healthcare

MDPI

Article

Reliability of Repeated LLM-Generated Exercise Prescriptions: A Multidimensional Evaluation of Semantic, Structural, and Safety-Content Consistency

Kihyuk Lee

Data Convergence Team, Office of Hospital Information, Seoul National University Bundang Hospital, Seongnam 13605, Republic of Korea; lkhlike@naver.com

Highlights

What are the main findings?

- Repeated LLM outputs showed high full-text semantic similarity (0.939–0.967), but resistance-exercise intensity was unclassifiable in 10–25% of outputs in the clinical scenarios.
- According to the AI evaluator, all predefined safety-related content categories were identified in every output, but the amount of safety-related content varied significantly across scenarios ($H = 86.18$, $p < 0.001$).

What are the implications of the main findings?

- High semantic similarity does not guarantee consistency in clinically actionable prescription components; quantitative elements such as exercise intensity require separate verification.
- Pre-deployment evaluation of generative exercise-prescription systems should assess repeated-generation reliability and quantitative consistency, rather than relying on single-output evaluation alone.

Abstract

Background/Objectives: Large language models (LLMs) are increasingly used to generate clinical decision-support outputs such as personalized exercise prescriptions, but high output quality does not by itself establish reproducibility across repeated generations. **Methods:** This study evaluated that reliability across semantic, structural, and safety-content dimensions. Six scenarios, covering patient groups and healthy individuals, were each generated 20 times under identical prompts using Gemini 2.5 Flash (total $n = 120$). Consistency was assessed via full-prescription SBERT embeddings constructed by pooling non-overlapping 254-content-token chunks (semantic), FITT-based AI-as-a-judge classification (structural), and inclusion rates and sentence counts for safety-related content (safety-content consistency). Output-level jackknife methods accounted for dependence among pairwise similarities. **Results:** Full-text semantic similarity was high across scenarios (mean cosine similarity: 0.939–0.967), with the highest descriptive mean in S6; the scenario comparison was not statistically significant under the primary Welch approximation ($p = 0.064$), and no Holm-adjusted pairwise comparison was significant. Frequency showed consistent patterns, but variability was observed in quantitative components, particularly exercise intensity; unclassifiable intensity expressions occurred in 10–25% of resistance training outputs. The AI evaluator identified safety-related expressions in 100% of outputs, although sentence counts varied significantly across scenarios ($H = 86.18$, $p < 0.001$), with clinical cases exceeding healthy adult cases. **Conclusions:** LLM-generated exercise prescriptions showed high semantic similarity and generally aligned with guideline-based structures,

Check for updates

Academic Editor: Daniele Giansanti

Received: 25 September 2026
Revised: 9 November 2026
Accepted: 23 November 2026
Published: 25 November 2026

Copyright: © 2026 by the author. 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.

but variability in quantitative components indicated that output stability may vary across prescription components even under identical prompting. Additional structural constraints and expert validation are needed before clinical deployment.

Keywords: large language models; exercise prescription; consistency; repeated generation; digital health

1. Introduction

Large language models (LLMs) are increasingly incorporated into clinical decision-support workflows, and their outputs are conventionally evaluated using accuracy or semantic similarity against a reference standard. A growing body of work suggests that such single-metric evaluation is insufficient to establish deployment readiness: models can score well on one axis while still exhibiting unstable reasoning, disagreement across independent verifier models, or outputs that do not hold up under repetition [1]. In generative systems specifically, the stability of outputs across repeated generations under identical conditions constitutes a distinct reliability dimension that single-generation accuracy alone cannot capture.

In the field of exercise prescription specifically, this advancement has opened new possibilities. LLMs are capable of generating structured exercise recommendations that account for individual health status, disease characteristics, and contraindications, and their potential as decision-support tools in settings where access to specialists is limited has been widely discussed [2]. A growing body of expert evaluation studies has examined LLM-generated prescriptions across various contexts, including resistance training programs, running training plans, and exercise prescriptions for patients with diabetes [3–5], and a recent scoping review confirmed that research in this area is transitioning from an exploratory phase to one of systematic validation [6].

Exercise prescription is fundamentally distinct from general health information provision, and in high-risk clinical contexts involving multiple chronic conditions, functional limitations, and contraindications, verification of not only output quality but also reproducibility is essential. Given these characteristics, exercise prescription in patients with comorbidity requires complex clinical reasoning that may lead to variability in professional judgment [7], and the application of AI-based systems demands even more rigorous validation of reproducibility. In this study, reproducibility refers to the capacity of an LLM to produce stable outputs under identical conditions, and consistency refers to the degree of similarity across repeated outputs as its measurable proxy. Output variability in this context carries direct clinical implications: the same patient profile may yield structurally different prescriptions across repeated generations, raising concerns about decision-making consistency in real-world deployment. Choi et al. [8] evaluated LLM-generated exercise prescriptions under a structured prompt framework and reported that prompt structuring improved safety and guideline adherence, but noted low inter-rater agreement (ICC (2,3) = 0.139) and difficulty securing statistical independence among repeated outputs, proposing systematic consistency analysis as a future direction.

