scientific reports

Check for updates

OPEN

Evaluation of LLMs accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval

Iman Azimi[1]✉, Mohan Qi[1], Li Wang[2], Amir M. Rahmani[3] & Youlin Li[1]

Large language models (LLMs) are fundamentally transforming human-facing applications in the health and well-being domains: boosting patient engagement, accelerating clinical decision-making, and facilitating medical education. Although state-of-the-art LLMs have shown superior performance in several conversational applications, evaluations within nutrition and diet applications are still insufficient. In this paper, we propose to employ the Registered Dietitian (RD) exam to conduct a standard and comprehensive evaluation of state-of-the-art LLMs, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, assessing both accuracy and consistency in nutrition queries. Our evaluation includes 1050 RD exam questions encompassing several nutrition topics and proficiency levels. In addition, for the first time, we examine the impact of Zero-Shot (ZS), Chain of Thought (CoT), Chain of Thought with Self Consistency (CoT-SC), and Retrieval Augmented Prompting (RAP) on both accuracy and consistency of the responses. Our findings revealed that while these LLMs obtained acceptable overall performance, their results varied considerably with different prompts and question domains. GPT-4o with CoT-SC prompting outperformed the other approaches, whereas Gemini 1.5 Pro with ZS recorded the highest consistency. For GPT-4o and Claude 3.5, CoT improved the accuracy, and CoT-SC improved both accuracy and consistency. RAP was particularly effective for GPT-4o to answer Expert level questions. Consequently, choosing the appropriate LLM and prompting technique, tailored to the proficiency level and specific domain, can mitigate errors and potential risks in diet and nutrition chatbots.

Keywords Large Language Models, Registered Dietitian, Nutrition, Prompt Engineering, Knowledge Retrieval

There is growing interest in leveraging conversational models, commonly known as chatbots, in healthcare, particularly in the areas of diet and nutrition[1–3]. The rise of large language models (LLMs) is significantly transforming human-machine interactions in this context, creating new opportunities for nutrition management applications and lifestyle enhancement that involve natural language understanding and generation[4–6]. These chatbots can serve as assistants to health providers (e.g., dietitian or nurses) or as ubiquitous companions for patients, providing preventive care, personalized meal planning, and chronic disease management[7].

Since the release of ChatGPT[8]in November 2022, numerous nutrition management studies have developed or employed LLM-based chatbots to target different health conditions, such as type 2 diabetes, obesity, liver diseases, kidney diseases, and cardiovascular diseases, to mention a few[1,7,9–16]. These studies highlight the potential of chatbots interventions to enhance diet and promote lifestyle behavior changes.

Due to the life-critical nature of these applications, they must provide high quality attributes, such as accuracy, consistency, safety, and fairness, before being deployed in real-world settings for end-users[17–19]. Recent studies have evaluated the LLM-based chatbots within nutritional and dietary contexts. For example, Sun *et al.*[20] and Barlas *et al.*[21]assessed the performance of ChatGPT in providing nutritional management support for diabetic

[1]Department of Engineering, iHealth Labs, Sunnyvale, CA 94085, United States. [2]Department of Clinical Research, iHealth Labs, Sunnyvale, CA 94085, United States. [3]School of Nursing and Department of Computer Science, University of California Irvine, Irvine, CA 92697, United States. ✉email: iman.azimi@ihealthlabs.com

patients. Other investigations focused on chatbots' reliability in delivering accurate calorie and macronutrient information[22,23]. For non-communicable diseases, the accuracy of dietary advice generated by ChatGPT's were assessed[10,24]. Other studies also examined ChatGPT's ability to address common nutrition-related inquiries, highlighting its strength and weakness in offering personalized and accurate nutritional information[25,26]. However, the existing evaluation studies on nutrition-related chatbots face three major challenges.

First, prior research on the LLMs application in nutrition has relied solely on ad-hoc or subjective evaluations. In these studies, domain experts designed a set of questions focused on specific diseases or nutrition topics. Subsequently, human evaluators were instructed to grade the responses in terms of accuracy, comprehensiveness, or attractiveness[20,21,27]. Human-in-the-loop evaluation is widely recognized as a popular and well-established strategy for assessing chatbots in the literature[18,19]. However, these evaluations are not comprehensive regarding nutrition problems and are prone to human errors or biases, as they depend on the opinion of an individual expert, especially when no standard guidelines are followed in the evaluation process. Additionally, they are time-consuming and costly. This limitation can be observed in the current nutrition chatbots evaluation, as their assessments are restricted to a few hundred interactions (i.e., prompts) at most.

Second, most of the nutrition and diet studies have focused only on ChatGPT-3.5 or ChatGPT-4[7,11,24,27]. The landscape of LLMs is rapidly evolving. New models and techniques are being released frequently, within weeks or months[28]. This rapid advancement requires the evaluation of a wide range of models to ensure the best possible solutions for diet and nutrition management applications. In addition, existing research on nutrition evaluation has ignored the impact of prompt engineering techniques[20,24,29]. They have been limited to zero-shot prompting methods with either no instructions or fixed instructions. The zero-shot prompting instructs LLMs to perform specific tasks without providing any prior examples. This technique is straightforward and is widely used. However, it might be insufficient for LLM response generation if problem-solving or contextual information is needed. Chain of thought and chain of thought with self-consistency prompting techniques have shown their potential to enhance the performance of LLMs in multiple non-nutrition studies by enabling chatbots to address complex reasoning tasks[30–32]. Retrieval augmented prompting models have also indicated their effectiveness in mitigating LLMs hallucination problems across generic scenarios. We hypothesize that these step-by-step reasoning techniques and retrieval models can surpass zero-shot prompting techniques, especially in breaking down complex nutritional questions, handling uncertainties by providing external information, and enabling better decision-making[33–36]. It is essential to investigate the impact of these prompting techniques on LLMs performance in handling various nutrition-related queries.

Third, previous work merely focused on the overall accuracy of LLMs responses. Their findings indicated that the models were generally accurate, but they still had errors[10,21,24,27]. These studies did not examine the errors, along with the strategies to enhance the LLMs' responses. Wang *et al.*[31]highlights this issue in the context of clinical medicine. Moreover, the non-deterministic behavior of chatbots was ignored[37]. The consistency and reliability of chatbots in answering nutrition-related questions must be evaluated to determine if their performance varies with identical or different prompts. In the nutrition context, to the best of our knowledge, only one study[22] has explored the consistency of ChatGPT-3.5 and ChatGPT-4 responses, using a zero-shot prompt for 222 food items across five repeated measurements.

In this paper, we thoroughly evaluate the accuracy and consistency of three leading LLMs chatbots, i.e., GPT-4o[38], Claude 3.5 Sonnet[39], and Gemini 1.5 Pro[40], in addressing nutrition-related inquiries. To achieve this, we leverage the Registered Dietitian (RD) exam[41] for the first time, as a standard certification examination that serves to assess whether dietitians meet the qualifications required to practice in the dietetics and nutrition field. Our evaluation includes 1050 multiple-choice questions with different proficiency levels, covering four nutrition domains: i.e., principles of dietetics, nutrition care, food service systems, and food and nutrition management. To investigate the impact of prompts, the questions are presented to the LLMs using four different prompting techniques. We then compare the responses with the ground truth answers, enabling an objective assessment of the model's performance. To examine the consistency of the responses, we perform repeated measurements by asking each model the same set of questions multiple times using each prompting technique. The responses for each technique and model are compared within and across groups.

Results

Accuracy

Overall performance

The results show that all the approaches obtained a score of over 88% in selecting the correct option for the 1050 RD exam questions, as indicated in Figure 1 and Table 1. Table 2 also summarizes the error counts for each prompting technique across the three chatbots. Overall, GPT-4o achieved the highest score (the blue markers in the figure) ranging between 91% and 95%, with the best score for CoT-SC. GPT-4o with CoT-SC obtained the lowest error count: i.e., 58 errors on average on the 1050 RD exam questions. On the other hand, Gemini 1.5 Pro (the green markers in Figure 1) had the lowest scores. The highest error count was for Gemini 1.5 Pro with CoT.

In both GPT-4o and Claude 3.5 Sonnet, the CoT and CoT-SC prompting techniques resulted in similar percentage scores, which were approximately 2.5 percent higher than the ZS prompting's scores. The error count of GPT-4o with ZS was approximately 25 higher than that observed for CoT and CoT-SC. However, the combination of Gemini with CoT or CoT-SC did not improve the accuracy but produced wider percentage scores across repeated measurements, with ranges of 1.9 and 1.2. The error count of Gemini with ZS was approximately 9 and 18 fewer than that observed for CoT and CoT-SC. Moreover, RAP obtained better scores (less error counts), compared to ZS, in GPT-4o; however it slightly decreased the performance of Claude and Gemini models.

