



Article

# DIETNERD: A Nutrition Question-Answering System That Summarizes and Evaluates Peer-Reviewed Scientific Articles

Shela Wu <sup>1</sup>, Zubair Yacub <sup>2</sup> and Dennis Shasha <sup>1,\*</sup>

<sup>1</sup> Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA; shela.wu@nyu.edu

<sup>2</sup> Department of Linguistics, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA; zyacub2@illinois.edu

\* Correspondence: shasha@cs.nyu.edu

**Abstract:** DIETNERD is a large language model-based system designed to enhance public health education in diet and nutrition. The system responds to user questions with concise, evidence-based summaries and assesses the quality and potential biases of cited research. This paper describes the system's workflow, back-end implementation, and the prompts used. Accuracy and quality-of-response results are presented based on an automated comparison against systematic surveys and against the responses of similar state-of-the-art systems through human feedback from registered dietitians. DIETNERD is among the highest-evaluated of these systems and is unique in combining safety features with sophisticated source analysis. Thus, DIETNERD could be a tool to bridge the gap between complex scientific literature and public understanding.

**Keywords:** large language models; generative AI; question-answering; nutrition; diet; PubMed



**Citation:** Wu, S.; Yacub, Z.; Shasha, D. DIETNERD: A Nutrition Question-Answering System That Summarizes and Evaluates Peer-Reviewed Scientific Articles. *Appl. Sci.* **2024**, *14*, 9021. <https://doi.org/10.3390/app14199021>

Academic Editor: Andrea Prati

Received: 2 August 2024

Revised: 2 October 2024

Accepted: 3 October 2024

Published: 6 October 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

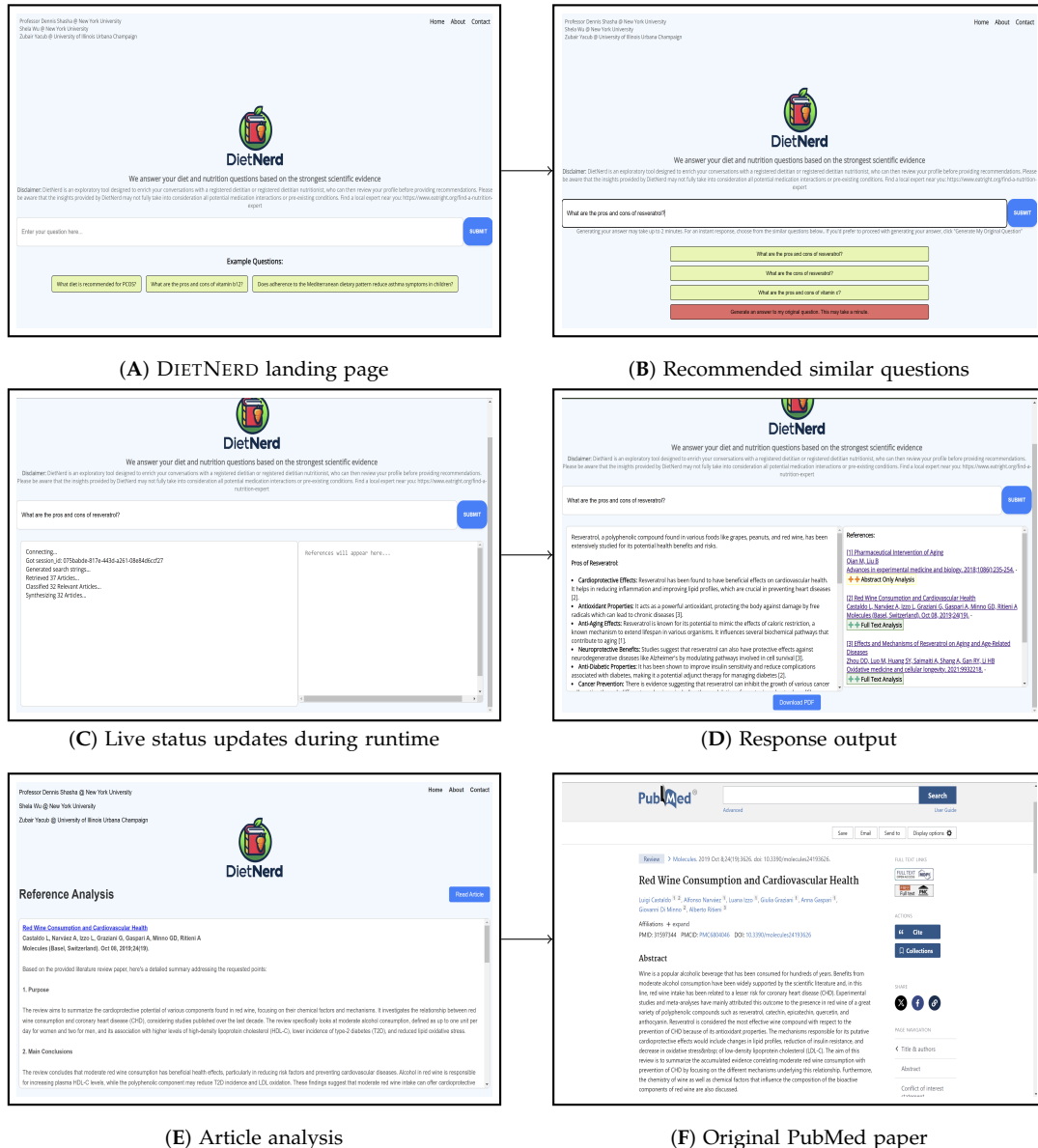
There is growing evidence of the close relationship between diet and health, on both a physical [1] and a mental health [2] level. As a result of that research and similar assertions by the popular media, many non-scientists search for dietary health best practices and craft a diet plan based on their health needs. A 2016 study involving 42,113 French adults showed that 85.1% used the internet for health and nutrition information [3]. The main problem with this approach is that there is a great deal of false or biased information in the media and on the web. That false information also infects large language models (LLMs) such as ChatGPT [4]. Moreover, given the ever-evolving nature of nutrition and dietary research, widely-held beliefs are continuously being upended by new discoveries.

Thus, there is a need for a consumer-friendly question-answering (Q&A) tool that leverages peer-reviewed research, offers individual article analysis, and prioritizes safety. That is what DIETNERD provides. The contributions of this research are:

- A combined parsing/prompt pipeline that takes a question, fetches relevant articles, and summarizes them.
- An analysis of each relevant article that extracts the purpose, the design of the experiment, the main conclusion, risks and benefits of the dietary change tested, study statistics, and any possible conflicts of interest.
- Accuracy tests based on systematic surveys that show a high semantic similarity of the responses of DIETNERD with those surveys for 64% of the articles and moderate semantic similarity for the rest.
- User tests with registered dietitians to assess the clarity, accuracy, and completeness of our summary responses compared to state-of-the-art systems show that DIETNERD ranked among the best in the state of the art.
- DIETNERD is unique among state-of-the-art systems in offering a safety analysis, showing the pros and cons of a diet intervention, based on peer-reviewed research.

### DIETNERD in Action

DIETNERD can be accessed at <https://dietnerd.org/>. When the user enters a question on the landing page, as shown in panel A of Figure 1, the system will suggest the top three most similar questions that DIETNERD has previously answered (panel B) as well as an option to generate an answer to the submitted question. Assuming the user chooses the generate option, the system gives status updates in the course of answering the question (panel C) and then generates a summary with citations (panel D). If the user selects a citation, the system displays an evaluation of the article (panel E). The actual article text can be accessed through the article summary page (panel F).



**Figure 1. User Interface of DietNerd:** User writes a question (panel A). System performs a similarity search and recommends similar questions above a 0.23 similarity threshold (panel B). If the user does not find any questions that are similar enough, the user can ask DIETNERD to generate an answer. The system shows the status of the generation as each execution step happens (panel C). The system presents a scrollable response accompanied by the article analyses and links to the original articles. (panels D–F).

## 2. Related Work

Various LLM-powered medical question-answering tools have emerged in recent years. These applications harness the generative capabilities of LLMs to provide increasingly sophisticated and context-aware interactions. As of the summer of 2024, the systems that are closest to DIETNERD in functionality are Clinfo.ai, Consensus, OpenEvidence, and Scholar GPT. Each of these systems, as characterized in Table 1, are well-regarded and widely popular.

**Table 1. System Functionality Comparisons:** DIETNERD, OpenEvidence and Clinfo.ai are single-turn question-answering tools. Consensus and Scholar GPT offer a conversational feature, while single-turn approaches like DIETNERD would require longer questions to achieve the same functionality. Additionally, DIETNERD and Clinfo.ai offer article summaries that highlight research quality attributes. DIETNERD’s major functional contributions have to do with (i) extracting information from full texts rather than abstracts alone, (ii) a Safety (risk/benefit) Analysis, and (iii) an in-depth article quality analysis that includes an analysis of statistics.

System	Full-Text Extraction (When Available)	Research Quality Analysis	Safety Analysis	Conversational	Multiple Databases Accessed
Clinfo.ai		✓			✓
Consensus		✓		✓	✓
DIETNERD	✓	✓	✓		
OpenEvidence					✓
Scholar GPT	✓			✓	✓

Clinfo.ai is an open source retrieval-augmented LLM developed by a research team at Stanford that answers medical questions using articles from either PubMed or Semantic Scholar, depending on user preference [5]. In addition to developing this tool, the team also developed PubMedRS-200, a medical question-answer dataset of 200 questions built off systematic reviews that can be openly leveraged for performance evaluation [6].

Consensus is a knowledge engine capable of performing various tasks related to scientific literature including article summarization, question answering, and reference searches [7]. Built on the Semantic Scholar database, it is fine-tuned to extract key takeaways from each paper [8]. Consensus employs vector and keyword searches against abstracts and titles while also considering citation count, citation velocity, study design, and publication date to highlight and rank relevant articles in response to user queries. It is available through the OpenAI GPT Marketplace, where it is one of the most used research GPTs, and offers a premium version through a paid subscription on its own in-browser platform.

Scholar GPT, available on the OpenAI GPT Marketplace, was developed by awesomegpts.ai [9]. With access to multiple research databases, Scholar GPT offers versatile capabilities such as scholarly article searches, access to papers and patents, critical reading and text summarization, data analysis, data visualization creation, and web page scraping. As of August 2024, it is the top featured research GPT. It should be noted that Scholar GPT typically requires explicit prompting to include resource hyperlinks when citing sources.

OpenEvidence is a medical search platform founded by Daniel Nadler and maintained by a collective of researchers and medical advisors from Harvard, MIT, Columbia, and Cornell [10]. It is designed to answer health professionals’ questions using research and studies. Relevant articles are displayed alongside icons indicating recent publication or top journal status. OpenEvidence was recently recognized as one of the top 100 AI companies globally [11] and is the first AI to have achieved a score exceeding 90% on the United States Medical Licensing Examination, surpassing performance of Google’s Med-PaLM 2, Anthropic’s Claude 2, ChatGPT, and GPT-4 [12].

Several other innovative systems have appeared as well. StatsPearls Semantic Search [13] and Elicit [14] are very comparable to ClinInfo.ai performance [5]. Similarly, Scholar AI, accessible in the OpenAI GPT Marketplace or as a separately web-hosted tool, presents a well-developed approach to academic search and synthesis. Because its functionality bears resemblance to Scholar GPT and Consensus, both of which are included in our comparative evaluation, to avoid redundancy and maintain a streamlined analysis, we have chosen not to include Scholar AI in our direct comparisons, though we note its significant contributions to the field.

**Enhancing LLM Performance.** Two prominent techniques to enhance LLM performance and adaptability are in-context learning, which enables models to quickly adapt to new tasks based on provided examples, and Retrieval-Augmented Generation (RAG) systems, which combine the generative power of LLMs with external knowledge retrieval for more informed and accurate outputs.

The emergence of in-context learning techniques allow for adaption to new tasks without fine-tuning [15]. Introduced in 2020 with GPT-3, this approach enables LLMs to learn from a few examples provided in the input prompt, effectively performing few-shot learning [16]. In-context learning has shown remarkable versatility across various Natural Language Processing tasks, including question answering, text classification, and even code generation. Its strength lies in its flexibility and efficiency, because it eliminates the need for task-specific model architectures or extensive fine-tuning procedures. The LLMs presented in our work make extensive use of carefully crafted prompts and few-shot examples.

A popular method for developing reliable LLMs are Retrieval-Augmented Generation (RAG) systems. By combining LLMs with external knowledge bases, RAG addresses key limitations of traditional language models. Not only are retrieval-augmented LLMs able to access up-to-date and domain-specific information [17], but they can generate responses that are factually grounded. This was demonstrated in a 2024 study where the addition of a RAG pipeline to the LLM, without embedding or fine-tuning, resulted in an increase of 0.048 in its recall and an increase of 0.103 in its F1 score [18], indicating that the system became better at retrieving relevant and more reliable information. Furthermore, RAG systems offer transparency and interpretability because the sources of information can be traced and verified, which is especially important for machine learning and artificial intelligence applications in healthcare and medicine [19]. Our work follows the spirit of the RAG approach.

**Generating Safe Outputs** With the popularity and increasingly widespread adoption of LLMs, there has been an increased focus on effective safety measures. To illustrate, Google's LaMDA adopts a multi-faceted safety strategy by combining fine-tuning with several additional layers of precaution that involve deploying classifiers to identify unsafe content, integrating a retrieval system for factual accuracy, and applying multi-turn safety checks to assess conversation-wide safety [20]. DeepMind's Sparrow model merges rule-based constraints with information retrieval techniques to govern behavior, encouraging the citation of sources to minimize errors and enhance factual integrity [21]. IBM's Project Debater, while not a traditional LLM, incorporates techniques such as argument mining, stance classification, and claim verification to attempt to promote balanced, safe, and accurate discourse [22]. Meta's BlenderBot introduces a persona-based approach to safety, assigning the model traits that promote safety-conscious interactions [23]. Other notable strategies include Reinforcement Learning from Human Feedback [24], Constitutional AI [25], and adversarial training [26].

While these strategies predominantly rely on fine-tuning, there is a burgeoning interest in harnessing in-context learning for safety applications — an approach which entails careful prompt engineering among other techniques. Although in-context learning methods may not achieve the same level of consistency as fine-tuned models, they allow for the dynamic adaptation of safety measures without the need for retraining, thus offering a more accessible and affordable alternative [16].

DIETNERD makes use of the in-context learning safety framework by chaining together several task-specific LLMs to produce safe responses in resource-constrained environments.

### 3. Materials and Methods

#### 3.1. Algorithm Overview

While the details of the algorithm depend on specific prompts and database accesses, the basic steps are shown diagrammatically in Figure 2. Algorithm 1 shows the method in text, along with references to the corresponding sections for a detailed description of those steps.

