JIM
JIM 2026; 3 (3): e1211
DOI: 10.61012/JIM_202608_1211
Category: *Original Article*

HOW RELIABLE ARE LARGE LANGUAGE MODELS IN METABOLIC DIETETICS? A CROSS-SECTIONAL BENCHMARKING EVALUATION

M. TOSI[1], M. MUSCARELLA[2], G. GUGELMO[3], A. PINTO[4], M. SUÁREZ GONZÁLEZ[5,6], S. POLONI[7], A. HÖLLER[8,9,10], V. PERICO[11], A. FAVARO[12], A. TAVIAN[13], A. BELVEDERE[14], A. DIANIN[12]

• • •

[1]Metabolic Diseases Unit, Department of Pediatrics, Vittore Buzzi Children's Hospital, Milan, Italy
[2]Dietetic Unit, Meyer Children University Hospital IRCCS, Florence, Italy
[3]Division of Metabolic Diseases, Department of Medicine, University Hospital of Padua, Padua, Italy
[4]Department of Dietetics, Birmingham Women's and Children's Hospital, Birmingham, UK
[5]Department of Pediatrics, Hospital Universitario Central de Asturias, Oviedo, Spain
[6]Pediatric Research, Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo, Spain
[7]Nutrition and Dietetics Service, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
[8]Division of Nutrition and Dietetics, University Hospital Innsbruck, Innsbruck, Austria
[9]Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL-University for Health Sciences and Technology, Hall in Tirol, Austria
[10]Digital Health Information Systems, Center for Health & Bioresources, AIT Austrian Institute of Technology, Innsbruck, Austria
[11]Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy
[12]Regional Center for Newborn Screening, Diagnosis and Treatment of Inherited Metabolic Diseases and Congenital Endocrine Diseases, Pediatric Unit C, University Hospital of Verona, Verona, Italy
[13]SOSD Technical Health Care Professions – Dietitian and Regional Coordinating Center for Rare Diseases, Azienda Sanitaria Universitaria Friuli Centrale, Udine, Italy
[14]Independent AI Researcher, Verona, Italy

Corresponding Author
Martina Tosi, Ph.D; e-mail: martina.tosi@unimi.it

***ABSTRACT* – Objective:** The aim of this study was to evaluate the performance of free-tier large language models (LLMs) in providing dietary recommendations for inherited metabolic disorders (IMDs), with a focus on scientific accuracy, completeness, and response consistency of generated responses across models under real-world conditions.

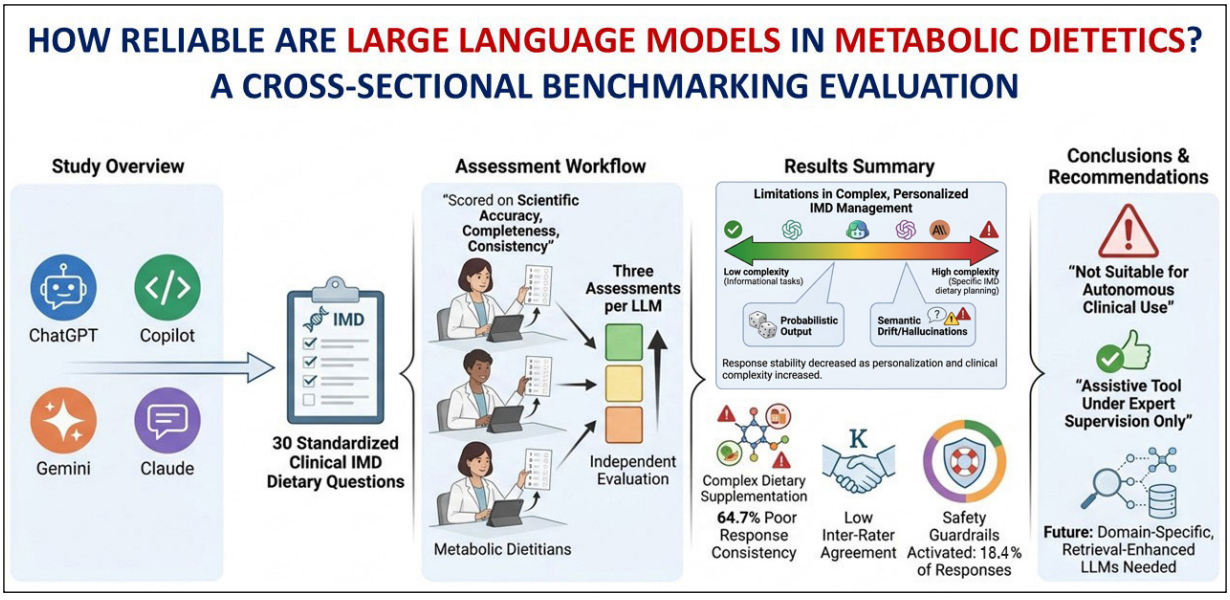
Materials and Methods: This cross-sectional benchmarking study assessed four LLMs – ChatGPT, GitHub Copilot, Gemini and Claude – using 30 standardized clinical questions on dietary management in IMDs. Models were tested under real-world free-tier conditions, without fine-tuning or prompt engineering. Each question was submitted three times to each model. Responses were independently evaluated by three metabolic dietitians using a 6-point ordinal scale for scientific accuracy and completeness, while response consistency was assessed across repeated outputs. Inter-rater agreement was assessed using the intraclass correlation coefficient (ICC).

Results: A total of 359 unique responses were collected for evaluation. Safety guardrails were activated in 18.4% of responses, with substantial variability across models and clinical domains. Of these, 80.3% resulted in a complete block and were excluded from the analysis. Consequently, the final sample consisted of 306 responses. Model performance was heterogeneous: ChatGPT achieved the highest and most consistent scores, with no severe inaccuracies or omissions. Copilot, Gemini and Claude showed intermediate performance, with relevant variability. Across clinical domains, responses related to diet planning and clinical monitoring were generally more stable, whereas dietary supplementation and nutritional requirements showed greater variability and clinically relevant errors. Poor response consistency was observed in 64.7% of responses, indicating limited consistency across repeated identical prompts. Inter-rater agreement is defined as poor for accuracy, completeness, and response consistency.

Conclusions: The performance of free-tier LLMs in IMD dietary management is heterogeneous, with substantial limitations in response consistency and in clinically complex scenarios. These findings indicate that current models are not yet suitable for autonomous clinical use and can be dangerous to use without a healthcare professional's oversight.

Keywords: Artificial intelligence, Large language models, Dietetics, Inherited metabolic disorder, Clinical decision support.

Abbreviations: AI, artificial intelligence; CKD, chronic kidney disease; IMD, inherited metabolic disorder; IMDs, inherited metabolic disorders; ICC, intraclass correlation coefficient; Q1, first quartile; Q3, third quartile; LLM, large language model; LLMs, large language models; MMA, methylmalonic acidemia; RAG, retrieval-augmented generation.



Graphical Abstract. Comparative evaluation of four free-tier large language models in answering patient-like questions on dietary management of inherited metabolic disorders, assessing scientific accuracy, completeness, and response consistency under real-world conditions, and demonstrating heterogeneous performance with limitations for autonomous clinical use.

