healthcare

MDPI

Article

ChatGPT-4o and 4o1 Preview as Dietary Support Tools in a Real-World Medicated Obesity Program: A Prospective Comparative Analysis

Louis Talay [1,*], Leif Lagesen [2], Adela Yip [3], Matt Vickers [4] and Neera Ahuja [5]

[1] Faculty of Arts and Social Sciences, University of Sydney, Sydney, NSW 2050, Australia
[2] Dietitians Australia, Phillip, ACT 2606, Australia; leif.apd@gmail.com
[3] Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA
[4] Eucalyptus, Sydney, NSW 2000, Australia
[5] Hospital Medicine Division, Department of Medicine, Stanford University, Stanford, CA 94305, USA; nkahuja@stanford.edu
* Correspondence: louis.talay@sydney.edu.au

Abstract: Background/Objectives: Clinicians are becoming increasingly interested in the use of large language models (LLMs) in obesity services. While most experts agree that LLM integration would increase access to obesity care and its efficiency, many remain skeptical of their scientific accuracy and capacity to convey human empathy. Recent studies have shown that ChatGPT-3 models are capable of emulating human dietitian responses to a range of basic dietary questions. Methods: This study compared responses of two ChatGPT-4o models to those from human dietitians across 10 complex questions (5 broad; 5 narrow) derived from patient–clinician interactions within a real-world medicated digital weight loss service. Results: Investigators found that neither ChatGPT-4o nor Chat GPT-4o1 preview were statistically outperformed ($p < 0.05$) by human dietitians on any of the study's 10 questions. The same finding was made when scores were aggregated from the ten questions across the following four individual study criteria: scientific correctness, comprehensibility, empathy/relatability, and actionability. Conclusions: These results provide preliminary evidence that advanced LLMs may be able to play a significant supporting role in medicated obesity services. Research in other obesity contexts is needed before any stronger conclusions are made about LLM lifestyle coaching and whether such initiatives increase care access.

Keywords: ChatGPT-4o; dietetics; health coaching; digital weight loss services; weight loss medications

check for updates

Academic Editor: Daniele Giansanti

Received: 13 February 2025
Revised: 13 March 2025
Accepted: 14 March 2025
Published: 16 March 2025

Citation: Talay, L.; Lagesen, L.; Yip, A.; Vickers, M.; Ahuja, N. ChatGPT-4o and 4o1 Preview as Dietary Support Tools in a Real-World Medicated Obesity Program: A Prospective Comparative Analysis. *Healthcare* **2025**, *13*, 647. https://doi.org/10.3390/healthcare13060647

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Study Context

Obesity has arguably become the most serious global health problem [1]. Over forty percent of the world's adults were overweight in 2022, and sixteen percent were living with obesity [1]. These figures tend to be even higher in Anglo-Saxon countries such as Australia, where adult overweight and obesity rates reached sixty-six and thirty-two percent in 2022, respectively [2]. Large analyses have revealed that women, people residing in households with upper-middle incomes, and people in the Americas, Polynesia, or Micronesia are overrepresented in obesity statistics [3,4]. As a chronic complex illness, obesity requires continuous multidisciplinary treatment [5]. Patients with significant work and/or family commitments have historically struggled to adhere to such treatment in face-to-face (F2F)

settings, where they often have to travel between clinics and compete for limited feasible appointment times [6,7]. Digital weight-loss services (DWLSs) have recently emerged as a promising solution to these access barriers to obesity care [8–10]. Certain DWLSs offer asynchronous options, which allow patients to access care at a time and place convenient to them [10]. However, a drawback of these asynchronous care models is that patients have to wait for a member of their multidisciplinary team (MDT) to respond to their input. Against the backdrop of this issue, stakeholders have become interested in the potential of large language models (LLMs) to deliver lifestyle coaching in DWLSs [11,12]. Lifestyle coaches (typically dietitians or nutritionists) should have the most contact with patients out of all MDT members in continuous obesity programs. If LLMs could deliver high-quality lifestyle coaching support within a DWLS, patients would not only be able to access care at a time and place of their convenience, but they would also receive real-time responses to their consultation input [12]. Moreover, the use of LLMs would likely reduce the cost of DWLSs, and it would make them more accessible to lower socioeconomic groups [13]. However, as weight loss is emotive and scientific, soliciting obesity treatment advice from non-human agents carries a considerable risk to patient safety. This risk is magnified if the DWLS prescribes modern weight loss medications, such as glucagon-like peptide-1 receptor agonists (GLP-1 RAs), which consistently yield side effects and are understudied in real-world settings [10,14].

1.2. Related Studies

ChatGPT is an LLM that uses artificial intelligence to generate human-like responses to a user's text-based prompts [15]. The model has been studied across a range of healthcare functions, including language translation, clinical diagnoses, and scientific writing [16–18]. Recent research has also explored the safety and utility of ChatGPT dietary coaching [19]. In addition to the cost and efficiency benefits discussed above, scholars have argued that ChatGPT can be an excellent supporting tool for dietitians who need to obtain quick secondary opinions or generate fast summaries for patients seeking additional information [11]. A 2025 systematic review of ChatGPT's reliability in providing dietary recommendations found that fifteen studies had hitherto focused on the LLM's performance in isolation, four had provided descriptive insights, and five had compared the LLM with human dietitians [20]. Two of the comparative studies evaluated the accuracy with which ChatGPT models could adhere to established dietary guidelines, including the US Department of Agriculture's dietary reference intake [21] and the Mayo Clinic Renal Diet Handbook for chronic kidney disease patients [22]. The former found that the LLM had difficulty catering to vegans [21], and the latter concluded that while ChatGPT-4 had outperformed its predecessor, it still made errors identifying potassium and phosphorous content in various foods [22]. Another study found that ChatGPT-4 accurately calculated protein content in 60.4% of food items selected from a United Nations report [23]. The remaining two studies from the systematic review [20] compared ChatGPT and human responses to a list of obesity patient questions [13,19]. The first of these qualitatively analyzed responses to ten distinct prompts and concluded that the frequency with which ChatGPT-3.5 provided misleading information outweighed its access benefits [13]. The latter, a 2023 study by Kirk et al., quantitatively compared ChatGPT-3.5 and human dietitian responses to eight questions using a three-item metric (scientific correctness, comprehensibility, and actionability). The study reported that the LLM outperformed human dietitians across five of the study's eight questions and achieved comparable scores for the remaining three questions [19].