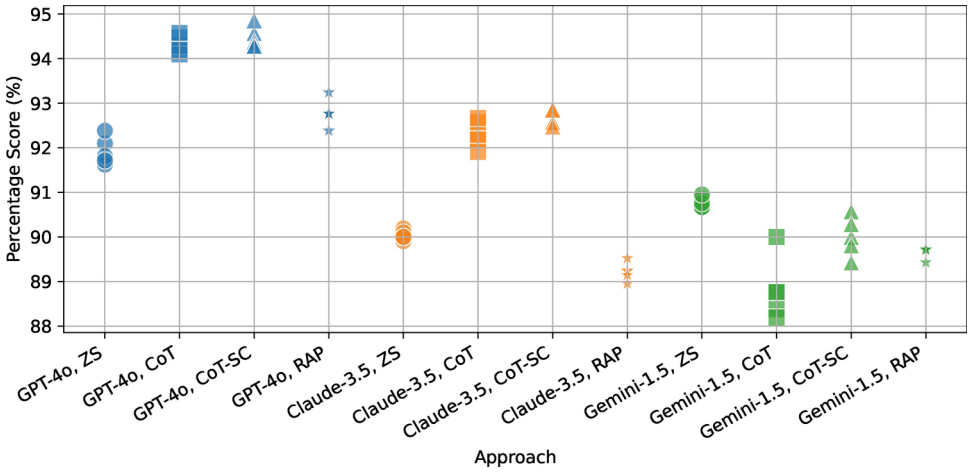


Fig. 1. Percentage Scores of the approaches on the RD exam. GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro are indicated with blue, orange, and green markers, respectively. The Zero Shot (ZS), Chain of Thought (CoT), Chain of Thought with Self Consistency (CoT-SC), and Retrieval Augmented Prompting (RAP) techniques are indicated with circle, square, triangle, and star markers, respectively.

Benchmark	Prompt	GPT-4o	Claude 3.5 S.	Gemini 1.5 P.
RD Exam	Zero Shot	**91.92% (0.28)**	90.04% (0.10)	90.78% (0.11)
	Chain of Thought	**94.32% (0.18)**	92.32% (0.27)	88.82% (0.63)
	Chain of Thought w. Self Consistency	**94.48% (0.22)**	92.67% (0.16)	90.02% (0.39)
	Retrieval Augmented Prompting	**92.78% (0.27)**	89.22% (0.18)	89.66% (0.11)

Table 1. The percentage scores (mean and standard deviation) of the LLMs' responses on the RD exam questions.

Benchmark	Prompt	GPT-4o	Claude 3.5 S.	Gemini 1.5 P.
RD Exam	Zero Shot	**84.8 (2.93)**	104.6 (1.02)	96.8 (1.17)
	Chain of Thought	**59.6 (1.85)**	80.6 (2.87)	117.4 (6.62)
	Chain of Thought w. Self Consistency	**58.0 (2.28)**	77.0 (1.67)	104.8 (4.12)
	Retrieval Augmented Prompting	**75.8 (2.86)**	113.2 (1.94)	108.6 (1.20)

Table 2. The error counts (mean and standard deviation) of the LLMs' responses on the 1050 RD exam questions.

Subgroup error analysis

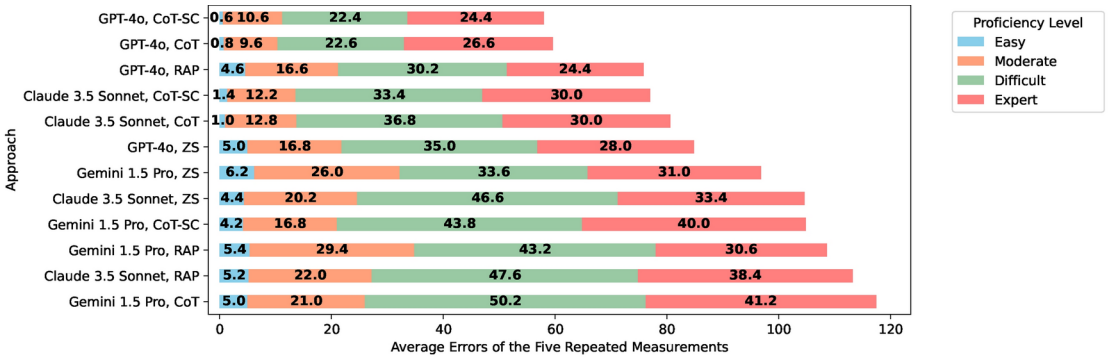
We categorize the RD exam questions into different subgroups, within which the LLMs' inaccurate responses are assessed. To achieve this, we analyze the errors obtained in terms of proficiency levels and four nutrition domains (i.e., topics).

Proficiency Levels: The approaches are evaluated based on the questions' proficiency levels, provided by the Academy of Nutrition and Dietetics, eatrightPREP for the RDN Exam[42]. The exam consists of 149 Easy, 352 Moderate, 392 Difficult, and 149 Expert levels questions. Figure 2a shows the average errors for each approach. In addition, the mean and standard deviation of the error counts of the approaches per proficiency level are indicated in Supplementary Table S.1.

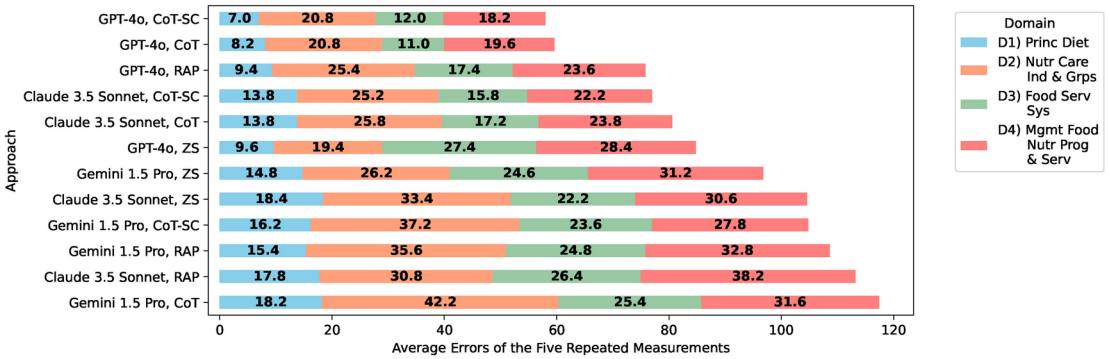
GPT-4o obtained the lowest overall average error counts. The model with CoT-SC resulted in the fewest errors across the proficiency levels, with the average errors of 0.6, 10.6, 22.4, and 24.4 for Easy, Moderate, Difficult, and Expert levels questions, respectively. Compared to ZS prompting, CoT and CoT-SC improved the model's performance at all levels, but RAP only enhanced the responses of the Difficult and Expert level questions.

Similar to the GPT-4o approaches, Claude 3.5 Sonnet performance was enhanced by CoT and CoT-SC. Claude 3.5 Sonnet with CoT and CoT-SC achieved similar average error rates. Conversely, using Claude 3.5 Sonnet, RAP recorded the highest error counts, particularly with 5 more errors (on average) for Expert questions, compared to the ZS prompting technique.

Gemini 1.5 Pro had the highest number of errors overall. The ZS prompting recorded the lowest average errors with Gemini. Compared to ZS, CoT and CoT-SC improved the responses of the Moderate questions but



(a) Average errors per approach by proficiency level. The exam includes 149 Easy, 352 Moderate, 392 Difficult, and 157 Expert levels questions.



(b) Average errors per approach by domain. The exam includes 237 principles of dietetics, 392 nutrition care for individuals and groups, 185 food service systems, and 236 management of food and nutrition programs and services questions.

Fig. 2. The LLMs' inaccurate responses based on the RD exam questions' proficiency levels and domains.

obtained higher average errors for the Difficult and Expert level questions. RAP obtained higher error rates for Moderate and Difficult questions.

Domains: The inaccurate responses collected by each approach is evaluated based on four domains: *D1) Principles of Dietetics*, *D2) Nutrition Care for Individuals and Groups*, *D3) Food Service Systems*, and *D4) Management of Food and Nutrition Programs and Service*. The exam consists of 237, 392, 185, and 236 questions for D1, D2, D3, and D4, respectively. As illustrated in Figure 2b, the impact of prompt engineering techniques varied across the domains for the three LLMs. The mean and standard deviation of the error counts of the approaches per domain are indicated in Supplementary Table S.2.

GPT-4o with CoT-SC reduced the average error counts in D3 from 27.4 to 12 and in D4 from 28.4 to 18.2, compared to GPT-4o with ZS. CoT and RAP also showed similar improvements in error rates although RAP recorded more errors for D2. Using GPT-4o, different prompting techniques resulted in small changes in the error rates observed in D1.

Claude 3.5 Sonnet showed that transitioning from ZS prompting to CoT-SC or CoT reduced the average errors across the four domains. On the other hand, RAP slightly improved D1 and D2 but obtained more errors in D3 and D4.

With Gemini 1.5 Pro, different prompts led to small variations in error counts, with changes of fewer than 4 errors on average in D1, D3, and D4. However, ZS prompting obtained the lowest error count in D2, with an average of 26.2 errors. Nevertheless, this outcome shows approximately 6 errors higher than the performance achieved by GPT-4o. In D2, Gemini and CoT obtained the highest error rates.