LLMs may produce variable outputs across repeated generations under identical prompts due to their probabilistic token generation mechanism [9], representing a direct concern for clinical reliability in exercise prescription. The LLM-as-a-Judge paradigm, in which a separate LLM evaluates outputs according to a predefined rubric, has been proposed as a scalable and reproducible complementary assessment approach [10,11]. However, limitations such as self-preference bias and length bias have been reported [12],

and more stable application is expected in domains where clear binary criteria exist, such as adherence to the FITT principle and inclusion of safety-related content.

In exercise prescription specifically, this problem is compounded by the numerical nature of FITT components: outputs that appear semantically similar may nonetheless specify different intensity, frequency, or duration values, elements that determine whether a prescription is clinically actionable.

Building on the structured prompt framework developed by Choi et al. [8], this study evaluates the reliability of repeated LLM-generated exercise prescriptions across three complementary dimensions: semantic similarity (SBERT-based cosine similarity), structural consistency based on the FITT principle, and safety-content consistency. By treating reproducibility as a reliability dimension distinct from single-generation accuracy and safety-focused assessments, this study provides a foundational basis for pre-deployment reliability assessment of LLM-based exercise prescription systems and for developing automated evaluation methodologies.

2. Materials and Methods

2.1. Study Design and Overview

This study was conducted to evaluate the intra-model consistency of exercise prescriptions generated by a large language model (LLM) under controlled conditions. A repeated generation design was applied, in which identical clinical scenarios and prompts were repeatedly submitted to produce multiple outputs, allowing for the assessment of within-model variability.

This study did not involve human participants, human data, or biological material; all clinical scenarios were hypothetical constructs developed without real patient information, and accordingly ethics approval and informed consent were not required. An overview of the study workflow is presented in Figure 1.

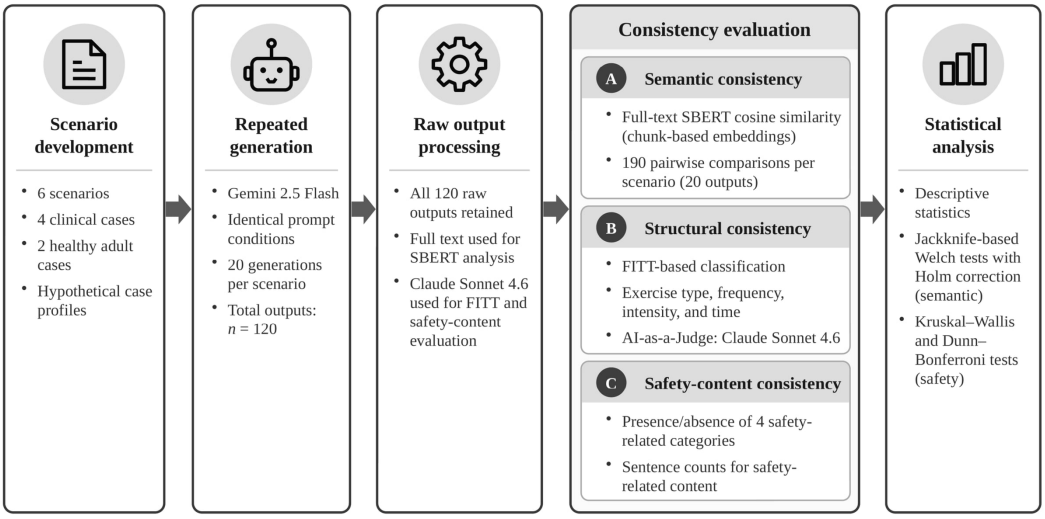


Figure 1. Overview of the study design and analytical workflow. FITT = frequency, intensity, time, and type; SBERT = Sentence-BERT.

2.2. Clinical Scenario Development

Six scenarios were used: three adopted from a previous study [8] representing high-risk conditions across metabolic, musculoskeletal, and oncological domains, and three newly developed to extend clinical variability by including a multimorbidity case with cardiovascular risk factors and two healthy adult cases targeting fat loss and muscle hypertrophy, respectively. All scenarios were hypothetical, constructed without real pa-

tient information, and designed to reflect clinically relevant factors including age, sex, disease characteristics, functional limitations, and exercise goals. Key characteristics are presented in Table 1, with a more detailed breakdown of clinical characteristics provided in Supplementary Table S1, and the corresponding prompt and case narratives are provided in Supplementary Material S1.

Table 1. Summary of Clinical Scenarios Used for Repeated Exercise Prescription Generation.

