

DOI 10.18372/2786-5487.1.21510

УДК 004.8:37.018.43:613.2

Yelyzaveta Chala 

Graduate student,

McGill University, School of Human Nutrition,

Quebec, Canada

yelyzaveta.chala@mail.mcgill.ca

Nadiia Adamchuk-Chala 

PhD in Biology, Research Associate,

McGill University,

Quebec, Canada

nadiia.adamchuk-chala@mcgill.ca

Nadezhda Yefymyshch,

PhD, Expert on language policy and national minority in Ukraine,

Doctoral school History Eastern European,

Faculty of Humanity,

Eötvös Loránd University,

Budapest, Hungary

komi@ua.fm

**ARTIFICIAL INTELLIGENCE IN EDUCATION:
ASSESSING THE ACCURACY OF AI-GENERATED ANSWERS
TO COMMON NUTRITION QUESTIONS**

***Annotation.** This article evaluates the accuracy of online nutrition information generated by ChatGPT, Microsoft Copilot, and Gemini by comparing their responses to common dietary questions with peer-reviewed scientific evidence. Forty questions across four nutrition domains were assessed for accuracy, completeness, clarity, additional*

value, and source quality, with statistical differences analyzed using One-Way ANOVA and post-hoc tests. The results show that while all tools provided partially reliable information, Gemini 2.5 Flash performed comparatively better in several areas, highlighting both the potential and limitations of AI in digital nutrition guidance.

Key words: Artificial intelligence, ChatGPT 4.0, Microsoft Copilot, Gemini 2.5 Flash, nutrition information, accuracy

Анотація. У цій статті оцінюється точність інформації про харчування, що надається в Інтернеті системами ChatGPT, Microsoft Copilot та Gemini, шляхом порівняння їхніх відповідей на поширені питання про харчування з рецензованими науковими даними. Сорок питань з чотирьох областей харчування були оцінені на точність, повноту, зрозумілість, додаткову цінність та якість джерел, а статистичні відмінності були проаналізовані за допомогою однофакторного дисперсійного аналізу та пост-хок тестів. Результати показують, що хоча всі інструменти надавали частково надійну інформацію, Gemini 2.5 Flash показав порівняно кращі результати в декількох областях, підкресливши як потенціал, так і обмеження штучного інтелекту в цифровому харчуванні.

Ключові слова: штучний інтелект, ChatGPT 4.0, Microsoft Copilot, Gemini 2.5 Flash, інформація про харчову цінність, точність

Introduction. The World Wide Web represents one of the largest repositories of nutrition-related information, with more than 4.2 billion active web pages available to the public (Hingoro & Nawaz, 2021) [1]. Although search engines provide rapid access to dietary content, the reliability and scientific accuracy of online information remain inconsistent (Edosomwan & Edosomwan, 2010) [2]. In recent years, artificial intelligence (AI) tools such as ChatGPT, Gemini, and Microsoft Copilot have emerged as widely used sources of nutrition advice, generating natural language responses to user queries. While these systems enhance accessibility and convenience, their clarity,

completeness, and reliability are not yet sufficiently validated (Myers, C., 2024; Solomon & Laye, 2025) [3; 4].

Empirical evidence indicates that AI-generated nutrient estimates may substantially deviate from established reference standards. Reported discrepancies of up to $\pm 50\%$ compared with data from the United States Department of Agriculture nutrient databases highlight the potential risk of inaccurate reporting of caloric and macronutrient values (Haman et al., 2024) [5]. Such variability raises concerns regarding the suitability of AI-generated outputs for clinical application and public health nutrition guidance.

Unlike licensed dietitians, AI-driven platforms do not systematically verify their responses against peer-reviewed scientific evidence. As a result, they may disseminate misinformation, oversimplified recommendations, or omit critical factors such as cultural dietary patterns, allergies, and medical conditions. Although AI applications in healthcare have been increasingly studied, relatively few investigations have directly compared AI-generated nutrition information with evidence-based dietary guidelines. Consequently, the credibility, safety, and clinical utility of AI tools in dietetic practice remain insufficiently explored.