---

**Algorithm 1: DIETNERD**  
**Algorithm Text:** Start from a user question  $Q$ , as also illustrated in Figure 2. The condition on previous question-answer pairs requires database access. The de novo generation of an answer entails several prompts as detailed in Sections 3.2.2, 3.2.3, 3.2.5 and 3.2.7–3.2.9 and their corresponding appendices.

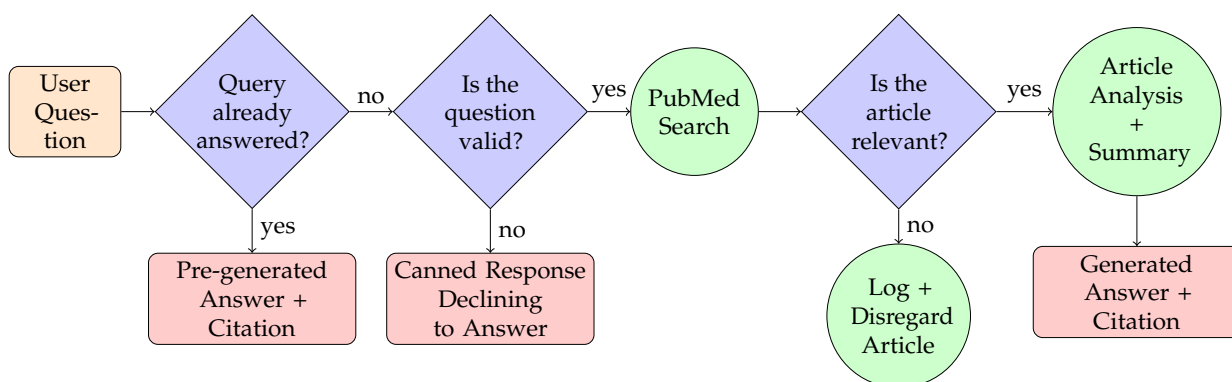
---

```

1  if user determines question  $Q$  is similar to some previously answered question  $Q'$  then
2  |   User selects  $Q'$ 
3  |   System returns the previously generated answer for  $Q'$ 
4  |   else
5  |       if question  $Q$  is valid (Appendix C.1) then
6  |           Parse the user question into keywords (Appendix C.2)
7  |           Perform a search in PubMed using those keywords
8  |           for Each returned article  $A$  do
9  |               if  $A$  is relevant to  $Q$  or includes safety considerations (Appendix C.3) then
10 |                   Summarize the article in a few sentences
11 |                   Perform an article analysis (Appendix C.5)
12 |               end
13 |           end
14 |           Only the strongest articles are used to answer the question
15 |           (Appendix C.6).
16 |       end
17 end

```

---



**Figure 2. DIETNERD Workflow Diagram:** A workflow showing the dynamics of DIETNERD starting with a user question. Purple diamonds are decision points. Green circles are actions. Red squares are outputs. This diagram is expressed in text in Algorithm 1.

#### 3.2. System Design and Architecture

DIETNERD utilizes a chain of LLMs integrated with the PubMed search index and two MySQL databases hosted on Amazon's Relational Database Service in the workflow/decision tree of Figure 2. Our system makes use of both the zero-shot and few-shot

reasoning capabilities of two OpenAI models (namely GPT-3.5-turbo-0125 and GPT-4-turbo) with the following temperature and nucleus sampling settings based on the task. For a more in-depth system architecture diagram, see Appendix D.

**Nucleus Sampling Settings:** For complex tasks that require a nuanced output, primarily text generation, GPT-4-turbo is used with a nucleus sampling of 1 (rich vocabulary). Otherwise, for categorization tasks with a finite set of output responses (e.g., “yes/no”), such as question relevance classification, GPT-3.5-turbo-0125 is used with a nucleus sampling set to 0.5 to restrict the diversity of words considered.

**Temperature Settings:** Tasks that require either a more creative output (e.g., text generation) or a nuanced interpretation of the input data (e.g., determining relevance of an article to a user’s question) have temperature settings greater than or equal to 0.5. Conversely, when the task outputs a finite set of responses or requires a simple copy-paste of the input text, the temperature setting is set close to 0. Specifically, the temperature is set to 0.2 for the question validity task and 0.1 for the full-text relevant section mapping task. See Appendix C for individual LLM implementation details.

DIETNERD’s core LLM functionality was built on the OpenAI library, which facilitates interaction with advanced language models. Bio and metapub were implemented to interface with the PubMed API and retrieve biomedical literature. For information retrieval, versatile web scraping, and full-text extraction, we utilized libraries such as requests, BeautifulSoup, and PyMuPDF. Database connectivity and RAG capabilities were achieved through mysql-connector-python, with environment variables managed using python-dotenv.

### 3.2.1. Question Match Against Question-Answer Database

When a user submits a question, DIETNERD first performs a similarity search against questions in a MySQL database that stores question-answer pairs. The similarity search algorithm begins by retrieving all existing questions from the database. These questions, along with the new user query, are then converted into Term Frequency-Inverse Document Frequency (TF-IDF) vectors using scikit-learn’s TfidfVectorizer. The cosine similarity is then computed between the TF-IDF vector of the user’s question and the vectors of all stored questions. The top three most similar questions above a threshold are selected. As shown in Figure 2, if a match is found, the user has the option to select from the top three closest matches. Alternatively, the user can ask the system to generate an answer based on a scan of the literature.

### 3.2.2. Question Validity

When a user asks the system to generate an answer, DIETNERD first passes the question through the Question Validity LLM, which utilizes few-shot prompts to determine if the user’s question is one that DIETNERD can address. These few-shot prompts are designed to help the LLM discern nuances between acceptable and unacceptable queries. While DIETNERD can recommend general diets that may be suitable for certain health conditions, it does not answer questions around recipe creation, in-depth meal prep, or questions that are asked on behalf of an animal. If a user question is found to be invalid based on these criteria, the output will either be “False—Recipe” or “False—Animal” and a canned response directs the user towards alternate resources. If a user question is valid, DIETNERD will continue to the subsequent downstream tasks. See Appendix C.1 for implementation details.

### 3.2.3. Query Generation for Safe Responses

If the user question passes the validity test, DIETNERD then generates five PubMed queries using two LLMs, each with their own few-shot prompting examples to inform PubMed-specific syntax and stylistic preferences (see Appendix C.2):

- **General Query Generation LLM:** Generates one PubMed query that represents the user’s question and is designed to retrieve articles that provide general context.

- Points of Contention Query Generation LLM:** Generates the remaining four PubMed queries by identifying the top four controversial debates surrounding the topic(s) in question and creating a custom query per point of contention. These queries are intended to surface safety considerations, because controversy in the health domain mostly pertains to safety.

By leveraging a multi-faceted article search strategy, DIETNERD considers a broader spectrum of perspectives, which helps prevent bias towards a single viewpoint. Furthermore, it establishes an additional layer of transparency by highlighting any safety concerns or ethical debates and explicitly acknowledging contentious issues derived from reliable sources.

For example, in Table 2, queries two and three highlight areas where some argue there is insufficient scientific evidence to support certain claims about St. John’s Wort. These queries are particularly valuable as they direct research towards contentious areas where empirical data is critically needed to resolve debates. Meanwhile, queries four and five draw attention to potential risks associated with using St. John’s Wort as a supplement for individuals already taking depression medication. This approach ensures a comprehensive exploration of both the contested efficacy claims and potential dangers of the treatment. These few-shot prompts enable the LLMs to generate targeted and contextually relevant queries without extensive training. All queries are then aggregated into a list.

**Table 2.** For every submitted user question, five PubMed queries are generated—each of which captures a different perspective on the question’s topic. Each point of contention query identified includes a summary explaining the debate. Shown below is an example of the five queries that result from the user question: “Is it safe to take St. John’s Wort if you are already on medication for depression?”.

Query ID	Query Type	Summary	Query
1	General Query	-	(“St. John’s Wort” OR Hypericum) AND (“depression medication” OR antidepressants) AND safety
2	Point of Contention	<p><b>Interaction of St. John’s Wort with Antidepressants</b></p> <p>The most debated argument is around the potential interaction of St. John’s Wort with antidepressants, particularly SSRIs. Some studies suggest a risk of serotonin syndrome, whereas others argue that the interaction might not be clinically significant.</p>	(“St. John’s Wort” OR Hypericum) AND antidepressants AND interaction
3	Point of Contention	<p><b>Efficacy of St. John’s Wort as a Complementary Treatment</b></p> <p>The debate revolves around whether St. John’s Wort can enhance the effects of standard antidepressants and improve treatment outcomes. While some studies suggest a potential benefit, others find no significant effect.</p>	(“St. John’s Wort” OR Hypericum) AND antidepressants AND efficacy

Table 2. Cont.

Query ID	Query Type	Summary	Query
4	Point of Contention	<p><b>Safety of St. John’s Wort in Polypharmacy</b></p> <p>There is a contention about the safety of St. John’s Wort in patients who are on multiple medications. Concerns about potential drug interactions and side effects are discussed.</p>	("St. John’s Wort" OR Hypericum) AND polypharmacy AND safety
5	Point of Contention	<p><b>Impact of St. John’s Wort on Antidepressant Withdrawal</b></p> <p>The debate is whether St. John’s Wort can help manage withdrawal symptoms from antidepressants. Some argue it may ease the process, while others caution against potential risks.</p>	("St. John’s Wort" OR Hypericum) AND antidepressants AND withdrawal

### 3.2.4. Article Retrieval

DIETNERD adopts a methodology akin to a Retrieval-Augmented Generation (RAG) system [17]. The five generated queries are utilized to retrieve relevant information from the PubMed database. This retrieval process is facilitated through Entrez Programming Utilities (E-utilities), an API provided by the National Center for Biotechnology Information (NCBI) that offers access to various NCBI databases, including PubMed [27].

The retrieval mechanism operates in two stages:

**Stage 1:** Using the API’s ESearch utility [28], the system runs each query through PubMed, identifies the ten most relevant articles per query, and stores a list of PubMed IDs (PMIDs). Therefore, for each question, the system can retrieve up to fifty articles.

**Stage 2:** Each of the five PMID lists is fed to the EFetch utility [28], which then retrieves article data in the form of Extensible Markup Language (XML) files.

This approach parallels the retrieval component of RAG systems, which typically access a knowledge base to enhance the contextual understanding and accuracy of language models. In our case, PubMed serves as a specialized knowledge base, providing up-to-date and peer-reviewed information. By integrating this retrieved information with the capabilities of a large language model, DIETNERD aims to generate responses that are both relevant to the user question and scientifically accurate based on the biomedical literature.

### 3.2.5. Relevance Classifier and Safety Considerations

Once all five queries have been processed, the retrieved articles are deduplicated and consolidated into a flattened list. This list is then concurrently processed through the Relevance Classifier LLM (see Appendix C.3) and categorized as either relevant or irrelevant, based on each article’s title, author, and abstract. An article is considered relevant if at least one of the following holds:

- It contains information that is helpful in answering the user’s question.
- It contains a safety aspect that would be important to include in the answer.
- It is NOT an animal-based study.

Again, an important risk-mitigation/safety element is established through the second criteria which qualifies articles as relevant if it includes safety-related information, even if it does not directly answer a user’s question. Doing so ensures that potential risks or safety concerns are not overlooked and that contraindications and warnings are considered. This strategy aligns with the principle of “do no harm” in healthcare by proactively surfacing safety information [29].

### 3.2.6. Article Match against Article Analysis Database

The article processing workflow is managed through two key functions. The first, `connect_to_reliability_analysis_db()`, establishes a connection to a MySQL database



containing previously analyzed articles, retrieving and normalizing this data into a pandas DataFrame. The second function, `article_matching()`, compares newly collected articles against this database. By matching articles through its unique PubMed ID (PMID), the system identifies articles that have already been analyzed and retrieves full records of the article, including previously-generated article analyses. Articles not found in the database are flagged for processing. This approach avoids duplicate work and speeds up the overall process. The functions work in tandem to ensure that only new articles undergo full analysis, while previously examined article analyses are quickly retrieved.

### 3.2.7. Full Text Parser

As noted in Table 1, several state-of-the-art nutrition systems reference only the abstract of a paper. Although an abstract functions well as an overview, it may omit details around design (e.g., bias mitigation tactics) and statistical results. Because DIETNERD builds a comprehensive profile for each article (as shown in Section 3.2.8 and Appendix C.5), we offer a thorough article summary. This summary can be used by both readers and subsequent LLMs to ascertain the level of evidence presented in the paper and thus give more credence to higher quality research.

To start, DIETNERD first checks whether a PubMed Central ID (PMCID) is available. If so, it retrieves the full text from PubMed Central, a free full-text archive [30]. To avoid hitting the context window ceiling, DIETNERD incorporates a Section Mapping LLM (as seen in Appendix C.4) which identifies the section names that closely map to a list of information-rich sections. Based on the attributes to be enumerated in each article analysis, the most relevant sections are: “Abstract”, “Background”, “Methods”, “Results”, “Discussion”, “Conclusion”, “Sources of Funding”, “Conflicts of Interest”, “References”, and “Table”.

In the case that the full text of an article is not available through PubMed Central, DIETNERD parses the additional links provided by PubMed to other journal databases that may freely host the full text. If a list of links is compiled, DIETNERD selects the most preferred of the available sources and captures its hyperlink. Preferred source is determined by referencing a ranked list of sources, where journal databases that are larger and encompass a wider variety of topics are prioritized. Then, depending on the source and provided hyperlink, the system will either:

- (a) Extract a unique article identifier from the hyperlink and concatenate it into URL strings that link to the full-text version of the article.
- (b) Navigate directly to the provided hyperlink.

Sources like Elsevier and Springer follow method (a), while sources such as JAMA Network and Wiley follow method (b). If none of these sources are available or if full-text retrieval fails, the system defaults to using the abstract.

### 3.2.8. Strength of Claim Analysis

After the article text is fetched, DIETNERD begins analyzing the articles (in parallel). For each article, a list of PubMed publication types [31] is extracted from the article’s XML and used to categorize the article as either a “Review” type (e.g., literature reviews, systematic reviews, meta-analyses) or a “Study” type (e.g., randomized controlled trials, observational studies). Depending on the assigned category, DIETNERD selects one out of two prompts and activates the Strength of Claim Analysis, a process led by an LLM tasked with profiling each article by extracting specified attributes and summarizing (see Appendix C.5). As demonstrated in Table 3, although the analyses of Review and Study paper types share core attributes, they diverge in some others. The analysis of “Study”-type papers look at characteristics like study size and statistics. The analysis of “Review”-type papers focus on attributes like selection criteria and a higher-level summary of statistical results (if applicable).

**Table 3. Attributes captured for an article’s analysis based on paper category:** “Study”-type papers have 13 attributes, while “Review”-type papers capture 9 attributes. Each of these listed attributes are specified and defined in the Strength of Claim LLM prompt.

