JMIR Publications
The Leading eHealth Publisher

Comparative Analysis of AI Systems and Human Nutrition Knowledge: Evaluating ChatGPT and Other AI Systems Against Dietetics Students and the General Population

Nicola Luigi Bragazzi, Stefania Monica, Federico Bergenti, Francesca Scazzina, Alice Rosi

Submitted to: Journal of Medical Internet Research
on: July 26, 2024

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Comparative Analysis of AI Systems and Human Nutrition Knowledge: Evaluating ChatGPT and Other AI Systems Against Dietetics Students and the General Population

Nicola Luigi Bragazzi[1]; Stefania Monica[2]; Federico Bergenti[1]; Francesca Scazzina[1]; Alice Rosi[1]

[1]University of Parma Parma IT
[2]University of Modena and Reggio Emilia Reggio Emilia IT

Corresponding Author:
Francesca Scazzina
University of Parma
Medical School Building A
Via Volturno 39
Parma
IT

Abstract

Background: Understanding the core principles of nutrition is essential in today's world of abundant, often contradictory dietary advice, empowering individuals to make informed dietary choices, crucial for having a proper diet and managing diet-related Non-Communicable Diseases (NCDs). The role of Artificial Intelligence (AI) systems in providing nutritional information is increasingly prominent, but their reliability in this domain is not well-established yet.

Objective: This study compares the nutrition knowledge of state-of-the-art AI systems (ChatGPT-4, Bard, Copilot, and ChatGPT-3.5) with human subjects having different levels of nutrition knowledge.

Methods: The "General Nutrition Knowledge Questionnaire–Revised" (GNKQ-R) was administered to four AI systems and human subjects. The AI systems were tested using zero-shot prompts. Responses were scored per the GNKQ's guidelines across four sections: "Dietary Recommendations"; "Food Groups"; "Healthy Food Choices"; "Diet, Disease and Weight Management". Human subjects were grouped based on their academic background (dietetics vs English students), age, sex/gender, education level, and health status.

Results: The average performance of AI systems across all LLMs was 77.3±5.1 out of 88, which comparable to the dietetics students and significantly higher than the English students. ChatGPT-4 scored highest among the AI systems (82/88), surpassing both groups of students (dietetics: 79.3/88, English: 67.7/88) as well as all other demographic groups. In "Dietary Recommendations", ChatGPT-4 and ChatGPT-3.5 nearly matched dietetics students. ChatGPT-4 excelled in "Food Groups", outperforming all human groups. In "Healthy Food Choices", ChatGPT-4 achieved a perfect score, indicating a deep understanding. ChatGPT-3.5 excelled in "Diet, Disease and Weight Management". Variations in the performances of the AI systems across different sections were observed, suggesting knowledge gaps in certain areas. AI systems, particularly ChatGPT-4 and ChatGPT-3.5, showed proficiency in nutrition knowledge, rivaling or surpassing dietetics students in certain sections. This indicates their potential utility in nutritional guidance. However, there are nuances and specific details where AI systems lack compared to specialized human education. The study highlights the potential of AI in public health and educational settings but also underscores the value of expert human judgment.

Conclusions: AI systems show promise in understanding complex subjects like nutrition and can be a valuable adjunct educational tool. However, specialized human education and expertise remain irreplaceable, emphasizing the need for a combined approach of AI systems insights with expert human judgment in nutrition and dietetics.

(JMIR Preprints 26/07/2024:64791)
DOI: https://doi.org/10.2196/preprints.64791

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Comparative Analysis of AI Systems and Human Nutrition Knowledge: Evaluating ChatGPT and Other AI Systems Against Dietetics Students and the General Population

Abstract

Background: Understanding the core principles of nutrition is essential in today's world of abundant, often contradictory dietary advice, empowering individuals to make informed dietary choices, crucial for having a proper diet and managing diet-related Non-Communicable Diseases (NCDs). The role of Artificial Intelligence (AI) systems in providing nutritional information is increasingly prominent, but their reliability in this domain is not well-established yet. This study compares the nutrition knowledge of state-of-the-art AI systems (ChatGPT-4, Bard, Copilot, and ChatGPT-3.5) with human subjects having different levels of nutrition knowledge.

Methods: The "General Nutrition Knowledge Questionnaire–Revised" (GNKQ-R) was administered to four AI systems and human subjects. The AI systems were tested using zero-shot prompts. Responses were scored per the GNKQ's guidelines across four sections: "Dietary Recommendations"; "Food Groups"; "Healthy Food Choices"; "Diet, Disease and Weight Management". Human subjects were grouped based on their academic background (dietetics vs English students), age, sex/gender, education level, and health status.

Results: The average performance of AI systems across all LLMs was 77.3±5.1 out of 88, which comparable to the dietetics students and significantly higher than the English students. ChatGPT-4 scored highest among the AI systems (82/88), surpassing both groups of students (dietetics: 79.3/88, English: 67.7/88) as well as all other demographic groups. In "Dietary Recommendations", ChatGPT-4 and ChatGPT-3.5 nearly matched dietetics students. ChatGPT-4 excelled in "Food Groups", outperforming all human groups. In "Healthy Food Choices", ChatGPT-4 achieved a perfect score, indicating a deep understanding. ChatGPT-3.5 excelled in "Diet, Disease and Weight Management". Variations in the performances of the AI systems across different sections were observed, suggesting knowledge gaps in certain areas. AI systems, particularly ChatGPT-4 and ChatGPT-3.5, showed proficiency in nutrition knowledge, rivaling or surpassing dietetics students in certain sections. This indicates their potential utility in nutritional guidance. However, there are nuances and specific details where AI systems lack compared to specialized human education. The study highlights the potential of AI in public health and educational settings but also underscores the value of expert human judgment.

