Original Paper

Assessment of Recommendations Provided to Athletes Regarding Sleep Education by GPT-4o and Google Gemini: Comparative Evaluation Study

Lukas Masur[1], MSci; Matthew Driller[2], PhD; Haresh Suppiah[2], PhD; Manuel Matzka[3], PhD; Billy Sperlich[3], PhD; Peter Düking[1], PhD

[1]Department of Sports Science and Movement Pedagogy, Technische Universität Braunschweig, Braunschweig, Germany
[2]School of Allied Health, Human Services, and Sport, La Trobe University, Melbourne, Australia
[3]Department of Sports Science, Integrative and Experimental Exercise Science & Training, University of Würzburg, Würzburg, Germany

Corresponding Author:

Peter Düking, PhD
Department of Sports Science and Movement Pedagogy
Technische Universität Braunschweig
Pockelsstraße 11
Braunschweig, 38106
Germany
Phone: 49 531 391 3432
Email: Peter.dueking@tu-braunschweig.de

Abstract

Background: Inadequate sleep is prevalent among athletes, affecting adaptation to training and performance. While education on factors influencing sleep can improve sleep behaviors, large language models (LLMs) may offer a scalable approach to provide sleep education to athletes.

Objective: This study aims (1) to investigate the quality of sleep recommendations generated by publicly available LLMs, as evaluated by experienced raters, and (2) to determine whether evaluation results vary with information input granularity.

Methods: Two prompts with differing information input granularity (low and high) were created for 2 use cases and inserted into ChatGPT-4o (GPT-4o) and Google Gemini, resulting in 8 different recommendations. Experienced raters (n=13) evaluated the recommendations on a 1-5 Likert scale, based on 10 sleep criteria derived from recent literature. A Friedman test with Bonferroni correction was performed to test for significant differences in all rated items between the training plans. Significance level was set to $P<.05$. Fleiss $\varkappa$ was calculated to assess interrater reliability.

Results: The overall interrater reliability using Fleiss $\varkappa$ indicated a fair agreement of 0.280 (range between 0.183 and 0.296). The highest summary rating was achieved by GPT-4o using high input information granularity, with 8 ratings >3 (tendency toward good), 3 ratings equal to 3 (neutral), and 2 ratings <3 (tendency toward bad). GPT-4o outperformed Google Gemini in 9 of 10 criteria ($P<.001$ to $P=.04$). Recommendations generated with high input granularity received significantly higher ratings than those with low granularity across both LLMs and use cases ($P<.001$ to $P=.049$). High input granularity leads to significantly higher ratings in items pertaining to the used scientific sources ($P<.001$), irrespective of the analyzed LLM.

Conclusions: Both LLMs exhibit limitations, neglecting vital criteria of sleep education. Sleep recommendations by GPT-4o and Google Gemini were evaluated as suboptimal, with GPT-4o achieving higher overall ratings. However, both LLMs demonstrated improved recommendations with higher information input granularity, emphasizing the need for specificity and a thorough review of outputs to securely implement artificial intelligence technologies into sleep education.

JMIR Form Res 2025;9:e71358; doi: 10.2196/71358

Keywords: individualization; personalization; artificial intelligence; education; recovery; monitoring

Introduction

Sleep is essential for the health and well-being of individuals across all age groups, as it supports cognitive function, mood, mental health, as well as cardiovascular, cerebrovascular, and metabolic health [1,2]. Short-term sleep deprivation, chronic sleep restriction, circadian misalignment, and untreated sleep disorders can significantly harm physical health, mental health, and mood [1,3,4].

One population frequently experiencing inadequate sleep or poor sleep quality is athletes [5-7]. Reasons for poor sleep may stem from multiple sport and nonsport factors including high training loads, long-haul travel, early morning training, family commitments, lifestyle choices including diet, or work or study commitments [6]. While it is beyond the scope of this paper to dive into the effects of inadequate sleep in detail, we refer the reader to existing papers on this topic [6,8]. Briefly, athletes' sleep is considered a primary mechanism facilitating both psychological and physiological recovery [9,10]. Poor sleep may be detrimental for athletes, with negative impacts on mental well-being, cognition, learning and memory consolidation, growth and repair of cells, glucose metabolism, and immune responses (eg, the resistance to respiratory infection) [6,11,12].

To improve aspects of sleep, a first step is to educate athletes and staff on the negative effects of poor sleep and factors affecting sleep [6], and there is evidence that sleep education improves the sleep behavior of team sport athletes [13]. Such sleep education might include (1) education of athletes on potential factors negatively impacting sleep [14] and (2) education on reducing the impact of factors negatively impacting sleep.

It was reported that even in elite athlete cohorts, there is a lack of sleep knowledge [15], and consequently, there is a need for sleep education. For this, artificial intelligence and, more specifically, publicly available large language models (LLMs) might offer a scalable solution. By simulating human-like conversations, LLMs leverage deep learning to process information and generate nuanced responses. LLMs such as ChatGPT (OpenAI) are rapidly gaining popularity among the general population [16] as well as in various scientific domains such as medical research and education [17-20], health care [21-23], or nutrition [24,25]. Thereby, it is likely that individuals turn to LLMs to receive responses to questions they face, for example, regarding sleep. However, it is currently unknown if recommendations regarding sleep generated by publicly available LLMs are appropriate and in line with recent scientific evidence and thus suitable for a specific athlete. Here, we aim (1) to investigate sleep recommendations provided to athletes generated by different publicly available LLMs as evaluated by experienced raters and (2) to investigate if sleep recommendations differ depending on the provided information by the user.