However, we contend that the questions in the Kirk et al. study were too simplistic to reflect real-world patient–clinician interactions; only one of the study's eight questions contained more than ten words and a single sentence [19]. Additionally, the as-

sessment rubric lacked an empathy component—a feature of LLMs that multiple experts consider a key limitation relative to human clinicians [24–26]. Empathy is defined as the ability to understand a person's emotions or to 'see the world through someone else's eyes' [27]. Several experts stress that the term encompasses the capacity to respond to another being's emotions in a nuanced and personable manner [24,27,28]. A 2024 systematic review of LLM capacity for empathy found that assessment measures in healthcare contexts varied considerably, ranging from qualitative analyses to the 10-item Jefferson Empathy Scale tool [27]. The investigators (Sorin et al.) encouraged future researchers to blind reviewers to response-givers (LLM vs. human) and to limit biases from lengthy responses. The Materials and Methods Section will detail how this study implemented Sorin et al.'s recommendations.

Although other recent studies have also made positive discoveries from LLM dietary coaching [29,30], negative conclusions tend to be more salient. In addition to empathy and relatability limitations, a tendency to provide inappropriate recommendations to specific or complex queries has been observed [12,31,32]. Studies have also noted that previous ChatGPT models have provided inconsistent responses to the same prompt over time, creating potential confusion, and that some of the sources they referenced were fake [31]. In an assessment of macro- and micro-nutrient advice, investigators highlighted the frequency with which ChatGPT misinterpreted decimal numbers, which led to consistent errors [13]. However, to the knowledge of the authors, all the above limitations pertain to ChatGPT-3.5 and earlier versions, rather than the latest model, ChatGPT-4o and a Beta version of its upgrade, ChatGPT-4o1 preview. Finally, the literature does not appear to have hitherto tested the competency of LLMs in providing lifestyle coaching in DWLSs that use GLP-1 RA medications. Given the large uptake of these services and GLP-1 RA medications for weight loss in general [33,34], coupled with ongoing skepticism of DWLS prescribing safety [34,35], assessing LLM coaching in this context could have significant public health implications.

1.3. Study Aims

This study aims to compare ChatGPT versions 4o and 4o1 preview with human dietitian responses to a set of questions from medicated DWLS patients. It is believed that the study findings will enrich the emerging literature on LLM dietetics by providing an indication of the extent to which the most advanced Chat GPT model to date can emulate dietitian responses to questions from real-world DWLS patients.

2. Materials and Methods

2.1. Study Design

Study questions were developed by a team of three dietitians from a large multinational DWLS, Eucalyptus [8]. All dietitians were accredited at Dietitians Australia and have at least two years' experience in DWLS dietary coaching. Investigators requested that the team develop a set of ten questions based on common themes from their experiences with managing Eucalyptus DWLS patients. Investigators also requested that the team divide these questions into five broad and five narrow subject frames, using standard patient register. Questions were discussed and refined over three videoconferencing sessions. Once the set of questions was finalized, investigators reproduced them on two separate sheets as follows: one in preparation for the two ChatGPT models and the other for human dietitians. The two models included ChatGPT-4o and ChatGPT-4o1 preview.

Investigators entered a detailed prompt into ChatGPT to describe the LLM's role as a dietitian, the characteristics of a mock patient, and the response guidelines (Appendix A). No example response was provided in the prompt, rendering it a 'zero-shot' prompt. Al-

though researchers have found that one- and multi-shot prompts (prompts that include single and multiple relevant examples) generate better ChatGPT responses, the investigators felt that such prompts would not reflect initial LLM-supported DWLSs in the real world. Moreover, as this is the first study on ChatGPT version 4o in the field of dietetics, investigators considered it necessary to test the potential of the model in its standard form before adding extra levels of sophistication. Existing studies support the use of zero-shot prompting as an initial evaluation method for LLMs in healthcare [36,37].

The study questions were finalized on 9 November 2024 (Appendix B) and sent to two Eucalyptus dietitians for testing. No further changes were made, and the final set of questions were entered into the two LLMs and sent to the two human dietitians via email on 12 November 2024. Both human dietitians came from the Eucalyptus DWLS; however, neither received any information about the study prior to receiving the questions. At the time of the study, both dietitians were accredited at Dietitians Australia and had over three years' experience in dietetics, including over 18 months experience in DWLS dietary coaching. Both received the same mock patient profile and response guidelines as the ChatGPT comparator, which included an instruction to limit responses to 200 words or less. They did not receive the detailed description of the dietitian role as this was expected as part of their employment.

Responses to all ten questions from ChatGPT and the two dietitians were forwarded to four independent dietitians for assessment. All four assessors were accredited at Dietitians Australia. Assessor 1 had over 5 years' experience as a dietitian; assessors 2 and 3 had over 2 years' experience; and assessor 4 was a recent graduate. The assessors were required to give a score from 0 to 10 on four criteria, with 0 representing the worst response imaginable and 10 reflecting a perfect response. The four assessment criteria were derived from the rubric used in the 2023 Kirk et al. study [19], which included scientific correctness, comprehensibility, and actionability. However, as noted in the introduction, we believed that this rubric lacked an empathy criterion, given that this response aspect has been regularly highlighted as a shortcoming of ChatGPT communication in healthcare [10,26]. Accordingly, the investigators added this criterion to the assessment rubric and made some minor modifications to the other criteria descriptions (Appendix C). We framed the empathy criterion as 'empathy/relatability', as we believed the latter word was necessary to emphasize the 'nuanced and personable' communication style discussed in the introduction. Assessors were also invited to discuss any potential strengths, weaknesses, or areas for improvement in a free-text section adjacent to each response. All assessors were blinded to the response-givers. In addition to this blinding, inter-rater variability was minimized through a clearly defined marking rubric (Appendix C), the encouragement of scoring justification (in the free-text section), and through the randomization of the order in which assessors received responses (e.g., ChatGPT-4o first, followed by dietitian 1, etc.).

2.2. Statistical Analysis

The median assessor scores were tabulated for all ten questions across the four respondents (two human dietitians and two LLMs). Levene's and Shapiro–Wilk tests were conducted to determine the normality of the distribution and variance [38], and thus, whether a parametric analysis of variance (ANOVA) or Kruskal–Wallis test (non-parametric equivalent) would be used to compare the question- and criterion-based differences [39]. To perform criterion-based analyses, the assessor scores were aggregated for each of the four criteria and the differences between respondents were compared using ANOVA or Kruskal–Wallis tests. All analyses and visualizations were conducted on R Studio (version 2023.06.1 +524).