Conclusions: AI systems show promise in understanding complex subjects like nutrition and can be a valuable adjunct educational tool. However, specialized human education and expertise remain irreplaceable, emphasizing the need for a combined approach of AI systems insights with expert human judgment in nutrition and dietetics.

Keywords: nutrition knowledge; artificial intelligence; conversational agent

Introduction

Dietary patterns have a significant impact on human health, and diet-related risk factors are among the main preventable risk factors linked to the incidence of Non-Communicable Diseases (NCDs) [1], causing 11 million deaths in 2017 [2]. Increasing the consumption of healthy diets can, therefore, improve public health and prevent chronic diseases. Proper nutrition knowledge is a crucial factor in promoting healthier food choices. Individuals who possess adequate nutrition knowledge are better equipped to make informed food choices, leading to improved health outcomes [3-5]. So, proper general nutrition knowledge is not just beneficial; it is a pivotal element in molding dietary behaviors and lifestyle choices, with far-reaching implications for individual well-being and overall public health [6].

In an era characterized by nutritional misinformation and disinformation, where dietary advice is abundant yet often contradictory [7,8], understanding the core principles of nutrition is more crucial than ever. This foundational knowledge empowers individuals to navigate the complex landscape of dietary choices, making well-informed decisions about their nutrition and health. Adequate knowledge of nutrition is key to preventing and managing prevalent NCDs like obesity, diabetes, heart diseases, and certain cancers, which are significantly influenced by dietary habits [9]. Further, it equips individuals to sift through myriads of diet trends and make choices that truly benefit their physical and mental health [10].

With an increasing number of individuals turning to digital platforms for dietary guidance [11], where nutritional misinformation and disinformation are particularly present, generative Artificial Intelligence (AI) systems, like conversational agents and similar tools, have rapidly gained interest. Conversational agents are emerging as key sources for personalized nutritional information [12-15]. In the rapidly evolving domain of AI, examining AI systems' understanding in specialized fields such as nutrition is becoming increasingly critical [16,17]. Yet, their capabilities to reliably address nutrition-related inquiries remain largely unexplored. Only a few studies [18-20] have been designed to evaluate the proficiency of these platforms in responding to questions about nutrition. However, these investigations generally make use of non-validated questionnaires and do not compare the different existing AI systems.

It is crucial to thoroughly assess and validate the effectiveness and accuracy of AI-based responses, especially from a comparative perspective that looks at contrasting various AI systems. Currently, a variety of AI systems exists from those developed by OpenAI (ChatGPT-3.5 and ChatGPT-4) to those devised by Microsoft (Copilot) and Google (Bard). These Large Language Models (LLMs) share foundational elements, built on the Transformer architecture [21] and trained on diverse datasets encompassing texts from various sources like books and websites. This allows them to proficiently comprehend and generate human-like text, excelling in natural language processing and conversational agents. However, these LLMs vary in model size and outcomes, including performance, accuracy, and coherence. Consequently, they may differ in their ability to handle complex or ambiguous queries and generate contextually relevant responses. Therefore, it is essential to compare these LLMs rather than focusing solely on a single AI system because understanding their strengths and weaknesses helps users choose the most appropriate tool for specific needs and promotes the advancement of AI technology. Analyzing each LLM's areas of expertise and limitations enables researchers to pinpoint opportunities for optimization, enhancement, and innovation in future iterations.

Currently, to the best of our knowledge, there exists no comparative study appraising various LLMs in the field of human nutrition. Therefore, the present study was undertaken with the aim of filling in this knowledge gap, presenting an in-depth comparison of nutrition knowledge between state-of-the-art AI systems (ChatGPT-4, Bard, Copilot, and ChatGPT-3.5) and human subjects, including students with diverse academic backgrounds, and various demographic groups (in terms of age, sex/gender, education level, and health status). Employing the revised version of the "General Nutrition Knowledge Questionnaire" (GNKQ-R) as a standard [22], overall assessments and scores across the four principal sections of the questionnaire (namely, Dietary Recommendations, Food Groups Healthy Food Choices, and Diet, Disease and Weight Management) were analyzed and contrasted.

The goal was to shed light on the proficiency of the tested AI systems in grasping and conveying nutritional information, compared with the nuanced and comprehensive understanding possessed by human individuals. We hypothesized that these AI systems would achieve high scores on the GNKQ-

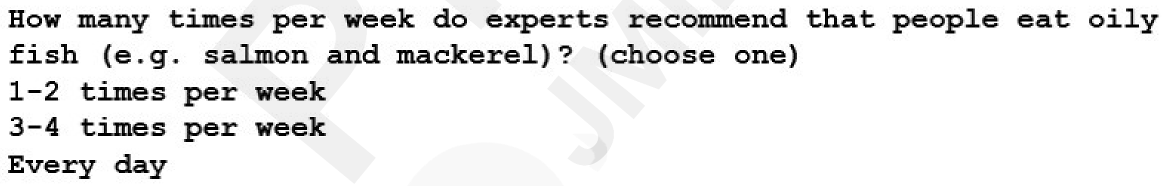
R, surpassing English students and demonstrating a level of nutrition knowledge comparable to or exceeding that of dietetics students. More specifically, these AI systems are expected to excel in sections related to basic dietary recommendations, food group classifications, and general healthy food choices. However, these AI systems are anticipated to show variability in their performance across different sections, with potential deficiencies in areas requiring deep, specialized knowledge and practical application of nutrition principles, such as diet-disease relationships and weight management strategies. Last, despite their high overall scores, these AI systems are expected to occasionally lack the nuanced understanding and context-specific expertise that human dietetics students possess.