“Study”-Type Papers	“Review”-Type Papers
Purpose & Design	Purpose
Main Conclusions	Main Conclusions
Risks	Risks
Benefits	Benefits
Type of Study	Search Methodology & Scope
Testing Subject	Selection Criteria
Size of Study	Quality Assessment of Included Studies
Length of Study	Synthesis & Analysis
Statistical Analysis of Results	Sources of Funding or Conflicts of Interest
Significance Level	
Confidence Interval	
Effect Size	
Sources of Funding or Conflicts of Interest	

### 3.2.9. Safety-Conscious Output Synthesis

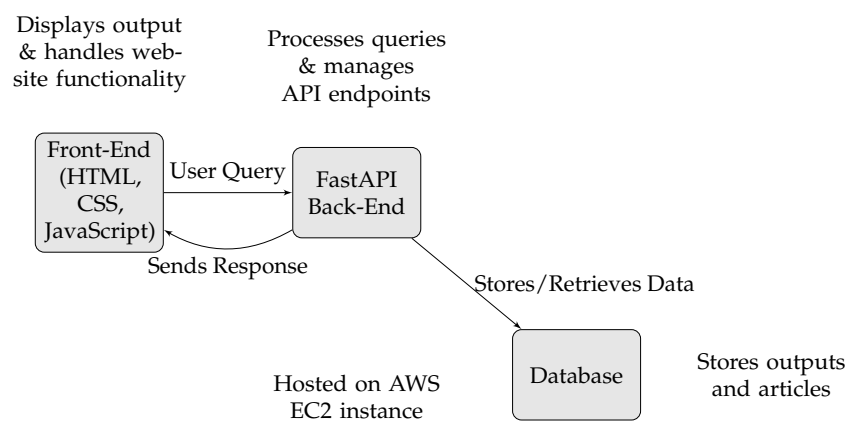
All relevant articles and their analyses are fed into the final LLM along with the original user question. DIETNERD is designed to answer the question using articles with the strongest level of evidence, address both benefits and risks for a balanced response, and be inclusive of as many demographics as possible. Moreover, the LLM is directed to include dosage information when available and explain technical concepts in layman’s terms to make the information accessible to a wider audience.

Multiple safety dimensions are incorporated throughout the prompt. The LLM is tasked with highlighting any risks or dangers associated with the topic. Safeguards against harmful queries are also included. If the user question is dangerous or malicious, DIETNERD will avoid answering and instead curb this behavior by suggesting safer or healthier alternatives. By emphasizing the importance of professional medical consultation, DIETNERD further reinforces responsible information dissemination. Collectively, these elements create a framework that prioritizes user safety, ethical considerations, and risk mitigation while delivering comprehensive responses to medical queries.

A few-shot fine-tuning approach is used to dictate format and tone, as seen through the inclusion of “example\_1\_question” and “example\_2\_question”. See Appendix C.6.

## 4. Website Interface Architecture

The website architecture comprises of a back-end built with FastAPI and a front-end utilizing HTML, CSS, and JavaScript. This section describes the structure and functionality of both components, illustrated in Figure 3.



**Figure 3.** Website architecture diagram showing the interaction between the front-end, FastAPI, back-end, and database components.

#### 4.1. Back-End Structure

The back-end architecture of our system consists of two main components: a Python-based API server using FastAPI and a JavaScript file handling front-end interactions and API calls. This structure allows for a separation of concerns between server-side processing and client-side operations.

The Python back-end is organized as follows:

1. **API Endpoints:** The FastAPI application defines several key endpoints, including:
  - `check_valid`: Validates user queries using the function defined in Section 3.2.2.
  - `process_query`: Initiates the full query processing workflow, using the functions defined in Section 3.2.3 through Section 3.2.9.
  - `sse`: Implements Server-Sent Events (sse) to give the user real-time updates while the generation workflow is running.
  - `db_sim_search`: Performs a similarity search on existing questions, using the algorithm defined in Section 3.2.1.
  - `db_get`: Retrieves stored answers from the database.
2. **Asynchronous Processing:** The back-end utilizes asynchronous programming techniques, specifically Python's `asyncio` library and FastAPI's asynchronous capabilities to handle concurrent requests efficiently.
3. **Background Tasks:** Long-running processes, such as query processing, are handled as background tasks to prevent blocking the main thread.
4. **Database Interactions:** The system interacts with two MySQL databases to store and retrieve processed question-answer pairs and article analyses.
5. **Hosting:** The FastAPI back-end is deployed on an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) `t2.small` instance.

#### 4.2. JavaScript Front-End Logic

The key functions of the JavaScript front-end include:

1. **API Communication:** Handles all API calls to the FastAPI back-end, including query submission, answer retrieval, and similarity search.
2. **Real-time Updates:** Implements an EventSource to receive real-time updates from the server during query processing, using the SSE endpoint defined in Section 4.1.
3. **Response Formatting:** Processes and formats the API responses for display, including handling references and citations.
4. **PDF Generation:** Incorporates functionality to generate PDF reports of the query results using the `jsPDF` library.
5. **User Interface Management:** Manages dynamic User Interface elements such as displaying similar questions and handling user interactions.

#### Algorithm Implementation

The implementation of the DIETNERD algorithm combines real-time processing with database retrievals:

1. **Question Submission:** When a user submits a question, the question is first checked for validity using the `check_valid` endpoint, defined in Section 4.1.
2. **Database Check:** If valid, the system checks if an answer already exists in the database using the `db_get` endpoint, defined in Section 4.1.
3. **Similar Questions:** If no exact match is found, the system performs a similarity search to suggest questions already answered by our system to the user's question using the `sim_search` endpoint in Section 4.1, and offers these questions as options.
4. **Real-time Processing:** If the user requests a new answer for their question:
  - A session ID is generated and returned to the client.
  - The `process_query` endpoint (Section 4.1) is called and the answer generation workflow starts.

- The client establishes an SSE connection to receive updates, using the SSE endpoint in Section 4.1.
  - The FastAPI back-end processes the question asynchronously, sending progress updates to the client.
5. **Result Delivery:** The final answer, along with references and citations, is sent to the client and displayed to the user.

This workflow achieves efficient handling of both new and previously answered questions, providing a responsive user experience while managing server resources effectively.

## 5. Quality Evaluation of DIETNERD

The assessment of DIETNERD comprises two main components:

1. Section 6 describes an automated accuracy evaluation of DIETNERD to assess its agreement with respect to systematic surveys.
2. Section 7 presents a blinded cross-tool comparison experiment involving domain experts and other state-of-the-art systems.

These two analyses allow us to assess DIETNERD's performance from multiple perspectives, combining quantitative metrics with qualitative expert feedback.

## 6. Automated Accuracy Evaluation

The automated evaluation compares the outputs of DIETNERD against a gold-standard benchmarking dataset to systematically assess the semantic similarity of the responses with those of systematic surveys.

### 6.1. Benchmark Dataset Development

We followed the process developed by Clinfo.ai to construct a benchmark dataset using systematic reviews as the gold standard [5]. From the systematic reviews available in PubMed, those meeting the following criteria were selected to compile a dataset of 100 systematic reviews:

- The article title contains a question.
- The question addresses a specific dietary intervention.
- The topic inquires about a clear health outcome.

To construct the benchmark dataset, questions were extracted from the titles of selected systematic reviews. Gold-standard responses were derived from the article abstracts, the results, and conclusion sections. In cases where abstracts lacked explicit section titles, these sections were inferred from the text. Minimal edits were made to the gold-standard responses to avoid undue penalization, primarily spelling out acronyms and removing references to specific study types or statistical results. The publication date of each article was included to ensure that the systematic review itself would not be cited as a resource in the comparative analysis. We could not constrain the search of other systems by date, so this benchmark test measures only the results of DIETNERD.

### 6.2. Generated Responses Dataset

The 100 questions from the benchmark dataset were processed through DIETNERD. Each generated query included a temporal constraint, instructing the system to retrieve only articles published up to one day prior to the listed publication date of the benchmark question. This constraint was implemented to simulate the information landscape just before the target systematic review was published, ensuring a fair comparison. In cases where questions failed to yield a response due to insufficient sources, they were replaced with newly sourced questions to maintain the benchmark dataset's size. Following generation, any acronyms in the responses were spelled out, mirroring the process applied to the benchmark dataset.

### 6.3. Metric Selection

As mentioned, the gold-standard reference text is based on systematic reviews, making it likely to incorporate technical language. DIETNERD, on the other hand, generates text that is meant to be consumer-friendly. Because of this difference in target audiences, the focus in our evaluation is on whether the final conclusions matched semantically rather than syntactically. Therefore, to assess the alignment between DIETNERD's generated text and the gold-standard reference text, two similarity metrics were employed: BERTScore [32] and Semantic Textual Similarity.

BERTScore analyzes textual similarity based on token-level contextual embeddings, focusing on finer lexical details while still taking into consideration semantic similarity. It outputs an F1 score that is rescaled so that values near or below 0 indicate poor similarity and values closer to 1 indicate higher similarity.

Semantic Textual Similarity evaluates the overall similarity in the conclusions reached by both texts by calculating the cosine similarity on a scale of  $-1$  to  $1$ , with  $1$  denoting absolute semantic equivalence. To calculate Semantic Textual Similarity, the model `all-mpnet-base-v2-negation` was used. Built upon the Sentence Transformer base model `all-mpnet-base-v2` [33], this model is fine-tuned on the Compilation of ANnotated, Negation-Oriented Text-pairs (CANNOT) dataset, making it much more sensitive to negations [34]. The Universal Sentence Encoder (USE) Q&A model [35] was considered but, since `all-mpnet-base-v2`'s underlying MPNet architecture incorporates permutation-based training, it potentially offers more versatility and flexibility when dealing with the varied sentence structures and complex language commonly found in medical texts.

While other evaluation metrics are widely used in LLM research, we determined that they were less suitable for our specific objectives. For instance, the Bilingual Evaluation Understudy (BLEU) [36], commonly used in text generation tasks, focuses on n-gram precision, which does not align well with our goal of assessing semantic understanding. BLEU can penalize semantically correct outputs that use different phrasing. Similarly, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [37] emphasizes lexical overlap, which does not capture the full range of acceptable variations in LLM outputs. Even ROUGE-WE (ROUGE with Word Embeddings) [38], an extension of ROUGE designed to better capture semantic similarities, was not a good fit because its use of static word embeddings limits its ability to account for the different meanings a word might have in different contexts. Our chosen metrics allow for greater flexibility in assessing semantically equivalent but lexically diverse responses, reflecting our research objectives.

Table 4 demonstrates the evaluation focus of BERTScore versus Semantic Textual Similarity. In this example, BERTScore's F1 score between the reference and generated texts was low at 0.043 reflecting a token-level difference. By contrast, Semantic Textual Similarity's score rated the texts significantly similar at a 0.788 level. The reason for that high score is that the two texts reach the same essential conclusion: PUFAs were shown to have benefits for some health markers but have no proven impact on others.

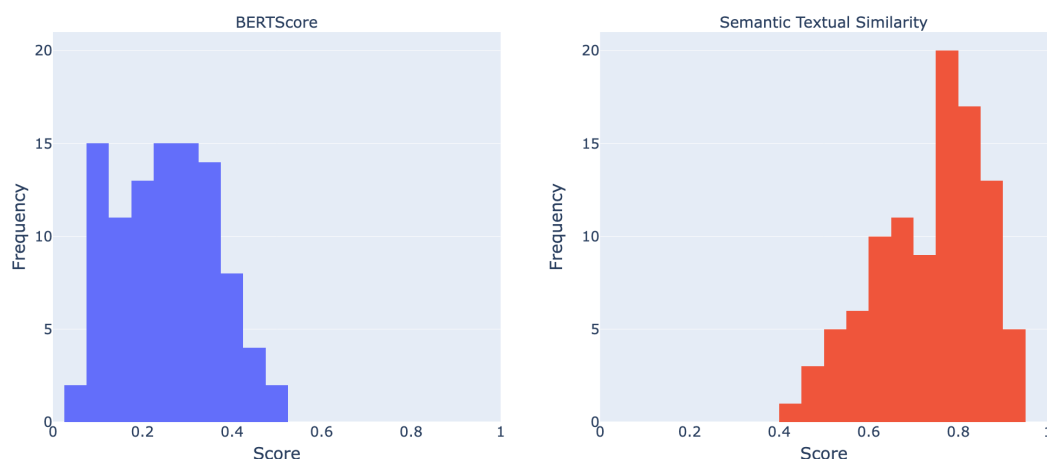
### 6.4. Quantitative Results

The results of our automated evaluation show poor lexical similarity. As seen in Figure 4, BERTScore's F1 scores ranged from  $-0.05$  to  $0.497$  with a mean of  $0.246$ , suggesting moderate token-level alignment. By contrast, Semantic Textual Similarity scores ranged from  $0.439$  to  $0.934$  with a mean of  $0.736$ . The Semantic Textual Similarity is more important for DIETNERD, because our primary goal is to deliver a semantically similar result in layman's terms to those of expert researchers. As noted by [39], similarity scores exceeding  $0.7$  indicate strong similarity, and our results show a concentration of scores in the upper ranges of  $0.75$  to  $0.85$ . Similarity scores above  $0.5$  indicate moderate agreement.

In summary, DIETNERD, when given primary literature, generates responses that are semantically similar to high-quality systematic review medical literature.

**Table 4.** For each of the 100 questions included in the automated evaluation, a gold-standard reference text and DIETNERD’s generated text are compared and analyzed. Below is an example where the BERTScore found a low similarity at the token level (0.043), while Semantic Textual Similarity showed high similarity at a semantic level (0.788). As one can see, both the generated text and the reference text arrive to similar conclusions, supporting the high semantic similarity score. By contrast, the reference text explicitly lists out each factor, while the generated text groups them together (e.g., “cardiovascular risk factors”), resulting in a syntactic difference. The generated text includes a safety-motivated recommendation to consult with a professional, contributing to a (benign) semantic difference between the texts.

Question	Reference Text	Generated Text
Does omega-3 PUFAs supplementation improve metabolic syndrome and related cardiovascular diseases?	Supplementation of omega-3 polyunsaturated fatty acids had a favorable effect on improving triglyceride, total cholesterol, high-density lipoprotein, systolic blood pressure, diastolic blood pressure, interleukin-6, tumor necrosis factor- $\alpha$ , C-reactive protein, and interleukin-1 levels, yet did not affect low-density lipoprotein, monocyte chemoattractant protein-1, intracellular adhesion molecule-1, and soluble endothelial selectin among patients with metabolic syndrome and related cardiovascular diseases.	In conclusion, omega-3 polyunsaturated fatty acids supplementation can be beneficial for reducing triglycerides and improving some cardiovascular risk factors in individuals with metabolic syndrome. However, the effects on cholesterol levels and other metabolic markers can vary. It is advisable for individuals considering omega-3 supplements to consult with healthcare providers to tailor the approach based on personal health needs and risks.