3. Results

Scores from all four assessors across all four criteria were grouped for each question, and Levene's and Shapiro–Wilk tests were conducted [38]. The latter tests revealed that homogeneity of variance was violated, and thus, a Kruskal–Wallis analysis was determined as the appropriate test for the study's endpoints [39].

First, data were aggregated from all criteria, and the median scores were calculated for each individual coach on each question. No median question score from any coach was below seven, and no statistical differences were observed between coach scores for any of the study's conventional questions (Q1—$p = 0.45$, $\eta^2 = -0.006$; Q2—$p = 0.67$, $\eta^2 = -0.024$; Q3—$p = 0.961$, $\eta^2 = -0.045$; Q4—$p = 0.055$, $\eta^2 = 0.077$; Q5—$p = 0.61$, $\eta^2 = -0.019$). However, the Kruskal–Wallis tests found statistically significant differences in dietitian responses to questions six, seven, eight, and ten (Table 1). The eta squared (η^2) value in questions six (0.097), seven (0.121), and ten (0.127) indicated a medium effect size, while the value in question eight (0.151) suggested a large effect.

Table 1. Median scores and correlation statistics for each question.

	Human Dietitian 1	Human Dietitian 2	GPT-4o	GPT-4o1 Preview	Eta Squared (η^2)	*p*-Value
Question 1	8 (CI 6.52, 8.7)	7 (CI 6.19, 7.72)	8 (CI 6.83, 8.97)	8 (CI 6.34, 8.65)	−0.006	0.449
Question 2	8.5 (CI 8.13,9.23)	8 (CI 7.24, 8.35)	7 (CI 6.14, 8.12)	8 (CI 7.42, 8.8)	−0.024	0.671
Question 3	7.5 (CI 6.42, 8.63)	7 (CI 6.54, 7.6)	7 (CI 6.23, 7.92)	8 (CI 6.94, 8.42)	−0.045	0.961
Question 4	7 (CI 6.22, 7.73)	5 (CI 4.60, 5.37)	7.5 (CI 6.55, 8.34)	8 (CI 6.71, 8.21)	0.077	0.055
Question 5	8 (CI 7.1, 9.03)	7.5 (CI 6.85, 8.3)	9 (CI 7.88, 9.45)	8.5 (CI 7.46, 9.04)	−0.019	0.608
Question 6	7 (CI 5.96, 8.14)	9 (CI 8.66, 9.41)	8 (CI 7.28, 9.11)	7 (CI 6.64, 8.49)	0.097	0.032 *
Question 7	7 (CI 6.35, 7.88)	8 (CI 7.04, 8.79)	9 (CI 8.14, 9.73)	7.5 (CI 5.32, 9.12)	0.121	0.017 *
Question 8	8 (CI 7.28, 9.04)	7.5 (CI 6.22, 8.65)	8 (CI 7.04, 9.19)	9 (CI 7.52, 9.41)	0.151	0.007 **
Question 9	9 (CI 8.43, 9.55)	8 (CI 7.2, 8.65)	10 (CI 10,10)	9 (CI 7.93, 9.84)	0.012	0.294
Question 10	7 (CI 6.3, 8.09)	8 (CI 7.08, 8.87)	9 (CI 8.7, 9.42)	7.5 (CI 6.87, 8.38)	0.127	0.014 *

Note: CI = 95% confidence interval; * $p < 0.05$, ** $p < 0.01$.

Post hoc Dunn tests revealed which coach scores differed significantly across these four questions (using adjusted *p*-values from the Benjamini–Hochberg method to mitigate false positives). The GPT-4o model received significantly higher scores than both human coaches in question ten (Table 2) (Figure 1), and one of the human coaches in question seven (Table 3). The GPT-4o1 preview model was found to have scored significantly better than one human coach on question eight (Table 4). On question six, the significant difference stemmed from a disparity between the two human coaches (Table 5).

Table 2. Post hoc Dunn test results—question 10 response score by coach.

Response Score	Coach	Levels	Z Score	*p-Adjusted* Value
		GPT-4o—GPT-4o1	0.97	0.396
		GPT-4o1—Human 1	−1.88	0.121
		GPT-4o—Human 1	−2.85	0.026 *
		GPT-4o1—Human 2	−1.52	0.191
		GPT-4o—Human 2	−2.50	0.038 *
		Human 1—Human 2	0.35	0.72

Note: * $p < 0.05$.