The purpose of the article is to comprehensively evaluate the quality, reliability, clarity, and completeness of responses generated by ChatGPT, Gemini, and Microsoft Copilot to the most frequently asked nutrition questions across four domains by comparing them with evidence-based peer-reviewed literature and statistically analyzing their performance using ANOVA, post-hoc tests, and effect size estimation.

Research Results. A One-Way ANOVA revealed domain-specific differences in the performance of ChatGPT, Microsoft Copilot, and Gemini across four nutrition domains and overall scores.

In the *Herbs, Foods and Phytochemicals* domain, no statistically significant differences were observed among ChatGPT ($89,5 \pm 0,0$), Copilot ($81,6 \pm 16,9$), and Gemini ($83,7 \pm 12,8$) ($p = 0,338$).

In *Nutrition and Behaviour*, significant differences were identified ($p = 0,043$). Gemini ($84,2 \pm 0,0$) scored significantly higher than Copilot ($82,1 \pm 2,7$), while ChatGPT ($83,7 \pm 1,7$) did not differ significantly from either engine.

For *Nutrition and Exercise*, differences were highly significant ($p < 0,001$). Gemini ($94,7 \pm 0,0$) significantly outperformed both ChatGPT ($78,9 \pm 0,0$) and Copilot ($73,7 \pm 0,0$), and ChatGPT significantly outperformed Copilot.

Similarly, in *Pediatric and Maternal Nutrition*, significant differences were detected ($p < 0,001$). Gemini ($99,0 \pm 3,3$) achieved the highest performance, followed by ChatGPT ($86,3 \pm 5,1$) and Copilot ($80,0 \pm 4,2$), with all pairwise differences statistically significant.

Overall performance across all domains also demonstrated significant differences ($p < 0,001$), with Gemini ($90,4 \pm 9,2$) scoring higher than ChatGPT ($84,6 \pm 4,7$) and Copilot ($79,3 \pm 9,1$); ChatGPT likewise significantly outperformed Copilot. Post-hoc comparisons were conducted using Tukey tests for domain-specific analyses and the Games–Howell procedure for overall comparisons ($p = 0,05$).

This study provides a systematic and multidimensional comparison of AI-generated nutrition information across four distinct domains using the ten most common search queries in each category. Unlike prior investigations that focused narrowly on calorie calculations or sample meal plans, the present research examined herbs and phytonutrients, nutrition and behaviour, nutrition and exercise, and pediatric and maternal nutrition, thereby increasing thematic coverage and ecological validity.

Methodologically, the study applied inferential statistics, including ANOVA, post-hoc testing, and effect size estimation, allowing for identification of both statistically and practically meaningful differences among AI systems. The evaluation framework incorporated multiple criteria—accuracy, completeness, clarity, quality of additional information, and reliability of sources—adapted from Solomon and Laye (2025) [4], thus avoiding reliance on single-indicator assessment.

All AI responses were cross-checked against peer-reviewed scientific literature and official clinical guidelines, including those issued by the United States Department of Agriculture, ensuring alignment with authoritative evidence. Furthermore, the use of real-world search engine queries enhanced relevance compared to earlier research employing artificial “zero-shot” prompts (Solomon & Laye, 2025) [4].

The findings may have limited generalizability beyond frequently searched nutrition questions. Additionally, zero-variance scores were observed, as each AI engine consistently produced identical evaluation totals across repeated assessments (ChatGPT-4o: 15/19; Microsoft 365 Copilot: 14/19; Google Gemini 2.5 Flash: 18/19), which constrained the robustness of inferential comparisons despite adherence to the methodological framework of Solomon et al. (2025) [4].

Potential bias may also have arisen from the selection of search queries and the single-reviewer evaluation process, rather than multiple independent raters typically recommended in scientific research. Finally, although AI systems demonstrate potential as supportive informational tools, their application in nutrition should remain subject to professional oversight to ensure safety, accuracy, and clinical appropriateness.

This study compared the performance of three large language models—ChatGPT, Gemini, and Microsoft Copilot—in responding to nutrition-related queries across multiple thematic domains. The evaluation framework incorporated accuracy, completeness, clarity, quality of additional information, and source reliability, with all responses benchmarked against peer-reviewed nutrition literature and official dietary guidelines. In the herbs and foods domain, all three systems performed consistently when their answers were aligned with established scientific sources, suggesting that for general and widely documented topics, contemporary large language models can reproduce mainstream evidence with reasonable fidelity.