Consistency

Inter-rater analysis

The inter-rater reliability of the responses from the approaches was analyzed to investigate their agreement. To achieve this, Cohen's Kappa coefficient was calculated for each pair of approaches to determine if they selected the same choices, whether accurate or inaccurate. Our study includes 12 distinct approaches (3 LLMs multiplied by 4 prompting techniques), so the tests were performed for each of the 12 pairwise comparisons. Since each approach is repeated five times, one set of measurements per approach is randomly selected to assess the inter-rater reliability. Figure 3presents the Cohen's Kappa coefficients, where dark blue indicates high levels of agreement, and light blue represents lower agreement levels. Additionally, we utilized the McNemar-Bowker test[43] to determine if each pair of responses are statistically different. Table 3 indicates the test results for responses collected from the three chatbots with different prompts. The detailed statistical data for the

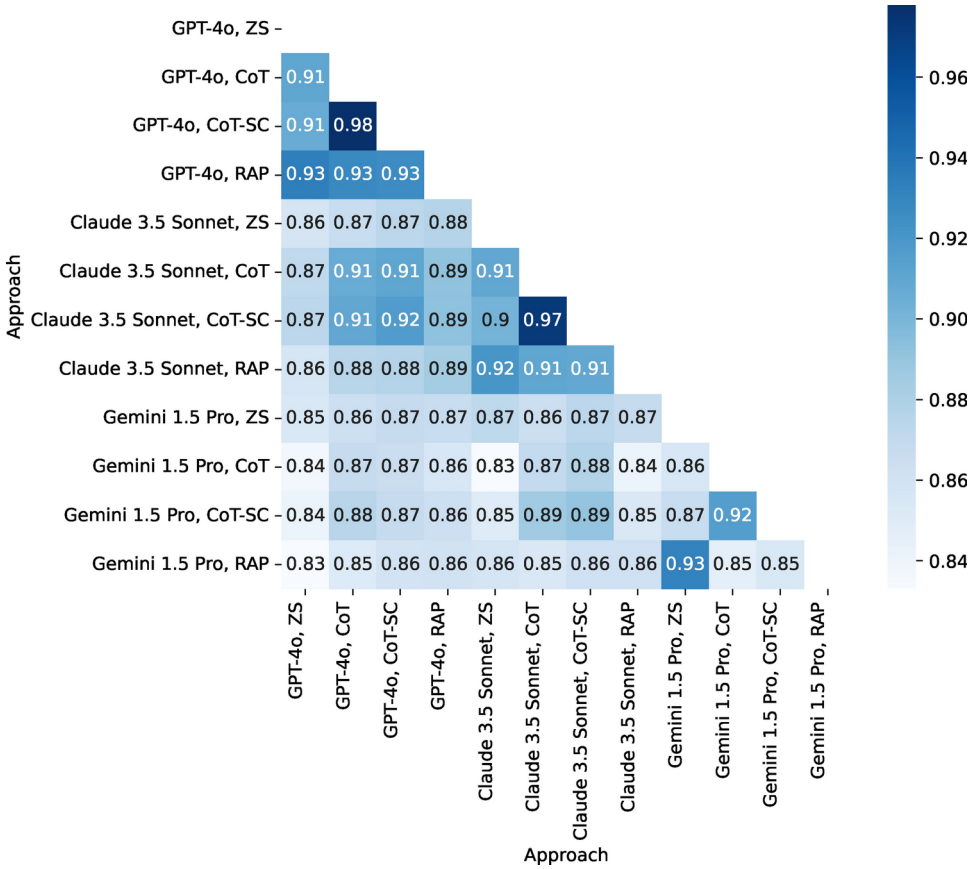


Fig. 3. The Cohen's Kappa coefficients measured for each of the 12 pairwise comparisons using the RD exam. The dark blue indicates high levels of agreement, while the light blue represents lower agreement levels.

LLMs with different prompts	Test Statistic	P value
GPT-4o, ZS vs CoT	7.20	0.303
GPT-4o, ZS vs CoT-SC	6.83	0.337
GPT-4o, ZS vs RAP	9.20	0.162
GPT-4o, CoT vs CoT-SC	7.53	0.274
GPT-4o, CoT vs RAP	10.23	0.115
GPT-4o, CoT-SC vs RAP	13.05	0.042*
Claude 3.5 S., ZS vs CoT	17.24	0.008*
Claude 3.5 S., ZS vs CoT-SC	19.82	0.003*
Claude 3.5 S., ZS vs RAP	2.26	0.894
Claude 3.5 S., CoT vs CoT-SC	10.13	0.119
Claude 3.5 S., CoT vs RAP	20.56	0.002*
Claude 3.5 S., CoT-SC vs RAP	23.22	0.001*
Gemini 1.5 P., ZS vs CoT	7.92	0.244
Gemini 1.5 P., ZS vs CoT-SC	8.07	0.233
Gemini 1.5 P., ZS vs RAP	5.55	0.475
Gemini 1.5 P., CoT vs CoT-SC	2.10	0.910
Gemini 1.5 P., CoT vs RAP	13.09	0.042*
Gemini 1.5 P., CoT-SC vs RAP	13.67	0.034*

Table 3. McNemar-Bowker test results for responses collected from the LLMs with different prompts. * indicates that the differences are statistically significant ($p < 0.05$).

LLM	Prompt	Fleiss' Kappa	95% CI
GPT-4o	Zero Shot	0.980	0.973 – 0.987
	Chain of Thought	0.969	0.960 – 0.977
	Chain of Thought w. Self Consistency	0.977	0.970 – 0.985
	Retrieval Augmented Prompting	0.985	0.978 – 0.991
Claude 3.5 S.	Zero Shot	0.987	0.981 – 0.992
	Chain of Thought	0.975	0.967 – 0.983
	Chain of Thought w. Self Consistency	0.982	0.975 – 0.988
	Retrieval Augmented Prompting	0.977	0.970 – 0.985
Gemini 1.5 P.	Zero Shot	0.996	0.993 – 0.999
	Chain of Thought	0.902	0.887 – 0.917
	Chain of Thought w. Self Consistency	0.938	0.926 – 0.951
	Retrieval Augmented Prompting	0.991	0.987 – 0.996

Table 4. The Fleiss Kappa coefficients of the 12 approaches. Each approach was repeated 5 times.

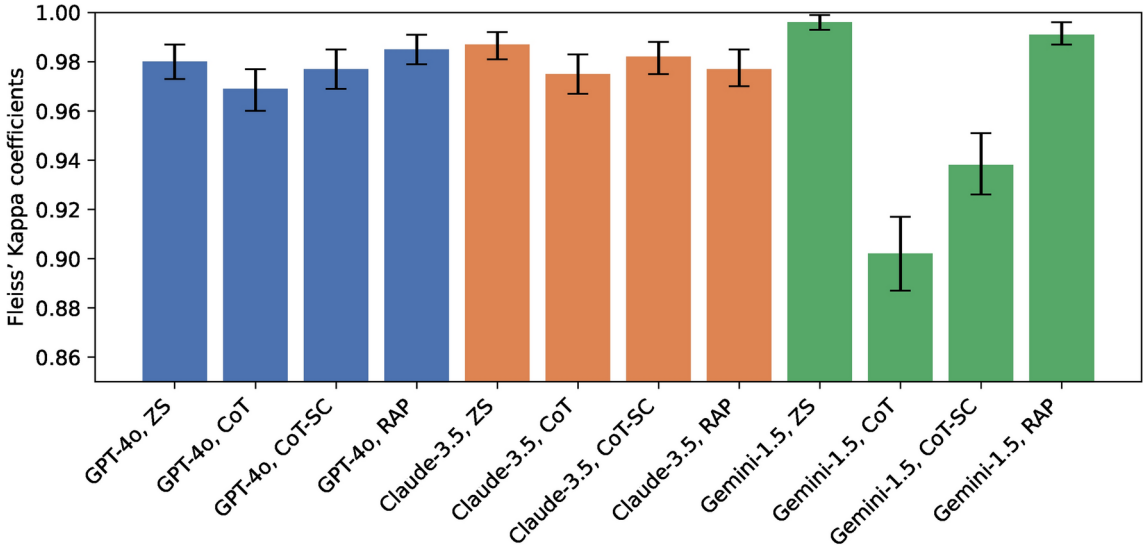


Fig. 4. The Fleiss Kappa coefficients of the 12 approaches.

Cohen's kappa and McNemar-Bowker tests, including 95% confidence intervals and P-values, are presented in Supplementary Table S.3.