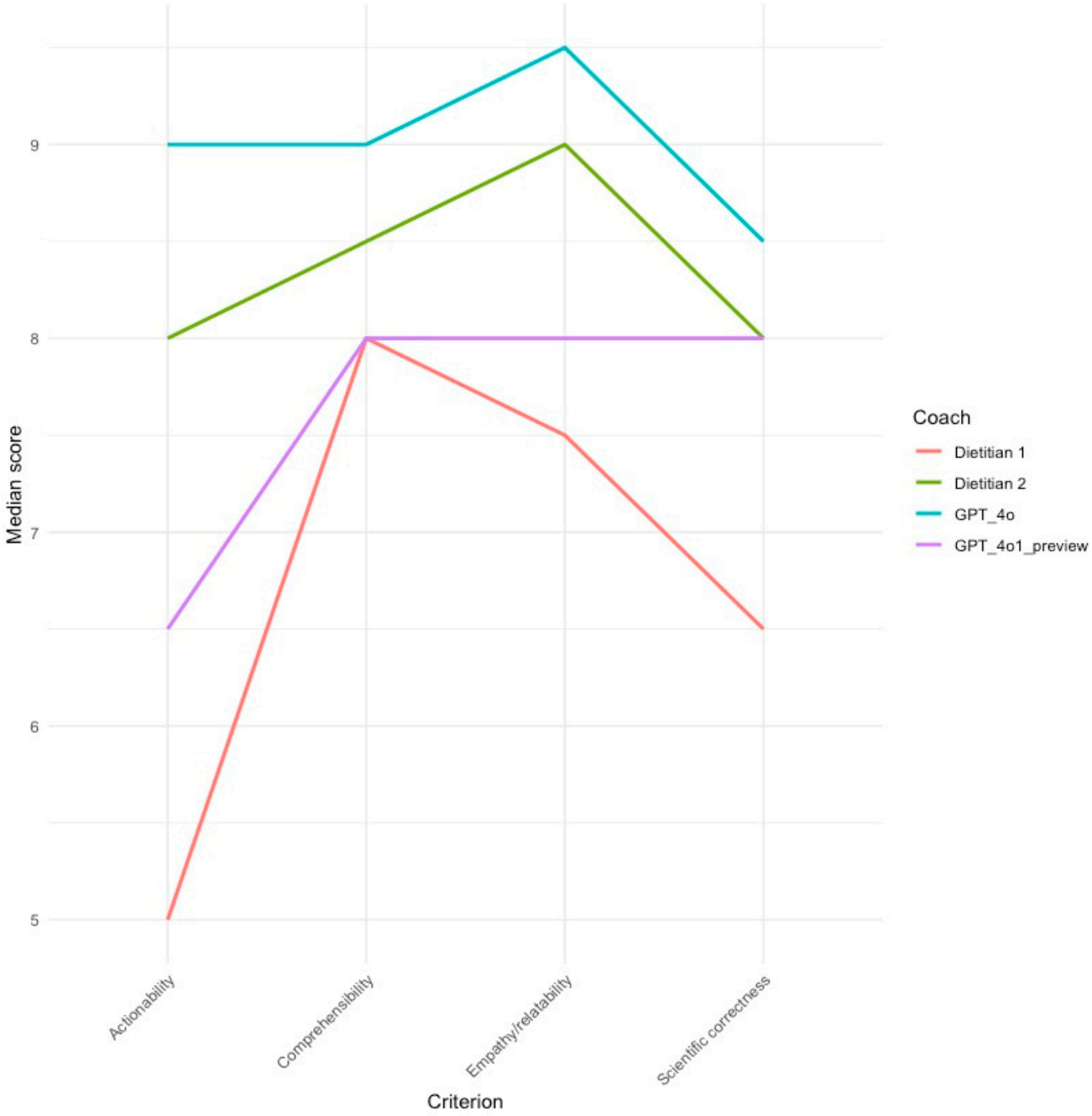


Figure 1. Question 10 scores by coach and criteria.

Table 3. Post hoc Dunn test results—question 7 response score by coach.

Response Score	Coach	Levels	Z Score	*p-Adjusted* Value
		GPT-4o—GPT-4o1	1.98	0.144
		GPT-4o—Human 1	3.12	0.011 *
		GPT-4o1—Human 1	1.14	0.305
		GPT-4o—Human 2	1.25	0.314
		GPT-4o1—Human 2	−0.72	0.469
		Human 1—Human 2	−1.86	0.125

Note: * $p < 0.05$.

Table 4. Post hoc Dunn test results—question 8 response score by coach.

Response Score	Coach	Levels	Z Score	*p-Adjusted* Value
		GPT-4o—GPT-4o1	2.12	0.102
		GPT-4o—Human 1	−1.32	0.225
		GPT-4o1—Human 1	−3.43	0.004 **
		GPT-4o—Human 2	0.14	0.891
		GPT-4o1—Human 2	−1.98	0.095
		Human 1—Human 2	1.45	0.219

Note: ** $p < 0.01$.

Table 5. Post hoc Dunn test results—question 6 response score by coach.

Response Score	Coach	Levels	Z Score	*p-Adjusted* Value
		GPT-4o—GPT-4o1	0.35	0.724
		GPT-4o—Human 1	1.11	0.4
		GPT-4o1—Human 1	0.76	0.538
		GPT-4o—Human 2	1.75	0.159
		GPT-4o1—Human 2	2.11	0.106
		Human 1—Human 2	−2.86	0.025 *

Note: * $p < 0.05$.

Response scores were also concatenated into the following four individual rubric criteria: comprehensibility, empathy/relatability, scientific correctness, and actionability. Levene's tests revealed that homogeneity of variance was violated across all four criteria, and subsequently, Kruskal–Wallis tests were used to assess the relationship between marker-specific scores and coaches (Table 6). These tests only detected a statistically significant association in the actionability category, (X^2[3, N = 40 = 12.726, $p < 0.01$]), whose η^2 value (0.27) indicated a large effect size. A pairwise post hoc Dunn test showed that this association stemmed from the significantly higher scores received by the GPT 4o model compared to the two human coaches (Table 7) (Figure 2).

Table 6. Kruskal–Wallis test results—difference in coach scores by criterion.

Variable	N	η^2	*p*-Value
Comprehensibility	40	−0.05	0.74
Empathy/Relatability	40	−0.05	0.79
Scientific correctness	40	0.04	0.22
Actionability	40	0.27	0.005 **

Note: ** $p < 0.01$.

Table 7. Post hoc Dunn test results—actionability score by coach.

Response Score	Coach	Levels	Z Score	*p-Adjusted* Value
		GPT-4o—GPT-4o1	1.78	0.149
		GPT-4o—Human 1	3.26	0.007 **
		GPT-4o1—Human 1	1.48	0.209
		GPT-4o—Human 2	2.85	0.013 *
		GPT-4o1—Human 2	1.07	0.341
		Human 1—Human 2	−0.41	0.683

Note: * $p < 0.05$, ** $p < 0.01$.