Methods

Procedure

Questions from the GNKQ-R [22], the revised and updated version of the "General Nutrition Knowledge Questionnaire", initially developed by Parmenter and Wardle in the nineties in the United Kingdom [23], and, subsequently, tested and validated in diverse populations [3], were submitted to Bard, Copilot, ChatGPT-3.5, and ChatGPT-4 using zero-shot prompts, which means that the LLMs were queried without any prior specific training or examples given for the task at hand. In other words, these tools were asked to perform a task they have not explicitly been trained to do, using only their pre-existing knowledge and capabilities. In detail, Bard was queried on January 10, 2024, using the available public web interface, and Copilot was queried on the same day using the available public web interface. Similarly, ChatGPT-3.5 and ChatGPT-4 were queried on January 10, 2024, using the available public web interface. Figure 1 shows an example of the prompts used to administer the GNKQ-R to the tested LLMs. Note that the questions of the GNKQ-R were rephrased to avoid references to visual elements (images and labels) and to the explicit request for ticking an answer.

Figure 1. Example of the prompts used to query the tested AI systems. The shown prompt was used to feed the AI systems with the fifth question of Section 1 of the "General Nutrition Knowledge Questionnaire–Revised" questionnaire.



```
How many times per week do experts recommend that people eat oily
fish (e.g. salmon and mackerel)? (choose one)
1-2 times per week
3-4 times per week
Every day
```

The full list of used prompts is available in the supplementary material 1 (Multimedia Appendix 1) available from the website of the journal.

Replies were collected, transcribed into an Excel spreadsheet, and scored according to the instructions of the developers of the questionnaire. Points were assigned on a per-item basis, with each correct response earning one point, summing up to a potential maximum of 18 points for Section 1 ("Dietary Recommendations"), 36 points for Section 2 ("Food Groups"), 13 points for Section 3 ("Healthy Food Choices"), and 21 points for Section 4 ("Diet, Disease and Weight Management"), and up to 88 points for the total questionnaire. The full list of obtained answers is available in the supplementary material 2 (Multimedia Appendix 2) available from the website of the journal.

Scores from the same questionnaire were collected from UK university students in a previous work [22]. Briefly, 96 students from nutrition or dietetics courses (named "dietetics students") and 89 students from English courses (named “English students”) from all over UK had completed the GNKQ-R questionnaire online on a single occasion. The two student groups were similar in terms of socio-demographic (age, sex/gender, and socio-economic status) and clinical (health status) related parameters but were expected to be different in their nutrition knowledge.

In addition, data from students were analyzed together with the scores from 266 UK respondents for a total sample of 451 participants who completed the GNKQ-R [22]. Participants were categorized into three age groups: i) 18-35 years of age (n=195), ii) 36-50 years (n=108), and iii) 50 years and older (n=148). Concerning sex/gender, female participants were the majority (n=335). Participants were also divided based on their highest level of education: i) no degree (including individuals with secondary school education, O levels to A levels, or a certificate/diploma, generally with lower nutrition knowledge scores, n=239), and ii) degree or higher (including participants with a degree or postgraduate degree, n=212). Finally, health status was self-reported and categorized into three groups: i) poor/fair health (n=148), ii) good health (n=183), and iii) very good/excellent health (n=120).

Statistical Analysis

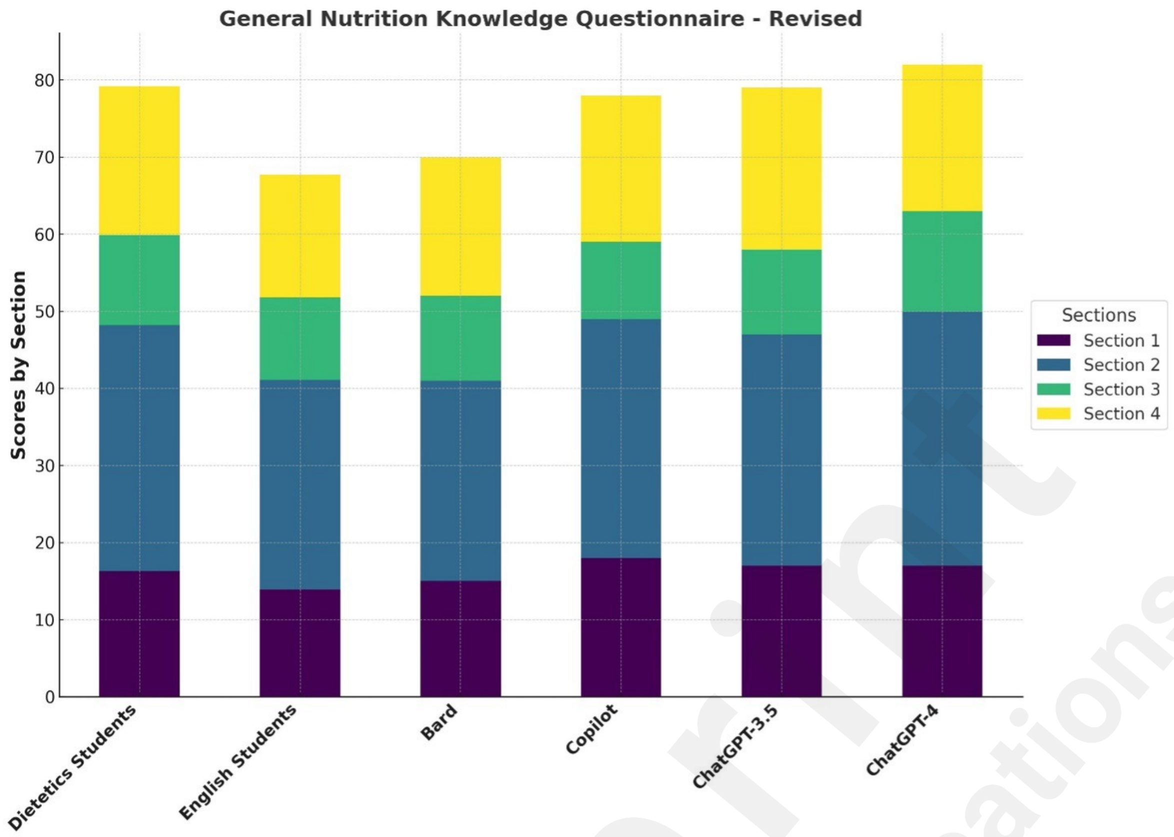
Descriptive statistics of the overall assessments and the scores broken down according to each section for each AI system were carried out. Then, the performances of the four AI systems were averaged for statistical comparison with the performances of dietetics and English students and the various demographics (age, sex/gender, education, and health status groups). These data were extracted from [22] and compared by carrying out an Analysis of Variance (ANOVA) from summary statistics and using the Tukey Honest Significant Difference (HSD) test for the post-hoc analysis. All statistical analyses were done using the commercial software “Statistical Package for Social Sciences” (SPSS version 28 for Windows, IBM, Armonk, NY, USA).