**Figure 4.** A histogram of the results of a 100-question comparative analysis using BERTScore (F1) and Semantic Textual Similarity (cosine similarity). At a token level (reflected in BERTScore), DIETNERD had only moderate success in matching the systematic surveys. At the semantic level (reflected in the Semantic Textual Similarity score), DIETNERD enjoyed much greater success with 64% of the questions showing a similarity score of 0.7 or higher. Scores between 0.5 and 0.7 indicate moderate semantic similarity.

### 7. Domain Expert Cross-Tool Comparison Experiment

We conducted a blinded experiment comparing the five systems listed in Table 1, which includes DIETNERD and the GPT Marketplace version of Consensus, with eight registered dietitians/registered dietitian nutritionists. Feedback and ratings were collected across four questions, each containing five anonymous and randomly ordered responses. Domain experts assessed the quality across five dimensions: relevance, accuracy, completeness, clarity, and overall quality.

For each question, the five system responses were identified only by letters and the order of the responses (and thus the correspondence between letter and system) differed for each question. For example, in question 1, DIETNERD was given letter A and appeared first, while in question 2, DIETNERD was given letter C and appeared third. Thus the authorship

of the response was hidden from the domain experts and the position of each system's response varied from question to question.

### 7.1. Questions and Feedback Criteria

**Question Pool Aggregation and Selection:** The diet questions we used came from questions asked on online forums and from diet/nutrition-focused frequently asked question pages. One of the primary feedback participants (who did not participate in the later evaluations) ranked these questions according to patient relevance. The two highest-ranked questions were selected for feedback, with the remaining two questions chosen randomly.

**Response Generation and Standardization:** For each of the four questions, each system generated a response. Responses for each system were then compiled in random order into a single portable document format (PDF) per question, minimally standardized to preserve system anonymity.

**Feedback Survey:** For each feedback question, a survey form containing five sections—one for each response—was provided. Each section included five questions analyzing a response across four criteria and an overall evaluation:

1. **Relevance:** How relevant was the response to the question(s) asked?
2. **Accuracy:** Was any of the information misleading or factually incorrect?
3. **Completeness:** Did the response miss significant information or context?
4. **Clarity:** Was the response clearly presented?
5. **Overall Quality:** How would you rate the overall quality of this response?

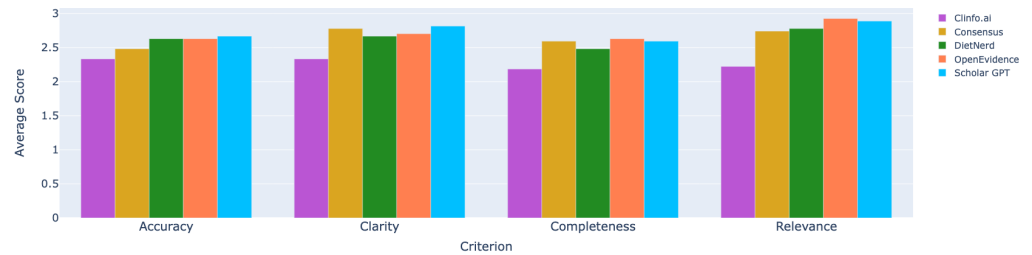
For each value in the "Overall Quality" question, participants were provided the following definitions:

- **Excellent** You would use the tool that generated this response again and would recommend it to others.
- **Pretty Good:** The response was reasonable, but not quite good enough to recommend.
- **Okay:** This response was more or less accurate, but incomplete or unclear.
- **Pretty Bad:** This response had significant problems in either accuracy or clarity.
- **Horrible:** This response was just unacceptable.

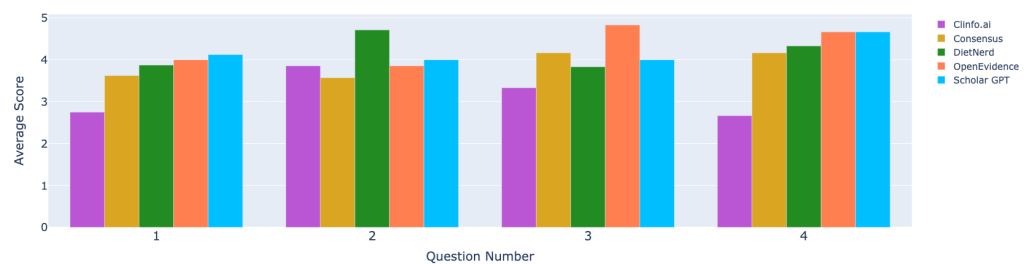
### 7.2. Quantitative Results of Expert Analysis

Using the Overall Quality score provided by participants (Figure 5), we employed a non-parametric paired test [40] between the results of each pair of systems. The series of pairwise comparisons of the five systems revealed several significant differences in performance. Given a significance threshold of 0.05, the analysis indicated that OpenEvidence, Scholar GPT, and DIETNERD performed similarly with no statistically significant differences observed among them. These three systems did however show a statistically significant advantage when compared to Clinfo.ai and, a lesser (and statistically insignificant) advantage over Consensus. Specifically, the comparison between Clinfo.ai and the three systems (DIETNERD, OpenEvidence, and Scholar GPT) resulted in  $p$ -values  $< 0.001$ , indicating significant differences in performance.

When looking at individual quality criteria (Figure 6), a non-parametric statistical analysis suggests a stratification in the performance of the evaluated systems. DIETNERD, Scholar GPT, and OpenEvidence exhibit statistically similar performance levels. Consensus showed some differences in performance compared to OpenEvidence and Scholar GPT, with a trend suggesting potential differences with DIETNERD. Clinfo.ai demonstrated a statistically significant disadvantage when compared to DIETNERD, Scholar GPT, and OpenEvidence.



**Figure 5.** Average Overall Quality Score across the four questions based on user feedback. OpenEvidence, Scholar GPT, and DIETNERD are the three best and are statistically tied.



**Figure 6.** Average of all Relevance, Accuracy, Completeness, and Clarity scores across the four questions based on user feedback. OpenEvidence, Scholar GPT, Consensus, and DIETNERD are the four best and are statistically tied.

## 8. Feedback from Non-Professional Users

Because DIETNERD is meant to be used by subject matter experts and laymen alike, we show here some anecdotal feedback from non-professional users.

- “Wow! Thank you for sending it to me! I have already had advice for my arthritis!!!”
- “The output was pretty consistent with what I’ve learned over my years living with psoriatic arthritis and the various diets I have tried and information I have read. That was a nice summary”!
- “I was really impressed with the results! The information provided was precise, thorough, and covered the issue from various perspectives”.
- “It was a smooth user experience. The similar questions feature was very helpful and provided an immediate informative response. The PDF download is also very convenient”.
- “The answers matched the findings I had come across when researching this topic on my own time—there were even quite a few new things I learned. I especially appreciated the inclusion of dosages and exact numbers. These were things I tried to source but was having trouble easily finding within the studies and research articles I read. Also, having the articles analyses and the full-text vs abstract-only indicator makes the vetting process much easier when I’m trying to filter out which research and journals I can trust”.
- “The responses were indeed helpful. I found my intentions were identified accurately and the responses were quite relevant to my questions. I particularly like the references presented besides the response—at least I get to know the information source and have access to verify the information that DIETNERD provides to me. It could be helpful to include more details of what this tool can provide on the website, like a user manual, and to have a type of “history” section that helps users easily retrieve previous questions and answers when needed. Lastly, it would be very helpful if the system could support follow-up questions so that users may narrow down their questions as they get more responses”.

These anecdotal responses show the system’s usability for a wide variety of users (for instance, the first arthritis user is close to 90). Users praised the relevance of the information provided, the ease-of-use, and appreciated the analysis of the references. Some users



suggested potential improvements, especially the capability of dealing with follow-up questions. We will address this in future work.

## 9. Conclusions

DIETNERD is a system that gives nutritional and diet advice based on peer-reviewed scientific articles. Based on an experiment measuring semantic similarity against systematic surveys, DIETNERD achieves high accuracy. Compared to other state-of-the-art systems, DIETNERD is among the best based on expert (i.e., registered dietitian) reviews. In addition, DIETNERD embeds multiple safety checkpoints within its design and provides detailed individual article analysis functionality, making it unique among sister systems. Last but not least, non-professional users find the system easy to use and informative.

These findings suggest that DIETNERD could help bridge the gap between complex primary literature and consumer-friendly content, thus contributing towards public health education. This has the potential to democratize access to medical knowledge, enabling consumers to better understand and engage with current research findings. Moreover, DIETNERD's ability to aggregate information from multiple studies in a manner that maintains high semantic similarity to systematic reviews suggests it could help consumers gain a more comprehensive and balanced understanding of medical topics, mirroring the rigor of systematic reviews but in a more digestible format. This could enable non-medical professionals to have better informed conversations with healthcare professionals and encourage stronger patient-doctor/dietitian communication as individuals may ask more pertinent questions and better understand medical advice.

The value of tools like DIETNERD is particularly pronounced in the current healthcare landscape, which suffers from shortages of specialized professionals. According to the Bureau of Health Workforce, there is a growing deficit of registered dietitians, a trend that is projected to worsen over time [41]. This shortage has far-reaching implications for patient care and health outcomes. As the gap between supply and demand widens, patients are experiencing progressively longer wait times for appointments, potentially delaying crucial nutritional interventions and guidance. In response to this challenge, some clinics have implemented interim solutions, such as offering consultations with health coaches to address basic questions and provide general wellness advice [42]. However, these stopgap measures may not fully bridge the knowledge gap or provide the depth of information that patients require.

In this context, DIETNERD can play a role in augmenting existing healthcare resources. By providing scientifically grounded, easily digestible information, our tool can answer fundamental questions during the waiting period between appointments. This is not to suggest that DIETNERD should replace professional consultation; rather, it serves as a complementary resource that can enhance the patient experience and preparedness.

Moreover, DIETNERD can support continuity of care by providing consistent, up-to-date information between appointments. This is particularly valuable for patients managing chronic conditions or implementing long-term dietary changes, where ongoing education and reinforcement are crucial. By empowering patients with knowledge, DIETNERD can potentially improve adherence to dietary recommendations and foster a more collaborative relationship between patients and healthcare providers.

## 10. Limitations and Future Work

There are four limitations that we plan to overcome next. First, we would like to build out access to other scientific databases to further enrich our knowledge base and provide a more comprehensive foundation for DIETNERD. The current system in fact already accesses the databases of specific publishers to get full text articles, but this could be more fully built out to widen the pool of available information.

Second, we would like to offer the functionality of recipe creation by leveraging resources such as FoodKG, a food knowledge graph [43], as well as other nutrient databases to generate nutritionally balanced recipes tailored to a user's dietary preferences and restrictions.

Third, some of our non-professional users have requested conversational functionality. While we primarily viewed this as a convenience feature rather than a scientific contribution, it holds potential for an improved user experience.

Fourth, the system could be designed to elicit clarifying information by posing questions to the user. This would help DIETNERD tailor responses to a user's medical profile, for example, by taking into account food allergies and other preexisting conditions.

**Author Contributions:** Conceptualization, D.S., S.W. and Z.Y.; formal analysis, S.W.; investigation, D.S. and S.W.; methodology, S.W.; software, Z.Y. and S.W.; supervision, D.S.; validation, S.W. and Z.Y.; visualization, S.W.; writing—original draft preparation, D.S., S.W. and Z.Y.; writing—review and editing, D.S., S.W. and Z.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was made possible through the support of NYU Wireless.

**Data Availability Statement:** The data presented in this study are openly available in the DietNerd repository at <https://github.com/shela-wu/DietNerd.git>.

**Acknowledgments:** We extend our gratitude to the group of registered dietitians and registered dietitian nutritionists who generously contributed their time and expertise to provide feedback and participate in the comparative analysis against other systems: Jessica Foug, MS, RD; Amelia Ti, MS, RD, CDN, CDCES; Mariana Chagas; Samantha DeVito, MS, RD, CDN; Christine Tseng, MS, RD, CNSC, CDN; Vicky Wang, RD, CDN; Jackie Topol, MS, RD, CDN; and Amy Gionta, MS, RD. We would like to give special recognition to Jessica Foug, MS, RD and Amelia Ti, MS, RD, CDN, CDCES, who were instrumental in providing both early-stage and final-stage feedback, significantly contributing to the refinement and development of our system. We are also grateful to Stephanie Wu, MD, MBA for her support and insights throughout the project.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A. Automated Accuracy Evaluation Results

Included are 10 sample question-answer sets used for our automated accuracy evaluation, followed by a table showing BERTScore and Semantic Textual Similarity evaluation results for all 100 question-answer sets.

### Question 1:

- **Question Text:** Are anti-inflammatory foods associated with a protective effect for cutaneous melanoma?
- **Reference Text:** Food components considered as anti-inflammatory, such as vitamin D, vitamin A, folic acid, niacin, vitamin C, omega-3 fatty acids, and carotenoids ( $\beta$ -carotene, lutein, zeaxanthin, and lycopene), were not associated with a protective effect for melanoma. Other anti-inflammatory food items, such as tea, fruits, and vegetables, except for citrus fruits that were borderline associated with an increased risk, were not associated with cutaneous melanoma. In conclusion, the only anti-inflammatory food item that was consistently associated with a protective effect for cutaneous was coffee in particular caffeinated coffee.
- **System Generated Text:** While these findings are promising, they are primarily based on observational studies, reviews of existing literature, and a limited number of clinical trials. The mechanisms by which these foods and compounds exert their effects are complex and not fully understood. Additionally, the studies vary in their methodologies, populations studied, and the types of anti-inflammatory compounds examined.

### Question 2:

- **Question Text:** Does omega-3 PUFAs supplementation improve metabolic syndrome and related cardiovascular diseases?

- **Reference Text:** Supplementation of omega-3 polyunsaturated fatty acids had a favorable effect on improving triglyceride, total cholesterol, high-density lipoprotein, systolic blood pressure, diastolic blood pressure, interleukin-6, tumor necrosis factor- $\alpha$ , C-reactive protein, and interleukin-1 levels, yet did not affect low-density lipoprotein, monocyte chemoattractant protein-1, intracellular adhesion molecule-1, and soluble endothelial selectin among patients with metabolic syndrome and related cardiovascular diseases.
- **System Generated Text:** In conclusion, omega-3 polyunsaturated fatty acids supplementation can be beneficial for reducing triglycerides and improving some cardiovascular risk factors in individuals with metabolic syndrome. However, the effects on cholesterol levels and other metabolic markers can vary. It is advisable for individuals considering omega-3 supplements to consult with healthcare providers to tailor the approach based on personal health needs and risks.

