than average), (user has Rarely or never check nutrition labels), (user has Takes few or no supplements), (user has Uses lots of salt in preparation), (user has Claims to have a poor diet), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats weight loss diet), (user has obesity), **(user has opioid_misuse)**, (user has hypertension), **(user has diabetes)**, **(user has Low fat/Low cholesterol diet)**, (obesity need low_calorie), **(opioid_misuse contradict low_protein)**, (opioid_misuse need low_sugar), (opioid_misuse contradict high_sodium), (hypertension contradict high_sodium), (diabetes need low_sugar), **(diabetes match low_carb)**, **(Low fat/Low cholesterol diet contradict high_cholesterol)**, **(Low fat/Low cholesterol diet match low_saturated_fat)** *[Note prompt]* Important Note: Your output must be a Yes or No followed by strictly a list of nutrients with high or low as prefix among these options: carb, protein, sugar, sodium, cholesterol, saturated fat, calorie. For example, the output is: Yes, because the food is high in carb, low in protein, high in sugar.
True Answer	No, because the food is high in sodium, high in cholesterol, low in protein.
LLM Answer	No, because the food is low in carb, low in protein, high in cholesterol.
LLM Reasoning	The user has diabetes, and a low-carb diet is suitable for diabetes management. Therefore, the food being low in carb is a positive aspect. However, the user also has opioid misuse, which contradicts with the food being low in protein. Protein is essential for overall health and recovery. Additionally, the user follows a Low fat/Low cholesterol diet. The food being high in cholesterol contradicts this requirement.
Observation	ToG's retrieved subgraph doesn't contain the important triple "(hypertension contradict high_sodium)", hence it fails to include this information in the reasoning process. Instead, it incorrectly includes the food tag low_carb into the final answer, despite successfully identifying that this is a positive aspect of the food.

Table 20: Case Study 6

Configurations	Task: Multi-label Classification Question: Complex Model: Llama-3.1-70B-instruct Method: ToG
Prompt	*[Question]* Based on the nutrients the food provides and the user needs, please answer what nutrient tags are used to determine whether the food **"Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal)"** is healthy or unhealthy for the user? *[Method prompt]* Below are the extra information you use to answer the question, note that you should not use your general knowledge and the answer is among this information. You will be given the textual description of a directed graph. *[Textualized graph] Note: Bolded triples are those that appear in ToG's retrieved subgraph.* (Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to Pasta mixed dishes, excludes macaroni and cheese), **(Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to low_carb)**, **(Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to low_sugar)**, **(Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to high_sodium)**, (Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to low_protein), **(Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to low_cholesterol)**, (Lasagna with cheese and meat sauce, reduced fat and sodium (diet frozen meal) belongs to low_saturated_fat), (user has Eats lots of shellfish), (user has Eats lots of fish), (user has Eats little to no frozen food), (user has Eats few to no meals outside home), (user has Eats few to no ready to eat meals), (user has Takes few or no supplements), (user has Uses lots of salt in preparation), (user has Claims to have a poor diet), (user has Ate more food than usual), (user has Ate less food than usual), (user has Eats weight loss diet), **(user has obesity)**, (user has opioid_misuse), **(user has hypertension)**, **(user has diabetes)**, **(user has Low fat/Low cholesterol diet)**, (obesity need low_calorie), (opioid_misuse contradict low_protein), (opioid_misuse match low_sugar), (opioid_misuse contradict high_sodium), **(hypertension contradict high_sodium)**, **(diabetes match low_sugar)**, **(diabetes match low_carb)**, **(Low fat/Low cholesterol diet match low_cholesterol)**, (Low fat/Low cholesterol diet match low_saturated_fat) *[Note prompt]* Important Note: Your output must be separated by comma, a list of nutrients with high or low before the nutrients among these options: carb, protein, sugar, sodium, cholesterol, saturated_fat, calorie. For example, the output is: high_carb, low_protein, high_sugar. Provide reasonings for your output.
True Answer	low_carb, low_sugar, high_sodium, low_cholesterol, low_saturated_fat, low_protein
LLM Answer	high_sodium, low_sugar, low_cholesterol, low_carb
LLM Reasoning	- high_sodium: Although the meal is labeled as "reduced sodium", it still belongs to the high_sodium category. This is a concern for the user who has hypertension, as high sodium intake can exacerbate this condition. - low_sugar: The meal belongs to the low_sugar category, which is good for the user who has opioid_misuse. - low_cholesterol: The user has Low fat/Low cholesterol diet, which matches low_cholesterol of the food. - low_carb: The food is low in carb, which is suitable for the user who has diabetes.
Observation	The model is able to correctly identify 4 out of 6 important tags in this case. Although it is missing half of the tags, one good thing to note is that ToG is able to pick out both the matches (low_cholesterol, low_sugar, low_carb) and the contradictions (high_sodium).