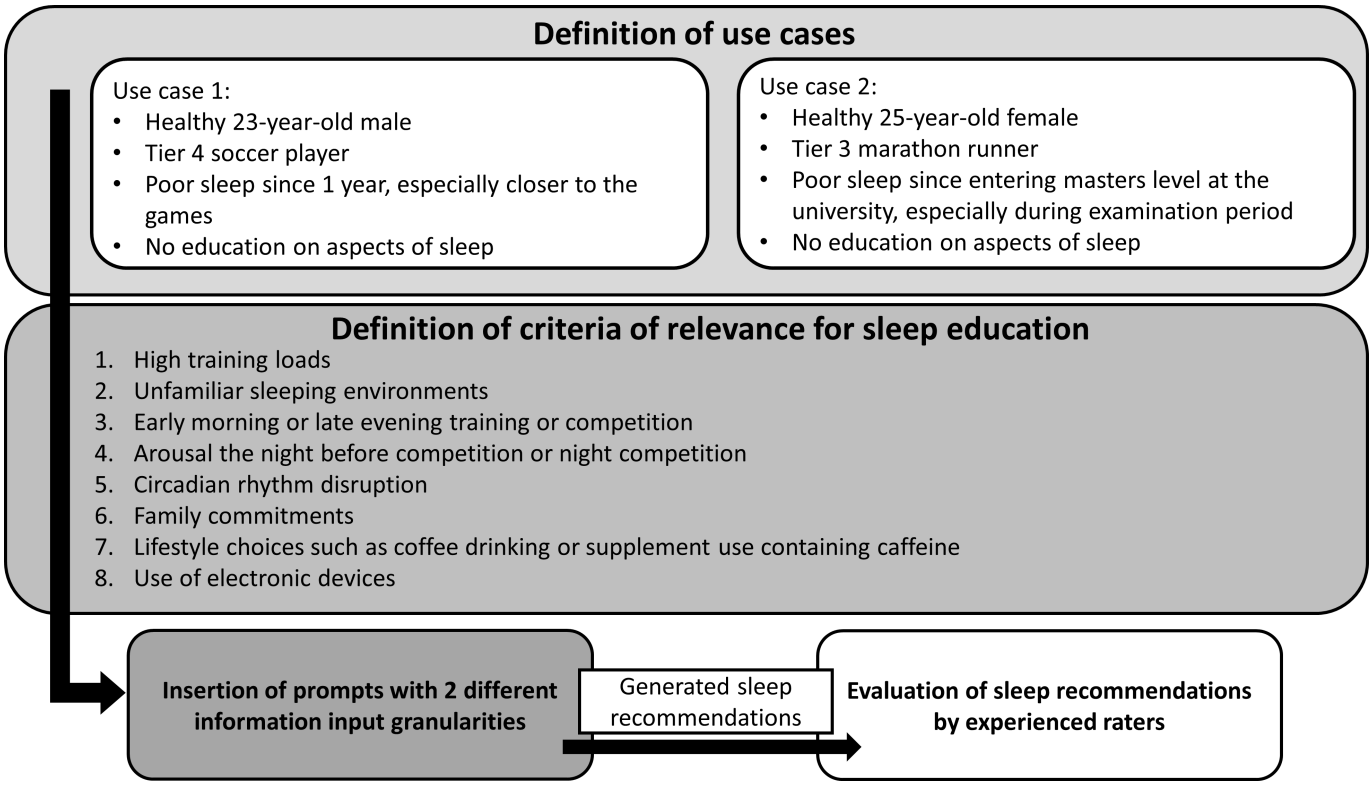
Methods

General Design

To evaluate sleep recommendations provided by LLMs, we followed methodologies of similar papers in the medical field [26-29] or in the exercise science literature [30-33] and adjusted these methodologies to the aim of our research. Figure 1 depicts the experimental workflow of the study.

For this, we (1) define a specific use case, (2) define criteria of relevance for sleep education in this use case, (3) define information input into publicly available LLMs, and (4) involve experienced raters in the topic of sleep within the athletic population to rate outcomes of LLM-generated responses.

Figure 1. Experimental workflow of the study.



Ethical Considerations

The ethics committee of the Faculty of Exercise Science and Training at the University of Würzburg approved the study (reference: EV2025/5-0606). Raters were informed about procedures and gave their consent to participate in the study. No compensation was given to the raters. After receiving the ratings, these were deidentified.

Definition of 2 Use Cases

Use Case 1: Male Tier 4 Soccer Player

For use case 1, we define a healthy, 23-year-old male tier 4, elite soccer player [34] who trains 5 times a week and has 1-2 competitive games per week at a national level. The soccer player experiences poor sleep for approximately a year, and sleep disturbance is closer to important games. The individual in our use case has no formal education on, for example, factors affecting sleep or on the effectiveness of countermeasures to improve sleep and no access to experienced and educated personnel who could educate him on sleep. We define that the major reasons for impaired sleep are high training loads, arousal the night before competition, and unfamiliar sleeping environments in the case of away games.

Use Case 2: Female Tier 3 Marathon Runner

For use case 2, we define a healthy, 25-year-old female tier 3, highly trained or national-level marathon runner [34] who runs around 100 km per week following a pyramidal training intensity distribution [35]. The female runner competes at the national level. In addition to her running training, the runner is enrolled at a master level at university.

The female runner experiences poor sleep since entering master level at university (approximately 6 months ago), and sleep is compromised especially during examination periods, which also affects running training and performance. The individual in our use case has no formal education on, for example, factors affecting sleep or on the effectiveness of countermeasures to improve sleep and no access to experienced and educated personnel who could educate her on sleep. We define that the major reasons for impaired sleep are training loads and disturbances stemming from university obligations.

In line with the aims of this research, we challenge the publicly available LLMs (1) to identify the reasons for sleep disturbances and (2) to give evidence-based guidance on how to reduce the impact of factors negatively impacting sleep.