### Question 3:

- **Question Text:** Is dairy consumption associated with depressive symptoms or disorders in adults?
- **Reference Text:** Findings were mixed, with one study reporting a positive association; five studies reporting no association; and seven studies reporting mixed associations depending on dairy type, gender or population group. We found conflicting and inconsistent associations in studies that were generally of fair quality. Future longitudinal and intervention studies that employ more rigorous dietary assessment methods are warranted.
- **System Generated Text:** There is some evidence suggesting that dairy consumption might be associated with lower depressive symptoms, particularly in specific populations such as those with multiple sclerosis or post-myocardial infarction. More comprehensive studies are required to establish a clear causal relationship between dairy consumption and depressive symptoms in the general adult population. Given the mixed evidence, individuals should consider maintaining a balanced diet and consult healthcare professionals for personalized dietary advice, especially if experiencing depressive symptoms.

### Question 4:

- **Question Text:** Does taking vitamin, mineral and fatty acid supplements prevent cognitive decline?
- **Reference Text:** There was little evidence of a beneficial effect from taking B vitamins or antioxidant supplements on global cognitive function in later life. Larger-scale randomized controlled trials of longer duration in selected age groups are needed.
- **System Generated Text:** The benefits of Vitamin B are potentially lowering homocysteine levels, which might reduce the risk of cognitive decline and dementia. The benefits of polyunsaturated fatty acids are that they can enhance cognitive functions and may counteract the negative effects of poor dietary habits. The benefits of combination therapies are potential synergistic benefits in cognitive function from combinations like estrogen, statins, and polyunsaturated fatty acids. While the benefits are promising, the studies often call for more research to fully understand the effectiveness and safety of long-term supplementation. Dosage and the specific combinations of supplements need to be carefully considered to avoid potential adverse effects. Individuals, particularly those at risk of cognitive decline or with existing mild cognitive impairment, might consider discussing with healthcare providers the possibility of incorporating certain supplements into their regimen. However, it is crucial to do so under professional guidance to tailor the supplement type and dosage to individual health needs and conditions.

**Question 5:**

- **Question Text:** Does ginger supplementation lower blood pressure?
- **Reference Text:** Pooled analysis suggested that ginger supplementation can reduced systolic blood pressure and diastolic blood pressure. When studies were categorized based on participants' mean age, ginger dosage and duration of intervention, systolic blood pressure and diastolic blood pressure were significantly decreased only in the subset of studies with mean age less than or equal to 50 years, follow-up duration of less than or equal to 8 weeks and ginger doses greater than or equal to 3 g per day. Our findings revealed that ginger supplementation has favorable effects on blood pressure. Nonetheless, further studies are warranted before definitive conclusions may be reached.
- **System Generated Text:** In conclusion, while there is some evidence supporting the potential of ginger to aid in blood pressure management, particularly through its vasodilatory properties and possible synergistic effects with anti-hypertensive medications, the evidence is not uniformly strong or conclusive. Further research, particularly large-scale human clinical trials, is needed to better understand the role of ginger in blood pressure regulation and its long-term health impacts.

**Question 6:**

- **Question Text:** Can improved legume varieties optimize iron status in low- and middle-income countries?
- **Reference Text:** Evidence shows that, in addition to repletion of biomarkers of iron status, consumption of improved legumes is associated with both clinical and functional outcomes. The prevalence of iron deficiency decreases with consumption of improved legumes, with increases of less than or equal to 3.0 g/L in Hb concentrations. Improvement in cognition and brain function in women has been reported as well. However, further research is necessary in more at-risk groups and also to show if the reported improvements in status markers translate to improved health outcomes. Evidence from the included studies shows potential from consumption of improved legumes suggesting them to be a sustainable solution to improve iron status.
- **System Generated Text:** Improved varieties of legumes that incorporate higher levels of bioavailable iron and lower inhibitors like phytic acid and polyphenols can significantly contribute to better iron status, particularly in regions where legumes are a dietary staple. Biofortification appears to be a sustainable and effective strategy to combat iron deficiency in low- and middle-income countries, potentially reducing the global disease burden associated with micronutrient deficiencies. While the current evidence is promising, ongoing research and careful implementation are necessary to optimize the benefits of biofortified legumes and ensure they reach the populations most in need.

**Question 7:**

- **Question Text:** Are Atkins and other low-carbohydrate diets a hoax or an effective tool for weight loss?
- **Reference Text:** The apparent paradox that ad-libitum intake of high-fat foods produces weight loss might be due to severe restriction of carbohydrate depleting glycogen stores, leading to excretion of bound water, the ketogenic nature of the diet being appetite suppressing, the high protein-content being highly satiating and reducing spontaneous food intake, or limited food choices leading to decreased energy intake. Long-term studies are needed to measure changes in nutritional status and body composition during the low-carbohydrate diet, and to assess fasting and post-prandial cardiovascular risk factors and adverse effects. Without that information, low-carbohydrate diets cannot be recommended.

- **System Generated Text:** In conclusion, low-carbohydrate diets like the Atkins diet are not a hoax and can be effective for weight loss and improving certain metabolic health markers. However, individuals considering these diets should be aware of the potential challenges related to long-term adherence and the need for further research into their long-term safety and efficacy. As with any diet, it is advisable to consult healthcare professionals before making significant changes to dietary habits.

#### Question 8:

- **Question Text:** Is vitamin d supplementation useful for weight loss programs?
- **Reference Text:** The meta-analyzed mean differences for random effects showed that cholecalciferol supplementation decreases the body mass index by  $-0.32 \text{ kg/m}^2$  and the waist circumference by  $-1.42 \text{ cm}$ , but does not statistically affect weight loss. This analysis lays the foundation for defining the potential clinical efficacy of vitamin D supplementation as a potential therapeutic option for weight loss programs, but further studies are needed to confirm the validity of these findings and delineate potential underlying mechanisms.
- **System Generated Text:** Vitamin D supplementation can be beneficial for improving metabolic markers in obese individuals, which might support weight loss efforts indirectly. Direct effects of vitamin D on weight loss are not strongly supported by current evidence. Obese individuals might require higher doses of vitamin D to achieve adequate serum levels. Combining vitamin D supplementation with other weight loss strategies and consulting healthcare providers for personalized advice is recommended.

#### Question 9:

- **Question Text:** Can vitamin E supplementation affect obesity indices?
- **Reference Text:** There was no significant effect of vitamin E supplementation on weight, body mass index and waist circumference. However, vitamin E supplementation might be associated with increasing body mass index in people with normal body mass index (18.5–24.9).
- **System Generated Text:** In conclusion, while vitamin E may offer benefits for conditions associated with obesity, such as liver health, direct evidence supporting its role in reducing obesity indices is lacking. More research is needed to clarify its effects on obesity itself and to establish appropriate dosages and safety profiles.

#### Question 10:

- **Question Text:** Does milk and dairy consumption during pregnancy influence fetal growth and infant birthweight?
- **Reference Text:** A beneficial fetal growth-increase was most pronounced for increasing maternal milk intake in the lower end of the consumption range. Evidence from prospective cohort studies is limited but suggestive that moderate milk consumption relative to none or very low intake, is positively associated with fetal growth and infant birthweight in healthy, Western populations.
- **System Generated Text:** The reviewed studies collectively suggest that milk and dairy consumption during pregnancy can have a beneficial impact on fetal growth and infant birthweight. While the evidence supports the positive effects of milk intake on increasing birth size and weight, further research is warranted to fully understand the scope and mechanisms of these benefits. Pregnant women should consider incorporating an appropriate amount of milk and dairy products into their diet as part of a balanced nutritional intake, but as always, it is advisable to consult with a healthcare provider or a registered dietitian to tailor dietary choices to individual health needs and conditions.

**Table A1.** BERTScore's F1 scores and Semantic Textual Similarity's cosine similarity scores across all 100 question-answer sets.

Question ID	BERTScore	Semantic Textual Similarity Score
1	-0.05000561476	0.4619983435
2	0.04257367551	0.7880448103
3	0.05669873953	0.6459783316
4	0.09011174738	0.4866028428
5	0.112316452	0.8278418779
6	0.1420929581	0.782550931
7	0.1335987002	0.5825534463
8	0.09624969959	0.7047141194
9	0.1313056201	0.6806269884
10	0.1734389067	0.8265659809
11	0.1661651731	0.7380071878
12	0.1925661564	0.5359122753
13	0.1281296313	0.7117017508
14	0.1761084497	0.8007249832
15	0.2037310153	0.6995931864
16	0.1045408845	0.7638986111
17	0.1353729814	0.5902849436
18	0.1489517093	0.6677079201
19	0.1576663405	0.6904413104
20	0.09014988691	0.7761998773
21	0.220869258	0.7093069553
22	0.197660163	0.6442862749
23	0.1274724007	0.6228430271
24	0.1643767655	0.6819694638
25	0.1949167997	0.8983559608
26	0.09089576453	0.8193558455
27	0.2125929147	0.7806860805
28	0.2025323808	0.7680702806
29	0.07705853134	0.6463490129
30	0.1016499028	0.6045994163
31	0.207694903	0.5410217047
32	0.2101981193	0.8683655262
33	0.2119964212	0.759580493
34	0.2190889716	0.501537323
35	0.2211948782	0.8102539182
36	0.2251022607	0.7882441282
37	0.2355180681	0.7829395533
38	0.2355689257	0.8067421913
39	0.292994827	0.5838332176
40	0.2401511967	0.6783252954
41	0.2437100112	0.7428564429
42	0.2452854663	0.6010507941
43	0.2286607176	0.8102938533
44	0.09178114682	0.5382189751
45	0.2444279939	0.7610019445
46	0.2511603236	0.7856425643
47	0.2512062192	0.742957592
48	0.2558661997	0.807508707
49	0.2530550361	0.7625072002
50	0.2962216735	0.7891664505
51	0.2587529421	0.5696384311
52	0.2611985803	0.654742837
53	0.264541626	0.4389564097
54	0.2786568105	0.8205165863
55	0.2795863152	0.7718273997

Table A1. Cont.

Question ID	BERTScore	Semantic Textual Similarity Score
56	0.2803887129	0.9213430285
57	0.2814955115	0.6929318309
58	0.2833760977	0.7235119343
59	0.09830898792	0.572914362
60	0.1070303246	0.5689245462
61	0.2842413485	0.8167885542
62	0.2847544849	0.6485208273
63	0.2884128988	0.7158298492
64	0.2979023457	0.8750890493
65	0.298217386	0.8679254055
66	0.3084926307	0.8777817488
67	0.3093059659	0.4605668187
68	0.3115234673	0.8180803061
69	0.4968356788	0.9338030219
70	0.330129087	0.7585834861
71	0.328743279	0.6795406938
72	0.3350864351	0.7533646822
73	0.4113920033	0.914045155
74	0.337575525	0.6367307901
75	0.3420274854	0.8524702787
76	0.3301834762	0.8520085812
77	0.342875421	0.8358888626
78	0.3463438153	0.8626087308
79	0.347088635	0.756069839
80	0.381343931	0.8439093828
81	0.3507251441	0.847517252
82	0.350726217	0.784525454
83	0.3604282737	0.8269191384
84	0.3648406863	0.7977852821
85	0.09125140309	0.6629832387
86	0.09588064998	0.6290306449
87	0.09667455405	0.5115911365
88	0.0865451619	0.6453903913
89	0.3744304478	0.8571535945
90	0.3821184337	0.8458015919
91	0.3994865716	0.9185526371
92	0.4133167565	0.8013061285
93	0.3931872249	0.7404608727
94	0.3941654861	0.6610897183
95	0.4802063406	0.8897250295
96	0.415214628	0.9021199942
97	0.4389337003	0.8694944382
98	0.4423240721	0.8887104988
99	0.4559440911	0.7602285743
100	0.4427044094	0.854303658

### Appendix B. Domain Expert Feedback

The following is the collected feedback for one of the four questions from the registered dietitians and registered dietitian nutritionists who participated in our cross-tool comparison experiment. The responses to the other three questions can be found on a website that will be available upon final publication.

*Appendix B.1. Question 1: Any Advice on the Best Food/Diet to Lose Weight and Feel High Energy?*

**Table A2.** Question 1 of the Cross-Tool Comparison Experiment. Eight participants responded.

Evaluator	System	Criterion	Feedback	Rating
0001	DietNerd	Relevance	The response was fairly relevant, however, it appeared to have a strong weight loss focus and less of a focus on high energy	2
0001	DietNerd	Accuracy	The conclusion discusses the importance of a balanced diet, but the response separates out a “Balanced Diet” into bulletpoint #6, which may lead readers to distinguish a balanced diet from the other diets mentioned when a balanced diet can/should be incorporated into all of these dietary patterns.	2
0001	DietNerd	Completeness	Discussion on how each of the dietary patterns support energy levels and not just weight loss. Additionally, the conclusion discusses consulting a “registered dietitian” OR “nutritionist” when there is no regulation regarding the term “nutritionist”. I would recommend suggesting a consultation with a “registered dietitian” or “registered dietitian nutritionist” as those are the two credentials for licensed practitioners.	2
0001	DietNerd	Clarity	The pros and cons discusses the specific diets listed, but the way that this answer is formatted, it reads like the Pros/Cons list should be discussing all of the diets mentioned. It may be more effective to list the points in the pro/con list under each respective diet instead of under the entire list of dietary patterns.	2
0001	DietNerd	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0001	Scholar GPT	Relevance	I felt that this response was definitely more relevant than Response 1A, as it provided a few strategies as well as dietary patterns for both weight management and energy maintenance	3
0001	Scholar GPT	Accuracy	The point regarding avoiding processed foods may be better described as reducing “added sugars and unhealthy fats” or as “ultra high processed foods” in the diet as many foods can be minimally processed and part of a healthy diet. (for example, washing an apple is technically “processing” the apple). It also misses a discussion that ultra processed foods often include these added sugars and trans fats—so the point isn’t avoiding processed foods, but the additives in them.	2
0001	Scholar GPT	Completeness	As not all dietary patterns are appropriate for all individuals, it is crucial to prompt the reader to discuss their individual situation with a healthcare professional, notably a Registered Dietitian/Registered Dietitian Nutritionist.	2
0001	Scholar GPT	Clarity	Yes	3
0001	Scholar GPT	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4



Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0001	Consensus	Relevance	This response focuses very heavily on weight loss with little discussion on energy levels.	2
0001	Consensus	Accuracy	The low energy density diet is misleading—it does not provide an adequate discussion on maintaining a balanced diet which would assist with energy levels, and solely discusses how the low energy density diet would help with weight loss. The point about decreasing overall energy intake “while still allowing for larger volumes of food” is also unclear. Does this mean that after eating foods lower in kcal that the individual can eat ad libitum? I fear that this point could be potentially dangerous. Additionally, the “Balanced Macronutrient Diet” isn’t exactly a diet but a dietary behavior that should be the goal in any diet that is pitched to a patient/individual. Modified Fat Diet—this may also be misleading as it may lead the reader to increase consumption of fats in general in their diet. Fat contains the most kcal/gram so an increase in fat consumption in the diet likely won’t aid in weight loss? Additionally, the push for calorie counting may also increase disordered eating habits when not guided by a medical professional.	1
0001	Consensus	Completeness	As not all dietary patterns are appropriate for all individuals, it is important to include a prompt to seek individual recommendations from a healthcare professional, most notably, a Registered Dietitian/Registered Dietitian Nutritionist	2
0001	Consensus	Clarity	The response was short and formatted in a bulleted fashion which can increase readability.	3
0001	Consensus	Overall	Pretty Bad—This response had significant problems in either accuracy or clarity.	2
0001	Clinfo.ai	Relevance	This response doesn’t really discuss the high energy component of the question. Additionally, it mentions anti-obesity and weight loss medication which is not food/diet at all.	1
0001	Clinfo.ai	Accuracy	The TLDR section discusses the “best” diet for wt loss, however, this is largely not true from person to person. I feel that phrasing this summary in this way may be dangerous to readers where a reduced energy high protein diet combined with wt loss meds is inappropriate. Additionally the recommendation of a reduced energy diet to promote energy levels can be confusing and misleading to individuals.	1
0001	Clinfo.ai	Completeness	It mentions the need for individualized counseling but does not direct the reader to who they should discuss this with. Additionally, this response doesn’t really answer the question (i.e., foods/diets that help with both energy maintenance and wt loss).	1
0001	Clinfo.ai	Clarity	It may be difficult for the general reader to understand the significance of a systematic review vs. other types of research. The large chunk of text may also reduce readership. This may be more of a personal thing but I don’t really like that the second paragraph summarizing the information is titled “TLDR” just seems kind of unprofessional.	1

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0001	Clinfo.ai	Overall	Horrible—This response was just unacceptable.	1
0001	OpenEvidence	Relevance	Relevant! I enjoyed the discussion of adherence along with the nutritional component of sustained energy levels and weight management.	3
0001	OpenEvidence	Accuracy	I'm hesitant about greenlighting the Low-Fat Vegan Diet, as the description describes more of a plant-forward diet rather than a Vegan Diet which restricts all animal products. While this may play a role in weight loss, the potential for micronutrient deficiencies may be high if conducted without guidance, thus leading to potential lower energy levels. I feel that the descriptions of each of these dietary patterns could use a little more explanation of what they entail and whether they promote both energy levels and weight management or if its only one or the other.	2
0001	OpenEvidence	Completeness	As not all dietary patterns are appropriate for all individuals, it is important to include a prompt to seek individual recommendations from a healthcare professional, most notably, a Registered Dietitian/Registered Dietitian Nutritionist	2
0001	OpenEvidence	Clarity	I thought so-	3
0001	OpenEvidence	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0002	DietNerd	Relevance	Extremely relevant and accurate	3
0002	DietNerd	Accuracy	No	3
0002	DietNerd	Completeness	No	3
0002	DietNerd	Clarity	Yes it was	3
0002	DietNerd	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0002	Scholar GPT	Relevance	The reponse was very relevant. The advices contributed a lot for the lose weight need.	3
0002	Scholar GPT	Accuracy	No, it is clear.	3
0002	Scholar GPT	Completeness	No, it didn't.	3
0002	Scholar GPT	Clarity	Yes, it was.	3
0002	Scholar GPT	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0002	Consensus	Relevance	It was relevant and complete.	3
0002	Consensus	Accuracy	No.	3
0002	Consensus	Completeness	No.	3
0002	Consensus	Clarity	Yes, totally.	3
0002	Consensus	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0002	Clinfo.ai	Relevance	Less relevant then the prior ones.	2
0002	Clinfo.ai	Accuracy	The vegan diet could have be better discussed on order to increase results on losing weight.	2
0002	Clinfo.ai	Completeness	Yes. Vegan diet could have be deeper explored.	2
0002	Clinfo.ai	Clarity	Yes, it was.	3

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0002	Clinfo.ai	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0002	OpenEvidence	Relevance	It was very relevant.	3
0002	OpenEvidence	Accuracy	No.	3
0002	OpenEvidence	Completeness	No.	3
0002	OpenEvidence	Clarity	Yes, it was.	3
0002	OpenEvidence	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0003	DietNerd	Relevance	The responses are all relevant.	3
0003	DietNerd	Accuracy	For the low-carb diets response, I would specify “healthy fats” since some fats are not as beneficial (such as, trans fats and saturated fats).	2
0003	DietNerd	Completeness	Intermittent fasting has many caveats such as, including a moderate caloric deficit. It may also not necessarily be energy promoting if fasting windows are too long. Would likely not recommend this to patients/clients.	2
0003	DietNerd	Clarity	Yes	3
0003	DietNerd	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0003	Scholar GPT	Relevance	The responses are all relevant.	3
0003	Scholar GPT	Accuracy	No	3
0003	Scholar GPT	Completeness	No. Would also probably include the disclaimer of consulting with a medical professional or registered dietitian to get tailored recommendations for people with medical conditions.	2
0003	Scholar GPT	Clarity	Yes	3
0003	Scholar GPT	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0003	Consensus	Relevance	Pretty relevant	3
0003	Consensus	Accuracy	No, none of the information in that response was factually incorrect or misleading.	3
0003	Consensus	Completeness	Would probably re-phrase the low-energy-density diets to something like “moderate caloric deficit”. Having a diet entirely of lower calorie foods could result in lower energy levels and is not sustainable.	2
0003	Consensus	Clarity	Yes	3
0003	Consensus	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0003	Clinfo.ai	Relevance	Pretty relevant	3
0003	Clinfo.ai	Accuracy	No	3
0003	Clinfo.ai	Completeness	No	3
0003	Clinfo.ai	Clarity	It is fairly presented. The formats of the previous responses are a little more clear but I like the TLDR section. This response would probably be more suitable for people with higher literacy levels and health literacy.	2
0003	Clinfo.ai	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0003	OpenEvidence	Relevance	Very	3
0003	OpenEvidence	Accuracy	Would probably swap out the low-fat vegan diets response to a “plant forward diet”. A low-fat vegan diet can be a little too restrictive and less sustainable for long term healthy eating habits. Sometimes vegan “diet foods/products” can be more pricey and not available to some people who can’t afford them.	2
0003	OpenEvidence	Completeness	No, not necessarily.	3
0003	OpenEvidence	Clarity	Yes. I particularly liked the mention of behavioral and supportive interventions.	3
0003	OpenEvidence	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0004	DietNerd	Relevance	Not totally relevant.	2
0004	DietNerd	Accuracy	No.	3
0004	DietNerd	Completeness	Did not specifically mention or address high energy levels. Detail seemed to be lacking.	2
0004	DietNerd	Clarity	No—the pros and cons section was not relevant to the question and was a bit confusing since various dietary patterns were listed.	1
0004	DietNerd	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0004	Scholar GPT	Relevance	Very—I like that the response covered the specific foods/food components and then the recommended dietary patterns. The response also addressed both weight loss and energy levels.	3
0004	Scholar GPT	Accuracy	No	3
0004	Scholar GPT	Completeness	No	3
0004	Scholar GPT	Clarity	Yes	3
0004	Scholar GPT	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0004	Consensus	Relevance	Pretty relevant.	3
0004	Consensus	Accuracy	No.	3
0004	Consensus	Completeness	Did not mention/address energy levels; only addressed/focused on weight loss.	2
0004	Consensus	Clarity	Yes.	3
0004	Consensus	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0004	Clinfo.ai	Relevance	Low relevancy—this response mentioned both diet and medications however the question asked for ‘best food/diet’.	1
0004	Clinfo.ai	Accuracy	No.	3
0004	Clinfo.ai	Completeness	Yes—minimal information/detail provided on the foods/dietary patterns to promote weight loss and high energy levels.	1
0004	Clinfo.ai	Clarity	It was clear enough.	3
0004	Clinfo.ai	Overall	Pretty Bad—This response had significant problems in either accuracy or clarity.	2

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0004	OpenEvidence	Relevance	Mostly relevant.	2
0004	OpenEvidence	Accuracy	No.	3
0004	OpenEvidence	Completeness	Did not mention anything about energy levels.	2
0004	OpenEvidence	Clarity	Yes.	3
0004	OpenEvidence	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0005	DietNerd	Relevance	Very relevant	3
0005	DietNerd	Accuracy	No	3
0005	DietNerd	Completeness	I don't think so	3
0005	DietNerd	Clarity	yes	3
0005	DietNerd	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5
0005	Scholar GPT	Relevance	Relevant	3
0005	Scholar GPT	Accuracy	Low glycemic index foods are no longer relevant (it's outdated) for weight loss.	2
0005	Scholar GPT	Completeness	No	3
0005	Scholar GPT	Clarity	Yes	3
0005	Scholar GPT	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0005	Consensus	Relevance	Relevant	3
0005	Consensus	Accuracy	Low glycemic index foods may mislead audiences to believe weight loss is achieved by choosing such foods, but in reality, is choosing low calorie foods (which are likely a low GI food)	2
0005	Consensus	Completeness	No	3
0005	Consensus	Clarity	Yes	3
0005	Consensus	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0005	Clinfo.ai	Relevance	Somewhat	2
0005	Clinfo.ai	Accuracy	The references are too old	1
0005	Clinfo.ai	Completeness	N/A— references are too old	1
0005	Clinfo.ai	Clarity	Yes, but not to the general public, This format is easy for clinicians who are used to reading literature summaries, but may be too difficult for others.	2
0005	Clinfo.ai	Overall	Pretty Bad—This response had significant problems in either accuracy or clarity.	2
0005	OpenEvidence	Relevance	Relevant	3
0005	OpenEvidence	Accuracy	No but try to keep references within 5 years	2
0005	OpenEvidence	Completeness	No	3
0005	OpenEvidence	Clarity	Yes	3
0005	OpenEvidence	Overall	Excellent—You would use the tool that generated this response again and would recommend it to others.	5

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0006	DietNerd	Relevance	Very relevant	3
0006	DietNerd	Accuracy	I feel that intermittent fasting offers the least flexibility of any of the diets. I would also include information on needing to be in a calorie deficit in order to lose weight.	2
0006	DietNerd	Completeness	For the minimal information listed, it was well represented. The balanced diet is very similar to the Mediterranean Diet. That could be included as a bullet point rather than a diet itself.	2
0006	DietNerd	Clarity	Yes	3
0006	DietNerd	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0006	Scholar GPT	Relevance	I feel like when asking the question of “best food” I would be looking for specifics ie. banana, chicken, asparagus, etc. I think framing it with “tips” or “guidance” is better suited. The word diet is listed in ways that are more so “meal patterns”	2
0006	Scholar GPT	Accuracy	Eating an abundance of low glycemic foods can still raise blood sugar, I’d change the work can prevent to may prevent. I would also give examples in the food groups listed.	2
0006	Scholar GPT	Completeness	no	3
0006	Scholar GPT	Clarity	Yes	3
0006	Scholar GPT	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0006	Consensus	Relevance	relevant, I like how it says dietary strategies verses diets.	3
0006	Consensus	Accuracy	1. low energy density is misleading because we are looking for nutrient rich/dense meals, I see that this is worded to focus on lower calories but is confusing.	2
0006	Consensus	Completeness	the second study in high protein diets says “Triglyceride change was negatively correlated with animal-protein intake” I would change how this reads that these diets improve lipid profiles, especially when consuming higher saturated fat from animal sources.	2
0006	Consensus	Clarity	5. is not presented clearly, I would want to see more explanation on “careful food selection”	2
0006	Consensus	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0006	Clinfo.ai	Relevance	I dont think including medications fits the question	2
0006	Clinfo.ai	Accuracy	I think adding in weight loss medications muddies the point of the question. the answer seems to be pointing towards the use of medications as a necessary accompaniment	2
0006	Clinfo.ai	Completeness	a lot of information is missing in terms of comparing high protein and vegan diets, I think more explanation could be added on animal vs plant proteins	2
0006	Clinfo.ai	Clarity	yes	3
0006	Clinfo.ai	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0006	OpenEvidence	Relevance	relevant	3
0006	OpenEvidence	Accuracy	2. I would say can improve TG and HDL	2
0006	OpenEvidence	Completeness	4. I would say why it's beneficial, labs? wt loss? ... the other diets promote moderate animal products and dairy, and mention the benefits	2
0006	OpenEvidence	Clarity	5. this isn't a diet, I would include this in the summary as well as to consult or work with a dietitian. these responses all lead the user to continue to do more research	2
0006	OpenEvidence	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0007	DietNerd	Relevance	Relevant, but much too brief, not detailed enough.	2
0007	DietNerd	Accuracy	I would not recommend intermittent fasting if there is a goal of "high energy". There is no research to support that. Need to define what "balanced diet" means.	2
0007	DietNerd	Completeness	I just felt the explanation of each diet was much too summarized.	2
0007	DietNerd	Clarity	see above.	2
0007	DietNerd	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0007	Scholar GPT	Relevance	Relevant but again, too brief. What does "low glycemic" mean? The average American may not know this. What does "appropriate portions" mean? This is vague.	2
0007	Scholar GPT	Accuracy	See above.	2
0007	Scholar GPT	Completeness	No mention of pros/cons (which I liked from Q1) or that they should seek personalized guidance from a registered dietitian or health professional. This is very important, since some of these answers contradict one another so may be confusing—i.e., recommending Plant Based, but then recommending Low Carb, High Protein.	2
0007	Scholar GPT	Clarity	I think these answers are too brief and not nuanced enough. (Why one diet may be better for someone versus someone else)	2
0007	Scholar GPT	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0007	Consensus	Relevance	Relevant but again, way too summarized. need to explain terms and concepts more.	2
0007	Consensus	Accuracy	see above. and see my comment to previous question.	2
0007	Consensus	Completeness	No mention of seeking guidance from RD or health professional.	2
0007	Consensus	Clarity	No, especially the conclusion. This line is extremely vague: For effective weight loss and sustained high energy levels, consider a diet low in energy density, high in protein, and balanced in macronutrients with an emphasis on low-GI foods and healthy fats	2
0007	Consensus	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3

Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0007	Clinfo.ai	Relevance	Relevant, but I think the way the information is presented would be overwhelming for the average American. Probably would be easier to read with bulletpoints/subheadings. This is also just looking at ONE systematic review. (i'm not clear how comprehensive it was)	2
0007	Clinfo.ai	Accuracy	The information is accurate, though I do not think it's accurate to say "The best diet for weight loss and high energy appears to be a reduced-energy, high-protein diet" after what was previously stated.	2
0007	Clinfo.ai	Completeness	pros and cons of each diet and meds.	2
0007	Clinfo.ai	Clarity	Again, I feel that this is all very summarized. I understand the need for that, but then there should be links to more information. (i.e., what intermittent fasting is)	2
0007	Clinfo.ai	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0007	OpenEvidence	Relevance	Relevant and I appreciate that behavioral support and the importance of meal plans reflecting "personal and cultural preferences" was included	3
0007	OpenEvidence	Accuracy	Mention a registered dietitian... "professional guidance" by who? It's important to point people in the right direction.	2
0007	OpenEvidence	Completeness	see above. Also again, terminology should be defined or at least linked to articles that provide more information.	2
0007	OpenEvidence	Clarity	It's still very summarized, but better than some of the others.	2
0007	OpenEvidence	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0008	DietNerd	Relevance	Pretty Good	3
0008	DietNerd	Accuracy	Low-carb diets would likely not help w/energy levels	2
0008	DietNerd	Completeness	With energy levels and wt loss, the composition of the diet is important (what was noted in the response), but where said calories are coming from is also important, which is more focused on what foods are being consumed rather than their macronutrient breakdown	2
0008	DietNerd	Clarity	Somewhat	2
0008	DietNerd	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0008	Scholar GPT	Relevance	Mostly relevant	3
0008	Scholar GPT	Accuracy	Do not recommend keto diet for high energy	2
0008	Scholar GPT	Completeness	Much more balanced than 1A, I would add something about movement/exercise too	2
0008	Scholar GPT	Clarity	Yes	3
0008	Scholar GPT	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0008	Consensus	Relevance	Good	3
0008	Consensus	Accuracy	No	3



Table A2. Cont.