Results

AI systems versus humans

Findings from the comparative performance analysis on the GNKQ-R questionnaire are presented in Figure 2 for the total score and by sections.

Figure 2. Comparative performance analysis on the “General Nutrition Knowledge Questionnaire–Revised” (GNKQ-R). This chart illustrates the scores of dietetics and English students versus various AI systems (Bard, Copilot, ChatGPT-3.5, and ChatGPT-4) across the four key sections of the GNKQ-R: Dietary Recommendations (Section 1), Food Groups (Section 2), Healthy Food Choices (Section 3), and Diet, Disease and Weight Management (Section 4).



In the validation study dietetics and English students had an average score of 79.3/88 and 67.7/88, respectively [22]. Among the AI systems, ChatGPT-4 led with 82/88, indicating a robust understanding of nutrition topics. ChatGPT-3.5 demonstrated a competitive understanding with a score of 79/88. Bard and Copilot showed intermediate performances with scores of 70/88 and 78/88.

Concerning “Dietary recommendations” (Section 1), ChatGPT-4 and ChatGPT-3.5 both scored 17/18, nearly mirroring the dietetics students (16.3/18), thus reflecting strong foundational knowledge. Copilot excelled with 18/18, suggesting a superior understanding in this section. Bard and English students scored lower, 15/18 and 13.9/18 respectively, indicating knowledge gaps.

In terms of the knowledge of “Food groups” (Section 2), ChatGPT-4 outperformed all groups with 33/36, including dietetics students (31.9/36), demonstrating a high level of proficiency. Copilot and ChatGPT-3.5 had close scores of 31/36 and 30/36. Bard and English students lagged behind, scoring 26/36 and 27.2/36, respectively.

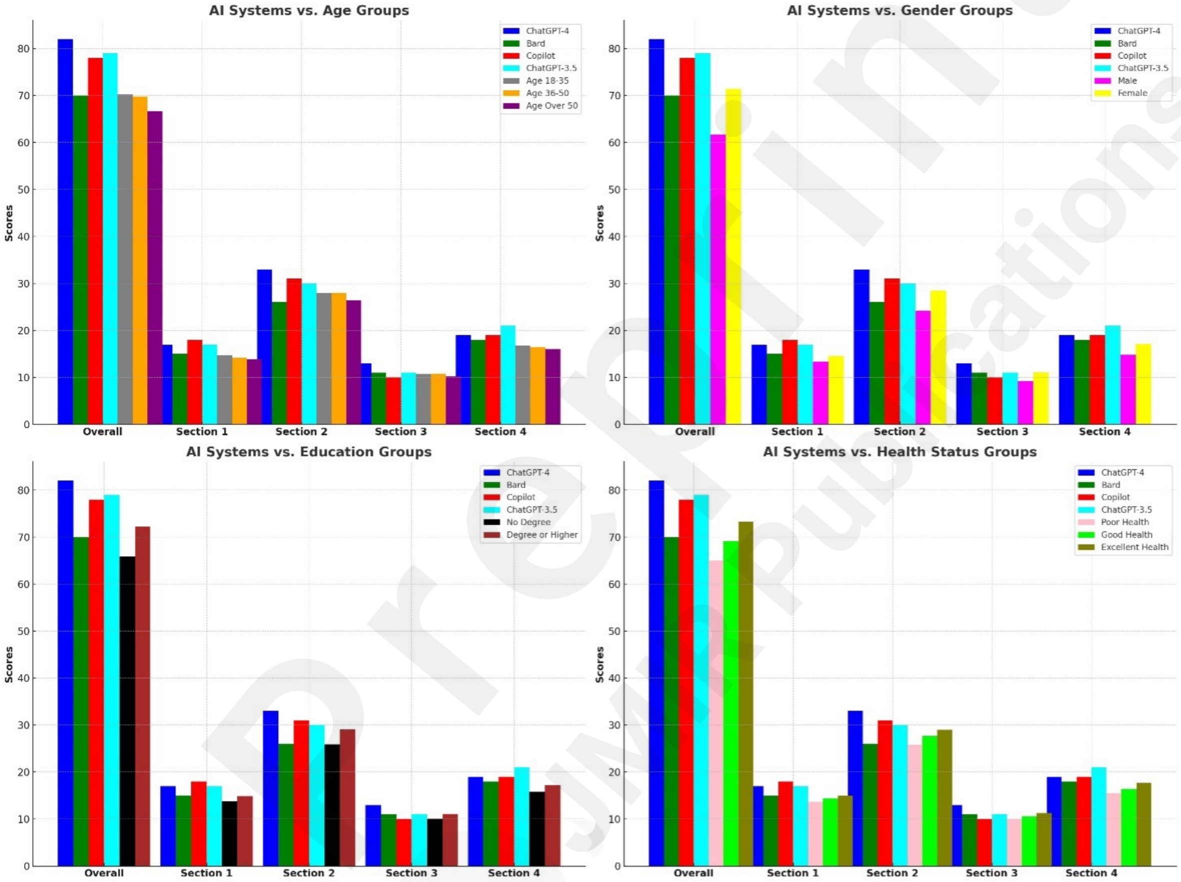
Regarding Section 3, ChatGPT-4 achieved a perfect score of 13/13, indicating a deep understanding of “Healthy food choices”, while dietetics students scored 11.7/13, showing strong, but not flawless, knowledge. ChatGPT-3.5, Bard, and English students had similar scores (11/13, 11/13, and 10.7/13), suggesting moderate understanding. Copilot scored the lowest in this category with 10/13.