Criteria of Relevance for Sleep Education

There are many sport and nonsport factors that impact the sleep of athletes. However, factors influencing sleep that are most commonly mentioned in the literature include high training loads [6,14], unfamiliar sleeping environments [6], early morning or late evening training or competition [6,14], arousal the night before competition or the night after competition [6,14], circadian rhythm disruption (eg, due to long-haul travel) [6,14], family commitments [6], lifestyle choices such as coffee drinking or supplement use containing caffeine [6,14], and use of electronic devices (eg, smartphone use) [14].

Prompts Inserted Into the LLMs

Given the chatbot nature of LLMs, we assume that the input provided by individuals seeking sleep education will vary,

like any other conversation. Depending on factors such as previous knowledge about sleep or personal experiences, we assume that some individuals may provide minimal information, while others may be more detailed. To accommodate this diversity in the input information, we developed 2 distinct input information scenarios for both our use cases. Scenario 1 resembles an individual who inserts little to no information and only asks superficial check-backs once a response is given by LLMs. Scenario 2 resembles an individual who inserts more information and asks more detailed check-backs once a response is given by LLMs. The complete conversations with LLMs are available in Multimedia Appendices 1 and 2. Prompts of the scenarios were designed by the authors who are frequently in conversations with athletes on aspects of sleep.

For use case 1, the initial prompts were as follows:

Scenario 1:

I feel tired in the morning and after waking up. I do not know why. This is especially worse close to soccer games. Can you give me advice on how to improve my sleep?

Scenario 2:

I am a 23 year old male highly trained/national level soccer player. I train 5 times a week and plus a game per week. I feel tired due to poor sleep at night since approximately a year. Especially the night before a game I sleep poorly. Can you give me advice on factors which might affect my sleep and how I can reduce the impact of these factors? Use only scientific literature and specifically, for each advice, state this literature and provide a reference list at the end.

For use case 2, the initial prompts were as follows:

Scenario 1:

I feel tired in the morning and after waking up. I do not know why. Maybe it is due to my running training or due to entering master level at university. Can you give me advice on how to improve my sleep?

Scenario 2:

I am a 25 year old female tier 3, highly trained/national level marathon runner player and I run approx. 100km per week with a pyramidal intensity distribution. I feel tired in the morning since approximately entering my master course at university. Can you give me advice on factors which might affect my sleep and how I can reduce the impact of these factors? Use only scientific literature and specifically, for each advice, state this literature and provide a reference list at the end.

Here, we used GPT-4o and Google Gemini Advanced without any use of plug-ins, as both are publicly available and thereby can be used by individuals who seek sleep education. Prompts were inserted on May 15, 2024.

Raters

We reached out to well-educated and experienced raters on sleep and athletes to assess the provided recommendations on the outlined aspects relevant for sleep education on a 1 to 5 Likert scale. We included a total of 13 experienced raters (age span: 28-42 years; n=6 with a PhD, n=7 with a master degree in human physiology or sleep science) working for an average of 9 (SD 7) years with athletes on matters of sleep and indicating a mean sleep research experience of 4 (SD 6) years participated in this study.

Statistical Analysis

We calculated descriptive statistics for the Likert scores on all rated items for each question. To test for significant differences in all rated items between the training plans, a Friedman test with Bonferroni correction was performed. Significance level was set to $P<.05$. Fleiss κ was calculated to assess interrater reliability [36]. Interpretation of Fleiss κ results was conducted according to the classification by Landis and Koch [37]. Fleiss κ values were interpreted as follows: a value of 0.00-0.20 as "slight," 0.21-0.40 as "fair," 0.41-0.60 as "moderate," 0.61-0.80 as "substantial," and >0.80 as "almost perfect" [37]. All statistical analyses were performed in SPSS (version 28; IBM Corp).

Results

Overview

Table 1 represents Fleiss κ values for the different LLMs and use cases. The analysis of Fleiss κ indicated a fair agreement of the overall interrater reliability (0.280), while results ranged between 0.183 and 0.296 (Table 1).

Descriptive statistics of the evaluated sleep recommendations are presented in Table 2.

Table 1. Fleiss κ results.

Large language model, use case, and scenario	Fleiss κ
Google Gemini	
Use case 1—scenario 1 (Gem_C1-S1)	0.270
Use case 1—scenario 2 (Gem_C1-S2)	0.198
Use case 2—scenario 1 (Gem_C2-S1)	0.296
Use case 2—scenario 2 (Gem_C2-S2)	0.183
GPT-4o	
Use case 1—scenario 1 (GPT_C1-S1)	0.258
Use case 1—scenario 2 (GPT_C1-S2)	0.264
Use case 2—scenario 1 (GPT_C2-S1)	0.256
Use case 2—scenario 2 (GPT_C2-S2)	0.229

Table 2. Descriptive analysis of Likert-scale ratings[a].