Evaluator	System	Criterion	Feedback	Rating
0008	Consensus	Completeness	No	3
0008	Consensus	Clarity	Yes	3
0008	Consensus	Overall	Pretty Good—The response was reasonable, but not quite good enough to recommend.	4
0008	Clinfo.ai	Relevance	Meh	2
0008	Clinfo.ai	Accuracy	No	3
0008	Clinfo.ai	Completeness	It noted IF being still under investigation, and assuming this is a broad scale recommendation (to males and females), I'd note that it works differently in males and females based on hormonal fluctuations. I also don't love the idea of noting wt loss medications as a route to take when seeking out weight loss, as those should be recommended by a doctor given very specific circumstances	2
0008	Clinfo.ai	Clarity	Mostly	2
0008	Clinfo.ai	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3
0008	OpenEvidence	Relevance	Good	3
0008	OpenEvidence	Accuracy	No	3
0008	OpenEvidence	Completeness	Low carb diets will likely not yield high energy levels	2
0008	OpenEvidence	Clarity	Mostly	2
0008	OpenEvidence	Overall	Okay—This response was more or less accurate, but incomplete or unclear.	3

## Appendix C. LLM Implementation Details

### Appendix C.1. Question Validity

#### Functionality Overview:

- Determines if the user's question is one that we can answer. While DIETNERD can recommend general diets that may be suitable for certain health conditions, DIETNERD does not answer questions around recipe creation and questions that are asked on behalf of an animal.

#### Input Variables:

- query (str): The user's question.

#### Output Variables:

- question\_validity (str): A string indicating whether the question is valid or not. Possible responses can only either be "True", "False—Recipe", or "False—Animal".

#### LLM Settings:

- gpt-4-turbo
- Prompt Strategy: few-shot learning
- temperature = 0.2
- top\_p = 0.5

**Prompt:**

You are an expert in classifying user questions. Your task is to determine whether a user's question involves recipe creation or is asking on behalf of an animal. Recipe creation questions involve detailing specific ingredients, cooking methods, and detailed instructions for preparing a dish. Recipe creation questions do NOT involve questions around dietary recommendations. If the user's question is about recipe creation, return "False - Recipe". If the question is asking on behalf of an animal, return "False - Animal". If the question does not involve any of these topics, return "True". Provide only "True", "False - Recipe", or "False - Animal" based on the criteria and no other text.

Here are some examples:

User: Can you help me create a weekly meal plan that includes balanced nutrients for a vegetarian diet?

AI: False - Recipe

User: How do I make a low-carb lasagna?

AI: False - Recipe

User: What are some ideas for healthy snacks I can prepare for my kids?

AI: True

User: What are some meals for someone with diabetes?

AI: True

User: What are the health benefits of intermittent fasting?

AI: True

User: What is the best diet for my cat?

AI: False - Animal

User: Can dogs eat raw meat?

AI: False - Animal

*Appendix C.2. Query Generation***Functionality Overview:**

- One prompt generates a keyword query to PubMed about the general topic of the user query and another prompt generates four keyword searches to PubMed to elicit points of contention. These points of contention are fundamental to our safety analysis.
  - \* The **General Query Generation LLM** builds one query directly from the user's question. This query retrieves articles that provide general context.
  - \* The **Point of Contention Query Generation LLM** generates four distinct queries that attempt to represent the top points of contention around the user's question. We have found that points of contention often reveal safety issues.

**Input Variables:**

- query (str): The user's question

**Output Variables:**

- general\_query (str): The broad query that will retrieve articles related to a specific topic.
- query\_contention (str): A list of 4 queries to represent the top points of contention around the topic.
- query\_list (list): A list of all 5 queries generated.

**LLM Settings—General Query Generation:**

- gpt-4-turbo
- Prompt Strategy: few-shot learning
- temperature = 0.7
- top\_p = 0.1

**Prompt—General Query Generation:**

You are an expert in generating precise and effective PubMed queries to help researchers find relevant scientific articles. Your task is to create a broad query that will retrieve articles related to a specific topic provided by the user. The queries should be optimized to ensure they return the most relevant results. Use Boolean operators and other search techniques as needed. Format the query in a way that can be directly used in PubMed's search bar. Return only the query and no other text.

Here are some examples:

User: Is resveratrol effective in humans?

AI: (resveratrol OR "trans-3,5,4'-trihydroxystilbene") AND human

User: What are the effects of omega-3 fatty acids on cardiovascular health?

AI: (omega-3 OR "omega-3 fatty acids") AND "cardiovascular health"

User: What does the recent research say about the role of gut microbiota in diabetes management?

AI: ("gut microbiota") AND ("diabetes management") AND ("recent"[Publication Date])

**LLM Settings—Points of Contention Query Generation:**

- gpt-4-turbo
- Prompt Strategy: few-shot learning
- temperature = 0.6
- top\_p = 1

**Prompt—Points of Contention Query Generation:**

You are an expert in generating precise and effective PubMed queries to help researchers find relevant scientific articles. Your task is to list up to 4 of the top points of contention around the given question, making sure each point is relevant and framed back to the original question.

Each point should be as specific as possible and have a title and a brief summary of what the conversation is around this point of contention.

The points should be ranked in order of how controversial the point is (how much debate and conversation is happening), where 1 is the most controversial. For each and every point of contention provided, generate 1 broad PubMed search query. Use Boolean operators and other search techniques as needed. Format each query in a way that can be directly used in PubMed's search bar.

Format the response like the following and do not include any other words:

\* Point of Contention 1: <title>

Summary: <summary>

Query: <search\_query>

Here is an example:

User: Is resveratrol effective in humans?

AI:

\* Point of Contention 1: Efficacy of resveratrol in humans

Summary: The debate revolves around the effectiveness of resveratrol supplements in humans. Some studies suggest that resveratrol may have various health benefits, such as cardiovascular protection and anti-aging effects, while others argue that the evidence is inconclusive or insufficient

Query: (resveratrol OR "trans-3,5,4'-trihydroxystilbene") AND human

\* Point of Contention 2: Dosage and Timing of Resveratrol Intake

Summary: This point of contention focuses on the optimal dosage and timing of resveratrol intake for life span extension. Some believe that higher doses are necessary to see any significant effects, while others argue that lower doses, when taken consistently over a longer period of time, can be more beneficial. Additionally, there is debate about whether resveratrol should be taken in a fasting state or with food to maximize its absorption and potential benefits.

Query: (resveratrol OR "trans-3,5,4'-trihydroxystilbene") AND dose

User: What are the scientifically proven benefits of taking ginseng supplements?

AI:

\* Point of Contention 1: Efficacy of Ginseng in Cognitive Function

Summary: The debate revolves around the effectiveness of ginseng supplements in enhancing cognitive function. Some studies suggest that ginseng may have various cognitive benefits, such as improving memory and concentration, while others argue that the evidence is inconclusive or insufficient.

Query: (ginseng OR "Panax ginseng") AND cognition

\* Point of Contention 2: Ginseng for Immune System Enhancement

Summary: This point of contention focuses on the role of ginseng in immune system enhancement. Some believe that ginseng can significantly boost immune system function, while others argue that the evidence is not strong enough to make such claims.

Query: (ginseng OR "Panax ginseng") AND immune

\* Point of Contention 3: Ginseng for Energy and Stamina

Summary: The efficacy of ginseng in increasing energy and stamina is a common point of debate. While some studies suggest that ginseng can help to combat fatigue and increase physical performance, others argue that these effects are not consistently observed across studies.

Query: (ginseng OR "Panax ginseng") AND energy

\* Point of Contention 4: Safety and side effects of Gingko supplements

Summary: The safety of Gingko supplements is a point of contention, with some concerns raised about potential side effects such as dizziness, upset stomach, and increased bleeding risk. While some studies suggest that Gingko supplements are generally safe, others argue that caution should be exercised, especially when combined with certain medications or in individuals with specific health conditions.

Query: (Gingko OR "Gingko Biloba") AND (safety OR "side effects")

*Appendix C.3. Relevance Classifier***Functionality Overview:**

- Classifies an article as relevant or irrelevant based on its abstract. An article is considered relevant if it contains information that is helpful in answering the question, it contains a safety aspect that would be important to include in the answer, or it is NOT an animal-based study.

**Input Variables:**

- article (dict): A dictionary containing the fetched PubMed article data.
- user\_query (str): The original query from the user.

**Output Variables:**

- pmid (str): PubMed ID of the article.
- article\_is\_relevant (bool): Whether the article is relevant or not.
- article (dict): The input article dictionary.

**LLM Settings:**

- gpt-3.5-turbo
- Prompt Strategy: zero-shot learning
- temperature = 0.8
- top\_p = 0.5

**Prompt:**

You are an expert medical researcher whose task is to determine whether research articles and studies are relevant to the question or that may be useful to know for safety reasons. Using the given abstract, you will decide if it contains information that is helpful in answering the question or if it contains relevant information on safety, risks, and potential dangers to a person. Please answer with a yes/no only. If the article is about an animal (e.g., hamster, mice), you must answer with "no".

*Appendix C.4. Full-Text Section Mapping***Functionality Overview:**

- This function is only used if the article's full text is available directly in PubMed.
- Captures only the most relevant sections from an article's full text to be cognizant of token size and context windows.
- Does a case-sensitive check to see which of the section titles provided within a given article best matches the required section titles.

**Input Variables:**

- list\_of\_strings (list): A list of all of an article's section titles to search through.
- required\_titles (list): A list of titles that are deemed to be the most relevant and helpful to include.

**Output Variables:**

- sections\_to\_pull (list): A list of matched section titles.

**LLM Settings:**

- gpt-3.5-turbo
- Prompt Strategy: zero-shot learning
- temperature = 0.1
- top\_p = 1

**Prompt:**

Of the given list of sections within the research paper, choose which sections most closely map to an "Abstract", "Background", "Methods", "Results", "Discussion", "Conclusion", "Sources of Funding", "Conflicts of Interest", "References", and "Table" section. Only use section names provide in the list to map. Multiple sections can map to each category. If there are multiple sections, separate them using the character |.

```

Format must follow:
Abstract: <sections>
Background: <sections>
Methods: <sections>
Results: <sections>
Discussion: <sections>
Conclusion: <sections>
Sources of Funding: <sections>
Conflicts of Interest: <sections>
Table: <sections>
References: <sections>

```

#### Appendix C.5. Strength of Claim Assessment

##### Functionality Overview:

- Handles multiple article types and sources
- Performs in-depth content analysis and summarization
- Extracts critical metadata and bibliographic information
- Adapts analysis based on article type (review vs. study)
- Focuses on technical details and statistical metrics

##### Input Variables:

- article (dict): A dictionary containing the raw article data.

##### Output Variables:

- article\_json (dict): A dictionary containing processed article information, including title, publication type, URL, abstract, "is relevant" flag, citation, PMID, PMCID, "full text" flag, and strength of claim analysis.

##### LLM Settings:

- gpt-4-turbo
- Prompt Strategy: zero-shot learning
- temperature = 0.6
- top\_p = 1

##### Prompt for Review Type Articles:

Given the following literature review paper, extract the following information and summarize it, being technical, detailed, and specific, while also explaining concepts for a layman audience. Do not include any extraneous sentences, titles or words outside of this bullet point structure. As often as possible, directly include metrics and numbers (especially significance level, confidence intervals, t-test scores, effect size). Follow the instructions in the parentheses:

1. Purpose (What is the review seeking to address or answer? What methods were used? If relevant and mentioned, include dosages.):
2. Main Conclusions (What are the conclusions and main claims made? What are its implications?):
3. Risks (Are there any risks mentioned (e.g., addiction, death)?):
4. Benefits (Are there any benefits purported?):
5. Search Methodology and Scope (What was the search strategy used to identify relevant literature? Assess the breadth and depth of the literature included. Is the scope clearly defined, and does it encompass relevant research in the field?):
6. Selection Criteria (Evaluate the criteria used for selecting the studies included in the review. What types of studies were included and which were excluded? Were diverse perspectives incorporated? Are contradictory findings or alternative theories addressed?):

7. Quality Assessment of Included Studies (Were quality assessment methods applied? How were the methodologies, results, and reliability of the studies assessed?):
8. Synthesis & Analysis (Evaluate how the findings from different studies are synthesized and analyzed. Is there a clear structure and methodology for synthesizing the literature? What statistical tests were used and for what purpose? Include all mention of statistical metrics and interpret what they mean, especially significance levels/p-values, confidence intervals, t-test scores, or effect size):
9. Sources of Funding or Conflict of Interest (Identify any sources of funding and possible conflicts of interest.):

#### Prompt for Study Type Articles:

Given the following research paper, extract only the following information enumerated below and summarize it, being technical, detailed, and specific, while also explaining concepts for a layman audience. Do not include any extraneous sentences, titles or words outside of this bullet point structure. As often as possible, directly include metrics and numbers (especially significance level, confidence intervals, t-test scores, effect size). Follow the instructions in the parentheses:

1. Purpose & Design (What is the study seeking to address or answer? What methods were used? Were there any exclusions or considerations? Include dosages if mentioned.):
2. Main Conclusions (What claims are made?):
- Risks (Are there any risks mentioned (e.g., addiction, death)?):
3. Benefits (Are there any benefits purported?):
- Type of Study ((e.g., observational, randomized). If randomized, mention if it was placebo controlled or double-blinded.):
4. Testing Subject (Human or animal; include other adjectives and attributes):
5. Size of Study (May be written as "N="):
6. Length of Study:
7. Statistical Analysis of Results (What tests were conducted? Include the following attributes with a focus on mentioning as many metrics):
8. Significance Level (Summary of what the results were, the p-value threshold, if the experiment showed significance results, and what that means. Mention as many significant p-value numbers as available.):
9. Confidence Interval (May be expressed as a percentage):
10. Effect Size (Did the study aim for a certain effect size? May be expressed as Cohen's d, Pearson's r, or SMD. Include % power if mentioned):
11. Sources of Funding or Conflict of Interest (Identify any sources of funding and possible conflicts of interest.):

#### Appendix C.6. Final Output Synthesis

##### Functionality Overview:

- Generate the final response to the user's question based on the strongest level of evidence in the provided article summaries

##### Input Variables:

- all\_relevant\_articles (list): List of all relevant article summaries.
- query (str): User question.