Table 21: Case Study 7

Question	Based on the nutrients the food provides and the user needs, please answer whether the food **"Cheese pastry puffs"** is healthy for the user? Please answer with a short sentence explaining why.
User Profile	Eats lots of shellfish, Eats lots of fish, Light cigarette smoker, Drinks alcohol less than average, Eats little to no fast food, Eats little to no frozen food, Often checks nutrition labels, Takes more supplements, Uses little to no salt in preparation, Drinks little water, Eats weight-gain diet, Has hypertension
Food Profile	low_carb, low_sugar, high_sodium, high_calorie, low_protein, high_cholesterol, high_saturated_fat
True Answer	No, because the food is high in sodium.
KAPING	- **Answer:** No, because the food is high in sodium, high in calorie, high in saturated fat, high in cholesterol. - **Reasoning:** Although the user often checks nutrition labels and eats little to no fast food or frozen meals *(factual hallucination - irrelevant information)*, their hypertension contradicts with the high sodium content found in Cheese pastry puffs. Moreover, being a light smoker might increase cardiovascular risks; consuming foods high in calorie, saturated fat, and cholesterol like Cheese pastry puffs will not help. *(contextual hallucination - unaware of user's needs)*
CoT-Zero	- **Answer:** No, because the food is high in sodium, low in protein, high in cholesterol, high in saturated fat, high in calorie. - **Reasoning:** The presence of 'high_sodium' category contradicts with 'hypertension', as excessive sodium can worsen hypertension. High saturated fat, calorie and cholesterol from Cheese and Margarine are also concerns for heart health. Low protein likely won't support muscle maintenance or satisfaction between meals. *(factual hallucination - irrelevant information)*
ToG	- **Answer:** No, because the food is high in sodium. - **Reasoning:** Cheese pastry puffs is not healthy because it is high in sodium, and the user has hypertension.

Table 22: Error Analysis 1

Question	Based on the nutrients the food provides and the user needs, please answer whether the food **"Shepherd's pie with lamb"** is healthy for the user? Please answer with a short sentence explaining why.
User Profile	Drinks lots of milk, Eats little to no shellfish, Eats little to no fish, Drinks alcohol more than average, Eats lots of fast food, Eats few to no ready-to-eat meals, Rarely or never checks nutrition labels, Takes more supplements, Claims to have a poor diet, Has obesity, Has diabetes.
Food Profile	low_carb, low_sugar, high_sodium, low_protein, low_cholesterol, low_saturated_fat
True Answer	Yes, because the food is low in carb, low in sugar.
KAPING	- **Answer:** No, because the food is high in sodium, low in protein. - **Reasoning:** While Shepherd's pie with lamb contains some nutritious ingredients like vegetables and lean meat, its high sodium content can be detrimental for someone with obesity and diabetes, as it may contribute to hypertension and other health issues. Additionally, the low protein content may not be sufficient for satiety and muscle maintenance. *(contextual & factual hallucinations - using external knowledge to make assumptions about user's health & include irrelevant information)*
CoT-Zero	- **Answer:** No, because the food is high in sodium, low in cholesterol, low in saturated fat. - **Reasoning:** Given the user's obesity and diabetes, a diet low in calories, sugar, and carbohydrates is essential. While Shepherd's pie with lamb is low in sugar and carbs, its high sodium content may pose a risk for those with diabetes and obesity, making it less suitable for the user's health needs. *(contextual hallucination)*
ToG	- **Answer:** Yes, because the food is low in sugar, low in carb. - **Reasoning:** Given that the user has diabetes, Shepherd's pie with lamb is healthy because it is low_sugar and low_carb.