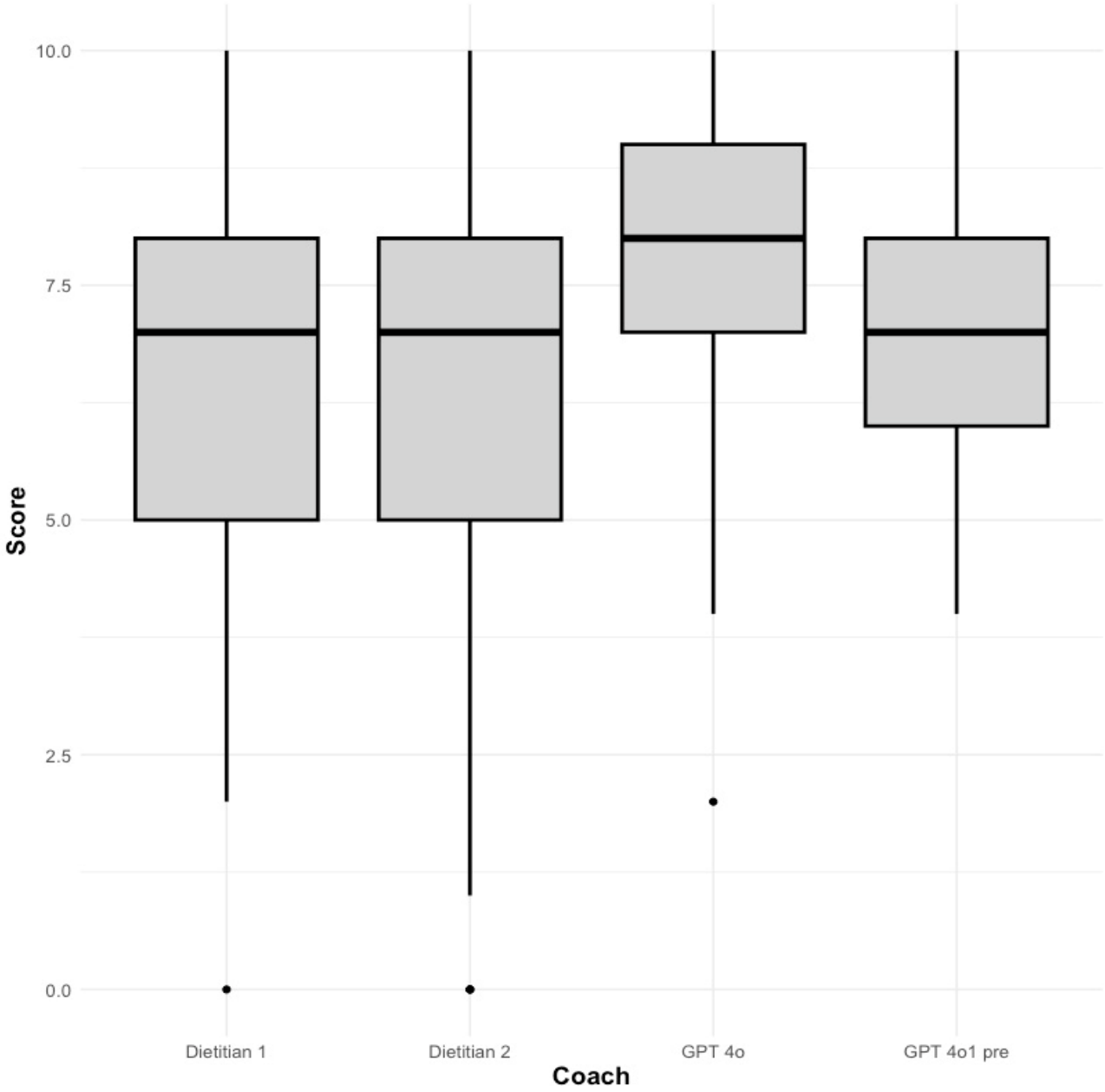


Figure 2. Box plot of actionability scores by coach.

4. Discussion

Rising global obesity rates and ongoing advances in artificial intelligence have engendered increasing scholarly interest in the use of LLMs in dietetics and obesity services. Prior to this study, some research had shown that LLMs could provide sound dietary counseling across a range of basic questions [19]. Moreover, it had been argued that LLM deployment in obesity services had the potential to increase care access by way of enhancing scalability and reducing consumer costs, and by enabling immediate responses to patient queries, thus minimizing care barriers of consultation scheduling and wait times [11–13,29]. To the knowledge of the authors, this study was the first to assess the capabilities of a ChatGPT-4o model in an obesity service context. It also appears to represent the first study to assess this LLM's capabilities in assisting patients from a medicated DWLS—a care model that is becoming increasingly popular in the current healthcare services landscape.

The analysis found that the study's human dietitians did not achieve a statistically higher score than either ChatGPT-4o model on any of the study's ten questions or the four

assessment criteria. The same outcome was observed in the 2023 Kirk et al. study [19], which used the Chat GPT-3.5 model; however, in that study, all eight questions used simple syntax (seven of the eight questions consisted of a single sentence, using ten words or less). By contrast, all ten questions in this study were based on patient communication in a large medicated DWLS, and they ranged from forty to ninety-two words in length. Moreover, this assessment contained an even mix of broad and narrow questions, some of which involved GLP-1 RA medications and thus an additional layer of complexity. The fact that neither Chat GPT-4o nor Chat GPT-4o1 preview was outscored by either human coach on any question suggests that these LLMs (and more advanced versions) have the potential to play a significant supporting role in medicated DWLSs.

This assertion is arguably reinforced by the finding that neither LLM was outscored by a human coach on the aggregated empathy/relatability criterion. While there has been general consensus around the utility of LLMs within a narrow healthcare scope, many experts have expressed concerns about their capability to convey empathy in patient consultations [18,24,27], which various studies have supported [28,29]. This study indicates that LLMs can in fact exhibit this capability across a range of complex queries. Moreover, the discovery that the standard Chat GPT-4o model achieved statistically higher scores for the aggregated actionability criterion may reflect greater reliability in this area of lifestyle coaching, but this needs to be further investigated before any conclusions are drawn. A possible explanation for this finding is that LLMs follow instructions without exception [15,17], whereas humans are prone to intuition and the omission of secondary information (which in certain contexts could be the inclusion of specific actions/objectives). This study was also unable to generate a clear explanation as to why the GPT-4o model outperformed human coaches more often than the more recent GPT-4o1 preview model. A feasible reason is that the GPT-4o1 preview model, while expected to be a more refined iteration of Chat GPT-4o, may not have received enough feedback at the time of the study to correct its additional features [35].

These findings could have several public health implications. Firstly, they provide preliminary evidence that advanced LLMs have the potential to support human lifestyle coaches in medicated DWLSs and thus help relieve workforce shortages and increase service scalability. Secondly, integrating LLMs into such services would enable patients to seek immediate advice at a time and place of their preference and likely reduce the temporal barrier to obesity care that many patients face [10,40]. Thirdly, the use of LLMs in DWLSs would feasibly lower their costs and enable more patients from lower socio-economic groups to access them, as these patients tend to be overrepresented in overweight and obesity statistics in Western countries [41,42]. Finally, these findings suggest that advanced LLMs could be utilized more broadly across health systems to improve efficiency. It must be stressed, however, that this study does not provide evidence that Chat GPT-4o models can replace dietitians in real-world DWLSs. Ethical considerations concerning patient privacy, potential LLM algorithmic bias, and general safety necessitate continued human oversight in all real-world weight loss interventions. As Aydin et al. (2025) assert, LLMs offer multiple benefits to healthcare interventions, but clear regulatory boundaries are needed to ensure they "serve as supportive rather than standalone resources" [43].