The approaches based on GPT-4o showed a high degree of agreement, indicated by a Cohen's Kappa coefficient of 0.98 between CoT and CoT-SC and a coefficient of 0.93 between RAP and the other three prompting techniques. The McNemar-Bowker tests also indicated that there were no statistically significant differences in the paired responses, except for CoT-SC and RAP ($P\ value = 0.042$). For the Claude 3.5-based approaches, the Cohen's Kappa coefficients were slightly lower compared to GPT-4o. However, the statistical tests showed that the responses of ZS and RAP were significantly different from CoT and CoT-SC (see Table 3). The Gemini 1.5 Pro's approaches recorded relatively lower Cohen's Kappa coefficients, despite maintaining high overall agreement. The Cohen's Kappa coefficients of the Gemini-based approaches were from 0.84 to 0.93. The agreement level between RAP and CoT / CoT-SC were 0.85. The statistical tests also showed that the responses of RAP were significantly different from CoT and CoT-SC. Interestingly, among the prompting techniques, the approaches (even with different LLMs) using CoT and CoT-SC obtained higher levels of agreement. The McNemar-Bowker tests also indicated that there were no statistically significant differences between the CoT and CoT-SC responses in the three LLMs.

Intra-rater analysis

In this study, each approach was repeated five times, resulting in five sets of responses. The intra-rater reliability of the responses was evaluated by measuring the repeatability of the approaches, determining how consistently they agreed with themselves when receiving the same questions. For this purpose, Fleiss Kappa was employed to assess the intra-rater agreements. Table 4 and Figure 4 indicate the Fleiss Kappa coefficients, and Supplementary Table S.4 includes the detailed statistical data.

Gemini 1.5 Pro combined with the ZS prompting achieved the highest agreement among all combinations, whereas the Gemini with CoT produced the lowest agreement. The approaches based on Claude 3.5 Sonnet demonstrated the highest overall agreement. For the three LLMs, the ZS prompting technique consistently resulted in the highest agreement, as indicated by Fleiss's Kappa coefficients of 0.996, 0.987, and 0.980 for Gemini 1.5 Pro, Claude 3.5 Sonnet, and GPT-4o, respectively. Similarly, the coefficients of the LLMs with RAP were high. The CoT-SC recorded the third highest agreement, while the CoT obtained the lowest.

Discussion

Our findings indicated that all the approaches, combining three LLMs with four prompt engineering techniques, successfully passed the RD exam and obtained a score of above 88%. In our tests, the combination of GPT-4o with CoT-SC prompting outperformed the other approaches in terms of accuracy, while Gemini 1.5 Pro with ZS prompting showed the highest consistency. On the other hand, the lowest average percentage score was 89.22% for Gemini 1.5 Pro with CoT, which also showed the lowest agreement in repeated measurements, with a coefficient of 0.902. GPT-4o recorded the highest accuracy overall (see Table 1).

In practice for dietitians taking the RD exam, the exam is scored from 1 to 50, and the minimum score to pass is 25[44]. In the exam, the questions might be weighted differently, and the score is calculated based on the candidate's performance as well as the difficulty levels of the questions. According to the RD Exam Pass/Fail Statistics published by the Commission on Dietetic Registration, the grand total first-attempt pass rate for the RD exam from January to June 2024 was 61.5% and the total first-attempt pass rate in 2023 was 88.4%[45].

Despite the success of the approaches to pass the RD exam, the three leading LLMs had different performance levels in terms of the number of inaccurate responses and consistency. Particularly, the prompting techniques had considerable impacts on the results. Such prompting impacts were also explored in other evaluation studies, for example, in clinical medicine[31], mental health[46], and radiology[47]. We observed that GPT-4o showed more consistent behavior in changing prompts, as indicated by its higher overall Cohen's Kappa values and only one significant difference in responses between CoT-SC and RAP. This is also supported by its high Fleiss Kappa values, which indicate its consistency in answering the same questions multiple times. The percentage scores of GPT-4o also indicate its robust accuracy when handling different prompt types. Claude 3.5 Sonnet demonstrated similar performance, with slightly lower Cohen's Kappa values and percentage scores. However, the responses of ZS and RAP were significantly different from the responses of CoT and CoT-SC. In contrast, changing the prompts had the most impact on Gemini's performance, with lowest Cohen's Kappa coefficients. Although Gemini with ZS was the most consistent approach, the model with CoT showed the lowest Fleiss Kappa values and percentage scores.

In addition to our findings, an overview of the models' performance on existing *non-nutrition-focused* knowledge and reasoning benchmarks are indicated in Table 5. The performance scores of these three benchmarks were collected from the following references[38,39,48]. The GPQA benchmark[49]includes 448 multiple-choice questions on biology, physics, and chemistry. The MMLU benchmark[50]contains multiple-choice questions from 57 topics, such as elementary mathematics, US history, computer science, and law; and the DROP benchmark[51] consists of 96,567 questions focusing on discrete reasoning over the content of paragraphs, including addition, counting, and sorting. Claude 3.5 Sonnet outperformed the other LLMs in all scenarios, except for MMLU using the ZS prompting.

Our findings presented in Table 1contrasts with these previous non-nutrition research, except in MMLU[50]with ZS prompting. Claude 3.5 with CoT obtained a 59.4% score on GPQA[49]. However, the three LLMs using CoT on the RD exam achieved percentage scores above 90%. This difference might be due to the different difficulty levels of the exams. Particularly, 14.9% of the questions in the RD Exam are at the Expert level. However, as reported by Rein *et al.*[49], the GPQA questions are *"extremely difficult,"*from which PhD students achieved a 65% score while non-expert individuals achieved a 34% score. Moreover, DROP[51] demonstrated that Claude 3.5 with Three Shot prompting outperformed in reasoning over text. Conversely, our results indicated that GPT-4o performed better using the reasoning process of CoT prompting.

Prior *nutrition-focused* research indicated that ChatGPT was accurate in most nutrition instances, but the chatbot also recorded errors that could potentially harm and negatively impact the end-users. Therefore, achieving general accuracy is insufficient for practical real-world applications. For example, Sun *et al.*[20] indicated that ChatGPT-3.5 and ChatGPT-4 passed the Chinese RD exam (included 200 questions) and the food recommendations were acceptable despite the presence of mistakes for specific foods, such as root vegetables and dry beans. Mishra *et al.*[52] tested ChatGPT in eight medical nutritional therapy scenarios and discussed that ChatGPT should be avoided for complex scenarios. Naja *et al.*[29]evaluated ChatGPT's accuracy and quality in providing nutrition management for Type 2 diabetes. They highlighted that ChatGPT exhibited errors in responses, for example, in weight loss recommendation and the adoption of specific dietary interventions.

Benchmark	Prompt	GPT-4o	Claude 3.5 S.	Gemini 1.5 P.
MMLU (Undergraduate Level Knowledge)	Zero Shot	**88.70%**	88.30%	-
	Five Shot	-	**88.70%**	85.90%
GPQA (Graduate Level Reasoning)	Chain of Thought	53.60%	**59.40%**	46.20%
DROP (Reasoning)	Three Shot	83.40%	**87.10%**	74.90%

Table 5. The performance of the LLMs on the MMLU[50], GPQA[49], and DROP[51]benchmarks, collected from the following references[38,39,48].

Similarly, other studies[10,24]discussed that ChatGPT has great potential for nutritional management focusing on non-communicable diseases, but the model might be potentially harmful by providing inaccurate responses, particularly in complex situations. Another study[22]leveraged ChatGPT-3.5 and ChatGPT-4 to provide nutritional information for eight menus. Their results indicated that responses had no significant differences compared to nutritionists' recommendations in terms of energy, carbohydrate, and fat contents, but the difference was statistically significant for protein. The potential of ChatGPT to generate dietary advice for patients with allergic to food allergens were also investigated[27]. It was shown that although the model was generally accurate, it produced harmful diets. These studies highlight the need for further investigation into LLM responses within the context of food and nutrition.

Our results confirmed previous findings about the overall accuracy of ChatGPT and the instances of inaccurate responses. However, unlike the existing work, our study is not merely restricted to ChatGPT or the ZS prompting technique. We also focused on examining errors across various subcategories and mitigate them by employing prompting techniques (reasoning and ensemble) and external knowledge retrieval.

In summary, we observed that GPT-4o with CoT-SC and CoT obtained the best performance across all the proficiency levels. GPT-4o with CoT-SC resulted in 0.6 errors (on average) for the 149 easy questions, 22.4 errors for the 392 difficult questions, and 24.4 errors for the 157 expert-level questions. CoT obtained the least errors (i.e., 9.6) for the 352 moderate-level questions. GPT-4o was also the most accurate model across all domains. GPT-4o with CoT-SC recorded only 7 errors in the *D1) principles of dietetics* questions and 18.2 errors in the *D4) food and nutrition management* questions. GPT-4o with CoT obtained the fewest errors (i.e., 11.0) in the *D3) food service systems* questions, and GPT-4o with ZS had the fewest errors (i.e., 19.4) in the *D2) nutrition care* questions. Figure 2 and Supplementary Tables S.1 and S.2 indicate each model's performance across proficiency levels and domains. In the following, we discuss how the prompting technique influenced the LLMs responses in more detail.