Relevant aspects when deploying sleep recommendations	Gem_C1-S1[b]	GPT_C1-S1[c]	Gem_C1-S2[d]	GPT_C1-S2[e]	Gem_C2-S1[f]	GPT_C2-S1[g]	Gem_C2-S2[h]	GPT_C2-S2[i]
General aspects, median (IQR)								
Overall training plan	3 (3-3)	3 (3-4)	3 (3-4)	4 (4-4)	3 (3-4)	3 (3-4)	4 (3-4)	4 (3.75-4)
Training load	0 (0-3)	0 (0-0)	3 (0-4)	4 (3-4)	3 (2-4)	4 (3-4)	4 (3-4)	4 (4-5)
Unfamiliar sleeping environments	0 (0-3)	2 (0-3)	4 (3-4)	3 (0-4)	3 (0-3)	3 (0-4)	3 (0-4)	3 (0-4)
Early morning or late evening training or competition	3 (0-4)	3 (3-3)	4 (4-4)	4 (4-4)	4 (3-4)	3 (0-4)	0 (0-4)	4 (3-4)
Arousal the night before competition or training	0 (0-3)	3 (3-4)	4 (3-4)	3 (3-4)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-4)
Circadian rhythm disruptions or consistency of sleep schedule	3 (3-4)	4 (3-4)	4 (4-4)	4 (3-5)	4 (3-4)	3 (3-4)	4 (3-4)	3 (3-4)
Family commitments	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)
Lifestyle choices	3 (0-3)	3 (0-4)	2 (0-3)	0 (0-3)	3 (2-4)	3.5 (2.25-4)	3 (0-3)	3 (3-3)
Use of electronic devices	3 (3-3.25)	4 (3-5)	0 (0-0)	4 (4-5)	3 (0-4)	4 (3-4)	4 (3-4)	4 (3-5)
Nutrition	3 (3-3)	4 (3-4)	4 (3-4)	3 (3-5)	4 (3-4)	3 (3-4)	4 (3-4)	4 (3-5)
Summary rating, n (%)								
<3	4 (40)	3 (30)	3 (30)	2 (20)	2 (20)	2 (20)	3 (30)	2 (20)
3	6 (60)	4 (40)	2 (20)	3 (30)	5 (50)	5 (50)	2(20)	3 (30)
>3	0 (0)	3 (30)	5 (50)	5 (50)	3 (30)	3 (30)	5 (50)	5 (50)
Scientific sources, median (IQR)								
Is the real and existing literature stated (no “fake citations”)?	0 (0-0)	0 (0-0)	4 (2-4)	5 (5-5)	0 (0-0)	0 (0-0)	4 (3-4)	5 (5-5)
Appropriateness of provided scientific evidence	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)	0 (0-0)	0 (0-0)	4 (4-5)	4 (4-5)
Quality of provided scientific evidence in this specific context	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)	0 (0-0)	0 (0-0)	4 (3-4)	4 (4-4)
Summary rating, n (%)								
<3	3 (100)	3 (100)	0 (0)	0 (0)	3 (100)	3 (100)	0 (0)	0 (0)
3	0 (0)	0 (0)	0 (0)	1 (33.3)	0 (0)	0 (0)	0 (0)	0 (0)
>3	0 (0)	0 (0)	3 (100)	2 (66.6)	0 (0)	0 (0)	3 (0)	3 (100)

[a]Results range from 1=bad to 5=good with 0=not applicable.
[b]Gem_C1-S1: Google Gemini, use case 1—scenario 1.
[c]GPT_C1-S1: ChatGPT, use case 1—scenario 1.
[d]Gem_C1-S2: Google Gemini, use case 1—scenario 2.
[e]GPT_C1-S2: ChatGPT, use case 1—scenario 2.
[f]Gem_C2-S1: Google Gemini, use case 2—scenario 1.
[g]GPT_C2-S1: ChatGPT, use case 2—scenario 1.
[h]Gem_C2-S2: Google Gemini, use case 2—scenario 2.
[i]GPT_C2-S2: ChatGPT, use case 2—scenario 2.

Differences Regarding Input Information Granularity and Between Google Gemini and GPT-4o

Results for significance testing regarding different input information granularities and differences between Google Gemini and GPT-4o are presented in Table 3.

Significance testing of the comparison between identical prompts across different LLMs shows that GPT-4o attained significantly higher Likert-scale scores in 9 of 10 criteria of relevance for sleep education ($P<.001$ to $P=.045$). Higher input information granularity, independent of the LLM used, exhibits significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education ($P<.001$ to $P=.049$).

In this paper, we compare the output of different LLMs when the same information was inserted. We do not show comparisons of different LLMs and different information input (eg, Gem_C1-S1 vs GPT_C1-S2) but provide these to the interested reader in Multimedia Appendix 3.

Table 3. Results of the significance testing comparing training plans: between Google Gemini and GPT-4o for the same input information granularity and within Google Gemini or GPT-4o for different input information granularity.

Relevant aspects when deploying sleep recommendations	Significance testing (*P* value) (Google Gemini versus GPT-4o; same prompt, different LLM)[a]				Significance testing (*P* value) (input information granularity; different prompt, same LLM)			
	Gem_C1-S1[b] versus GPT_ C1-S1[c]	Gem_C1-S2[d] versus GPT_ C1-S2[e]	Gem_C2-S1[f] versus GPT_ C2-S1[g]	Gem_C2-S2[h] versus GPT_ C2-S2[i]	Gem_C1-S1 versus Gem_C1-S2	GPT_C1-S1 versus GPT_C1-S2	Gem_C2-S1 versus Gem_C2-S2	GPT_C2-S1 versus GPT_C2-S2
General aspects								
Overall training plan	.92	*.02*	.16	.07	.47	*.003*	.38	*.005*
Training load	.19	.21	.06	.21	*.003*	*<.001*	*.005*	*.03*
Unfamiliar sleeping environments	.63	.21	.76	.72	*.006*	.28	.15	.37
Early morning or late evening training or competition	.88	.88	.08	*.004*	*.01*	*.01*	*.009*	*.04*
Arousal the night before competition or training	*.005*	.85	.11	*.02*	*.001*	.54	.14	.35
Circadian rhythm disruptions or consistency of sleep schedule	.16	.96	.43	.96	*.01*	.25	.71	.71
Family commitments	>.99	*.049*	>.99	>.99	*.049*	>.99	>.99	>.99
Lifestyle choices	.74	.96	.74	.55	.37	.21	.28	.41
Use of electronic devices	.07	*<.001*	.27	.38	*<.001*	.22	.48	.64
Nutrition	*.03*	.70	.96	.28	*.046*	.83	.74	.14
Scientific sources								
Is real and existing literature stated (no “fake citations”)?	.62	*<.001*	.22	*<.001*	*<.001*	*<.001*	*<.001*	*<.001*
Appropriateness of provided scientific evidence	.92	.52	.17	.40	*<.001*	*<.001*	*<.001*	*<.001*
Quality of provided scientific evidence in this specific context	*.04*	.72	.25	.08	*<.001*	*<.001*	*<.001*	*<.001*

[a]LLM: large language model.
[b]Gem_C1-S1: Google Gemini, use case 1—scenario 1.
[c]GPT_C1-S1: ChatGPT, use case 1—scenario 1.
[d]Gem_C1-S2: Google Gemini, use case 1—scenario 2.
[e]GPT_C1-S2: ChatGPT, use case 1—scenario 2.
[f]Gem_C2-S1: Google Gemini, use case 2—scenario 1.
[g]GPT_C2-S1: ChatGPT, use case 2—scenario 1.
[h]Gem_C2-S2: Google Gemini, use case 2—scenario 2.
[i]GPT_C2-S2: ChatGPT, use case 2—scenario 2.