Finally, concerning “Diet, disease and weight management” (Section 4), ChatGPT-3.5 excelled with a perfect score of 21/21. ChatGPT-4 and Copilot both scored 19/21, showing strong competency, similar to the one of dietetics students (19.3/21). Bard scored 18/21, indicating good, but not perfect, knowledge. English students scored the worst (15.9/21).

In general, as shown in Figure 2, ChatGPT-4 outperformed both groups of students in the overall score and in most individual sections. ChatGPT-3.5 had scores similar to ChatGPT-4 in some sections and also demonstrated a strong overall performance, similar to the one of the dietetics students. Bard and Copilot showed competitive but varying performance across different sections, and Bard displayed an overall performance like not trained students.

Finally, the comparative performance analysis on the GNKQ-R questionnaire total score and section sub-scores is presented in Figure 3 for age groups, sex/gender groups, education level groups, and health status groups. ChatGPT-4 outperformed all demographic groups in the overall score (82/88 versus 61.7/88 in males and 71.4/88 in females; versus 65.8/88 in the no degree group and 72.3/88 in the group with degree or higher; versus 70.3/88 in the group aged 18-35 years, 69.8/88 in the group aged 36-50 years, and 66.7/88 in the group aged 50 years and greater; versus 65.1/88 in the poor health group, 69.1/88 in the good health group, and 73.2/88 in the very good/excellent health group).

Figure 3. Comparative analysis of the “General Nutrition Knowledge Questionnaire–Revised” (GNKQ-R) scores. This set of four bar charts illustrates the performance of the AI systems (ChatGPT-4, Bard, Copilot, and ChatGPT-3.5) against different demographic groups categorized by age, gender, education level, and health status. Each chart provides a detailed breakdown of scores across ‘Overall’ performance and individual sections (1 to 4) of the – GNKQ-R, highlighting the proficiency levels in each category.



ChatGPT-4 outperformed as well in all individual sections of the GNKQ-R questionnaire.

ChatGPT-3.5 had scores similar to ChatGPT-4 in some sections and also demonstrated a strong overall performance in comparison to all demographic groups, being slightly outperformed only in Section 3 by females (11/13 versus 11.1/13) and the very good/excellent health group (11.3/13).

Finally, Bard and Copilot showed competitive but varying performance across different sections. In terms of the overall score, Bard was outperformed by female participants (71.4/88 versus 70/88), the group with degree or higher (72.3/88), the very good/excellent health status group (73.2/88), and the group aged 18-35 years (70.3/88). While exhibiting strong scores in Section 1 and Section 4, Bard was weaker in Section 2 and Section 3. Copilot demonstrated a very good overall score, and outperformed all demographics in sections 1, 2, and 4, demonstrating a weaker performance for Section 3.

AI systems overall performance

For statistical purposes, the performances of the four AI systems were averaged to be compared against those of dietetics and English students. AI systems overall mean score was 77.3±5.1. For the “Dietary Recommendations” section the mean score was 16.8±1.3, in the “Food Groups” category the mean was 30.0±2.9 and the “Healthy Food Choices” section had a mean score of 11.3±1.3. Lastly, the “Diet, Disease and Weight Management” category showed a mean of 19.3±1.3. At the ANOVA, mean scores of the performances of all AI systems and dietetics students did not differ in a statistically significant way (both overall and for each section), while they differed from the scores achieved by English students (overall, with a p-value of 0.0292, and for the sections 1 and 4, with p-values of 0.0105 and 0.0060, respectively) (Table 1).

Table 1. Findings from ANOVA and post-hoc analyses comparing the average performance of the AI systems against nutrition and English students.

GNKQ-R	Overall ANOVA	Post-hoc analyses	
		Nutrition students *vs* Averaged AI	English students *vs* Averaged AI
Overall GNKQ-R	F=59.3, p=0.0000	Diff=-2.1, 95%CI=-10.8 to 6.7, p=0.8451	Diff=9.6, 95%CI=0.8 to 18.3, p=0.0292
Section 1	F=38.2, p=0.0000	Diff=0.5, 95%CI=-1.8 to 2.7, p=0.8884	Diff=2.9, 95%CI=0.6 to 5.1, p=0.0105
Section 2	F=36.9, p=0.0000	Diff=-1.9, 95%CI=-6.4 to 2.6, p=0.5769	Diff=2.8, 95%CI=-1.7 to 7.3, p=0.3060
Section 3	F=8.6, p=0.0003	Diff=-0.5, 95%CI=-2.4 to 1.5, p=0.8519	Diff=0.6, 95%CI=-1.4 to 2.5, p=0.7878
Section 4	61.4, p=0.0000	Diff=-0.1, 95%CI=-2.6 to 2.5, p=0.9948	Diff=3.4, 95%CI=0.8 to 5.9, p=0.0060

Concerning the other demographics, the average performance of the AI systems passed the performance of male participants in terms of the overall score (p=0.0177), Section 1 (p=0.0018), Section 2 (p=0.0175), Section 4 (p=0.0004), and Section 3 in a borderline way (p=0.0898). Performance was comparable with female participants, slightly outperforming in Section 1 (p=0.0668) and Section 4 (p=0.1441) (Table 2).

Table 2. Findings from ANOVA and post-hoc analyses comparing the average performance of the AI systems against male and female participants.