- **CoT** guided LLMs to perform a reasoning process when answering a question. Our findings showed that CoT, compared to ZS prompting, enhanced the accuracy of GPT-4o and Claude 3.5 Sonnet but led to diminished consistency. CoT did not consistently generate the same reasoning paths (lower Fleiss Kappa coefficients), even with identical prompts (see Figure 4). This variability indicates randomness in the selection of reasoning paths. We observed that the reasoning steps of CoT considerably reduced the LLMs' (particularly GPT-4o and Claude 3.5 Sonnet) mistakes for the questions with easy, moderate, and difficult proficiency levels. This demonstrates the effectiveness of this prompting technique in enhancing the chatbots' performance to handle a wide range of nutrition question complexities by breaking down problems into reasoning steps. However, this improvement was less for the expert-level questions. At the expert proficiency level, where questions required deeper understanding and reasoning, only a small number of errors were corrected, indicating the limitations of CoT in complex nutrition scenarios. It should be noted that the combination of Gemini 1.5 Pro with CoT showed different patterns, where both accuracy and consistency decreased. Gemini with CoT was unable to select a choice from the given multiple-choice options for 20 out of 1050 questions (on average). Although the errors on easy and moderate levels questions slightly decreased, the errors on difficult and expert levels questions notably increased. Additionally, CoT notably improved the questions about *D3) food service systems*, which involved calculations for food cost and portion estimation/forecasting. CoT also enhanced the accuracy of *D4) food and nutrition management*, which included theoretical and conceptual questions requiring an understanding of implicitly stated relationships. These improvements by CoT are consistent with existing literature, indicating CoT enhances LLMs' performance in arithmetic and commonsense tasks by establishing logical connections[53]. Although CoT reduced errors in questions requiring calculations, our observations indicate that CoT responses still include miscalculations and rounding errors. This issue may arise due to the inherent characteristics of Transformer models, designed to generate text based on tokens rather than numerical values. Potential solutions to address these issues include agentic approaches[54,55], which integrate LLMs with calculator tools or symbolic computing systems.
- **CoT-SC** guided LLMs to perform multiple independent reasoning processes, then the responses were merged using a majority voting method. Our findings revealed that CoT-SC (compared to CoT) improved accuracy, particularly in Gemini 1.5 Pro. However, in GPT-4o and Claude 3.5, this improvement was small, as it only led to the correction of a few errors. This small difference can also be observed in their high inter-rater coefficient agreement, as illustrated in Figure 3. This finding does not support the literature suggesting that CoT-SC considerably enhances the accuracy of CoT[56]. Similar to CoT, the reasoning steps of CoT-SC notably reduced the GPT-4o errors for all the proficiency levels. GPT-4o with CoT-SC recorded fewest errors for the easy-, difficult-, and expert-levels questions. Claude 3.5 with CoT-SC improved the easy- moderate-, and difficult-levels questions, compared with CoT. Gemini with CoT-SC was considerably better than CoT but worse than ZS. It is worth noting that, in our analysis, we observed that the impact of CoT-SC (compared with CoT) was more on consistency (intra-rater agreement) rather than accuracy. The ensemble process enabled by CoT-SC mitigates the randomness in the selection of reasoning paths (which was observed in CoT). For GPT-4o and Claude 3.5 Sonnet, the Fleiss' Kappa agreements of CoT-SC were as robust as the agreements of ZS prompting. The Gemini's inability to select a choice from the given multiple-choice options also improved, reducing them from 20 in CoT to 6 in CoT-SC. This highlights the importance of employing such ensemble techniques to enhance the consistency of LLM's reasoning process by combining multiple reasoning paths rather than relying on a single path.
- **RAP** integrated external relevant information from multiple references into the input prompts. GPT-4o effectively leveraged the retrieved information to reduce error rates, particularly for Difficult and Expert questions that required more comprehensive understanding. Like CoT-SC, RAP recorded the fewest errors in the expert-level questions. Moreover, similar to CoT and CoT-SC, RAP improved *D3) food service systems* and *D4)*

food and nutrition management questions. Although relevant information was provided in our knowledge base, RAP (compared to ZS) has recorded higher error rates for *D2) nutrition care*. D2 questions are mostly related to medical nutrition therapy, dietary guidelines, counseling skills, and nutrition care process. In contrast to GPT-4o, Gemini 1.5 Pro and Claude 3.5 Sonnet with RAP showed opposite behavior, as the accuracy for the Difficult and Expert questions reduced. We noticed that, in some cases, Gemini was prioritizing external information over its own internal knowledge, even when that external information was irrelevant to the question. This resulted in incorrect interpretations and answers. For example, for two questions, the model generated "*The provided text does not contain the answer to the question as it pertains to dietary restrictions for patients on Linezolid*." and "*The provided text focuses on Body Mass Index (BMI) but does not contain information about when weight and BMI peak*."This issue was particularly observed in D2, where error rates increased from 26.2 (ZS) to 35.6 (RAP). It was anticipated that the LLMs' performance improved when using the external information. However, our findings showed that RAP did not consistently enhance accuracy across the three models. This higher error rates with RAP might arise from irrelevant retrieval, where the retrieval model fetches extraneous information[57]. As previously mentioned, we observed that the external information led Gemini and Claude to select a wrong option or fail to select an option. Additionally, the complexity or ambiguity of the queries might contribute to this problem making it challenging for the retrieval model to find the most relevant chunks.It is worth noting that the prompting techniques had less impact, whether positive or negative, on *D1) principles of dietetics* questions compared to the other domains. D1 questions primarily focus on general food science, nutrients, biochemistry, and related research (e.g. *which fruit has the highest fructose?*), compared to the other domains that are more specialized in dietetics or involve more domain knowledge. For D1, GPT-4o achieved the best accuracy.

Conversational models including LLM chatbots are expected to be broadly used for various nutrition-related tasks, such as diet recommendations and recipes generation[1,3]. Our findings offer insights into the readiness, potential, and limitations of these models, since the RD exam consists of a wide range of topics designed to assess a dietitian's competency. The performance of the chatbots on the RD exam can reflect their ability to comprehend nutrition queries, understand fundamentals of dietetics, and transform the knowledge into nutrition management, care, and reliable advice. These findings show the potential of LLM chatbots to be effectively used to support decision-making in practice and enhance dietitian support applications, including patient assessment, nutrition plan generation, and providing accurate answers to nutrition inquiries. Additionally, our results highlight the role of prompt engineering on improving the accuracy and consistency of the models. It can help clinicians interact with the models more effectively and provide insights for developers when designing nutrition chatbots. A high-performing and consistent LLM can enhance users trust and satisfaction while boosting safety by ensuring that generated responses are accurate and reliable. Consequently, using RD exam to benchmark LLMs can be a valuable approach to evaluate LLMs capabilities in the nutrition domain and their potential to be used in real-world dietitian support scenarios.

However, the LLM evaluation using the RD exam might have multiple limitations. First, our findings cannot show LLMs' performance in handling open-ended questions. In our evaluation, the model was instructed to choose one answer from provided options. Second, the results cannot indicate the models contextual understanding and personalization aspects. The accuracy, trustworthiness, and safety of a response might be highly dependent to user's profile, medical history, and personal situation. Another limitation of this test is the lack of assessment of nutrition literacy and the clarity of the responses. Additionally, the RD exam cannot assess bias and cultural insensitivity in responses that could lead to misunderstandings or mistrust, even though generated responses are accurate.

Future work in this direction will involve evaluating LLMs on open-ended questions and by leveraging patient-centric questions, answers, and conversations. Our evaluation has primarily concentrated on the accuracy and consistency of the models. Given the sensitivity of health and nutrition applications, ensuring high accuracy and consistency is essential. However, it is important to assess LLMs holistically and from multiple perspectives, such as safety, bias, privacy, and emotional support, to mention a few[18,19,58].

In addition, future research should concentrate on the performance of open-source LLMs in the diet and nutrition field. Our study is limited to the leading proprietary LLM models. These models are user-friendly and highly powerful. Our results also confirm their significant potential in food and nutrition applications. Yet, growing concerns are being raised about their lack of openness and limited access. The exact architecture and training data of these LLMs are not publicly known. In contrast, open-source LLMs are emerging rapidly, offering benefits, such as improved data security and privacy, decreased reliance on vendors, and the ability to customize models. Examples of the state-of-the-art open-source LLMs are Llama 3[59], Falcon 2[60], and Yi-34B[61].

Another avenue for future research should evaluate the impact of various information retrieval, fine-tuning, and agentic approaches in nutrition management applications. Recent studies have explored the role of fine-tuning[32,62,63]and agentic methods[12,23,64] in healthcare and nutrition applications. A solid benchmarking approach is required to comprehensively evaluate the effectiveness of these approaches in nutrition chatbots.