Discussion

Principal Findings

We aimed (1) to investigate the quality of sleep recommendations provided to athletes generated by different publicly available LLMs as evaluated by experienced raters and (2) to investigate if the quality of sleep recommendations differs depending on the provided information by the user.

Our main results are as follows:

- The highest Likert-scale rating was achieved by GPT-4o using the prompt with high input granularity (use case 2, scenario 2) with 8 ratings >3 (tendency toward good), 3 ratings equal 3 (neutral), and 2 ratings <3 (tendency toward bad) on a 1-5 Likert Scale. This indicates that even the highest-ranked recommendations provided by the herein investigated LLMs are not optimal.
- Sleep recommendations by GPT-4o received higher Likert-scale ratings compared to those by Google Gemini (9 of 10 significant differences, with $P<.001$ to $P=.04$). This suggests a tendency that GPT-4o outperforms Google Gemini in the investigated sleep deficiency scenarios.
- Quality of sleep recommendations enhances with higher input information granularity (significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education; $P<.001$ to $P=.049$); however, some criteria of relevance for sleep education were partly or completely omitted, irrespective of the input information granularity.

Ratings of Generated Recommendations Regarding Sleep

Our results indicate that sleep recommendations of publicly available LLMs are not rated optimally, even when inserting a prompt with a high input information granularity. Although prompting GPT-4o with high input information granularity (user case 2, scenario 2) gained the highest Likert-scale ratings by the experienced raters (n=8 >3 Likert-scale rating), the summarized ratings demonstrate that the sleep recommendations exhibit deficiencies (n=3 ratings equal 3 [neutral]; n=2 ratings <3 [tendency toward bad]).

The insufficiency in providing optimal recommendations aligns with previous studies assessing LLMs in other research fields. For example, a systematic review and meta-analysis on ChatGPT's performance in answering medical questions showed that ChatGPT has an overall accuracy of 56%, suggesting potential but inadequacy for independent clinical decision-making [38]. Research in the field of nutrition reported inappropriate recommendations generated by ChatGPT, indicating that personalized dietary recommendations by ChatGPT involve unpredictable errors [32,39]. Therefore, LLMs should not be relied on to provide current nutritional advice without nutrition professionals [32,39].

Our results revealed further limitations pertaining to the negligence of criteria that are relevant for sleep education by the LLMs. In particular, the criterion "family commitments" did not exceed a Likert-scale median of 0 (IQR 0-0) across all LLMs and levels of input information granularities. Regarding the evaluation of all LLMs and input information granularities, "arousal the night before competition" attained a median of 0 (IQR 0-0 to 0-4) in 5 of 8 Likert-scale ratings, while "lifestyle" gained a maximum median of 3.5 (IQR 2.25-4), reflecting a neutral rating. These criteria were highlighted by recent research to influence both sleep quality and quantity [6,14]. Nastasi et al [40] reported similar results in assessing ChatGPT's ability to provide appropriate responses to medical questions within care contexts. While the authors noted appropriate responses corresponding to clinical guidelines, they indicated insufficient recommendations in regard to personalized medical advice, especially neglecting social factors [40].

Collectively, our results reveal that tailored sleep recommendations generated by GPT-4o and Google Gemini exhibit deficiencies according to received Likert-scale ratings by experienced raters, particularly in neglecting relevant criteria for sleep education. Therefore, sleep recommendations by LLMs should be carefully reviewed by a qualified coach or sleep professional before being applied in sleep educational settings.

Differences in Ratings of GPT-4o and Google Gemini

Our results demonstrate that sleep recommendations by GPT-4o generally were rated higher compared to those by Google Gemini. Of the 10 significant differences between the 2 LLMs in Likert-scale ratings, 9 favored GPT-4o ($P=.001$ to $P=.04$). Although the remaining scenarios (n=42) did not display significant results, our findings suggest a better quality of recommendations for sleep education by GPT-4o compared to Google Gemini.

Similar to our work, different authors compared different publicly available LLMs in different scenarios. For example, Günay et al [41] assessed GPT-4o and Google Gemini on 40 electrocardiogram cases and their responses to the most likely diagnosis. The results revealed that GPT-4o achieved higher accuracy compared to Gemini in electrocardiogram diagnostics. Carlà et al [42] evaluated ChatGPT and Google Gemini on 4 retinal detachment cases related to planned surgeries and revealed that ChatGPT received higher ratings for accuracy and precision compared to Gemini. Hieronimus et al [43] analyzed the completeness and accuracy of dietary reference intake in meal plans for different dietary patterns created by ChatGPT and Google Bard (subsequently rebranded Gemini) and revealed higher quality in the response of ChatGPT. In hand surgery, it was shown that Google Gemini outperformed ChatGPT in classifying injuries, while ChatGPT provided more sensitive recommendations regarding surgical interventions [44].

Collectively, it appears that different LLMs differ in output quality, irrespective of the use case, and there seems to be a tendency that versions of ChatGPT outperform Google

Gemini (which is in line with the results of our study), even though such statements need further investigation.