While the study generated important knowledge, it had multiple limitations. Firstly, LLM responses were only compared with responses from two human dietitians. Although both dietitians had had over eighteen months lifestyle coaching experience at medicated DWLSs, their responses may not be representative of coaches across this service spectrum. Secondly, questions were developed by a team of dietitians from the Eucalyptus DWLS and therefore may have missed content that features more commonly in other medicated DWLSs. Thirdly, as is the case in any subjective scoring system, this study's scoring metric

may have engendered inter-rater variability (e.g., one assessor's interpretation of a 6, may have been considered a 7 by another). Investigators opted for a continuous 0–10 metric to align with the Kirk et al. study [19] and enable more meaningful statistical analyses. Fourthly, the study did not solicit specific feedback on empathy, and thus, the assessor scores were not enriched by qualitative assessments. And finally, as this was the first assessment of Chat GPT-4o in an obesity service context, investigators wanted to test the model in its most rudimentary state—through 'zero-shot' prompting. Consequently, responses did not reflect the back-and-forth communication style that would typically be observed in patient–clinician interactions of a real-world DWLS. Future investigations should seek to build upon this study's findings by investigating the competency of LLM lifestyle coaches to engage in back-and-forth conversation after receiving multi-shot prompting. Real-world DWLSs that have already integrated AI may consider training LLMs with exemplary responses to weight-loss patient queries prior to testing. Researchers should also consider comparable analyses of LLM responses to questions developed from other medicated DWLSs.

5. Conclusions

Recent research has demonstrated the potential of Chat GPT-3 models as dietary support tools across a series of basic prompts. This study aimed to compare the responses of the more advanced Chat GPT-4o models to human dietitian responses across a range of complex questions based on patient–clinician interactions at a real-world medicated DWLS. The analysis found that neither Chat GPT-4o nor Chat GPT-4o1 preview were statistically outperformed by human dietitians on any of the study's ten questions. It also generated the same findings when scores were aggregated from the ten questions across the following four individual study criteria: scientific correctness, comprehensibility, empathy/relatability, and actionability. These findings provide preliminary evidence that advanced LLMs may be able to play a significant supporting role in medicated obesity services. They should not, however, be interpreted as evidence that such LLMs can safely replace dietitians or clinical decision makers in DWLSs, as the study did not simulate the back-and-forth dialogue and nuanced decision making that characterize real-world patient–clinician interactions. To better approximate these interactions, future studies should incorporate multi-shot LLM prompting, fine-tuned models trained on patient-dietitian exchanges, and reinforcement learning techniques that adapt to complex consultations. Such studies should compare multiple modern LLMs to generate insights regarding potential differences. Future researchers should also seek to examine the effect of LLM integration on DWLS access. Real-world DWLSs should only use LLMs as a supportive tool until clear regulatory mechanisms are established.

Author Contributions: Conceptualization, L.T., L.L., and A.Y.; methodology, L.T. and L.L.; software, L.T. and L.L.; validation, L.T., L.L., A.Y., M.V., and N.A.; formal analysis, L.T., L.L., A.Y., and N.A.; investigation, L.T. and L.L.; resources, L.L. and M.V.; data curation, L.T. and L.L.; writing—original draft preparation, L.T.; writing—review and editing, L.T., L.L., A.Y., and N.A.; visualization, L.T.; supervision, L.L. and M.V.; project administration, L.T., L.L., and M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: The authors would like to thank all dietitians who were involved in this study, including those who responded to the mock questions and those who assessed responses.

Conflicts of Interest: L.T and M.V are paid a salary by Eucalyptus. L.L was paid a salary by Eucalyptus during the data collection phase of the study. N.A is a paid advisor at Eucalyptus.

Abbreviations

The following abbreviations are used in this manuscript:

DWLS	Digital Weight Loss Service
LLM	Large Language Model
GLP-1 RAs	Glucagon-Like Peptide-1 Receptor Agonists
ANOVA	Analysis of Variance

Appendix A. Prompt, Instructions, and Description of Mock Patient for the LLM

LLM Prompt

You are an experienced, accredited dietitian with a strong background in supporting patients on weight loss journeys.

You work for a digital health company specializing in weight loss by combining lifestyle guidance with GLP-1 therapy. Most communication with patients is conducted asynchronously through digital platforms.

Please use evidence-based best practices to inform your responses. Provide practical, actionable advice that is directly relevant to the patient's individual needs and circumstances. Communicate in a clear, empathetic, and patient-centered manner.

Support the patient in achieving their weight loss goals through sustainable behavior change. Empower the patient with knowledge and strategies to make informed choices.

Motivate and encourage the patient, fostering a positive mindset toward their weight loss journey.

Instructions

When answering the patient's questions, ensure your response aligns with the above guidelines.

Provide answers that address all aspects of the patient's query.

You can solicit further information if you feel you need it.

Keep your answers to 200 words or less.

Patient profile

Name: Sarah

Age: Fifty-two

Gender: Female

Ethnicity: Caucasian

Body Mass Index (BMI): 35

Medical History: Hypertension + Sleep Apnea

Sarah is a middle-aged woman facing the dual challenges of hypertension and sleep apnea, both exacerbated by her current weight. Sarah works as a project manager in a corporate office, which often involves long hours and a high level of stress. Sarah is married with children and a dog. Sarah is motivated to improve her health but needs guidance tailored to her lifestyle and constraints. By addressing her dietary habits, increasing physical activity, and managing stress, Sarah aims to make sustainable changes that will enhance her quality of life.