In conclusion, this study assessed the accuracy and consistency on the GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro in responding to diet and nutrition questions of the RD exam. In contrast to the previous LLM evaluation studies focusing on nutritional management, our experiments were not restricted to ChatGPT or ZS prompting. We evaluated the models using the RD exam and analyzed their errors across various questions complexities and nutrition domains. Our findings highlighted the strengths and weaknesses of the three LLMs, showing the influence of different prompting techniques on their responses to the RD exam questions. GPT-4o with CoT-SC prompting outperformed other approaches, while Gemini 1.5 Pro with ZS indicated the highest consistency. For GPT-4o and Claude 3.5, the application of CoT improved accuracy, while CoT-SC enhanced both accuracy and consistency. RAP particularly improved GPT-4o performance in addressing difficult-

expert-level questions. Consequently, selecting the appropriate LLM and prompt engineering, tailored to the proficiency level and specific domain, can considerably reduce errors and mitigate potential risks in diet and nutrition chatbot applications.

Methods

In this study, we use the RD exam questions to benchmark the performance of three leading LLMs chatbots, i.e., GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, in addressing nutrition-related inquiries. The RD exam was selected due to its comprehensive coverage across multiple nutrition topics. We define accuracy and consistency as key performance metrics in this experiment. The assessment is conducted by employing four distinct prompting techniques: 1) Zero Shot prompting (ZS), 2) Chain of Thought (CoT), 3) Chain of Thought with Self Consistency (CoT-SC), and 4) Retrieval Augmented Prompting (RAP) enabled by external nutrition knowledge. The responses are then analyzed, by comparing them to the ground truth, to evaluate the performance of each LLM and prompting technique. We perform inter-rater and intra-rater analysis to identify the strengths and weaknesses of each approach. In the following, we, first, provide more details about the RD exam. Then, we briefly describe the LLMs and prompting techniques used. Finally, we outline response collection and analysis in this benchmarking.

Registered dietitian exam

The Registration Examination for Dietitians is a required exam for individuals seeking to obtain the registered dietitian credential. To take the exam, candidates must successfully complete the eligibility requirements provided by the Commission on Dietetic Registration (CDR)[65]. The examination consists of 125 to 145 four-choice questions[44], covering four major domains: D1) Principles of Dietetics (21%), D2) Nutrition Care for Individuals and Groups (45%), D3) Food Service Systems (13%), and D4) Management of Food and Nutrition Programs and Services (21%)[44]. D1 covers topics related to i) food, nutrition, and supporting sciences, ii) education, communication and technology, and iii) research applications. D2 consists of the topics related to i) screening and assessment, ii) diagnosis, iii) planning and intervention, and iv) monitoring and evaluation. D3 includes topics related to i) menu development, ii) procurement, production, distribution, and service, iii) sanitation and safety, and iv) equipment and facility planning. D4 includes topics related to i) functions of management, ii) human resource management, iii) financial management, iv) marketing and public relations; and v) quality management and regulatory compliance[44].

The RD exam consists of a wide range of topics designed to assess a dietitian's professional competency. Therefore, we posit that the exam can be a valuable nutrition benchmark for evaluating the ability of the LLMs to respond nutrition queries. Nevertheless, it is worth noting that the RD exam might not consist of complex clinical scenarios encountered in real-world dietetic practice, including personalized dietary recommendations, comorbidity management, and food-drug interactions. In addition, it cannot address open-ended conversations, as the exam only include multiple-choice questions. In real-life scenarios, conversations might include detailed and long responses describing complex nutrition or health issues.

Large language models

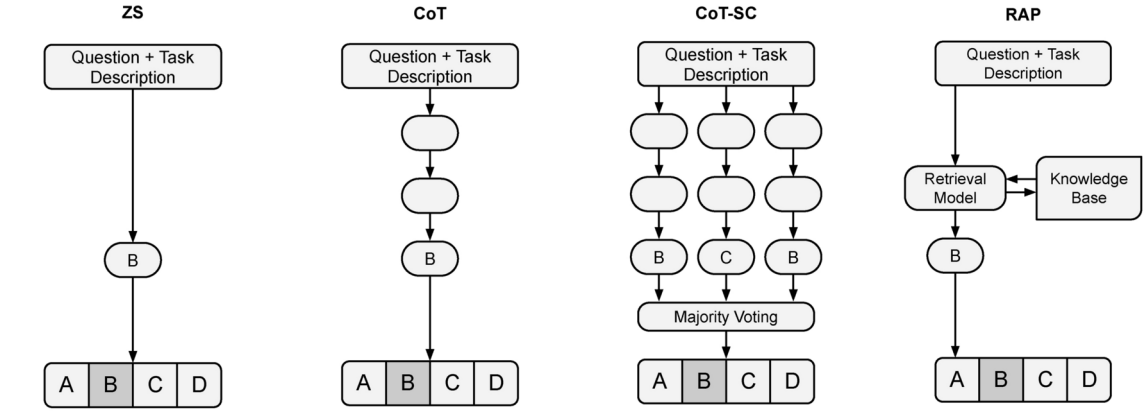
Within the state-of-the-art nutrition and diet studies, OpenAI models (i.e., ChatGPT-3.5 and ChatGPT-4) have been mostly employed and evaluated[7,11,24,27]. In this study, we propose to employ the RD exam to conduct a standard and comprehensive evaluation of the leading OpenAI model (i.e., GPT-4o). Therefore, our findings would be comparable to the existing literature in this field. Moreover, we extend the evaluation to include the recently proposed gold standard chatbots. In addition to GPT-4o, Claude 3.5 Sonnet and Gemini 1.5 Pro are other chatbots that showed top performance on industry benchmarks for graduate-level reasoning, undergraduate-level knowledge, text-based reasoning, and math problem-solving.

Moreover, OpenAI, Anthropic, and Google generative LLM chatbots holds the largest market share (more than 90% in total) as of October 2024, highlighting the usability of these chatbots[66–68]. OpenAI released GPT-4o, their new flagship model, on May 13, 2024[38], Claude 3.5 Sonnet was launched, by Anthropic, as their strongest vision model yet, on Jun 20, 2024[39], and Google announced Gemini 1.5 Pro as their next-generation model on February 15, 2024[40]. An overview of the models' performance on other benchmarks are indicated in Table 5. Find more details in the following references[38,39,48].

In this study, we set the temperature setting to 0 in all the experiments to better evaluate the chatbots' knowledge and decision-making in nutrition and diet applications, minimizing the effect of external variables on consistency. The temperature parameter, ranging from 0 to 2 for GPT-4o and Gemini 1.5 Pro and from 0 to 1 for Claude 3.5 Sonnet, regulates the uncertainty or randomness in the output[69]. A higher temperature value leads to more randomness in the output. In other words, it raises the likelihood of selecting tokens other than the most probable ones. However, with a temperature setting of 0, chatbots generate responses by selecting the next words with the highest probability. Therefore, this selection leads to more deterministic behavior of chatbots.

Prompt engineering

Prompt engineering enables the design and optimization of input prompts to instruct LLMs in performing specific tasks. Studies show that various prompt engineering techniques can differently affect on the performance of LLMs and the quality of the outcome[33,34]. These techniques can be applied in different tasks across multiple fields, including healthcare and nutrition[31,70]. For example, they can be leveraged in diet recommendation chatbots to provide conversations with users while incorporating relevant nutritional information into their responses. In this study, four prompting techniques are utilized for the models evaluation.



(a) Zero Shot (ZS) prompting includes a task description to guide the LLM in completing the task.

(b) Chain of Thought (CoT) prompting guides LLMs to carry out the task through a multi-step reasoning process.

(c) Chain of Thought with Self Consistency (CoT-SC) prompting guides the LLM to carry out the task through multiple paths, each of which is based on CoT.

(d) Retrieval Augmented Prompting (RAP) fetches relevant information from a knowledge base in real-time and integrates it into the input prompt.

Fig. 5. Schematic illustrations of the four prompting techniques used in the evaluation. The inputs include multiple-choice questions, and the generated output includes the selected choice.