INTRODUCTION

Internet searches for health-related information represent a rapidly expanding phenomenon in modern healthcare communication. Advances in artificial intelligence (AI) have led to the development of large language models (LLMs), such as ChatGPT, which use natural language processing and machine learning techniques to generate contextually appropriate, human-like responses to user queries[1]. These systems are increasingly able to synthesize information and provide coherent, interactive answers, contributing to growing interest in their use as accessible tools for preliminary medical and health-related guidance[2]. These tools may offer advantages by providing real-time, personalized information and enhancing patient engagement[3]. More broadly, AI and machine learning methods are increasingly being applied in nutri-

tion research to address high-dimensional datasets, enable predictive modeling, and support personalized dietary recommendations[4]. These approaches range from classical statistical learning techniques to advanced deep learning architectures, including models such as LLMs that can process unstructured clinical and nutritional text data[5]. However, despite their rapid adoption, the effectiveness and reliability of LLMs in providing domain-specific guidance remain insufficiently characterized. Recent scoping evidence indicates that, although a growing number of studies have explored applications of LLMs in nutrition and dietetics, most investigations are based on hypothetical scenarios rather than real-world clinical implementations, with limited use of domain-specific or retrieval-augmented approaches[6]. In particular, their utility in nutritional counseling and dietary management has not been fully evaluated, especially in the context of rare and complex conditions such as inherited metabolic disorders (IMDs). Dietary management in IMDs is particularly challenging because recommendations are disease-specific, age-dependent, and frequently require continuous adjustment according to biochemical monitoring and metabolic control. The aim of this study was to evaluate the scientific accuracy, completeness, and response consistency of responses generated by four freely available LLMs (ChatGPT, GitHub Copilot Free, Gemini, and Claude) to standardized patient-oriented questions concerning dietary management in IMDs. Responses were independently assessed by expert metabolic dietitians according to current clinical guidelines.

MATERIALS AND METHODS

Study Design

To address the current lack of evidence regarding the reliability of freely available LLMs in providing dietary recommendations for IMDs, we conducted a cross-sectional comparative study. Four LLMs (ChatGPT, GitHub Copilot Free, Gemini, and Claude) were assessed under real-world conditions using their publicly accessible free versions, without fine-tuning, retrieval-augmented generation (RAG), or domain-specific prompt engineering. The overall study design is summarized in Figure 1. The study consisted of three sequential phases: (i) development of standardized patient-style questions, (ii) response generation by each LLM under identical conditions, and (iii) blinded evaluation of responses by expert metabolic dietitians.

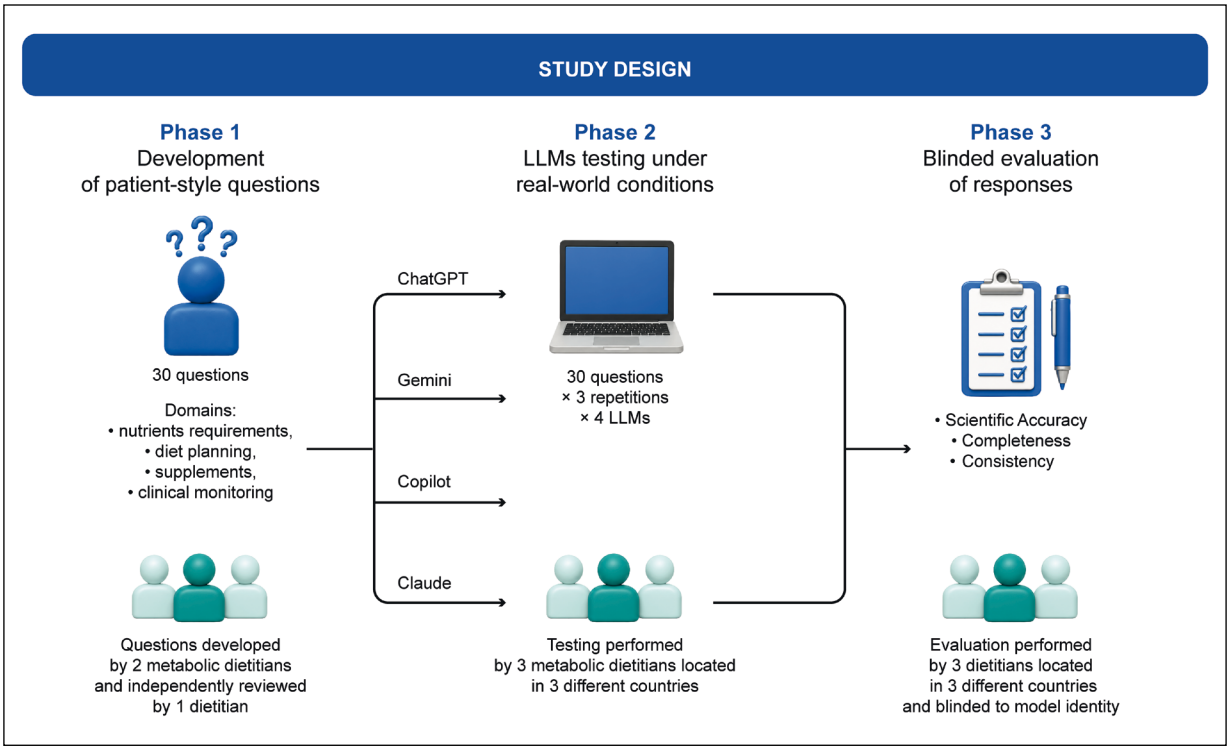


Figure 1. Study design illustrating the evaluation of free-tier LLM responses to patient-style dietary management queries for IMDs.

Questions Development

A standardized set of 30 clinically relevant questions was developed to cover the main domains of dietary management in IMDs (Appendix 1): nutrient requirements and restrictions (n = 10), diet planning (n = 5), dietary supplements (n = 5), and clinical considerations and monitoring (n = 10).

The complete set of questions is available in Appendix 1, while Table 1 provides an overview of the investigated dietary management domains together with representative questions submitted to the LLMs. Questions were developed by two experienced metabolic dietitians based on current dietary guidelines for IMDs and independently reviewed by a third metabolic dietitian. As the study involved experts from different countries, all questions were standardized and submitted to the LLMs in English. Content validity was established through alignment with current international guideline recommendations[7-18] and representative real-world clinical scenarios. The question set was designed to encompass different levels of complexity, ranging from basic nutritional knowledge to advanced clinical decision-making. Any disagreements regarding the wording or content of the questions were resolved by consensus. Investigators involved in question development did not participate in LLM testing or response evaluation.

Table 1. Main investigated topics and categories with representative questions submitted to the LLMs.