Differences in Quality Regarding Prompt Information Granularity

Our results indicate that higher input granularity leads to better-rated sleep recommendations, independent of the LLM. In the combined results of both use cases and LLMs, scenario 2 (higher input information granularity) received significantly higher Likert-scale ratings in 28 of 52 criteria items of relevance for sleep education (P=.001 to P=.049) compared to scenario 1 (low input information granularity), while scenario 1 attained higher Likert-scale ratings in 2 of 52 items (P=.001 to P=.009) compared to scenario 2.

Our finding is in line with research examining LLMs in other contexts [30,31,45-48]. For example, Kunze et al [48] inserted 20 knee complaints necessitating triage into ChatGPT-4 and investigated the accuracy and suitability rated by orthopedic sports medicine physicians. The authors stated that when providing additional input information, the accuracy of information output by ChatGPT-4 improved, particularly with enhancements in conservative management, surgical approaches, and related treatments [48].

In the context of endurance sports, Düking et al [30] provided 3 prompts with different levels of input information granularities to ChatGPT, resulting in 3 training plans aimed to improve running performance. Following an evaluation of these training plans by coaching experts, the authors indicated an increased quality with more input information provided [30].

Collectively, when using LLMs, it seems important for users to input a sufficient amount of information into LLMs to improve output quality, at least for the herein investigated scenario.

An additional point of interest is the quality and appropriateness of the scientific sources. Our results demonstrate that more input information leads to significantly higher ratings in items pertaining to the used scientific sources (P<.001), irrespective of the analyzed LLM. While higher input information granularity leads to improved sleep recommendations, the higher quality and appropriateness of the scientific sources may also contribute to these improvements. Since our analysis did not involve a differentiation between the effect of input information granularity and the use of scientific resources, future research should investigate both factors separately.

Conclusively, higher input information granularity leads to improved sleep recommendations and might be influenced by higher quality and appropriateness of the scientific sources. When supplying LLMs for sleep education, coaches and athletes should be highly aware of the appropriate input granularity and conduct a thorough review of the received output, potentially in consultation with an individual with strong experience in the field of sleep and athletic populations.

Strengths, Limitations, and Future Research

Strengths of our study include the assessment of different publicly available LLMs (ie, GPT-4o and Google Gemini) in different use cases and with different input information granularity, allowing a detailed analysis of provided sleep recommendations to the athletic population. Another strength of our study is the involvement of experienced raters evaluating the recommendations of LLMs.

Our results are limited to GPT-4o and Google Gemini on the versions available on May 15, 2024. As LLMs show fast developments (eg, by being able to search the internet), transferring our results to newer versions should be performed with caution. Due to such fast developments, it appears necessary to develop assessment methods or frameworks to evaluate the quality of LLMs in different scenarios to inform practitioners of the quality of currently available LLMs. Additionally, our results are, strictly speaking, only valid for the herein tested prompts, and other prompts might yield different results.

Despite the fact that all raters were experienced or well-educated in the field of sleep, the interrater reliability of our study ranges between slight and fair agreements (0.183 to 0.296). The tendency toward low interrater reliability is in line with previous research [30]. It may indicate that no single approach is universally optimal or perfect regarding athlete sleep recommendations. This suggests that different experts might rate the recommendations provided by LLMs in the respective use case differently, for example, based on personal preferences or personal experiences. While our results hold practical insights for athletes seeking sleep recommendations from LLMs (eg, to insert detailed information), individual sleep coaches might disagree with the recommendations provided by LLMs.

It seems important to note that even though it was shown that sleep education approaches can result in enhanced sleep behavior [13,49-51], to the best of our knowledge, there is no scientific evidence available that sleep recommendations generated by LLMs improve aspects of sleep in the athletic population. Future studies should evaluate the effectiveness of sleep education provided by publicly available LLMs to improve aspects of sleep.

Conclusions

Our study indicates that sleep recommendations generated by GPT-4o or Google Gemini are not rated optimally, independently of the level of input information granularity. However, our results demonstrate that GPT-4o provides better sleep recommendations to athletes compared to Google Gemini and that sleep recommendations improved with more detailed input information for both herein investigated LLMs. Collectively, for LLMs to be used in practice, it seems essential to insert detailed information into LLMs and to thoroughly review the provided sleep recommendations for athletic populations.

Acknowledgments

The authors acknowledge the use of large language models for assisting with grammar and language checks during the final preparation of this manuscript. The authors checked the final manuscript and approved it for publication. JMIR Publications has provided article processing fee (APF) support for the publication of this paper. The authors also acknowledge funding for the open-access publication fees by TU Braunschweig.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Conversation with ChatGPT4o.
[DOCX File (Microsoft Word File), 24 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Conversation with Google Gemini.
[DOCX File (Microsoft Word File), 29 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Results of the significance testing comparing training plans generated by Google Gemini and GPT-4o in response to different input information granularity.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 3]