1. *Zero Shot (ZS) prompting* leverages precisely formulated prompts, including a task description, to guide the LLM in completing the task. The prompting technique is straightforward and do not require any prior examples, as the model leverages its internal knowledge to generate responses[33]. However, the accuracy of the LLM's outputs may be diminished when handling more complex tasks. To the best of our knowledge, existing evaluations of LLM chatbots focusing on nutrition and diet have utilized ZS prompting for their assessments[20,24,29]. In our study, the ZS prompt consists of a question, multiple choices, and a fixed task description.
2. *Chain of Thought (CoT) prompting,* proposed by Wei *et al.*[53], guides LLMs to carry out the task through a multi-step reasoning process. This reasoning process might increase the accuracy of the LLM's outputs in complex tasks, such as arithmetic and symbolic reasoning. However, the model might be susceptible to error propagation in scenarios where multiple reasoning paths are available, and it cannot explore all the paths and select the most accurate outcome[71]. CoT has been widely used in medical studies[30,31]. In our study, the CoT prompt includes a question, multiple choices, and a description to the model to answer the question through intermediate reasoning steps.
3. *Chain of Thought with Self Consistency (CoT-SC)* guides the LLM to carry out the task through multiple paths, each of which is based on CoT. Subsequently, the outcomes are aggregated[56]. Although CoT-SC might improve the performance of the model, it increases computational costs and response generation latency[56]. CoT-SC has been leveraged in clinical studies[32]. In our study, the CoT-SC consists of three independent reasoning paths (CoT) and a majority voting method for the aggregation.
4. *Retrieval Augmented Prompting (RAP)* fetches relevant information from a knowledge base in real-time and integrates it into the input prompt[57,72]. In contrast to the other prompting techniques, using RAP, the model generates responses by relying not only on its internal knowledge but also on external information, which might mitigate hallucination. However, its implementation is more complex and introduces dependencies on external data sources. In our study, the knowledge base includes 125 documents (such as articles, books, and guidelines) recommended by the Academy of Nutrition and Dietetics[42], as references for the RD exam. The full list of the references used for RAP is provided in Supplementary Table S.7. For the implementation, we leveraged a conventional Retrieval Augmented Generation (RAG) framework[57]. To achieve this, the references were divided into 512-token chunks, using the Amazon Titan Text Embeddings v2 model[73] for text embeddings. Then, the Cosine Similarity method[74] was utilized to identify the most similar chunks.Schematic illustrations of the four techniques are shown in Figure 5. Supplementary Table S.5 also includes a brief summary of the four techniques, along with their strengths and weaknesses. Additionally, the instructions used for the prompting techniques in this study are presented in Supplementary Table S.6.

Data collection

For response collection, the questions were first stored in JSON (JavaScript Object Notation) format and saved in a MySQL Workbench database. Then, the questions were delivered to the three LLMs via their respective Application Programming Interfaces (APIs). The data collection was performed in Python using OpenAI v.1.26.0[75], google-generativeai v.0.5.4[76], Boto3 v. 1.34.114[77], and lxml.etree v.5.2.2[78] libraries. Details about the code, along with its documentation, are available in the GitHub repository at https://github.com/iHealthLab/DietitianExamEval.

Using this setup, the 1050 RD exam questions were delivered to the three models through the four prompting techniques. Each question was asked five times. Consequently, we collected 60 (i.e., $3 \times 4 \times 5$) sets of 1050

responses. As previously mentioned, the questions include four choices. We observed that sometimes the LLMs were unable to select an option from the multiple choices and provided responses such as, *"None of the above," "Since no option is correct, we cannot provide a final answer within the requested tags,"* or *"Cannot be determined with the given information."* In summary, this issue occurred once for GPT-4o with CoT, once for GPT-4o with CoT-SC, 15 times for Claude 3.5 with RAP, 100 times for Gemini 1.5 with CoT, 30 times for Gemini 1.5 with CoT-SC, and 63 times for Gemini 1.5 with RAP. For these responses, we added another option, labeled "Others."

It is worth noting that five repeated measurements was chosen as it has been identified in recently published studies as an acceptable number of repeated measurements[22,24,31,79]. For examples, Wang et al.[31] evaluated the consistency of the agreement LLMs with the American Academy of Orthopedic Surgeons osteoarthritis evidence-based guidelines. Each question was posed five times in this study. Hoang et al.[22] investigated the consistency of ChatGPT-3.5 and ChatGPT-4 in providing the energy and macronutrient content of 222 food items across five repeated measurements. A high number of repeated measurements increases the reliability of evaluations. However, it also significantly increases the costs and latency of response collection, particularly for techniques such as the CoT-SC, which includes multiple multi-step reasoning paths, and retrieval-augmented prompting.

The collected responses were compared with the ground truth answers provided by the Academy of Nutrition and Dietetics, eatrightPREP[42]. It should be noted that we used a new chat session for each query to minimize bias in the evaluation caused by information leakage from other questions.

Statistical analysis

The chatbots were evaluated in terms of accuracy and consistency. The responses consists of choices (i.e., four categories) from the same set of questions. Accuracy measures how close a set of responses aligns with the ground truth answers (i.e., if they are equal). To this end, we calculate the percentage score and error count. The percentage score indicates how well an LLM can detect the correct option and is obtained as follows.

$$Score = \frac{Correct\ Responses}{All\ Responses} * 100 \tag{1}$$

The error count is the number of times the selected choice does not match the ground truth. As previously mentioned, each measurement is repeated five times. The five repeated measurements in each test are grouped, and the mean and standard deviation of the scores and error counts are calculated. We also assess the performance of the LLMs by considering the proficiency levels and domains of the questions. To do this, we compute the average error counts for each subgroup.

Consistency refers to the degree to which responses produce the same results. To assess consistency, we perform inter-rater and intra-rater analysis approaches[80]. For the former, the agreement between the responses obtained from different models / prompting techniques are evaluated. To this end, Cohen's Kappa[81] was utilized to measure the degree of agreement between paired sets of responses. For example, the agreement between responses obtained from GPT-4o with ZS prompting and GPT-4o with CoT prompting are calculated. Cohen's Kappa was selected as the analysis included paired categorical responses, requiring a method that takes into account the possibility of agreement by chance. We also utilize the McNemar-Bowker test[43] to investigate if the collected paired sets of responses are statistically different. The test indicates if the contingency table is symmetric, evaluating whether there are significant differences in patterns between two sets of responses. The McNemar-Bowker test was selected, as it allows the analysis of paired categorical data involving more than two categories (i.e., an extension of McNemar's test).

Furthermore, for the intra-rater analysis, Fleiss Kappa test[82] was used to indicate the degree of overall agreement between the repeated measurements under fixed conditions. For instance, we assess whether GPT-4o with ZS prompting provides the same choices in repeated measurements. Fleiss Kappa was selected, since the analysis consisted of multiple raters, requiring a method to measure agreement across multiple sets of responses.

Note that the statistical analysis was conducted in R Programming using the irr v.0.84.1[83]library to perform Cohen's Kappa and Fleiss Kappa tests, and the boot v.1.3–30[84] library to compute Bootstrap confidence intervals. Details about the code, along with its documentation, are available in the GitHub repository at https://github.com/iHealthLab/DietitianExamEval.

Data availability

The RD exam questions used in this study are not publicly available and can be accessed via https://www.eatrightprep.org.

Received: 15 August 2024; Accepted: 30 December 2024
Published online: 09 January 2025

References

1. Singh, B. *et al.* Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *npj Digital Medicine* **6**, 118 (2023).
2. Webster, P. Six ways large language models are changing healthcare. *Nature Medicine* **29**, 2969–2971 (2023).
3. Ma, P. *et al.* Large language models in food science: Innovations, applications, and future. *Trends in Food Science & Technology* 104488 (2024).
4. Clusmann, J. et al. The future landscape of large language models in medicine. *Communications medicine* **3**, 141 (2023).