Appendix B. Mock Patient Questions

Broad Questions		
	Question	**Theme**
1.	I've been on Ozempic for the past three months, and in the beginning, I was really pleased with the progress—I was losing weight steadily each week. However, for the past few weeks, I've noticed that my weight loss has come to a complete standstill, even though I've been eating healthy and exercising as directed. I'm feeling really demotivated and frustrated because I'm putting in the effort but not seeing any results anymore. It's starting to affect my motivation to stick with the program. Can you please provide me with some advice?	Weight-loss plateau
2.	I've been on Ozempic for a few weeks now, and I've noticed that my appetite has decreased significantly. Sometimes I feel low on energy or even a bit lightheaded, and I'm worried that I might not be getting all the nutrients my body needs. Could you please offer some advice on how I can ensure I'm eating enough even when I'm not feeling particularly hungry?	Appetite
3.	I'm having trouble meeting my protein goals. Can you give me some tips or suggestions on how to add more protein to my diet? I'd love to find ways to increase my protein intake without making my meals too complicated.	Protein intake
4.	I've started my weight loss journey with Ozempic and noticed that my suggested daily calorie targets seem quite high for someone trying to lose weight. I'm worried that eating this many calories might slow down my progress or even cause weight gain. Can you help me understand why my calorie targets are set this way and how they fit into my weight loss plan?	Calories
5.	I'm struggling to fit in the five weekly exercise sessions you've recommended. My long hours at work, combined with weekends with the family, leave me with almost no time or energy for structured workouts. I even tried walking to work as you suggested, but arriving sweaty and red-faced to morning meetings was uncomfortable and impractical. I'm really at a loss here—how can I realistically manage to make progress without sacrificing my work or family responsibilities?	Time constraints
Narrow Questions		
6.	I'm considering trying intermittent fasting to enhance my weight loss while taking Ozempic. I'm pre-diabetic and curious about how intermittent fasting might interact with the medication's effects on appetite and hunger cues, including any impact on emotional well-being. Also, are there specific fasting windows that work better than others?	Intermittent fasting
7.	After you recommended a daily protein shake for breakfast, I researched supplements and found BCAAs and creatine to be essential for health, so I started taking them a few weeks ago. However, I've been experiencing regular nausea and noticed my weight loss has plateaued since adding these. Is this just some weird reaction with Ozempic? What should I do?	Supplementation

Broad Questions	
Question	**Theme**
8. I'm aiming to lower my carbohydrate intake to support my weight loss, but I'm confused about how to make balanced choices. For example, rice is in some of my recommended meals, yet the recipe says it contains 50g of carbs per serving. I thought I was supposed to be on a low-carb diet?	Carbohydrates
9. I've been weighing myself every day on these supposedly 'accurate' scales, but I'm not seeing any real progress in the numbers, even though I look and feel better and am moving more than I have in ages. It's honestly disheartening and frustrating when all the hard work I'm putting in doesn't seem to be recognized by the one tool that's supposed to validate it! Is there something wrong with these scales, or am I just wasting my time with this routine? What can I actually rely on to track my progress?	Weight tracking
10. As you suggested, I've been swapping sugar-sweetened drinks and snacks with options that contains sugar substitutes, but my friend told me that Sucralose and Erythritol cause cancer. I tried a few days without sweet stuff, but I felt like I couldn't control my cravings and I ended up overeating. I also don't want to give them up because the taste of sweetness is one of life's great pleasures. Wouldn't sugar be better than stuff that causes cancer, even if it contains a few more calories? I'm so frustrated and confused.	Artificial sweeteners

Appendix C. Scoring Matrix

Scoring Criterion	Description
Scientific correctness	How accurately each answer reflects the current state of knowledge in the scientific domain to which the question belongs. Where relevant, this includes a consideration of patient conditions and/or medications.
Comprehensibility	How well the answer could be expected to be understood by the layman.
Actionability	The degree to which the answers to the questions contain information that is useful and can be acted upon by the hypothetical layman asking the question. For example, while a different home-cooked, protein-rich dinner every weeknight might be an effective weight-loss strategy, this would not be a helpful suggestion for someone with limited spare time.
Empathy/Relatability	The degree to which the answers convey empathy and understanding of the emotional state of the patient.

References

1. World Health Organization. Obesity and Overweight. 1 March 2024. Available online: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed on 29 September 2024).
2. Australian Government. Overweight and Obesity. 17 June 2024. Available online: https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity/contents/overweight-and-obesity (accessed on 29 September 2024).

3. NCD Risk Factor Collaboration. Worldwide trends in underweight and obesity from 1990 to 2022: A pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. *Lancet* **2024**, *403*, 1027–1050. [CrossRef] [PubMed]
4. *World Obesity Atlas 2022*; The World Obesity Federation: London, UK, 2022.
5. World Health Organization. *Health Service Delivery Framework for Prevention and Management of Obesity*; WHO: Geneva, Switzerland, 2023.
6. Deslippe, A.; Soanes, A.; Bouchaud, C.; Beckenstein, H.; Slim, M.; Plourde, H.; Cohen, T.R. Barriers and facilitators to diet, physical activity and lifestyle behavior intervention adherence: A qualitative systematic review of the literature. *Int. J. Behav. Nutr.* **2023**, *20*, 14. [CrossRef]
7. Helland, M.; Nordbotten, G. Dietary changes, motivators, and barriers affecting diet and physical activity among Overweight and Obese: A mixed methods approach. *Int. J. Environ. Res. Public. Health* **2021**, *18*, 10582. [CrossRef]
8. Hinchliffe, N.; Capehorn, M.; Bewick, M.; Feenie, J. The potential role of digital health in obesity care. *Adv. Ther.* **2022**, *39*, 4397–4412. [CrossRef]
9. Idris, I.; Hampton, J.; Moncrieff, F.; Whitman, M. Effectiveness of a digital lifestyle change program in obese and type 2 diabetes populations: Service evaluation of real-world data. *JMIR Diabetes* **2020**, *5*, e15189. [CrossRef] [PubMed]
10. Talay, L.; Vickers, M.; Ruiz, L. Effectiveness of an email-based, Semaglutide-supported weight-loss service for people with overweight and obesity in Germany: A real-world retrospective cohort analysis. *Obesities* **2024**, *4*, 256–269. [CrossRef]
11. Chatelan, A.; Clerc, A.; Fonta, P. ChatGPT and future artificial intelligence chatbots: What may be the influence on credentialed nutrition and dietetics practitioners? *J. Acad. Nutr. Diet.* **2023**, *11*, 1525–1529. [CrossRef]
12. Arslan, S. Exploring the potential of Chat GPT in personalized obesity treatment. *Ann. Biomed. Eng.* **2023**, *51*, 1887–1888. [CrossRef]
13. Agne, A.; Gedrich, K. Persnalized dietary recommendations for obese individuals—A comparison of ChatGPT and the Food4Me algorithm. *Clin. Nutr. Open Sci.* **2024**, *56*, 192–201. [CrossRef]
14. Golovaty, I.; Hagan, S. Direct-to-consumer platforms for New Antiobesity Medications—Concerns and potential opportunities. *N. Engl. J. Med.* **2024**, *390*, 677–680. [CrossRef]
15. Open AI. Chat GPT. Optimizing Language Models for Dialogue. Available online: https://openai.com/blog/chatgpt (accessed on 20 September 2024).
16. Brewster, R.; Gonzalez, P.; Khazanchi, R.; Butler, A.; Selcer, R.; Chu, D.; Aires, B.P.; Luercio, M.; Hron, J.D. Performance of ChatGPT and Google Translate for Pediatric Discharge instruction translation. *Pediatrics* **2024**, *154*, e2023065573. [CrossRef]
17. Stoneham, S.; Livesey, A.; Cooper, H. ChatGPT versus clinician: Challenging the diagnostic capabiltiies of artificial intelligence in dermatology. *Clin. Exp. Dermatol.* **2024**, *49*, 707–710. [CrossRef] [PubMed]
18. Sallam, M. ChatGPT utility in healthcare education, research and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
19. Kirk, D.; van Eijnatten, E.; Camps, G. Comparison of answers between Chat GPT and human dietitians to common nutrition questions. *J. Nutr. Metab.* **2023**, *1*, 5548684.
20. Guo, P.; Liu, G.; Xiang, X.; An, R. From AI to the table: A systematic review of ChatGPT's potential and performance in meal planning and dietary recommendations. *Dietetics* **2025**, *4*, 7. [CrossRef]
21. Hieronimus, B.; Hammann, S.; Podszun, M. Can the AI tools ChatGPT and Bard generate energy, macro- and micro-nutrient sufficient meal plans for different dietary patterns? *Nutr. Res.* **2024**, *128*, 105–114. [CrossRef]
22. Qarajeh, A.; Tangpanithandee, S.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Aiumtrakul, N.; Valencia, O.A.G.; Miao, J.; Qureshi, F.; Cheungpasitporn, W. AI-powered renal diet support: Performance of ChatGPT, Bard AI, and Bing Chat. *Clin. Pract.* **2023**, *13*, 1160–1172. [CrossRef] [PubMed]
23. Bayram, H.; Ozturkcan, A. AI showdown: Info accuracy on protein quality content in foods from ChatGPT 3.5, ChatGPT 4, bard AI and bing chat. *Br. Food J.* **2024**, *126*, 3335–3346. [CrossRef]
24. Carlbring, P.; Hadjistavropolous, H.; Kleiboer, A.; Andersson, G. A new era in Internet interventions: The advent of Chat=GPT and AI-assisted therapist guidance. *Internet Interv.* **2023**, *32*, 100621. [CrossRef]
25. Nashwan, A.; Abujaber, A.; Choudry, H. Embracing the future of physiscian-patient communication: GPT-4 in gastroenterology. *Gastroenterol. Endosc.* **2023**, *1*, 132–135. [CrossRef]
26. Ilicki, J. A framework for critically assessing ChatGPT and other large language artificial intelligence model applications in health care. *Mayo Clin. Proc. Digit. Health* **2023**, *1*, 185–188. [CrossRef]
27. Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; Klang, E. Large Language Models and Empathy: Systematic Review. *J. Med. Internet Res.* **2024**, *26*, e52597. [CrossRef] [PubMed]
28. Chuen, C.; Tan, L.; Khanh Le, M.; Tang, B.; Liaw, S.Y.; Tierney, T.; Ho, Y.Y.; Lim, B.E.E.; Lim, D.; Ng, R.; et al. The development of empathy in the healthcare setting: A qualitative approach. *BMC Med. Educ.* **2022**, *22*, 245.