Table 23: Error Analysis 2

MIXED DISHES	Code	Description	No. FNDDS Codes
Mixed Dishes – Meat, Poultry, Seafood	3002	Meat mixed dishes	233
	3004	Poultry mixed dishes	133
	3006	Seafood mixed dishes	96
Mixed Dishes – Bean/Vegetable-based	3102	Bean, pea, legume dishes	23
	3104	Vegetable dishes	34
Mixed Dishes – Grain-based	3202	Rice mixed dishes	132
	3204	Pasta mixed dishes, excludes macaroni & cheese	174
	3206	Macaroni and cheese	16
	3208	Turnovers and other grain-based items	36
Mixed Dishes – Asian	3402	Fried rice and lo/chow mein	44
	3404	Stir-fry and soy-based sauce mixtures	70
	3406	Egg rolls, dumplings, sushi	25
Mixed Dishes – Mexican	3502	Burritos and tacos	48
	3504	Nachos	7
	3506	Other Mexican mixed dishes	52
Mixed Dishes – Pizza	3602	Pizza	91
Mixed Dishes – Sandwiches	3702	Burgers	62
	3703	Frankfurter sandwiches	29
	3704	Chicken fillet sandwiches	21
	3706	Egg/breakfast sandwiches	47
	3720	Cheese sandwiches	14
	3722	Peanut butter and jelly sandwiches	22
	3730	Seafood sandwiches	20
	3740	Deli and cured meat sandwiches	63
	3742	Meat and BBQ sandwiches	20
	3744	Vegetable sandwiches/burgers	11
Mixed Dishes - Soups	3804	Soups, broth-based	47
	3806	Soups, cream-based	13
	3808	Ramen and Asian broth-based soups	15

Figure 9: FNDDS Category Code - Mixed Dishes.

Table of conditions for nutrient content claims

COMPONENT	CLAIM	CONDITIONS (not more than)
Energy	Low	40 kcal (170 kJ) per 100 g (solids) or 20 kcal (80 kJ) per 100 ml (liquids)
	Free	4 kcal per 100 ml (liquids)
Fat	Low	3 g per 100 g (solids) 1.5 g per 100 ml (liquids)
	Free	0.5 g per 100 g (solids) or 100 ml (liquids)
Saturated Fat[2]	Low	1.5 g per 100 g (solids) 0.75 g per 100 ml (liquids) and 10% of energy from saturated fat
	Free	0.1 g per 100 g (solids) 0.1 g per 100 ml (liquids)
Cholesterol[2]	Low	0.02 g per 100 g (solids) 0.01 g per 100 ml (liquids)
	Free	0.005 g per 100 g (solids) 0.005 g per 100 ml (liquids) and, for both claims, less than:1.5 g saturated fat per 100 g (solids) 0.75 g saturated fat per 100 ml (liquids) and 10% of energy from saturated fat
Sugars	Free	0.5 g per 100 g (solids) 0.5 g per 100 ml (liquids)
Sodium	Low	0.12 g per 100 g
	Very Low	0.04 g per 100 g
	Free	0.005 g per 100 g
COMPONENT	**CLAIM**	**CONDITIONS (not less than)**
Protein	Source	10% of NRV per 100 g (solids) 5% of NRV per 100 ml (liquids) or 5% of NRV per 100 kcal (12% of NRV per 1 MJ) or 10% of NRV per serving
	High	2 times the values for “source”
Vitamins and Minerals	Source	15% of NRV per 100 g (solids) 7.5% of NRV per100 ml (liquids) or 5% of NRV per 100 kcal (12% of NRV per 1 MJ) or 15% of NRV per serving
	High	2 times the value for “source”
Dietary Fibre	Source	3 g per 100 g[3] or 1.5 g per 100 kcal or 10 % of daily reference value per serving[4]
	High	6 g per 100 g[3] or 3 g per 100 kcal or 20 % of daily reference value per serving[4]

Figure 10: Guidelines for use of nutrition and health claims.

Vitamins	
Vitamin A (µg RAE or RE)	800
Vitamin D (µg)	5 - 15*
Vitamin C (mg)	100
Vitamin K (µg)	60
Vitamin E (mg)	9
Thiamin (mg)	1.2
Riboflavin (mg)	1.2
Niacin (mg NE)	15
Vitamin B6 (mg)	1.3
Folate (µg DFE)	400
Vitamin B12 (µg)	2.4
Pantothenate (mg)	5
Biotin (µg)	30
Minerals	
Calcium (mg)	1 000
Magnesium (mg)	310
Iron (mg)**	14 (15% dietary absorption; Diversified diets, rich in meat fish, poultry, and/or rich in fruit and vegetables) 22 (10% dietary absorption; Diets rich in cereals, roots or tubers, with some meat, fish, poultry and/or containing some fruit and vegetables)
Zinc (mg)**	11 (30% dietary absorption; Mixed diets, and lacto-ovo vegetarian diets that are not based on unrefined cereal grains or high extraction rate (>90%) flours) 14 (22% dietary absorption; Cereal-based diets, with >50% energy intake from cereal grains or legumes and negligible intake of animal protein)
Iodine (µg)	150
Copper (µg)	900
Selenium (µg)	60

Figure 11: Daily nutrition value from Codex Alimentarius.