GNKQ-R	Overall ANOVA	Post-hoc analyses	
		Males *vs* Averaged AI	Females *vs* Averaged AI
Overall GNKQ-R	F=33.5, p=0.0000	Diff=15.6, 95%CI=2.2 to 28.9,	Diff=5.9, 95%CI=-7.4 to 19.1,

		p=0.0177	p=0.5518
Section 1	F=20.2, p=0.0000	Diff=3.4, 95%CI=1.1 to 5.6, p=0.0018	Diff=2.2, 95%CI=-0.1 to 4.4, p=0.0668
Section 2	F=46.7, p=0.0000	Diff=5.8, 95%CI=0.8 to 10.8, p=0.0175	Diff=1.5, 95%CI=-3.4 to 6.4, p=0.7540
Section 3	F=42.6, p=0.0000	Diff=2.1, 95%CI=-0.2 to 4.3, p=0.0898	Diff=0.2, 95%CI=-2.1 to 2.4, p=0.9866
Section 4	F=47.2, p=0.0000	Diff=4.5, 95%CI=1.7 to 7.2, p=0.0004	Diff=2.2, 95%CI=-0.5 to 4.8, p=0.1441

Compared with participants with degrees or higher titles (Table 3), the average performance of the AI systems was similar, with the exception of Section 1 (p=0.0399) where it was higher. It outperformed participants with no degree in Section 1 (p=0.0003) and Section 4 (p=0.0595).

Table 3. Findings from ANOVA and post-hoc analyses comparing the average performance of the AI systems against participants of various education groups.

GNKQ-R	Overall ANOVA	Post-hoc analyses	
		No degree *vs* Averaged AI	Degree or higher *vs* Averaged AI
Overall GNKQ-R	F=19.4, p=0.0000	Diff=11.5, 95%CI=-2.0 to 24.9, p=0.1138	Diff=5.0, 95%CI=-8.5 to 18.4, p=0.6640
Section 1	F=35.3, p=0.0000	Diff=3.0, 95%CI=1.2 to 4.7, p=0.0003	Diff=1.9, 95%CI=0.1 to 3.6, p=0.0399
Section 2	F=20.2, p=0.0000	Diff=4.1, 95%CI=-2.3 to 10.5, p=0.2883	Diff=0.9, 95%CI=-5.5 to 7.3, p=0.9416
Section 3	F=25.3, p=0.0000	Diff=1.2, 95%CI=-0.6 to 2.9, p=0.2835	Diff=0.2, 95%CI=-1.6 to 1.9, p=0.9785
Section 4	F=13.9, p=0.0000	Diff=3.5, 95%CI=-0.1 to 7.0, p=0.0595	Diff=2.1, 95%CI=-1.5 to 5.6, p=0.3663

Finally, no differences could be found when comparing the AI systems against human participants in terms of age and health status (Table 4 and Table 5).

Table 4. Findings from ANOVA and post-hoc analyses comparing the average performance of the AI systems against participants of various age groups.

GNKQ-R	Overall ANOVA	Post-hoc analyses		
		18–35 years *vs* Averaged AI	36-50 years *vs* Averaged AI	>50 years *vs* Averaged AI

Overall GNKQ-R	F=0.0, p=0.9960	Diff=7.0, 95%CI=-192.3 to 206.2, p=1.0000	Diff=7.5, 95%CI=-193.4 to 208.3, p=1.0000	Diff=10.6, 95%CI=-189.3 to 210.4, p=0.9991
Section 1	F=0.0, p= 0.9943	Diff=2.1, 95%CI=-39.3 to 43.4, p=0.9992	Diff=2.6, 95%CI=-39.1 to 44.2, p=0.9985	Diff=2.9, 95%CI=-38.6 to 44.3, p=0.9978
Section 2	F=0.0, p= 0.9971	Diff=2.1, 95%CI=-92.2 to 96.4, p=1.0000	Diff=2.0, 95%CI=-93.0 to 97.0, p=1.0000	Diff=3.6, 95%CI=-91.0 to 98.2, p=0.9992
Section 3	F=0.0, p= 0.9979	Diff=0.6, 95%CI=-35.5 to 36.6, p=1.0000	Diff=0.5, 95%CI=-35.9 to 36.8, p=1.0000	Diff=1.1, 95%CI=-35.1 to 37.2, p=1.0000
Section 4	F=0.0, p= 0.9975	Diff=2.5, 95%CI=-53.6 to 58.5, p=0.9992	Diff=2.8, 95%CI=-53.8 to 59.3, p=0.9992	Diff=3.3, 95%CI=-53.0 to 59.5, p=0.9987

Table 5. Findings from ANOVA and post-hoc analyses comparing the average performance of the AI systems against participants of various health status groups.

GNKQ-R	Overall ANOVA	Post-hoc analyses		
		Poor health status *vs* Averaged AI	Good health status *vs* Averaged AI	Excellent health status *vs* Averaged AI
Overall GNKQ-R	F=0.1, p= 0.9764	Diff=12.2, 95%CI=-183.7 to 208.0, p=0.9984	Diff=8.2, 95%CI=-187.2 to 203.5, p=0.9992	Diff=4.0, 95%CI=-192.4 to 200.5, p=1.0000
Section 1	F=0.0, p= 0.9862	Diff=3.1, 95%CI=-37.4 to 43.5, p=0.9972	Diff=2.4, 95%CI=-38.0 to 42.7, p=0.9987	Diff=1.8, 95%CI=-38.8 to 42.3, p=0.9992
Section 2	F=0.0, p= 0.9857	Diff=4.2, 95%CI=-87.8 to 96.2, p=0.9992	Diff=2.3, 95%CI=-89.5 to 94.1, p=1.0000	Diff=1.0, 95%CI=-91.3 to 93.3, p=1.0000
Section 3	F=0.1, p=	Diff=1.3,	Diff=0.7,	Diff=-0.1,