##### Output Variables:

- final\_output (str): Final response to the user's question.

##### LLM Settings:

- gpt-4-turbo
- Prompt Strategy: few-shot learning
- temperature = 0.5

- top\_p = 1

**Prompt:**

You are an expert in evaluating research articles and summarizing findings based on the strength of evidence. Your task is to review the provided Evidence and Claims and use only those information that have strong evidence to answer the user's question. You must choose at least 6 articles and at most 15 articles and use only the information in these articles to answer the question. Strong evidence means the research is well-conducted, peer-reviewed, and widely accepted in the scientific community. Provide a direct, research-backed answer to the question and focus on identifying the pros and cons of the topic in question. The answer should highlight when there are potential risks or dangers present.

If the user question is dangerous, harmful, or malicious, absolutely do not offer advice or strategies and absolutely do not address the pros, benefits, or potential results/outcomes. You must only focus on deterring this behavior, addressing the risks, and offering safe alternatives. The answer should also try to include as many different demographics as possible. Absolutely NO animal studies should be referenced or included in the final response. Mention dosage amounts when the information is available. Medical terms and technical concepts must be explained to a layman audience. Be sure to emphasize that you should always go and see a registered dietitian or a registered dietitian nutritionist.

There must be a reference list with the AMA citation format. Articles must be cited in-line in Vancouver style using brackets. References listed must be numerically listed using brackets. Include section titles like "Conclusion" and organize sections as a bulleted list using an asterisk. List each and every one of the cited articles mentioned at the end using the citations in Evidence and Claims. Do not list duplicate references.

The output must follow this format:

<summary\_of\_evidence>

References:

[1] <AMA\_citation\_1>  
[2] <AMA\_citation\_2>  
[3] <AMA\_citation\_3>  
[4] <AMA\_citation\_4>  
[5] <AMA\_citation\_5>  
[6] <AMA\_citation\_6>  
[7] <AMA\_citation\_7>  
[8] <AMA\_citation\_8>  
[9] <AMA\_citation\_9>  
[10] <AMA\_citation\_10>  
...

Here are some examples:

User: {example\_1\_question}

AI: {example\_1\_response}

User: {example\_2\_question}

AI: {example\_2\_response}



### Appendix D. System Diagram

Here we present comprehensive diagram of DIETNERD's system in the context of a typical user request/system response.

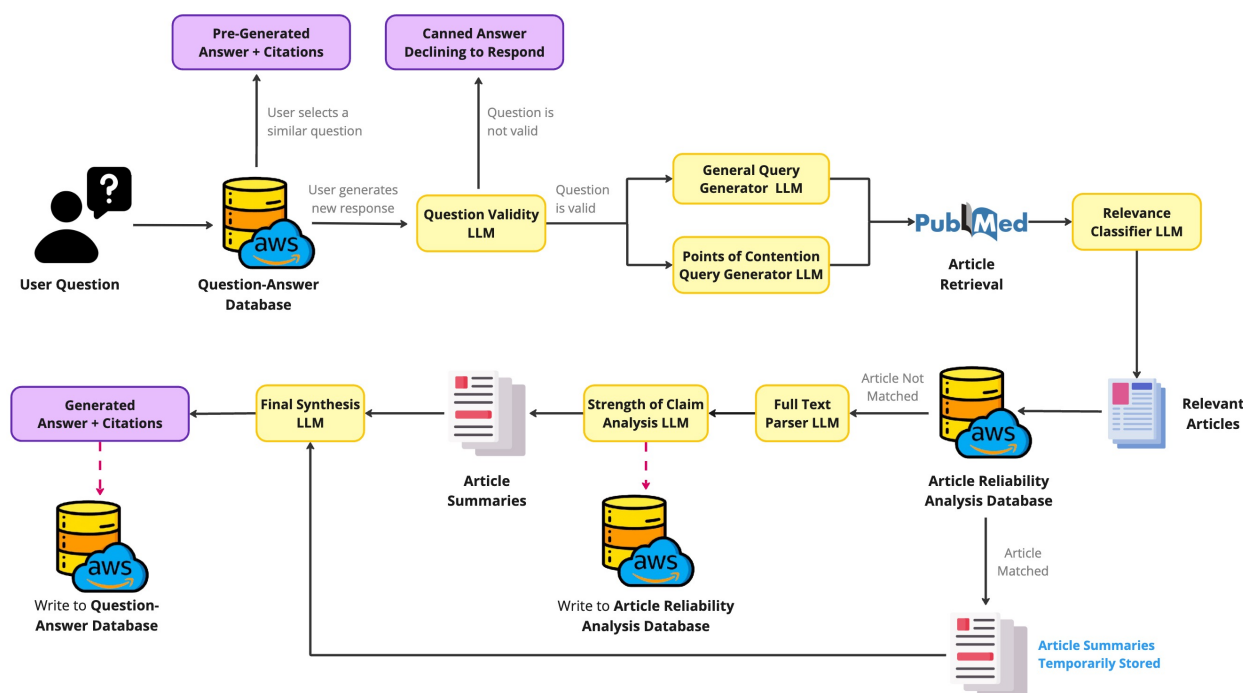


Figure A1. Architectural Components of DIETNERD when answering a typical user request.

### References

- Yu, E.; Rimm, E.; Qi, L.; Rexrode, K.; Albert, C.M.; Sun, Q.; Willett, W.C.; Hu, F.B.; Manson, J.E. Diet, Lifestyle, Biomarkers, Genetic Factors, and Risk of Cardiovascular Disease in the Nurses' Health Studies. *Am. J. Public Health* **2016**, *106*, 1616–1623. [CrossRef] [PubMed]
- Firth, J.; Marx, W.; Dash, S.; Carney, R.; Teasdale, S.B.; Solmi, M.; Stubbs, B.; Schuch, F.B.; Carvalho, A.F.; Jacka, F.; et al. The Effects of Dietary Improvement on Symptoms of Depression and Anxiety: A Meta-Analysis of Randomized Controlled Trials. *Psychosom. Med.* **2019**, *81*, 265–280. [CrossRef] [PubMed]
- Fassier, P.; Chhim, A.S.; Andreeva, V.A.; Hercberg, S.; Latino-Martel, P.; Pouchieu, C.; Touvier, M. Seeking health- and nutrition-related information on the Internet in a large population of French adults: Results of the NutriNet-Santé study. *Br. J. Nutr.* **2016**, *115*, 2039–2046. [CrossRef] [PubMed]
- Johnson, A. Here's How To Use AI—Like ChatGPT And Bard—For Everyday Tasks Like Creating A Budget, Finding Airfare Or Planning Meals. *Forbes* **2023**. Available online: <https://www.forbes.com/sites/ariannajohnson/2023/04/10/heres-how-to-use-ai-like-chatgpt-and-bard-for-everyday-tasks-like-creating-a-budget-finding-airfare-or-planning-meals> (accessed on 1 August 2024).
- Lozano, A.; Fleming, S.L.; Chiang, C.C.; Shah, N. Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In Proceedings of the Pacific Symposium on Biocomputing 2024, Kohala Coast, HI, USA, 3–7 January 2024; pp. 8–23.
- Lozano, A.; Fleming, S.L.; Chiang, C.C.; Shah, N. PubMedRS-200 Dataset. 2023. Available online: <https://github.com/somshahlab/Clinfo.AI/blob/main/PubMedRS-200/PubMedRS-200.csv> (accessed on 1 August 2024).
- Olson, E.; Salem, C. Consensus. Available at OpenAI's GPT Marketplace, 2022. Available online: <https://chatgpt.com/g/g-bo0FiWLY7-consensus> (accessed on 1 August 2024).
- Consensus. How It Works & Consensus FAQ's. Available online: <https://consensus.app/home/blog/welcome-to-consensus/> (accessed on 1 August 2024).
- Awesomegpts.ai. Scholar GPT. Available at OpenAI's GPT Marketplace, March 2023. Available online: <https://chatgpt.com/g/g-kZ0eYXlJe-scholar-gpt> (accessed on 1 August 2024).
- Nadler, D.; Ziegler, Z.; Wulff, J.; Smith, M.; Hernandez, E.; Lehman, E.; Xu, J.; Ferreira, F.; Esdaile, J.; Hu, H.; et al. OpenEvidence. 2021. Available online: <https://www.openevidence.com/> (accessed on 1 August 2024).
- CB Insights. AI 100: The Most Promising Artificial Intelligence Startups of 2024, 2024. Available online: <https://www.cbinsights.com/research/report/artificial-intelligence-top-startups-2024/> (accessed on 1 August 2024).

12. Business Insider. OpenEvidence AI Becomes the First AI in History to Score above 90% on the United States Medical Licensing Examination (USMLE), 2023. Available online: <https://markets.businessinsider.com/news/stocks/openevidence-ai-becomes-the-first-ai-in-history-to-score-above-90-on-the-united-states-medical-licensing-examination-usmle-1032446924> (accessed on 1 August 2024).
13. Hippocratic AI. StatPearls Semantic Search. Available at Hippocratic Medical Questions Website, 2024. Available online: <https://hippocratic-medical-questions.herokuapp.com/> (accessed on 1 August 2024).
14. Elicit. Elicit. Available at Elicit Website, 2023. Available online: <https://elicit.com> (accessed on 24 January 2023).
15. He, K.; Mao, R.; Lin, Q.; Ruan, Y.; Lan, X.; Feng, M.; Cambria, E. A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics. *arXiv* **2024**, arXiv:2310.05694. [[CrossRef](#)]
16. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [[CrossRef](#)]
17. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2021**, arXiv:2005.11401. [[CrossRef](#)]
18. Li, J.; Yuan, Y.; Zhang, Z. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. *arXiv* **2024**, arXiv:2403.10446. [[CrossRef](#)]
19. Funer, F. Accuracy and Interpretability: Struggling with the Epistemic Foundations of Machine Learning-Generated Medical Information and Their Practical Implications for the Doctor-Patient Relationship. *Philos. Technol.* **2022**, *35*, 5. [[CrossRef](#)]
20. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239. [[CrossRef](#)]
21. Glaese, A.; McAleese, N.; Trębacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *arXiv* **2022**, arXiv:2209.14375. [[CrossRef](#)]
22. Bar-Haim, R.; Kantor, Y.; Venezian, E.; Katz, Y.; Slonim, N. Project Debater APIs: Decomposing the AI Grand Challenge. *arXiv* **2021**, arXiv:2110.01029. [[CrossRef](#)]
23. Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E.M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. BlenderBot 3: A Deployed Conversational Agent That Continually Learns to Responsibly Engage. *arXiv* **2022**, arXiv:2208.03188. [[CrossRef](#)]
24. Christiano, P.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. *arXiv* **2023**, arXiv:1706.03741. [[CrossRef](#)]
25. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv* **2022**, arXiv:2203.02155. [[CrossRef](#)]
26. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv* **2020**, arXiv:1911.00536. [[CrossRef](#)]
27. Sayers, E. A General Introduction to the E-Utilities. 2009. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK25497/> (accessed on 17 November 2022).
28. National Library of Medicine. The 9 E-Utilities and Associated Parameters Available at National Library of Medicine. Available online: <https://www.nlm.nih.gov/dataguide/eutilities/utilities.html> (accessed on 11 September 2024).
29. Luxford, K. 'First, do no harm': Shifting the paradigm towards a culture of health. *Patient Exp. J.* **2016**, *3*, 5–8. [[CrossRef](#)]
30. PubMed Central. PubMed Central: An Archive of Biomedical and Life Sciences Journal Literature. Available at National Center for Biotechnology Information. Available online: <https://www.ncbi.nlm.nih.gov/pmc/> (accessed on 1 August 2024).
31. National Library of Medicine. Publication Types. Available at National Library of Medicine. Available online: <https://www.nlm.nih.gov/mesh/pubtypes.html> (accessed on 1 August 2024).
32. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675. [[CrossRef](#)]
33. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019.
34. Anshütz, M.; Lozano, D.M.; Groh, G. This is not correct! Negation-aware Evaluation of Language Generation Systems. *arXiv* **2023**, arXiv:2307.13989. [[CrossRef](#)]
35. Cer, D.; Yang, Y.; Kong, S.-Y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal Sentence Encoder. *arXiv* **2018**, arXiv:1803.11175. [[CrossRef](#)]
36. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [[CrossRef](#)]
37. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
38. Ng, J.-P.; Abrecht, V. Better Summarization Evaluation with Word Embeddings for ROUGE. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015, Lisbon, Portugal, 17–21 September 2015; pp. 1925–1930. [[CrossRef](#)]
39. Xia, P.; Zhang, L.; Li, F. Learning similarity with cosine similarity ensemble. *Inf. Sci.* **2015**, *307*, 39–52. [[CrossRef](#)]
40. Katari, M.S.; Shasha, D.; Tyagi, S. *Statistics Is Easy: Case Studies on Real Scientific Datasets*, 1st ed.; Synthesis Lectures on Mathematics & Statistics; Springer: Cham, Switzerland, 2021; Volume 1, p. XI, 62. [[CrossRef](#)]

41. Allied Health Workforce Projections, 2016–2030: Registered Dietitians Allied Health Workforce Project. Available online: <https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/data-research/registered-dieticians-2016-2030.pdf> (accessed on 10 September 2024).
42. Lipscomb, R. Health Coaching: A New Opportunity for Dietetics Professionals. *J. Am. Diet. Assoc.* **2006**, *106*, 801–803. [[CrossRef](#)]
43. Gharibi, M.; Zachariah, A.; Rao, P. FoodKG: A Tool to Enrich Knowledge Graphs Using Machine Learning Techniques. *Front. Big Data* **2020**, *3*, 12. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.