TC1S1 1	MULTUM THERAPEUT SUB-SUB-CLASS FOR TC1S1	3.0	NUM	92	94
	VALUE				UNWEIGHTED
	-9 NOT ASCERTAINED				2,549
	-1 INAPPLICABLE				111,353
	38 VIRAL VACCINES				40
	59 MISCELLANEOUS ANALGESICS				922
	60 NARCOTIC ANALGESICS				5,447
	61 NONSTEROIDAL ANTI-INFLAMMATORY AGENTS				8,158
	62 SALICYLATES				2,640
	63 ANALGESIC COMBINATIONS				524
	69 BENZODIAZEPINES				2,989
	70 MISC ANXIOLYTICS, SEDATIVES AND HYPNOTICS				3,851
	76 MISCELLANEOUS ANTIDEPRESSANTS				93
	(CONT'D ON NEXT PAGE)				

TC1S1 1	MULTUM THERAPEUT SUB-SUB-CLASS FOR TC1S1	3.0	NUM	92	94
	VALUE				UNWEIGHTED
	(CONT'D FROM PREVIOUS PAGE)				
	77 MISCELLANEOUS ANTIPSYCHOTIC AGENTS				538
	89 ANTICHOLINERGICS/ANTISPASMODICS				459
	102 CONTRACEPTIVES				4,496
	126 METHYLXANTHINES				96
	137 TOPICAL ANTI-INFECTIVES				86
	138 TOPICAL STEROIDS				1,948
	139 TOPICAL ANESTHETICS				327
	140 MISCELLANEOUS TOPICAL AGENTS				200
	141 TOPICAL STEROIDS WITH ANTI-INFECTIVES				114
	143 TOPICAL ACNE AGENTS				686
	144 TOPICAL ANTIPSORIATICS				69
	149 SPERMICIDES				3
	154 LOOP DIURETICS				3,862
	155 POTASSIUM-SPARING DIURETICS				1,012
	156 THIAZIDE DIURETICS				5,996
	157 CARBONIC ANHYDRASE INHIBITORS				21
	159 FIRST GENERATION CEPHALOSPORINS				730
	160 SECOND GENERATION CEPHALOSPORINS				110
	161 THIRD GENERATION CEPHALOSPORINS				224
	163 OPHTHALMIC ANTI-INFECTIVES				553
	164 OPHTHALMIC GLAUCOMA AGENTS				2,215
	165 OPHTHALMIC STEROIDS				348
	166 OPHTHALMIC STEROIDS WITH ANTI-INFECTIVES				148
	167 OPHTHALMIC ANTI-INFLAMMATORY AGENTS				308
	168 OPHTHALMIC LUBRICANTS AND IRRIGATIONS				263
	170 OTIC ANTI-INFECTIVES				62
	171 OTIC STEROIDS WITH ANTI-INFECTIVES				183
	173 HMG-COA REDUCTASE INHIBITORS				21,917
	178 SKELETAL MUSCLE RELAXANTS				3,866
	180 ADRENERGIC BRONCHODILATORS				5,916
	181 BRONCHODILATOR COMBINATIONS				3,058
	182 ANDROGENS AND ANABOLIC STEROIDS				359
	183 ESTROGENS				1,593
	185 PROGESTINS				233
	186 SEX HORMONE COMBINATIONS				194
	191 NARCOTIC ANALGESIC COMBINATIONS				8,625
	193 ANTIMIGRAINE AGENTS				856
	195 5HT3 RECEPTOR ANTAGONISTS				665
	196 PHENOTHIAZINE ANTIEMETICS				782
	197 ANTICHOLINERGIC ANTIEMETICS				729
	198 MISCELLANEOUS ANTIEMETICS				335
	201 BARBITURATE ANTICONVULSANTS				233
	203 BENZODIAZEPINE ANTICONVULSANTS				3,846
	204 MISCELLANEOUS ANTICONVULSANTS				14
	205 ANTICHOLINERGIC ANTIPARKINSON AGENTS				318
	208 SSRI ANTIDEPRESSANTS				10,282
	209 TRICYCLIC ANTIDEPRESSANTS				1,513
	211 PLATELET AGGREGATION INHIBITORS				2,631
	(CONT'D ON NEXT PAGE)				

Figure 12: Multum Lexicon Therapeutic Classification Scheme - Part of Level 3.