	Nutrient requirements and restrictions	Diet planning	Dietary supplements	Clinical considerations and monitoring
Disorders of protein or amino acid metabolism GA I, TYR II, UCD, MSUD, PKU, MMA, CTLN 1	• Low-protein foods • High in lysine foods • Low in Phe foods • Breastfeeding	• Meals at a restaurant	• PS recommended dose • Gastrointestinal symptoms following PS use	• Night-time DBS collection • DBS results • Pregnancy • Emergency feeds (vomiting)
Disorders of carbohydrate metabolism GAL, HFI, GSD I, GSD III, GSD V	• Lactose-free dairy products • Ketogenic diet • Fructose in vegetables	• Ketones levels • Daily snacks • Ideas for breakfast	• Multivitamin supplements	• Obesity and weight loss
Defects of fatty acid beta-oxidation VLCAD, LCHAD, CPT 2	• Low-fat foods	• Number of meals during the day and night	• Maltodextrin use and physical activity	• Sports • Hiking

CPT II, carnitine palmitoyltransferase II deficiency; CTLN1, citrullinemia type I; DBS, dried blood spot; GA I, glutaric aciduria type I; GAL, galactosemia; GSD I, glycogen storage disease type I; GSD III, glycogen storage disease type III; GSD V, glycogen storage disease type V; HFI, hereditary fructose intolerance; LCHAD, long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency; MMA, methylmalonic acidemia/aciduria; MSUD, maple syrup urine disease; Phe, phenylalanine; PKU, phenylketonuria; PS, protein substitute; TYR II, tyrosinemia type II; UCD, urea cycle disorders; VLCAD, very-long-chain acyl-CoA dehydrogenase deficiency.

LLMs Selection

Four freely available LLMs were evaluated: 1. ChatGPT (OpenAI, free version), accessed through OpenAI's publicly available web interface. The exact model version and inference configuration were not disclosed by the provider. 2. GitHub Copilot Free (GitHub/Microsoft), accessed through the GitHub Copilot interface. This platform dynamically routes user requests to different LLMs depending on system availability, without disclosing the specific model used for each interaction. 3. Gemini (Google, free

version), accessed through Google's publicly available web interface. Although access was provided to the Gemini family of models, the exact model variant and configuration were not specified. 4. Claude (Anthropic, free version), accessed through Anthropic's publicly available web interface. The precise model version and inference parameters were not disclosed. All models were evaluated using their publicly accessible free versions without subscription plans, custom instructions, fine-tuning, or external knowledge integration. Testing was conducted under routine user conditions, with the understanding that free-tier services may be subject to usage limits and dynamic model allocation policies.

Model Testing Procedure

All LLMs were evaluated using their default configurations without fine-tuning, custom instructions, or additional prompt engineering. Each model was accessed through a newly created, neutral free-tier account with no prior interaction history. To minimize potential personalization effects, each query was submitted in a separate chat session, thereby preventing conversational memory from influencing subsequent responses. Identical, clinically phrased prompts were submitted to all models. Each question was entered three times independently for each LLM to assess response consistency. Model testing was performed by three metabolic dietitians from three different countries (Brazil, Austria, and Italy), who were responsible for submitting the standardized questions and collecting the generated responses. These investigators were not involved in either question development or response evaluation. All interactions with the LLMs were conducted through their respective publicly available web interfaces between February and March 2026. Testing sessions were distributed across different days to reduce the potential influence of transient server load or temporary model updates. The exact version of the model and its inference configuration were not made available by the provider. Because LLM providers implement silent updates and automatic changes without notifying users, it was not possible for users of the free versions to determine or verify the specific model version generating the response. No follow-up questions or clarifications were provided after the initial prompt, and responses were recorded without modification.

Response Evaluation

Responses were independently evaluated by three metabolic dietitians from three different countries (Italy, the United Kingdom, and Spain), all with extensive experience in the clinical management of IMDs and a peer-reviewed publication record in the field. To minimize assessment bias, all responses were anonymized, randomized, and evaluated in a blinded manner with respect to model identity. Each evaluator independently assessed all responses using predefined scoring criteria. As all prompts and responses were generated in English, evaluations were also performed in English. All evaluators were proficient in English and had more than 10 years of experience in the clinical management of IMDs, ensuring reliable and consistent assessment.

Outcome Measures

- Scientific accuracy was defined as the extent to which each response was consistent with current evidence-based dietary guidelines for IMDs. Each response was rated using a 6-point ordinal scale: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; and 5 = completely accurate according to current evidence-based guidelines. Higher scores indicated greater scientific accuracy.
- Completeness was defined as the extent to which each response addressed all clinically relevant elements required to provide a comprehensive answer to the question. Responses were rated using a 6-point ordinal scale: 0 = fails to address the question; 1 = addresses very few relevant aspects; 2 = addresses some relevant aspects; 3 = addresses most relevant aspects; 4 = addresses all relevant aspects with minor omissions or limited depth; and 5 = fully addresses all relevant aspects of the question. Higher scores indicated greater completeness.
- Response consistency: Response reproducibility was defined as the consistency of responses generated by the same LLM when presented with identical prompts across repeated independent runs. For each question–model pair, three independently generated responses were compared descriptively to assess within-model response consistency. Responses were classified as fully consistent (score =

2), partially consistent (score = 1), or inconsistent (score = 0). A score of 2 indicated that all three responses conveyed essentially the same information with only negligible wording differences; a score of 1 indicated minor variations in wording or emphasis without changes in the overall content; and a score of 0 indicated substantial differences in the information or recommendations provided.
- Inter-rater agreement. Agreement among the three independent evaluators was assessed for each outcome domain (scientific accuracy, completeness, and response consistency) using the intraclass correlation coefficient (ICC). A two-way random-effects model with absolute agreement was used, as all responses were independently evaluated by the same three raters, who were considered representative of a broader population of expert metabolic dietitians. ICC values were interpreted according to commonly accepted criteria: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement. In addition, the percentage agreement among evaluators was calculated as a descriptive measure.

Statistical Analysis

All outcomes were analyzed using descriptive statistics. For scientific accuracy and completeness, scores assigned by the three independent evaluators were aggregated at the response level using two complementary approaches. First, the mean score across evaluators was calculated for each response, and the resulting values were summarized using the median and first and third quartiles (Q1-Q3). Second, the median score across evaluators was calculated for each response and subsequently summarized using the overall median and Q1-Q3. These two approaches were reported to provide complementary estimates of central tendency and variability, with the median-based aggregation providing a more robust measure less influenced by potential outlier ratings among evaluators. Response-level median scores were used to identify extreme performance outcomes. Severe inaccuracy and severe incompleteness were defined as median scores ≤1.5 on the 6-point ordinal scale. For response reproducibility, consistency scores derived from repeated runs of each LLM were summarized descriptively using median values and Q1-Q3. Poor response reproducibility was defined as a median consistency score <2. Inter-rater agreement among the three independent evaluators was assessed for each outcome domain (scientific accuracy, completeness, and response consistency) using the ICC, as described above. A two-way random-effects model with absolute agreement and average measures [ICC(2,*k*)] was applied. ICC values were interpreted according to commonly accepted criteria: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement. Safety guardrails identified in LLM-generated responses were characterized using descriptive analyses. No comparative inferential analyses between LLMs were performed.

Ethical Considerations

This study evaluated only AI-generated outputs; no human participants or patient data were included. Expert participation was voluntary.

RESULTS

A total of 360 AI-generated responses were initially collected across four free-tier LLMs. Before the blinded evaluation phase, one response was excluded due to a data collection error, as a duplicated response had been inadvertently recorded for a different query (diet planning domain, Question 2, Gemini – Model C). The final dataset, subjected to expert evaluation and statistical analysis, therefore consisted of 359 unique responses. When LLMs were unable to fully address specific medical queries, safety guardrails were activated in 18.4% of cases (66/359). Of these, 80.3% (53/66) resulted in a complete block with no scientific content and were excluded from the analysis. Consequently, the final sample consisted of 306 responses.