References

1. Ramar K, Malhotra RK, Carden KA, et al. Sleep is essential to health: an American Academy of Sleep Medicine position statement. J Clin Sleep Med. Oct 1, 2021;17(10):2115-2119. [doi: 10.5664/jcsm.9476] [Medline: 34170250]
2. Agarwal AK, Gonzales R, Scott K, Merchant R. Investigating the feasibility of using a wearable device to measure physiologic health data in emergency nurses and residents: observational cohort study. JMIR Form Res. Feb 22, 2024;8:e51569. [doi: 10.2196/51569] [Medline: 38386373]
3. Xu Q, Kim Y, Chung K, Schulz P, Gottlieb A. Prediction of mild cognitive impairment status: pilot study of machine learning models based on longitudinal data from fitness trackers. JMIR Form Res. Jul 18, 2024;8:e55575. [doi: 10.2196/55575] [Medline: 39024003]
4. Knowlden AP, Winchester LJ, MacDonald HV, Geyer JD, Higginbotham JC. Associations among cardiometabolic risk factors, sleep duration, and obstructive sleep apnea in a southeastern US rural community: cross-sectional analysis from the SLUMBRx-PONS Study. JMIR Form Res. Nov 8, 2024;8:e54792. [doi: 10.2196/54792] [Medline: 39514856]
5. Gupta L, Morgan K, Gilchrist S. Does elite sport degrade sleep quality? A systematic review. Sports Med. Jul 2017;47(7):1317-1333. [doi: 10.1007/s40279-016-0650-6] [Medline: 27900583]
6. Walsh NP, Halson SL, Sargent C, et al. Sleep and the athlete: narrative review and 2021 expert consensus recommendations. Br J Sports Med. Apr 2021;55(7):356-368. [doi: 10.1136/bjsports-2020-102025]
7. Driller MW, Suppiah H, Rogerson D, Ruddock A, James L, Virgile A. Investigating the sleep habits in individual and team-sport athletes using the Athlete Sleep Behavior Questionnaire and the Pittsburgh Sleep Quality Index. Sleep Sci. 2022;15(1):112-117. [doi: 10.5935/1984-0063.20210031] [Medline: 35662975]
8. Halson SL. Sleep in elite athletes and nutritional interventions to enhance sleep. Sports Med. May 2014;44(Suppl 1):S13-23. [doi: 10.1007/s40279-014-0147-0] [Medline: 24791913]
9. Cunha LA, Costa JA, Marques EA, Brito J, Lastella M, Figueiredo P. The impact of sleep interventions on athletic performance: a systematic review. Sports Med Open. Jul 18, 2023;9(1):58. [doi: 10.1186/s40798-023-00599-z] [Medline: 37462808]
10. Venter RE. Perceptions of team athletes on the importance of recovery modalities. Eur J Sport Sci. 2014;14 Suppl 1:S69-76. [doi: 10.1080/17461391.2011.643924] [Medline: 24444246]
11. Czeisler CA, Klerman EB. Circadian and sleep-dependent regulation of hormone release in humans. Recent Prog Horm Res. 1999;54:97-130. [Medline: 10548874]
12. Fullagar HHK, Skorski S, Duffield R, Hammes D, Coutts AJ, Meyer T. Sleep and athletic performance: the effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise. Sports Med. Feb 2015;45(2):161-186. [doi: 10.1007/s40279-014-0260-0] [Medline: 25315456]

13. Caia J, Scott TJ, Halson SL, Kelly VG. The influence of sleep hygiene education on sleep in professional rugby league athletes. Sleep Health. Aug 2018;4(4):364-368. [doi: 10.1016/j.sleh.2018.05.002] [Medline: 30031530]
14. Cook JD, Charest J. Sleep and performance in professional athletes. Curr Sleep Med Rep. 2023;9(1):56-81. [doi: 10.1007/s40675-022-00243-4] [Medline: 36683842]
15. Miles KH, Clark B, Fowler PM, Miller J, Pumpa KL. Sleep practices implemented by team sport coaches and sports science support staff: a potential avenue to improve athlete sleep? J Sci Med Sport. Jul 2019;22(7):748-752. [doi: 10.1016/j.jsams.2019.01.008] [Medline: 30685228]
16. Salah M, Alhalbusi H, Ismail MM, Abdelfattah F. Chatting with ChatGPT: decoding the mind of Chatbot users and unveiling the intricate connections between user perception, trust and stereotype perception on self-esteem and psychological well-being. Curr Psychol. Mar 2024;43(9):7843-7858. [doi: 10.1007/s12144-023-04989-0]
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). Mar 19, 2023;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
18. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ. Sep 4, 2023;9:e46482. [doi: 10.2196/46482] [Medline: 37665620]
19. Dagli MM, Oettl FC, Gujral J, et al. Clinical accuracy, relevance, clarity, and emotional sensitivity of large language models to surgical patient questions: cross-sectional study. JMIR Form Res. Jun 7, 2024;8:e56165. [doi: 10.2196/56165] [Medline: 38848553]
20. Skryd A, Lawrence K. ChatGPT as a tool for medical education and clinical decision-making on the wards: case study. JMIR Form Res. May 8, 2024;8:e51346. [doi: 10.2196/51346] [Medline: 38717811]
21. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic hepatitis B: cross-language study of English and Chinese. JMIR Med Inform. Aug 8, 2024;12:e56426. [doi: 10.2196/56426] [Medline: 39115930]
22. Kim J, Vajravelu BN. Assessing the current limitations of large language models in advancing health care education. JMIR Form Res. Jan 16, 2025;9:e51319. [doi: 10.2196/51319] [Medline: 39819585]
23. Yang X, Li G. Psychological and behavioral insights from social media users: natural language processing-based quantitative study on mental well-being. JMIR Form Res. Jan 20, 2025;9:e60286. [doi: 10.2196/60286] [Medline: 39832365]
24. Zheng J, Wang J, Shen J, An R. Artificial intelligence applications to measure food and nutrient intakes: scoping review. J Med Internet Res. Nov 28, 2024;26:e54557. [doi: 10.2196/54557] [Medline: 39608003]
25. Tagi M, Hamada Y, Shan X, et al. A food intake estimation system using an artificial intelligence-based model for estimating leftover hospital liquid food in clinical environments: development and validation study. JMIR Form Res. Nov 5, 2024;8:e55218. [doi: 10.2196/55218] [Medline: 39500491]
26. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 1, 2023;183(6):589. [doi: 10.1001/jamainternmed.2023.1838]
27. Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet. Dec 2023;308(6):1831-1844. [doi: 10.1007/s00404-023-07130-5] [Medline: 37458761]
28. Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. Aesthet Surg J. Sep 14, 2023;43(10):1126-1135. [doi: 10.1093/asj/sjad140] [Medline: 37158147]
29. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg. Jun 2023;33(6):1790-1796. [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]
30. Düking P, Sperlich B, Voigt L, Van Hooren B, Zanini M, Zinner C. ChatGPT generated training plans for runners are not rated optimal by coaching experts, but increase in quality with additional input information. J Sports Sci Med. Mar 2024;23(1):56-72. [doi: 10.52082/jssm.2024.56] [Medline: 38455449]
31. Dergaa I, Saad HB, El Omri A, et al. Using artificial intelligence for exercise prescription in personalised health promotion: a critical evaluation of OpenAI's GPT-4 model. Biol Sport. Mar 2024;41(2):221-241. [doi: 10.5114/biolsport.2024.133661] [Medline: 38524814]
32. Dergaa I, Ben Saad H, Ghouili H, et al. Evaluating the applicability and appropriateness of ChatGPT as a source for tailored nutrition advice: a multi-scenario study. NAJM. 2024;2(1):1-16. [doi: 10.61838/kman.najm.2.1.1]