More pronounced differences emerged in complex domains requiring integration of physiological knowledge, life-stage considerations, and applied dietary recommendations. Gemini outperformed both ChatGPT and Copilot in nutrition and

physical activity as well as pediatric and maternal health. These areas demand contextual reasoning about macronutrient timing, micronutrient safety thresholds, growth and developmental requirements, and risk-sensitive guidance. The superior results achieved by Gemini likely reflect structural and architectural differences in system design. According to publicly available documentation (Google, 2025; DeepMind, 2025) [6; 7], Gemini 2.5 Flash integrates directly with Google Search infrastructure, enabling dynamic retrieval of up-to-date web information. It also supports extended context windows, allowing the model to process longer prompts and synthesize more detailed responses. Furthermore, the Pro version incorporates Deep Research functionality designed to generate structured, evidence-oriented reports. Such features may enhance the system's ability to contextualize nuanced nutrition questions and synthesize multi-source information in real time.

In contrast, Microsoft Copilot operates by accessing and summarizing publicly available web content and frequently provides structured answers with citations (Microsoft, 2024) [8]. This citation-based approach can increase transparency and user confidence by indicating information sources. However, its coverage is constrained by accessibility limitations; Copilot cannot retrieve content from subscription-based academic journals or paywalled scientific databases. Given that a substantial proportion of high-quality nutrition research is published in such venues, this restriction may limit the depth and precision of evidence integration in more specialized domains. As a result, Copilot's comparatively lower performance in complex categories may be attributable not only to model reasoning capacity but also to the breadth and quality of retrievable data.

ChatGPT, meanwhile, does not inherently possess real-time access to scientific databases and relies primarily on static pretraining data rather than continuous live retrieval. This design introduces potential risks, particularly in rapidly evolving areas of nutrition science where recommendations are periodically updated. Without dynamic integration of newly published guidelines, there is an increased possibility of outdated or

generalized statements. Nevertheless, ChatGPT demonstrated competitive overall performance in this study. Its broad training across diverse textual corpora contributes to linguistic fluency, structured explanations, and logically coherent answers. Moreover, its strengths in clarity, contextualization, and the provision of supplemental explanatory details may enhance user comprehension and perceived credibility (Rahman Nazil, 2023; Financial Times, 2023) [9; 10]. These attributes likely contributed to its strong relative standing despite architectural constraints.

The present findings align with emerging evidence suggesting that iterative advancements in large language model development are associated with measurable improvements in factual reliability. For example, research indicates that newer ChatGPT versions, such as GPT-4, demonstrate enhanced nutrient estimation accuracy and generate more diverse and nutritionally balanced meal recommendations compared with earlier iterations (Guo, 2025) [11]. Such improvements reflect ongoing refinements in training data curation, model scaling, and alignment strategies. At the same time, these enhancements underscore the dynamic nature of AI performance; evaluations conducted at a single time point may not remain fully representative as models continue to evolve.

Despite encouraging results, important limitations remain within the broader evidence base. Many existing studies assessing AI-generated nutrition guidance are conducted under controlled experimental conditions, often using standardized prompts and structured evaluation criteria. While this methodological rigour supports internal validity, it may reduce generalizability to real-world settings in which users pose ambiguous, incomplete, or highly personalized questions. Furthermore, the interactive nature of AI systems—where follow-up prompts can refine or alter recommendations—is rarely captured in cross-sectional evaluation designs.

Crucially, there is a lack of longitudinal research examining the real-world consequences of relying on AI-derived dietary guidance. To date, no published studies have systematically investigated whether sustained use of ChatGPT, Gemini, or Copilot influences dietary adherence, nutrient adequacy, metabolic biomarkers, or broader health

outcomes over time. The absence of outcome-based validation limits definitive conclusions regarding clinical safety and effectiveness. While short-term accuracy assessments provide valuable insight into informational quality, they do not address whether AI-generated advice translates into improved or impaired health behaviours.