Overall Model Performance

Overall performance varied across platforms. ChatGPT (Model A) showed the highest overall performance, with a consensus median scientific accuracy score of 4.0 (4.0-4.0) and a median of averaged scores of 4.0 (3.3-4.3). Completeness was also high, with a consensus median score of 4.0 (4.0-4.0) and

a median of averaged scores of 4.0 (3.7-4.3). In contrast, GitHub Copilot (Model B) showed low performance in scientific accuracy, with a consensus median scientific accuracy score of 3.0 (2.0-3.0) and a median of averaged scores of 3.3 (3.0-3.7). Completeness followed a different pattern, with a consensus median score of 4.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.0-3.7). Gemini (Model C) and Claude (Model D) showed comparable performance. Gemini had a consensus median scientific accuracy score of 3.0 (2.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7); completeness was similar, with a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.3 (3.0-3.7). Claude had a consensus median scientific accuracy score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7). Completeness was slightly lower, with a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.0 (2.7-3.3). Table 2 summarizes the performance of the evaluated models.

Table 2. Summary of the overall model performance, reporting scientific accuracy and completeness, together with the number of severe inaccuracies and instances of severe incompleteness for each model.

Model	Scientific accuracy*	Completeness*	Severe inaccuracies % (n/total number of responses)	Severe incompleteness % (n/total number of responses)
Model A (ChatGPT)	4.0 (4.0-4.0)/ 4.0 (3.3-4.3)	4.0 (4.0-4.0)/ 4.0 (3.7-4.3)	0.0% (0/90)	0.0% (0/90)
Model B (Copilot)	3.0 (2.0-3.0)/ 3.3 (3.0-3.7)	4.0 (3.0-4.0)/ 3.3 (3.0-3.7)	2.4% (1/42)	0.0% (0/42)
Model C (Gemini)	3.0 (2.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-3.0)/ 3.3 (3.0-3.7)	6.0% (5/84)	6.0% (5/84)
Model D (Claude)	3.0 (3.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-3.0)/ 3.0 (2.7-3.3)	8.9% (8/90)	0.0% (0/90)
Total	**–**	**–**	**4.6% (14/306)**	**1.6% (5/306)**

*Data are presented as consensus median (Q1-Q3)/median of averaged scores (Q1-Q3). Severe inaccuracies and incompleteness are defined as a consensus median score ≤1.5. For Scientific Accuracy: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; 5 = completely accurate based on current evidence-based guidelines. For Completeness: 0 = fails to address the question; 1 = addresses very few aspects; 2 = addresses some aspects; 3 = addresses most aspects; 4 = addresses all aspects with minor omissions or limited depth; 5 = fully addresses all aspects of the question.

Performance by Clinical Domain

Performance varied across clinical domains, reflecting differences in task complexity. In the domain of "nutrient requirements and restrictions", models achieved a consensus median scientific accuracy score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (2.7-3.7). Completeness followed a similar distribution, with a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.0-3.7). "Diet planning" showed similar performance, reaching a consensus median score of 3.0 (3.0-3.0) and a median of averaged scores of 3.3 (3.3-3.3) for both scientific accuracy and completeness, reflecting absolute consistency in the averaged metrics.

The highest scientific accuracy was observed in the "dietary supplement" domain, which reached a consensus median score of 4.0 (4.0-4.0) and a median of averaged scores of 3.8 (3.8-3.8). Completeness in this domain presented a median score of 4.0 (4.0-4.0) and a median of averaged scores of 3.7 (3.7-3.7).

In contrast, the "clinical considerations and monitoring" domain showed comparable performance, with a slightly higher average accuracy. For scientific accuracy, this domain achieved a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.7 (3.0-4.0). For completeness, it reached a consensus median score of 3.0 (3.0-4.0) and a median of averaged scores of 3.3 (3.3-3.7). Table 3 summarizes the performance across the clinical domain.

Table 3. Summary of the performance across clinical domains, reporting clinical accuracy and completeness, together with the number of severe inaccuracies and instances of severe incompleteness for each domain.

Clinical domain	Scientific accuracy*	Completeness*	Severe inaccuracies % (n/total number of responses)	Severe incompleteness % (n/total number of responses)
Nutrient requirements and restrictions	3.0 (3.0-4.0)/ 3.3 (2.7-3.7)	3.0 (3.0-4.0)/ 3.3 (3.0-3.7)	4.6% (5/108)	1.9% (2/108)
Diet planning	3.0 (3.0-3.0)/ 3.3 (3.3-3.3)	3.0 (3.0-3.0)/ 3.3 (3.3-3.3)	2.1% (1/47)	0.0% (0/47)
Dietary supplements	4.0 (4.0-4.0)/ 3.8 (3.8-3.8)	4.0 (4.0-4.0)/ 3.7 (3.7-3.7)	10.0% (5/50)	4.0% (2/50)
Clinical considerations and monitoring	3.0 (3.0-4.0)/ 3.7 (3.0-4.0)	3.0 (3.0-4.0)/ 3.3 (3.3-3.7)	3.0% (3/101)	1.0% (1/101)
Total	–	–	**4.6% (14/306)**	**1.6% (5/306)**

*Data are presented as consensus median (Q1-Q3)/median of averaged scores (Q1-Q3). Severe inaccuracies and incompleteness are defined as a consensus median score ≤ 1.5. For Scientific Accuracy: 0 = serious scientific inaccuracies; 1 = major inaccuracies; 2 = some inaccuracies; 3 = mostly accurate with minor errors; 4 = accurate with only very minor or negligible errors; 5 = completely accurate based on current evidence-based guidelines. For Completeness: 0 = fails to address the question; 1 = addresses very few aspects; 2 = addresses some aspects; 3 = addresses most aspects; 4 = addresses all aspects with minor omissions or limited depth; 5 = fully addresses all aspects of the question.

Critical Failures: Severe Inaccuracies and Incompleteness

Overall, severe scientific inaccuracies were identified in 4.6% of responses (14/306), while severe incompleteness was observed in 1.6% (5/306). When stratified by LLM, the distribution of critical failures was uneven. Claude accounted for the highest rate of severe inaccuracies (8.9%, 8/90). Gemini and GitHub Copilot showed lower rates, observed in 6.0% (5/84) and 2.4% (1/42) of responses, respectively. Gemini was the only model presenting severe incompleteness (6.0%, 5/84). ChatGPT did not generate any responses classified as severely inaccurate or severely incomplete based on the consensus median score.

When the 14 severe scientific inaccuracies were stratified by clinical domain, a notable pattern emerged in the “dietary supplement” domain: despite having the highest overall median accuracy, this domain contained the highest number of severe errors (10.0%, 5/50). The remaining severe inaccuracies were distributed across “nutrient requirements and restrictions” (4.6%, 5/108), “clinical considerations and monitoring” (3.0%, 3/101), and “diet planning” (2.1%, 1/47).

Regarding severe incompleteness, the five identified cases were concentrated within specific clinical domains. The highest number of critical omissions occurred in the “dietary supplement” domain (4.0%, 2/50), followed by “nutrient requirements and restrictions” (1.9%, 2/108) and “clinical considerations and monitoring” (1.0%, 1/101). Table 2 summarizes the performance of the evaluated models with respect to scientific accuracy, completeness, number of severe inaccuracies, and severe incompleteness. Table 3 presents an analogous analysis within the clinical domain.