	0.9843	95%CI=-34.0 to 36.5, p=1.0000	95%CI=-34.5 to 35.8, p=1.0000	95%CI=-35.4 to 35.3, p=1.0000
Section 4	F=0.1, p= 0.9756	Diff=3.8, 95%CI=-49.5 to 57.0, p=0.9977	Diff=2.9, 95%CI=-50.3 to 56.0, p=0.9990	Diff=1.6, 95%CI=-51.9 to 55.0, p=1.0000

Discussion

The significance of proper nutrition knowledge in shaping dietary behaviors and health outcomes cannot be overstated in today's world, where nutritional misinformation and disinformation are rampant. With the rise of digital platforms as sources of dietary guidance, AI-based tools like conversational agents are becoming increasingly popular for providing personalized nutritional information [24,25]. However, the reliability of these AI systems in addressing nutrition-related inquiries is not well-established, with only a few studies having evaluated the proficiency of these tools in nutrition, often without validated methodologies or comprehensive comparisons across different AI tools.

To address this gap, this comparative analysis of the scores in each section of a validated questionnaire assessing the general nutrition knowledge is aimed to offer valuable insights into the present capabilities and potential roles of AI in the fields of nutrition and dietetics.

The performance of the AI systems, especially ChatGPT-4 and ChatGPT-3.5, was remarkable, often rivaling or surpassing the dietetics students in specific sections. This suggests that these AI systems have a robust, up-to-date nutrition knowledge, which can be valuable in disseminating nutrition information and making dietary recommendations, with ChatGPT-4 showing high competence in providing nutrition-related information and being better aligned with established nutritional guidelines.

This is in line with the few existing studies that have demonstrated that AI-based tools, such as ChatGPT, hold promise as valuable tools for addressing frequently posed nutrition queries to dietitians, offering supportive evidence for the potential utility of conversational agents in delivering nutritional guidance [18-20]. Haman and coworkers [18] showed that ChatGPT was highly accurate and consistent in providing nutritional estimations, particularly in energy values (97% within a 40% range of the data from the United States Department of Agriculture, USDA), and efficient in devising daily meal plans that aligned closely with USDA caloric values. Kirk and colleagues [19] solicited the most frequent nutrition-related queries from dietitians along with their expert responses. These queries were subsequently posed to ChatGPT, and both sets of replies were then evaluated by a panel of eighteen dietitians and nine subject matter experts for scientific accuracy, practical applicability, and clarity. ChatGPT's responses attained higher overall scores compared to those from dietitians in five out of eight questions. Specifically, ChatGPT outperformed on scientific accuracy in five instances, on practical applicability in four, and on clarity in five. In comparison, the dietitians' responses did not surpass ChatGPT's average scores for any question, whether overall or in individual grading criteria. Last, Ponzo and colleagues [20] tested the ability of ChatGPT in providing nutritional guidelines in relation to different NCDs, comparing the responses given by ChatGPT to the most up-to-date nutritional guidelines. ChatGPT showed to be accurate in giving nutritional recommendation for specific diseases, with better performance for non-alcoholic fatty liver disease, type 2 diabetes, and hypercholesterolemia/hypertriglyceridemia. In addition, authors examined the performance of ChatGPT in providing personalized dietary advice in patients with multiple NCDs. In this case, the weaknesses of the AI system emerged, underlining its useless in complex situations in which personalized nutritional strategies are needed. However, these studies were limited to ChatGPT alone and did not explore other AI systems.

To the best of our knowledge, this study is the first in which several AI systems are tested and compared using a validated nutrition knowledge questionnaire to test their ability to provide proper nutrition information and advice for the general population. Our findings show variation in AI performances across different nutrition-related sections, which indicates that while the tested AI systems are highly knowledgeable, there are areas where they can still improve or lack the nuanced understanding that specialized human education provides. The high scores in “Healthy Food Choices” and “Diet, Disease and Weight Management” sections by some AI systems suggest their potential utility in public health and educational settings, especially for providing general guidance and assist health professionals.

Some common mistakes among AI systems could be noted, with errors being mostly around specific food group classifications, and the relationship between diet and some health conditions. These results suggest that while AI can be highly knowledgeable in nutrition, there are specific situations and details that may be challenging.

On the other hand, the differences between the two tested versions of ChatGPT (3.5 and 4) indicate advancements in AI capabilities over time, with newer versions showing improved accuracy in nutrition-related knowledge.

Future directions

The rapid advancement and widespread adoption of AI-based tools like ChatGPT herald a new era in various fields, including nutrition and dietetics. As these technologies evolve, they are poised to become invaluable assistants for health professionals, offering quick information retrieval, draft text generation, and educational content creation. However, their limitations in accuracy, bias, and depth of understanding underscore the irreplaceable value of human expertise and judgment. Looking ahead, the integration of AI into professional practice must be approached with caution and a focus on complementing rather than replacing human knowledge, skills, and expertise. The future will likely see a collaborative synergy between AI tools and human health professionals, enhancing efficiency and accessibility while maintaining the quality and personalization of services, even for the most complex cases. Education and continuous training will play a crucial role in equipping professionals with the skills to effectively utilize these technologies, ensuring they remain at the forefront of their respective fields [15].

Strengths and limitations

The present study has some strengths, which should be properly acknowledged, including its novelty, its methodological rigor, and the systematic appraisal of major AI systems using a validated questionnaire. However, it suffers from some limitations: being a cross-sectional survey, it does not capture ongoing development and refinement in AI systems. Moreover, only general nutrition knowledge was tested: therefore, the findings cannot be generalized to situations in which highly specialized professional advice may be needed, especially in cases of medical nutrition therapy or specific dietary needs.