29. Liu, Z.; Zhang, L.; Wu, Z.; Yu, X.; Cao, C.; Dai, H.; Liu, N.; Liu, J.; Liu, W.; Li, Q.; et al. Surviving ChatGPT in healthcare. *Front. Radiol.* **2024**, *3*, 1224682. [CrossRef]
30. Javaid, M.; Haleem, A.; Pratap Singh, R. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Trans. Benchmarks Stand. Eval.* **2023**, *3*, 100105. [CrossRef]
31. Garcia, M. ChatGPT as a virtual dietitian: Exploring its potential as a tool for improving nutrition knowledge. *Appl. Syst. Innov.* **2023**, *6*, 96. [CrossRef]
32. Ponzo, V.; Goitre, I.; Favaro, E. Is ChatGPT an effective tool for providing dietary advice? *Nutrients* **2024**, *16*, 469. [CrossRef]
33. Mahase, E. GLP-1 agonists: US sees 700% increase over four years in number of patients without starting treatment. *BMJ* **2024**, *386*, q1645. [CrossRef]
34. Strumila, R.; Lengvenyte, A.; Guillaume, S.; Nobile, B.; Olie, E.; Courtet, P. GLP-1 agonists and risk of suicidal thoughts and behaviours: Confound by indication once again? A narrative review. *Eur. Neuropsychopharmacol.* **2024**, *87*, 29–34. [CrossRef]
35. De Winer, J.; Dodou, D.; Eisma, Y. System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam. *Computers* **2024**, *13*, 278. [CrossRef]
36. Sivarajkumar, S.; Kelley, M.; Samolyk-Mazzanti, A.; Visweswaran, S.; Wang, Y. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Med. Inf.* **2024**, *12*, e55318. [CrossRef] [PubMed]
37. Luo, X.; Muhammad Tahabi, F.; Marc, T.; Haunert, L.A.; Storey, S. Zero-shot learning to extract assessment criteria and medical services from the preventative healthcare guidelines using large language models. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1743–1753. [CrossRef] [PubMed]
38. Ruxton, G.; Wilkinson, D.; Neuhauser, M. Advice on testing the null hypothesis that a sample is drawn from a normal distribution. *Anim. Behaviour.* **2015**, *107*, 249–252. [CrossRef]
39. Nwobi, F.; Akanno, F. Power comparison of ANOVA and Kruskal-Wallis tests when error assumptions are violated. *Adv. Methodol. Stat.* **2021**, *18*, 53–71. [CrossRef]
40. Talay, L.; Vickers, M. Why people seek obesity care through digital rather than in-person services: A quantitative multinational analysis of patients from a large unsubsidized digital obesity provider. *Cureus* **2024**, *16*, e75603. [CrossRef] [PubMed]
41. Kim, T.; von dem Knesebeck, O. Income and obesity: What is the direction of the relationship? A systematic review and meta-analysis. *BMJ Open* **2018**, *8*, e019862. [CrossRef]
42. Australian Institute of Health and Welfare. *Inequalities in Overweight and Obesity and the Social Determinants of Health*; Australian Institute of Health and Welfare: Canberra, Australia, 2021.
43. Aydin, S.; Karabacak, M.; Vlachos, V.; Margetis, K. Navigating the potential and pitfalls of large language models in patient-centered medication guidance and self-decision support. *Sec. Regul. Sci.* **2025**, *12*, 1527864. [CrossRef]

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