Response Consistency

The models showed limited consistency when the same query was submitted multiple times. Overall, 64.7% of responses (198/306) were classified as inconsistent or only partially consistent across the three repetitions. Detailed results of the response consistency analysis, stratified by model and clinical domain, are presented in Tables 4 and 5.

Table 4. Response consistency across models, reporting the proportion of responses with poor response consistency (median consistency score <2) and those with full consistency (consistency score=2).

Model	Poor consistency (score <2), % (n/total number of responses)	Fully consistent (score=2), % (n/total number of responses)
Model A (ChatGPT)	60.0% (54/90)	40.0% (36/90)
Model B (Copilot)	92.9% (39/42)	7.1% (3/42)
Model C (Gemini)	82.1% (69/84)	17.9% (15/84)
Model D (Claude)	40.0% (36/90)	60.0% (54/90)
Total	**64.7% (198/306)**	**35.3% (108/306)**

Table 5. Response consistency across clinical domains, reporting the proportion of responses with poor response consistency (median consistency score <2) and those with full consistency (consistency score=2).

Clinical Domain	Poor consistency, (score <2)% (n)	Fully consistency (score = 2), % (n)
Nutrient requirements and restrictions	66.7% (72/108)	33.3% (36/108)
Diet planning	74.5% (35/47)	25.5% (12/47)
Dietary supplements	58.0% (29/50)	42.0% (21/50)
Clinical considerations and monitoring	66.4% (62/101)	33.6% (39/101)
Total	**64.7% (198/306)**	**35.3% (108/306)**

Inter-rater Agreement

Inter-rater agreement was poor across all outcome domains. For scientific accuracy, the ICC(2,*k*) was 0.468 (95% CI: 0.36-0.56; $p<0.001$), indicating poor agreement among evaluators. Similarly, completeness showed poor agreement, with an ICC(2,*k*) of 0.351 (95% CI: 0.20-0.47; $p<0.001$). Finally, response reproducibility also demonstrated poor agreement, with an ICC(2,*k*) of 0.312 (95% CI: -0.01-0.53; $p<0.001$) (Table 6).

Table 6. Inter-rater agreement for each evaluation metric, reporting ICC(2,*k*) values and corresponding 95% confidence intervals (CI) for scientific accuracy, completeness, and response reproducibility.

Metric	ICC(2,*k*)	95% CI	*p*-value	Interpretation
Scientific accuracy	0.468	0.36, 0.56	<0.001	Poor
Completeness	0.351	0.2, 0.47	<0.001	Poor
Response consistency	0.312	-0.01, 0.53	<0.001	Poor

Data interpretation: <0.50, poor; 0.50-0.75, moderate; 0.75-0.90, good; and >0.90, excellent agreement.

Safety Guardrails

Safety guardrails were activated in 18.4% of cases (66/359) when LLMs were unable to fully address specific medical queries. Among these, 80.3% (53/66) resulted in a complete response block with no scientific content, whereas 19.7% (13/66) generated a partial response that included a recommendation to seek clinical consultation while still providing scientific information.

Considering the median scientific accuracy and completeness scores across all 359 responses, an important association with the activation of safety filters emerged. Specifically, a median scientific accuracy score of 0.0 was observed in 7 of the 359 responses, and all of these cases were associated with the activation of safety mechanisms. Of these seven responses, 1/7 belonged to the "diet planning" domain and 6/7 to the "dietary supplement" domain. Looking at the model, 4/7 responses were generated by Copilot (Model B) and 3/7 by Gemini (Model C), with most cases occurring during Testing 3 (5/7) rather than Testing 2 (2/7). Similarly, 8 responses received a median completeness score of 0.0, all associated with safety filter activation. These cases involved the "diet planning" (2/8), "dietary supplement" (5/8), and "clinical considerations and monitoring" (1/8) domains. Copilot (Model B) and Gemini (Model C) each accounted for four responses, with most cases occurring during Testing 3 (6/8) rather than Testing 2 (2/8).

Excluding responses with a median score of 0.0 (i.e., fully blocked outputs), 59 responses remained for the analysis of scientific accuracy and 58 for completeness. For the 59 responses assessed for scientific accuracy, safety mechanisms led to complete blocks in 39 cases, partial blocks in 13 cases, and brief pathology-only explanations in seven cases. A similar distribution was observed for the 58 responses assessed for completeness, with 39 cases resulting in a complete block, 12 in a partial block, and seven providing only a brief pathological overview before termination.

For scientific accuracy, safety mechanism activation was highest in GitHub Copilot (Model B; 53.3%, 48/90), followed by Gemini (Model C; 20.2%, 18/89), while no activations were observed for ChatGPT (Model A) or Claude (Model D). A qualitative divergence was observed in the implementation of safety filters across LLMs. Model B (GitHub Copilot) exhibited a more rigid safety architecture, with restrictions manifesting almost exclusively as complete blocks (87.5%, 42/48) or, in a minority of cases, as brief pathology-only explanations (12.5%, 6/48). No mid-generation terminations were observed. In contrast, Model C (Gemini) showed a more reactive filtering behavior, with 72.2% (13/18) of safety interventions resulting in partial blocks characterized by mid-output interruption. Notably, Model C accounted for all partial blocks observed in the study, indicating a distinct approach to clinical content restriction compared with Model B.

Activation varied across clinical domains, peaking in "dietary supplement" (21.7%, 13/120) and "diet planning" (20.3%, 12/59), and being lowest in "nutrient requirements and restrictions" (15.0%, 23/120). Safety restriction patterns varied by clinical domain. Clinical considerations and monitoring elicited the most rigid responses, with 82.6% (19/23) resulting in complete blocks. In contrast, nutrient requirements and restrictions showed a more heterogeneous distribution (seven complete blocks, six partial blocks, and five brief pathology-only responses).

Substantial geographic and session-based variability was also observed, with activation occurring during Testing 2 in Brazil (18.3%, 22/120) and peaking during Testing 3 in Austria (37.0%, 44/129), while no activations were observed during Testing 1 in Italy.

Responses in which the safety mechanism was applied were not evaluated uniformly by the three evaluators. Regarding scientific accuracy (0.0-5.0 scale), 32.3% (64/198) received a score of 0.0, 13.1% (26/198) a score of 1.0, 41.9% (83/198) a score of 2.0, 7.1% (14/198) a score of 3.0, 4.0% (8/198) a score of 4.0, and only 1.5% (3/198) achieved the maximum score of 5.0. Evaluator 2 applied a stricter scoring approach, assigning 0.0 in 86.4% of cases (57/66), whereas Evaluator 1 and Evaluator 3 predominantly assigned scores of 2.0 (54.5%, 36/66 and 71.2%, 47/66, respectively). Out of 1,077 total evaluations, only 8 (0.7%) received a scientific accuracy score of 0.0 without safety mechanism activation, reflecting incorrect rather than safety-restricted responses.

DISCUSSION

In this study, the performance of free-tier LLMs applied to dietary management in IMDs varied considerably across all models in terms of accuracy, completeness, and consistency of responses.