33. Havers T, Masur L, Isenmann E, et al. Reproducibility and quality of hypertrophy-related training plans generated by GPT-4 and Google Gemini as evaluated by coaching experts. Biol Sport. Apr 2025;42(2):289-329. [doi: 10.5114/biolsport.2025.145911] [Medline: 40182716]
34. McKay AKA, Stellingwerff T, Smith ES, et al. Defining training and performance caliber: a participant classification framework. Int J Sports Physiol Perform. Feb 1, 2022;17(2):317-331. [doi: 10.1123/ijspp.2021-0451] [Medline: 34965513]
35. Casado A, González-Mohíno F, González-Ravé JM, Foster C. Training periodization, methods, intensity distribution, and volume in highly trained and elite distance runners: a systematic review. Int J Sports Physiol Perform. 2022;17(6):820-833. [doi: 10.1123/ijspp.2021-0435]
36. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378-382. [doi: 10.1037/h0031619]
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159. [doi: 10.2307/2529310]
38. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform. Mar 2024;151:104620. [doi: 10.1016/j.jbi.2024.104620] [Medline: 38462064]
39. Agne I, Gedrich K. Personalized dietary recommendations for obese individuals—a comparison of ChatGPT and the Food4Me algorithm. Clin Nutr Open Sci. Aug 2024;56:192-201. [doi: 10.1016/j.nutos.2024.06.001]
40. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Sci Rep. Oct 19, 2023;13(1):17885. [doi: 10.1038/s41598-023-45223-y] [Medline: 37857839]
41. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. Am J Emerg Med. Oct 2024;84:68-73. [doi: 10.1016/j.ajem.2024.07.043] [Medline: 39096711]
42. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. Sep 20, 2024;108(10):1457-1469. [doi: 10.1136/bjo-2023-325143] [Medline: 38448201]
43. Hieronimus B, Hammann S, Podszun MC. Can the AI tools ChatGPT and Bard generate energy, macro- and micro-nutrient sufficient meal plans for different dietary patterns? Nutr Res. Aug 2024;128:105-114. [doi: 10.1016/j.nutres.2024.07.002] [Medline: 39102765]
44. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in hand surgery: assessing large language models in the classification and management of hand injuries. J Clin Med. May 11, 2024;13(10). [doi: 10.3390/jcm13102832] [Medline: 38792374]
45. Dergaa I, Fekih-Romdhane F, Hallit S, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. 2023;14:1277756. [doi: 10.3389/fpsyt.2023.1277756] [Medline: 38239905]
46. Washif JA, Pagaduan J, James C, Dergaa I, Beaven CM. Artificial intelligence in sport: exploring the potential of using ChatGPT in resistance training prescription. Biol Sport. Mar 2024;41(2):209-220. [doi: 10.5114/biolsport.2024.132987] [Medline: 38524820]
47. Zaleski AL, Berkowsky R, Craig KJT, Pescatello LS. Comprehensiveness, accuracy, and readability of exercise recommendations provided by an AI-based chatbot: mixed methods study. JMIR Med Educ. Jan 11, 2024;10:e51308. [doi: 10.2196/51308] [Medline: 38206661]
48. Kunze KN, Varady NH, Mazzucco M, et al. The large language model ChatGPT-4 exhibits excellent triage capabilities and diagnostic performance for patients presenting with various causes of knee pain. Arthroscopy. May 2025;41(5):1438-1447. [doi: 10.1016/j.arthro.2024.06.021]
49. O'Donnell S, Driller MW. Sleep-hygiene education improves sleep indices in elite female athletes. Int J Exerc Sci. 2017;10(4):522-530. [doi: 10.70252/DNOL2901] [Medline: 28674597]
50. Fullagar HHK, Skorski S, Duffield R, Julian R, Bartlett J, Meyer T. Impaired sleep and recovery after night matches in elite football players. J Sports Sci. Jul 2016;34(14):1333-1339. [doi: 10.1080/02640414.2015.1135249] [Medline: 26750446]
51. Driller MW, Lastella M, Sharp AP. Individualized sleep education improves subjective and objective sleep indices in elite cricket athletes: a pilot study. J Sports Sci. Sep 2019;37(17):2121-2125. [doi: 10.1080/02640414.2019.1616900] [Medline: 31076021]

Abbreviations

GPT-4o: ChatGPT-4o
LLM: large language model

Edited by Amaryllis Mavragani; peer-reviewed by Marcus Schmidt, Shahab Haghayegh; submitted 16.01.2025; final revised version received 06.03.2025; accepted 11.03.2025; published 08.07.2025

Please cite as:
Masur L, Driller M, Suppiah H, Matzka M, Sperlich B, Düking P
Assessment of Recommendations Provided to Athletes Regarding Sleep Education by GPT-4o and Google Gemini: Comparative Evaluation Study
JMIR Form Res 2025;9:e71358
URL: https://formative.jmir.org/2025/1/e71358
doi: 10.2196/71358