Taken together, the results suggest that differences in system architecture—particularly regarding search integration, retrieval capacity, context processing, and data accessibility—play a meaningful role in shaping performance across nutrition domains. Gemini’s integration with search and research-oriented features appears advantageous in complex and evidence-sensitive contexts. Copilot’s citation transparency enhances interpretability but is limited by access constraints. ChatGPT, despite lacking live database connectivity, remains competitive due to broad pretraining and strengths in explanatory clarity.

These findings highlight both the promise and the current limitations of large language models in nutrition science. AI systems may serve as accessible supplementary informational tools, especially for general knowledge queries. However, their outputs should be interpreted cautiously, particularly in clinically sensitive domains such as maternal and pediatric nutrition. Until longitudinal outcome data and robust validation studies become available, AI-generated dietary guidance should complement—not replace—professional dietetic expertise and evidence-based clinical judgment.

Conclusions. AI-generated responses to nutrition queries exhibit variable but generally promising alignment with peer-reviewed scientific evidence. Gemini 2.5 Flash consistently outperformed ChatGPT and Copilot in complex domains, highlighting the importance of system design and information retrieval capabilities. Despite these strengths, AI tools cannot replace professional dietetic judgment, particularly in context-sensitive areas such as pediatric and maternal nutrition. With continued validation, integration of authoritative sources, and methodological standardization, AI systems have the potential to serve as reliable and accessible supplementary resources in digital nutrition guidance.

References

1. Hingoro, M. A., & Nawaz, H. (2021). A comparative analysis of search engine ranking algorithms. *International Journal of Advanced Trends in Computer Science and Engineering*, 10(2), 1247–1252. <https://doi.org/10.30534/ijatcse/2021/1081022021> (last accessed: 2026/01/12).
2. Edosomwan, J., & Edosomwan, T. O. (2010). Comparative analysis of some search engines. *South African Journal of Science*, 106(11/12), Article #169. <https://doi.org/10.4102/sajs.v106i11/12.169>
3. Myers, C. (2024, February 9). Should you be using ChatGPT for nutrition advice? New study says not. *EatingWell*. URL: <https://www.eatingwell.com/should-you-be-using-chatgpt-for-nutrition-advice-new-study-says-not-8575274> (last accessed: 2026/01/12).
4. Solomon, T. P. J., & Laye, M. J. (2025). The sports nutrition knowledge of large language model (LLM) artificial intelligence (AI) chatbots: An assessment of accuracy, completeness, clarity, quality of evidence, and test–retest reliability. *PLOS ONE*, 20(6), e0325982. <https://doi.org/10.1371/journal.pone.0325982>
5. Haman, M., Školník, M., & Lošťák, M. (2024). AI dietician: Unveiling the accuracy of ChatGPT’s nutritional estimations. *Nutrition*, 119, Article 112325. <https://doi.org/10.1016/j.nut.2023.112325>
6. Google. (2025). *Gemini Deep Research overview*. URL: <https://gemini.google/overview/deep-research> (last accessed: 2026/01/12).
7. DeepMind. (2025). *Gemini 2.5 technical report*. Google DeepMind. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf
8. Microsoft. (2024). *Create effective prompts*. Microsoft Learn. URL: <https://learn.microsoft.com/en-us/copilot/security/prompting-tips> (last accessed: 2026/01/12).
9. Rahman Nazil, A. (2023, December 24). Why ChatGPT creates scientific citations that do not exist. *DEV Community*. URL:

https://dev.to/ashikur_rahmannazil93/why-chatgpt-creates-scientific-citations-that-dont-exist-47co (last accessed: 2026/01/12).

10. Financial Times. (2023, May 29). Why ChatGPT gets things wrong. *Financial Times*. URL: <https://www.ft.com/content/7a4e7eae-f004-486a-987f-4a2e4dbd34fb> (last accessed: 2026/01/12).

11. Guo, P., Liu, G., Xiang, X., & An, R. (2025). From AI to the table: A systematic review of ChatGPT's potential and performance in meal planning and dietary recommendations. *Dietetics*, 4(1), 7. <https://doi.org/10.3390/dietetics4010007>