Among the investigated models, ChatGPT (Model A) showed the most consistent performance, suggesting greater robustness than other systems assessed, whereas Claude (Model D) accounted for the most severe inaccuracies. Copilot (Model B), Gemini (Model C), and Claude (Model D) achieved acceptable median scores but

showed relevant variability, especially across repeated or more complex clinical queries. The models generally demonstrated a consistent ability to handle basic nutritional information. Across clinical domains, "dietary supplements" received higher and more stable scores, but also the most severe inaccuracies. "Nutrient requirements and restrictions", "diet planning" " and "clinical considerations and monitoring" showed moderate performance, lower consistency and greater variability. This suggests that domains requiring more nuanced or guideline-dependent reasoning remain challenging for these models. The uneven distribution of severe errors also indicates that summary measures alone may mask high-risk outliers, and the broader distribution in the interquartile ranges, particularly in decision-sensitive clinical scenarios, highlights the inherent ambiguity and higher risk of inconsistent or incomplete advice. Response consistency analyses further showed inconsistency across prompt repetitions, indicating limited stability of free-tier models in generating clinical recommendations. Finally, the poor inter-rater agreement among expert evaluators reflects the inherent ambiguity of AI-generated clinical text and the difficulty of its consistent interpretation in specialized clinical contexts.

The results of this study show that safety guardrails play a central role in shaping both the completeness and the scientific accuracy of LLM outputs in clinical contexts. The strong parallelism observed between these two metrics suggests that when safety mechanisms are activated, they consistently limit not only the presence of clinical content but also its depth and correctness. Safety mechanism activation was not uniform across models, with Model B and Model C accounting for all observed interventions, while no activations were recorded for Model A and Model D. This highlights substantial heterogeneity in how different LLMs implement safety constraints in response to similar clinical queries. The nature of the safety intervention differed across models, ranging from complete response blocks to partial outputs or brief pathology-only explanations. In particular, Model B predominantly produced full blocks, whereas Model C more frequently generated partial responses, suggesting different operationalizations of content restriction. These patterns were also reflected in the variability observed across clinical domains, where more sensitive or high-risk topics elicited stronger safety responses. Taken together, these findings indicate that safety guardrails are not merely binary filters but rather structured mechanisms that significantly influence both the availability and the form of clinical information provided by LLMs. The variability observed across LLMs, geographic locations, and testing periods may also be partially influenced by platform-level safety and control mechanisms, rather than by intrinsic model differences alone. Potential contributing factors include iterative system updates, regional policy differences, and variations in medical safety filtering configurations across platforms[19].

Evidence from nutrition and clinical settings supports both the potential and the limitations of disease-specific LLM applications. In diabetes care, ChatGPT has generally shown acceptable performance when addressing common patient questions[2,20]; however, occasional inaccuracies and excessive or mixed-quality information raise concerns about response reliability[20,21]. In type 2 diabetes and metabolic syndrome, ChatGPT outputs were often clear and coherent but showed gaps in key components, including weight-loss strategies, energy-deficit recommendations, nutrient-specific guidance, and monitoring[22]. Although LLMs can generate plausible nutritional advice, they may deviate from established clinical guidelines or dietitian-designed references, particularly in more complex and multi-step interventions[23].

In chronic kidney disease (CKD), ChatGPT and other models achieved moderate-to-high accuracy in structured tasks such as identifying potassium and phosphorus content in foods; however, performance appears to depend on task complexity and is less reliable when generating personalized, stage-specific dietary recommendations, which require the integration of guidelines and patient-specific factors[24,25]. In addition, guideline-based evaluations of CKD dietary scenarios showed inconsistent energy, macro- and micronutrient targets, marked inter-model variability and no full compliance with established nutritional recommendations[26]. Expert assessments further confirm discrepancies in perceived accuracy, comprehensiveness, and clinical applicability across models, with performance depending on task formulation and model type[26].

In celiac disease, LLMs showed generally high accuracy and clarity; however, the non-negligible rates of misinformation (ranging from approximately 13% to 24%) have raised concerns about unsupervised patient use[27]. Mean quality scores of LLMs' responses may be high, around 4.3/5, yet variability persists across models and question types, especially for diagnostic content[28].

In IMDs such as methylmalonic acidemia (MMA), where treatment requires precise, individualized, and tightly controlled dietary interventions, ChatGPT displayed low appropriateness, with many dietary recommendations judged as inadequate[29].

In general nutrition advice, ChatGPT was described as an effective tool[1] and as providing answers similar to or even superior to those of dietitians, in terms of scientific correctness, practical applicability, and clarity[3]. However, limitations emerge with complex individualized nutritional care and in domains characterized by scientific uncertainty and heterogeneous evidence, such as dietary supplements. In-

deed, accuracy remains considerably lower than that observed in structured nutrition tasks, and vulnerability to misinformation increases along with inconsistent or controversial evidence[30].

Compared with more prevalent chronic conditions, IMDs involve highly specialized and strictly individualized dietary management, characterized by precise quantitative restrictions and disorder-specific metabolic requirements. This level of complexity may challenge LLMs, as it demands highly detailed, condition-specific metabolic knowledge that is less frequently encountered in general clinical nutrition literature, potentially limiting the consistency and completeness of generated recommendations.

Overall, LLM performance appears stronger in low-complexity informational tasks and progressively less reliable with increasing clinical specificity and personalization. This pattern may reflect the probabilistic architecture of LLMs, which generate outputs through statistical inference over learned representations rather than explicit rule-based reasoning[31]. Because knowledge is distributed across high-dimensional semantic spaces, model performance depends on the density and coherence of the activated semantic region and on the constraints provided by the prompt[32,33]. General queries usually align with well-represented semantic regions, whereas highly specialized clinical inputs may activate sparse or overlapping domains, increasing the risk of semantic drift and hallucinations[34,35]. Transparent reporting of model conditions is methodologically essential, as generation settings can introduce variability in model outputs, potentially affecting response consistency and comparability across runs. Thus, identical inputs may yield different responses across runs, especially in complex clinical scenarios where prompt ambiguity interacts with probabilistic decoding[36].

In real-world settings, the LLM response variability may be further amplified by underspecified or inconsistently formulated user queries. Additionally, professional output interpretation may differ, with clinicians prioritizing medical plausibility and dietitians emphasizing nutritional accuracy and practical applicability.

Collectively, these findings highlight the limited reliability of current LLMs for clinical use, with patient safety remaining critical, especially when inaccurate recommendations may have direct clinical consequences. While LLMs may offer advantages in accessibility, scalability, and patient engagement, they should be considered as assistive tools for clinical nutrition, operating under human oversight rather than autonomous decision-making systems[37]. A cautious and structured integration of LLMs into clinical nutrition, primarily for information access and educational support, rather than as standalone systems for dietary decision-making, is therefore desirable.

The development of domain-specific, retrieval-enhanced systems grounded in validated clinical databases represents a necessary step toward safer deployment in medical contexts[38]. Although generative AI systems are rapidly expanding in personalized dietary applications, most current implementations remain largely preference-driven rather than biologically grounded[39]. Clinically reliable AI-driven nutrition will require improved model architectures, standardized reporting frameworks, and stronger integration of biological data sources. General LLMs, which are not systematically trained on structured food-nutrition mappings or ontology-aligned datasets, have shown reduced performance in nutrient estimation, food classification, and semantic entity linking, whereas domain-specific models fine-tuned on large-scale, task-aligned datasets achieved higher accuracy and semantic consistency across nutritional profiling and structured classification tasks[40]. This could be achieved through supervised fine-tuning on curated clinical nutrition datasets, integration of structured food composition databases, and alignment with biomedical ontologies to improve consistency in nutrient representation and clinical reasoning.