5. Meskó, B. The impact of multimodal large language models on health care's future. *Journal of medical Internet research* **25**, e52865 (2023).
6. Bond, A., Mccay, K. & Lal, S. Artificial intelligence & clinical nutrition: What the future might have in store. *Clinical nutrition ESPEN* (2023).
7. Dao, D., Teo, J. Y. C., Wang, W. & Nguyen, H. D. LLM-Powered Multimodal AI Conversations for Diabetes Prevention. In *Proceedings of the 1st ACM Workshop on AI-Powered Q &A Systems for Multimedia*, 1–6 (2024).
8. OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ (2022). Accessed: August 2024.
9. Liu, Y. *et al.* Exploring the usability of a chatbot-based conversational dietary assessment tool among cardiovascular patients. *European Journal of Preventive Cardiology* **30**, zwad125–281 (2023).
10. Pugliese, N. et al. Accuracy, Reliability, and Comprehensibility of ChatGPT-generated Medical Responses for Patients with Nonalcoholic Fatty Liver Disease. *Clinical Gastroenterology and Hepatology* **22**, 886–889 (2024).
11. Kim, D. W. et al. Qualitative evaluation of artificial intelligence-generated weight management diet plans. *Frontiers in Nutrition* **11**, 1374834 (2024).
12. Abbasian, M. *et al.* Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients. In *the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, 2024).
13. Haman, M., Školník, M. & Lošťák, M. AI dietitian: Unveiling the accuracy of ChatGPT's nutritional estimations. *Nutrition* **119**, 112325 (2024).
14. Qarajeh, A. et al. AI-Powered Renal Diet Support: Performance of ChatGPT, Bard AI, and Bing Chat. *Clinics and Practice* **13**, 1160–1172 (2023).
15. Tsai, C.-H. *et al.* Generating Personalized Pregnancy Nutrition Recommendations with GPT-Powered AI Chatbot. In *20th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, vol. 2023, 263 (2023).
16. Zhou, P. *et al.* FoodSky: A Food-oriented Large Language Model that Passes the Chef and Dietetic Examination. arXiv:2406.10261 (2024).
17. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nature medicine* **29**, 1930–1940 (2023).
18. Abbasian, M. et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine* **7**, 82 (2024).
19. Liang, P. *et al.* Holistic evaluation of language models. arXiv:2211.09110 (2022).
20. Sun, H. et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. *Journal of Medical Internet Research* **25**, e51300 (2023).
21. Barlas, T., Altinova, A. E., Akturk, M. & Toruner, F. B. Credibility of ChatGPT in the assessment of obesity in type 2 diabetes according to the guidelines. *International Journal of Obesity* **48**, 271–275 (2024).
22. Hoang, Y. N. et al. Consistency and accuracy of artificial intelligence for providing nutritional information. *JAMA network open* **6**, e2350367–e2350367 (2023).
23. Yang, Z. et al. ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. *Smart Health* **32**, 100465 (2024).
24. Ponzo, V. et al. Is ChatGPT an Effective Tool for Providing Dietary Advice?. *Nutrients* **16**, 469 (2024).
25. Kirk, D., van Eijnatten, E. & Camps, G. Comparison of answers between ChatGPT and human dieticians to common nutrition questions. *Journal of Nutrition and Metabolism* **2023**, 5548684 (2023).
26. Szymanski, A., Wimer, B. L., Anuyah, O., Eicher-Miller, H. A. & Metoyer, R. A. Integrating Expertise in LLMs: Crafting a Customized Nutrition Assistant with Refined Template Instructions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22 (2024).
27. Niszczota, P. & Rybicka, I. The credibility of dietary advice formulated by ChatGPT: robo-diets for people with food allergies. *Nutrition* **112**, 112076 (2023).
28. Minaee, S. *et al.* Large language models: A survey. arXiv:2402.06196 (2024).
29. Naja, F. *et al.* Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome. *European Journal of Clinical Nutrition* 1–10 (2024).
30. Holmes, J. et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology* **13**, 1219326 (2023).
31. Wang, L. *et al.* Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* **7**, 41 (2024).
32. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
33. Sahoo, P. *et al.* A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv:2402.07927 (2024).
34. Chen, B., Zhang, Z., Langrené, N. & Zhu, S. Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv:2310.14735 (2023).
35. Wang, J. *et al.* Prompt engineering for healthcare: Methodologies and applications. arXiv:2304.14670 (2023).
36. Maharjan, J. et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports* **14**, 14156 (2024).
37. Ouyang, S., Zhang, J. M., Harman, M. & Wang, M. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. arXiv:2308.02828 (2023).
38. OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ (2024). Accessed: August 2024.
39. Anthropic. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet (2024). Accessed: August 2024.
40. Google. Introducing Gemini 1.5, Google's next-generation AI model. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/ (2024). Accessed: August 2024.
41. Commission on Dietetic Registration. Registered Dietitian Nutritionist. https://www.cdrnet.org/RDN. Accessed: August 2024.
42. Academy of Nutrition and Dietetics . https://www.eatrightprep.org/. Accessed: August 2024.
43. Fagerland, M., Lydersen, S. & Laake, P. *Statistical analysis of contingency tables* (Chapman and Hall/CRC, 2017).
44. Commission on Dietetic Registration. Candidate Handbook, RD Exam. https://admin.cdrnet.org/vault/2459/web//RD%20Handbook%20for%20Candidates%20-%206-2024.pdf (2024).
45. Commission on Dietetic Registration. RD Exam Pass / Fail Statistics. https://www.cdrnet.org/RDExamStats. Accessed: October 2024.
46. Grabb, D. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence* **6** (2023).
47. Russe, M. F., Reisert, M., Bamberg, F. & Rau, A. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren* (Georg Thieme Verlag KG, 2024).
48. Reid, M. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 (2024).
49. Rein, D. *et al.* GPQA: A Graduate-level Google-proof Q &A Benchmark. arXiv:2311.12022 (2023).
50. Hendrycks, D. *et al.* Measuring massive multitask language understanding. arXiv:2009.03300 (2020).
51. Dua, D. *et al.* DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. arXiv:1903.00161 (2019).

52. Mishra, V., Jafri, F., Abdul Kareem, N., Aboobacker, R. & Noora, F. Evaluation of accuracy and potential harm of ChatGPT in medical nutrition therapy-a case-based approach. *F1000Research* **13**, 137 (2024).
53. Wei, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems* **35**, 24824–24837 (2022).
54. Gou, Z. *et al.* ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving. arXiv:2309.17452 (2023).
55. Abbasian, M., Azimi, I., Rahmani, A. M. & Jain, R. Conversational health agents: A personalized LLM-powered agent framework. arXiv:2310.02374 (2023).
56. Wang, X. *et al.* Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 (2022).
57. Gao, Y. *et al.* Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997 (2023).
58. Sun, L. *et al.* TrustLLM: Trustworthiness in large language models. arXiv:2401.05561 (2024).
59. Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/ (2024). Accessed: August 2024.
60. Technology Innovation Institute. Falcon LLM. https://falconllm.tii.ae/ (2024). Accessed: August 2024.
61. 01.AI. Yi-34B. https://huggingface.co/01-ai/Yi-34B (2024). Accessed: August 2024.
62. Xu, L., Xie, H., Qin, S.-Z. J., Tao, X. & Wang, F. L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv:2312.12148 (2023).
63. Zhang, X. et al. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. *AMIA Summits on Translational Science Proceedings* **2024**, 478 (2024).
64. Li, Y. *et al.* Personal LLM agents: Insights and survey about the capability, efficiency and security. arXiv:2401.05459 (2024).
65. Commission on Dietetic Registration, Registered Dietitian (RD) Or Registered Dietitian Nutritionist (RDN) Certification. https://www.cdrnet.org/RDN. Accessed: August 2024.
66. Chatbot Arena. Chat with Open Large Language Models. https://chat.lmsys.org/?leaderboard. Accessed: August 2024.
67. Artificial Analysis. Model & API Providers Analysis. https://artificialanalysis.ai/. Accessed: August 2024.
68. Evan Bailyn. Top Generative AI Chatbots by Market Share - October 2024. https://firstpagesage.com/reports/top-generative-ai-chatbots/ (2024). Accessed: October 2024.
69. Peeperkorn, M., Kouwenhoven, T., Brown, D. & Jordanous, A. Is temperature the creativity parameter of large language models? arXiv:2405.00492 (2024).
70. Zaghir, J. et al. Prompt engineering paradigms for medical applications: Scoping review. *Journal of Medical Internet Research* **26**, e60501 (2024).
71. Saparov, A. & He, H. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. arXiv:2210.01240 (2022).
72. Li, Y. *et al.* ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* **15** (2023).
73. Sebastien Stormacq. Amazon Titan Text Embeddings V2 now available in Amazon Bedrock, optimized for improving RAG. https://aws.amazon.com/blogs/aws/amazon-titan-text-v2-now-available-in-amazon-bedrock-optimized-for-improving-rag/ (2024). Accessed: August 2024.
74. Manning, C. D., Raghavan, P. & Schütze, H. Scoring, term weighting & the vector space model. In *Introduction to Information Retrieval* (Cambridge University Press, 2008).
75. Libraries, OpenAI API. https://platform.openai.com/docs/libraries/python-library. Accessed: August 2024.
76. google-generativeai, Gemini API. https://pypi.org/project/google-generativeai/. Accessed: August 2024.
77. AWS SDK for Python (Boto3) Documentation. https://docs.aws.amazon.com/pythonsdk/. Accessed: August 2024.
78. Stefan Behnel. The lxml.etree Tutorial. https://lxml.de/tutorial.html. Accessed: August 2024.
79. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the performance of chatgpt in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmology science* **3**, 100324 (2023).
80. Hallgren, K. A. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* **8**, 23 (2012).
81. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37–46 (1960).
82. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin* **76**, 378 (1971).
83. Gamer, M., Lemon, J. & Singh, I. F. P. *irr: Various Coefficients of Interrater Reliability and Agreement* (2019). R package version 0.84.1, https://CRAN.R-project.org/package=irr.
84. Canty, A., Ripley, B. & Brazzale, A. R. *boot: Bootstrap Functions* (2024). R package version 1.3-30, https://CRAN.R-project.org/package=boot.

Acknowledgements

We are grateful for iHealth Labs to sponsor this research, which will be used in their LLM-based nutrition assistant product. The authors would like to thank Prof. Penny M. Kris-Etherton for her thorough review and suggestions. We acknowledge UCI Center for Statistical Consulting for their expert statistical consulting.

Author contributions

I.A. drafted the manuscript, contributed to the study design, analyzed and interpreted the LLM data, and prepared Figures 1-4. M.Q. contributed to the acquisition of LLM data, contributed to the study design, analyzed and interpreted nutrition information, and co-drafted the manuscript. L.W. contributed to the analysis and interpretation of the nutrition information and critically revised the manuscript. A.R. contributed to the interpretation of the LLM responses and critically revised the manuscript. Y.L. contributed to the study design, analyzed and interpreted both the LLM data and nutrition information, and critically revised the manuscript. All authors reviewed the manuscript.

Declarations

Competing Interests

The authors declare no competing interests. Moreover, the funders of the study had no role in study design, data collection and analysis, or interpretation of results and preparation of the manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-85003-w.

Correspondence and requests for materials should be addressed to I.A.