Conclusions

The present comparison highlights the advancements in AI in understanding complex subjects like nutrition and its potential role in providing general information, or as an adjunct educational tool, in an era where digital and AI-based sources are increasingly sought for health-related guidance. However, it also underscores the irreplaceable value of specialized human education and expertise, and the importance of combining AI tools with expert human judgment in the fields of nutrition and dietetics.

Acknowledgements

NB: conceptualization, methodology, and data analysis and data presentation. NB and AR: writing—original draft preparation. SM, FB, and FS: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for proposals No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Award Number: Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title "ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods". The sponsor did not have any role in the study design, the collection, analysis and interpretation of data, nor in the writing process.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence
GNKQ-R: General Nutrition Knowledge Questionnaire–Revised
LLMs: large language models
NCDs: non-communicable diseases

Multimedia Appendix 1

Prompts used to query each AI system.

Multimedia Appendix 2

Replies provided by AI systems.

References

1. Brauer, M, Roth, GA, Aravkin, AY, Zheng, P, Abate, KH, Abate, YH, et al. Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2024;403(10440):2162-203. PMID: 38762324
2. Afshin, A, Sur, PJ, Fay, KA, Cornaby, L, Ferrara, G, Salama, JS, et al. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2019;393(10184):1958-72. PMID: 30954305
3. Barbosa, LB, Vasconcelos, SML, Correia, LODS, Ferreira, RC. Nutrition knowledge assessment studies in adults: a systematic review. Cien Saude Colet 2016;21:449-62. PMID: 26910153
4. Spronk, I, Kullen, C, Burdon, C, O'Connor, H. Relationship between nutrition knowledge and dietary intake. Br J Nutr 2014;111(10):1713-26. PMID: 24621991
5. Wardle, J, Parmenter, K, Waller, J. Nutrition knowledge and food intake. Appetite 2000;34(3):269-75. PMID: 10888290
6. Scalvedi ML, Gennaro L, Saba A, Rossi L. Relationship Between Nutrition Knowledge and Dietary Intake: An Assessment Among a Sample of Italian Adults. Front Nutr 2021;8:714493. PMID: 34589511
7. Ayoob KT, Duyff RL, Quagliani D; American Dietetic Association. Position of the American Dietetic Association: food and nutrition misinformation. J Am Diet Assoc 2002;102(2):260-6.

PMID: 11846124.

8. Diekman C, Ryan CD, Oliver TL. Misinformation and Disinformation in Food Science and Nutrition: Impact on Practice. J Nutr 2023;153(1):3-9. PMID: 36913465.
9. Cena H, Calder PC. Defining a Healthy Diet: Evidence for The Role of Contemporary Dietary Patterns in Health and Disease. Nutrients 2020;12(2):334. PMID: 32012681
10. Brown R, Seabrook JA, Stranges S, Clark AF, Haines J, O'Connor C, Doherty S, Gilliland JA. Examining the Correlates of Adolescent Food and Nutrition Knowledge. Nutrients 2021;13(6):2044. PMID: 34203666
11. Tan, SSL, Goonawardene, N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res 2017;19(1), e9. PMID: 28104579
12. Côté M, Lamarche B. Artificial intelligence in nutrition research: perspectives on current and future applications. Appl Physiol Nutr Metab 2022; 47(1):1-8. PMID: 34525321
13. Arslan S. Exploring the Potential of Chat GPT in Personalized Obesity Treatment. Ann Biomed Eng 2023;51(9):1887-1888. PMID: 37145177
14. Khan U. Revolutionizing Personalized Protein Energy Malnutrition Treatment: Harnessing the Power of Chat GPT. Ann Biomed Eng 2024; 52(5):1125-1127. PMID: 37728811.
15. Chatelan A, Clerc A, Fonta PA. ChatGPT and Future Artificial Intelligence Chatbots: What may be the Influence on Credentialed Nutrition and Dietetics Practitioners? J Acad Nutr Diet 2023;123(11):1525-1531. PMID: 37544375
16. Bommasani R, Liang P, Lee T. Holistic Evaluation of Language Models. Ann N Y Acad Sci 2023;1525(1):140-146. PMID: 37230490.
17. Niszczota P, Rybicka I. The credibility of dietary advice formulated by ChatGPT: Robo-diets for people with food allergies. Nutrition 2023;112:112076. PMID: 37269717
18. Haman M, Školník M, Lošťák M. AI dietitian: Unveiling the accuracy of ChatGPT's nutritional estimations. Nutrition 2023;119:112325. PMID: 38194819
19. Kirk D, van Eijnatten E, Camps G. Comparison of Answers between ChatGPT and Human Dieticians to Common Nutrition Questions. J Nutr Metab 2023; 2023:5548684PMID: 38025546.
20. 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(4), 469. PMID: 38398794
21. Vaswani, A, Shazeer, N, Parmar, N, Uszkoreit, J, Jones, L, Aidan, NG, Kaiser, Ł, Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems 2017;30. ISBN: 9781510860964
22. Kliemann N, Wardle J, Johnson F, Croker H. Reliability and validity of a revised version of the General Nutrition Knowledge Questionnaire. Eur J Clin Nutr 2016;70(10):1174-1180. PMID: 27245211
23. Parmenter K, Wardle J. Development of a general nutrition knowledge questionnaire for adults. Eur J Clin Nutr 1999; 53(4):298-308. PMID: 10334656.
24. World Health Organization. Digital food environments: factsheet (No. WHO/EURO: 2021-2755-42513-59052). World Health Organization. Regional Office for Europe. Geneva. 2023. https://iris.who.int/bitstream/handle/10665/342072/WHO-EURO-2021-2755-42513-59052-eng.pdf?sequence=1
25. Detopoulou, P, Voulgaridou, G, Moschos, P, Levidi, D, Anastasiou, T, Dedes, V, et al. Artificial intelligence, nutrition, and ethical issues: A mini-review. Clin Nutr Open Sci; 2023;50:46-56. https://doi.org/10.1016/j.nutos.2023.07.001