This study has several limitations that should be acknowledged. First, the number of evaluators was limited, as this was a preliminary exploratory study designed to assess feasibility and generate initial insights rather than to provide definitive conclusions. The large number of evaluated responses also imposed a substantial cognitive and time burden, which may have contributed to inter-evaluator variability. Second, the study employed a predominantly descriptive and qualitative analytical approach, which, while appropriate for exploratory purposes, does not allow for formal inferential comparisons. Third, only four LLMs were included, which limits the generalizability of the findings across the rapidly evolving landscape of AI systems. In addition, given the fast pace of development in this field, model performance and configurations may change over time, meaning that results may not fully reflect future or updated versions of the same systems. Finally, all LLMs were accessed *via* free-tier interfaces with quotas, rate limits, and restricted availability, representing a further limitation. Exact model versions, configurations, and routing procedures were often undisclosed, limiting response consistency and direct comparison across platforms. Inter-evaluator variability may also have been influenced by heterogeneous response formats, including partial or safety-filtered outputs, and by varying levels of evaluator familiarity with specific clinical domains, which could have led to differences in scoring interpretation across conditions.

Future research should also define the levels of dietary education and counseling that can be safely supported by LLMs in IMDs. Transversal educational content may be more amenable to AI assistance, whereas counseling involving treatment complexity, metabolic control, clinical monitoring, or patient-specific dietary adaptation should remain highly personalized and supervised by specialized professionals.

CONCLUSIONS

Free-tier LLMs can generate clinically plausible responses in metabolic dietetics but remain insufficiently accurate, complete, and consistent for autonomous use in IMD dietary management. Their performance appears acceptable for general information retrieval and potentially for standardized educational support; however, it declines in domains requiring individualized interpretation, guideline integration, monitoring, and risk-sensitive decision-making. LLMs should therefore be considered assistive tools and used only under specialist supervision, as current models are not yet suitable for autonomous clinical use and can be dangerous to use without a healthcare professional's oversight. Future work should define which levels of dietary education and counseling may be safely delegated to AI and should prioritize domain-specific, retrieval-enhanced systems grounded in validated metabolic nutrition resources.

Acknowledgements:
We thank the Dietetic Working Group of the Italian Society for the Study of Hereditary Metabolic Diseases and Newborn Screening (SIMMESN).

Artificial Intelligence-Assisted Technologies:
Artificial intelligence-assisted technologies were used for language editing purposes, including grammar and syntax correction and improvement of readability.

Authors' Contributions:
Study conception and design: M.T., A.B., A.D.; acquisition, collection and interpretation of data: M.M., G.G., A.P., M.S.G., S.P., A.H., V.P., A.F., A.T.; manuscript drafting: M.T. and M.M.; manuscript editing: M.T., M.M., G.G., A.P., M.S.G., S.P., A.H., V.P., A.F., A.T., A.B. and A.D.; supervision: M.T. and A.D.; approval to submit: all the authors have approved the manuscript.

Availability of Data and Material:
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest:
The authors declare that they have no conflict of interest to disclose.

Ethics approval:
Not applicable.

Funding:
No funding was received for this study.

Informed Consent:
Not applicable.

ORCID ID:
Martina Tosi: 0000-0003-4859-5288
Marta Muscarella: 0009-0005-6203-5940
Giorgia Gugelmo: 0000-0002-3557-8078
Alex Pinto: 0000-0002-8206-0829
Marta Suárez González: 0000-0001-7552-4894
Soraia Poloni: 0000-0003-4506-6487
Alexander Höller: 0000-0002-8541-5904
Veronica Perico: 0000-0001-7741-0874
Aurora Favaro: 0009-0002-0624-6418
Alessandra Tavian: 0009-0006-0005-9661
Andrea Belvedere: 0009-0005-2539-8187
Alice Dianin: 0009-0000-9539-2649

REFERENCES

1. Ponzo V, Goitre I, Favaro E, Merlo FD, Mancino MV, Riso S, Bo S. Is ChatGPT an Effective Tool for Providing Dietary Advice? Nutrients 2024; 16: 469.
2. Bean AM, Payne RE, Parsons G, Kirk HR, Ciro J, Mosquera-Gómez R, Hincapié M S, Ekanayaka AS, Tarassenko L, Rocher L, Mahdi A. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nat Med 2026; 32: 609-615.
3. Kirk D, van Eijnatten E, Camps G. Comparison of Answers between ChatGPT and Human Dieticians to Common Nutrition Questions. J Nutr Metab 2023; 2023: 5548684.
4. Gugelmo G, Chesini F, Lenzini L, Vitturi N, Girelli D, Marchi G. Potential applications of digital health technologies in adults with inherited metabolic disorders. JIM 2025; 2: e1003.
5. Southey NL, Zhu R, Holscher HD. Machine Learning and Artificial Intelligence in Nutrition Research: Analytical Methods, Applications, and Key Considerations. J Nutr 2026; 156: 101528.
6. O'Hara C, Kent G, Leydon CL, Walsh NM, Gibney ER, Skoutas D, Flynn AC, Timon CM. Language models in nutrition and dietetics: a scoping review. Am J Clin Nutr 2026; 123: 101127.
7. van Wegberg AMJ, MacDonald A, Ahring K, Bélanger-Quintana A, Beblo S, Blau N, Bosch AM, Burlina A, Campistol J, Coşkun T, Feillet F, Giżewska M, Huijbregts SC, Leuzzi V, Maillot F, Muntau AC, Rocha JC, Romani C, Trefz F, van Spronsen FJ. European guidelines on diagnosis and treatment of phenylketonuria: First revision. Mol Genet Metab 2025; 145: 109125.
8. Chinsky JM, Singh R, Ficicioglu C, van Karnebeek CDM, Grompe M, Mitchell G, Waisbren SE, Gucsavas-Calikoglu M, Wasserstein MP, Coakley K, Scott CR. Diagnosis and treatment of tyrosinemia type I: a US and Canadian consensus group review and recommendations. Genet Med 2017; 19.
9. Boy N, Mühlhausen C, Maier EM, Heringer J, Assmann B, Burgard P, Dixon M, Fleissner S, Greenberg CR, Harting I, Hoffmann GF, Karall D, Koeller DM, Krawinkel MB, Okun JG, Opladen T, Posset R, Sahm K, Zschocke J, Kölker S; Additional individual contributors. Proposed recommendations for diagnosing and managing individuals with glutaric aciduria type I: second revision. J Inherit Metab Dis 2017; 40: 75-101.
10. Frazier DM, Allgeier C, Homer C, Marriage BJ, Ogata B, Rohr F, Splett PL, Stembridge A, Singh RH. Nutrition management guideline for maple syrup urine disease: an evidence- and consensus-based approach. Mol Genet Metab 2014; 112: 210-217.
11. Häberle J, Burlina A, Chakrapani A, Dixon M, Karall D, Lindner M, Mandel H, Martinelli D, Pintos-Morell G, Santer R, Skouma A, Servais A, Tal G, Rubio V, Huemer M, Dionisi-Vici C. Suggested guidelines for the diagnosis and management of urea cycle disorders: First revision. J Inherit Metab Dis 2019; 42: 1192-1230.
12. Forny P, Hörster F, Ballhausen D, Chakrapani A, Chapman KA, Dionisi-Vici C, Dixon M, Grünert SC, Grunewald S, Haliloglu G, Hochuli M, Honzik T, Karall D, Martinelli D, Molema F, Sass JO, Scholl-Bürgi S, Tal G, Williams M, Huemer M, Baumgartner MR. Guidelines for the diagnosis and management of methylmalonic acidaemia and propionic acidaemia: First revision. J Inherit Metab Dis 2021; 44: 566-592.
13. Welling L, Bernstein LE, Berry GT, Burlina AB, Eyskens F, Gautschi M, Grünewald S, Gubbels CS, Knerr I, Labrune P, van der Lee JH, MacDonald A, Murphy E, Portnoi PA, Õunap K, Potter NL, Rubio-Gozalbo ME, Spencer JB, Timmers I, Treacy EP, Van Calcar SC, Waisbren SE, Bosch AM; Galactosemia Network (GalNet). International clinical guideline for the management of classical galactosemia: diagnosis, treatment, and follow-up. J Inherit Metab Dis 2017; 40: 171-176.
14. Úbeda F, Santander S, Luesma MJ. Clinical Practice Guidelines for the Diagnosis and Management of Hereditary Fructose Intolerance. Diseases 2024; 12: 44.
15. John TA, Anastasopoulou C. Glycogen Storage Disease. [Updated 2025 Jan 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459277/
16. LoPiccolo MK, Vockley J. Carnitine Palmitoyltransferase II Deficiency. 2004 Aug 27 [Updated 2026 Apr 9]. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1253/
17. Van Calcar SC, Sowa M, Rohr F, Beazer J, Setlock T, Weihe TU, Pendyal S, Wallace LS, Hansen JG, Stembridge A, Splett P, Singh RH. Nutrition management guideline for very-long chain acyl-CoA dehydrogenase deficiency (VLCAD): An evidence- and consensus-based approach. Mol Genet Metab 2020; 131: 23-37.
18. Prasun P, LoPiccolo MK, Ginevic I. Long-Chain Hydroxyacyl-CoA Dehydrogenase Deficiency/Trifunctional Protein Deficiency. 2022. In: Adam MP, Bick S, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2026. Available from: https://www.ncbi.nlm.nih.gov/books/NBK583531/
19. Shvetsova O, Katalshov D, Lee SK. Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment. Appl Sci 2025; 15: 7298.
20. Hong J, Kikuta NT, Simos A, Tsai S, Lin B, Rodriguez F, Palaniappan L. Testing the Appropriateness of Diabetes Prevention and Care Information Given by the Online Conversational AI ChatGPT. Clin Diabetes 2023; 41: 549-552.
21. Huang C, Chen L, Huang H, Cai Q, Lin R, Wu X, Zhuang Y, Jiang Z. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions. J Transl Med. 2023; 21: 502.
22. Naja F, Taktouk M, Matbouli D, Khaleel S, Maher A, Uzun B, Alameddine M, Nasreddine L. Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome. Eur J Clin Nutr 2024; 78: 887-896.
23. Karakas PE, Calik A, Bilen AB, Kandemir K, Alphan ME. Large Language Models as Clinical Nutrition Decision Tools: Quantitative Bias and Guideline Deviation in Type 2 Diabetes Meal Planning. Healthcare (Basel) 2026; 14: 739.
24. Qarajeh A, Tangpanithandee S, Thongprayoon C, Suppadungsuk S, Krisanapan P, Aiumtrakul N, Garcia Valencia OA, Miao J, Qureshi F, Cheungpasitporn W. AI-Powered Renal Diet Support: Performance of ChatGPT, Bard AI, and Bing Chat. Clin Pract 2023; 13: 1160-1172.
25. Kairat M, Adilmetova G, Ibraimova I, Gaipov A, Varol HA, Chan MY. Benchmarking ChatGPT and Other Large Language Models for Personalized Stage-Specific Dietary Recommendations in Chronic Kidney Disease. J Clin Med 2025; 14: 8033.
26. Aydin Cil M, Karatas N, Mustafaoglu O, Sevinc C. Comparison of AI-generated renal diets by different large language models: a guideline-based evaluation with expert input. BMC Nephrol 2026; 27: 133.
27. Bertin L, Branchi F, Ciacci C, Lee AR, Sanders DS, Trott N, Zingone F. Efficacy of Large Language Models in Providing Evidence-Based Patient Education for Celiac Disease: A Comparative Analysis. Nutrients 2025; 17: 3828.

28. Peled N, Shouval DS, Gillett P, Szajewska H, Valitutti F, Shamir R, Guz-Mark A. Performance of large language models in answering frequently-asked questions on celiac disease. J Pediatr Gastroenterol Nutr 2026; 82: 1072-1077.
29. Yolcu F, Koç Yekedüz M, Köse E, Gülhan Samur F, Eminoğlu FT. Evaluation of scientific validity and appropriateness of artificial intelligence-assisted ChatGPT advices in dietary treatment of methylmalonic acidemia. Nutr Hosp 2025; 42: 910-915.
30. Liu M, Okuhara T, Shirabe R, Nishiie Y, Chang X, Okada H, Kiuchi T. Validity and reliability of ChatGPT's responses on dietary supplements in Japan: A quality assessment and content analysis. PEC Innov 2026; 8: 100461.
31. Ferraris AF, Audrito D, Di Caro L, Poncibò C. The architecture of language: Understanding the mechanics behind LLMs. Cambridge Forum on AI: Law and Governance 2025; 1: e11.
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv [preprint]. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03762
33. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering 2023; 25: 51-65.
34. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive mirage: A review of hallucinations in large language models. arXiv [preprint]. 2023. Available from: https://doi.org/10.48550/arXiv.2309.06794
35. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care 2023; 27: 120.
36. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering in large language models: A comprehensive review. Patterns 2025; 6.
37. Wang D, Zhang S. Large language models in medical and healthcare fields: Applications, advances, and challenges. Artif Intell Rev 2024; 57: 299.
38. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Hu S, Chen Y, Chan CM, Chen W, Yi J, Zhao W, Wang X, Liu Z, Zheng HT, Chen J, Liu Y, Tang J, Li J, Sun M. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 2023; 5: 220-235.
39. Abdur Rahman L, Dedousis V, Papathanail I, Poursoleymani R, Kafyra M, Kalafati IP, Mougiakakou SG. Generative AI in Precision Nutrition: A Review of Current Developments and Future Directions. Nutrients 2026; 18: 938.
40. Gjorgjevikj A, Martinc M, Cenikj G, Stojanov R, Drole J, Ispirova G, Menichetti G, Ogrinc N, Trajanov D, Džeroski S, Koroušić Seljak B, Eftimov T. Large language models in food and nutrition science: Opportunities, challenges, and the case of FoodyLLM. Curr Res Food Sci 2026; 12: 101351.