Clinical Cases	Sex	Age (Years)	Primary Condition (s)	Key Functional Limitation or Risk	Baseline Physical Activity Level	Primary Exercise Goal
Case 1 (T2DM + Obesity)	Male	55	Type 2 diabetes mellitus, obesity	Limited exercise experience, mild peripheral neuropathy	Low	Weight reduction and improved glycemic control
Case 2 (Knee OA + Fall Risk)	Female	70	Knee osteoarthritis	Knee pain during walking, prior fall incident	Low	Pain reduction, maintenance of walking ability, and fall prevention
Case 3 (Post-Colon CA Recovery)	Male	60	Post-colon cancer surgery	Deconditioning, limited walking endurance, fatigue	Low	Physical recovery, fatigue reduction, and improvement of lifestyle habits
Case 4 (HTN + T2DM + Obesity)	Female	68	Hypertension, type 2 diabetes mellitus, obesity	Dizziness, mild exertional dyspnea, on antihypertensive and antidiabetic medication	Low	Safe improvement of cardiovascular fitness, glycemic control, and gradual weight reduction
Case 5 (Fat Loss + Cardiovascular Endurance)	Female	28	None (healthy individual)	No clinical limitations; moderate exercise experience	Moderate to High	Body fat reduction and improvement of cardiovascular endurance
Case 6 (Muscle Hypertrophy + Strength)	Male	30	None (healthy individual)	No clinical limitations; prior resistance training experience	High	Increase muscle mass and improve overall strength

T2DM = type 2 diabetes mellitus; OA = osteoarthritis; CA = cancer; HTN = hypertension. Cases 1–3 were adopted from Choi et al. [8].

2.3. AI Model and Session Control

All outputs were generated automatically through the Vertex AI API (gemini-2.5-flash, Google LLC, Mountain View, CA, USA). Temperature was set to 1.0 to examine variability under stochastic generation rather than to simulate a low-temperature clinical deployment setting, and other generation parameters, including top-p, were kept at their default values. Unlike the previous study [8], which involved manual prompt entry through a web interface using a non-logged-in guest session, this study adopted an API-based automated approach to enhance the reproducibility and degree of control over experimental conditions. No additional reasoning elicitation techniques such as chain-of-thought were applied.

2.4. Output Generation Procedure

Exercise prescriptions were generated 20 times per scenario under identical prompt conditions, yielding a total of 120 outputs. This repetition count was determined with reference to repeated generation designs used to evaluate LLM response variability [9]. The prompt was based on the guideline-based instruction from the previous study [8], with three modifications: addition of %1RM for resistance exercise intensity notation, removal of the 'ACSM Guidelines for Cancer Survivors' as inapplicable to all scenarios, and removal of the 'if applicable' condition to ensure consistent inclusion of all four exercise components across outputs. The full prompt and case descriptions are provided in Supplementary Material S1. All prompts were kept identical across repetitions. Data collection was conducted between 5 April and 9 April 2026.

2.5. Data Processing and Preprocessing

All 120 generated outputs were retained in their raw form for semantic analysis to reflect the model's actual response characteristics. No formatting constraints were imposed during output generation. Raw outputs included greetings, closing remarks, tables, and bullet points. FITT and safety-content classifications were evaluated separately using Claude Sonnet 4.6, as described below.

2.6. Consistency Evaluation

Consistency was evaluated across three dimensions: (1) semantic consistency, (2) structural consistency, and (3) safety-content consistency.

Semantic consistency was assessed using all-MiniLM-L6-v2, a computationally efficient general-domain sentence embedding model selected to evaluate overall semantic similarity rather than clinical equivalence [13]. Each raw output was divided into non-overlapping chunks of up to 254 content tokens, with two additional special tokens per chunk, to accommodate the model's 256-token input limit. All chunks were encoded using the same model, and a token-count-weighted average of the chunk embeddings was calculated and L2-normalized to obtain a document-level embedding. Equal-chunk-weight pooling was examined in a sensitivity analysis. Within each scenario, cosine similarity was calculated for all 190 distinct pairs of 20 document embeddings, and the mean pairwise similarity was treated as an order-two U-statistic.

Structural consistency was evaluated based on four FITT components: exercise type, frequency, intensity, and time. Classification was performed using Claude Sonnet 4.6 as an independent LLM evaluator, with the generation and evaluation models separated to mitigate self-evaluation bias [10,12,14]. Intensity was classified for the initial prescription period (weeks 1–4) into three levels (low, moderate, high) using RPE, %HRmax, heart rate reserve (HRR), and MET for aerobic exercise [15,16] and %1RM for resistance exercise [17,18]. When the relevant metric was not explicitly stated, descriptive expressions were interpreted and classified into the low, moderate, or high categories according to the criteria above; outputs without a resistance-intensity value that could be mapped to the specified %1RM categories for weeks 1–4 were labeled "unclassifiable," even if an alternative qualitative or RPE-based instruction was present (Table 2).

Safety-content consistency was assessed by evaluating the binary presence (1) or absence (0) of four categories: contraindications, precautions, symptom monitoring, and risk warnings. Contraindications and precautions were based on the safety evaluation rubric of the previous study [8], symptom monitoring referenced the ACSM Guidelines for Exercise Testing and Prescription [19], and risk warnings were operationally defined with reference to the same guidelines. The inclusion rate per scenario served as the primary safety-content consistency indicator. Additionally, sentence-level counts per category were

computed to quantify the extent of safety expression. Each qualifying sentence or distinct clause was assigned to its single most relevant safety category without double counting; category totals were summed per output. This measure reflects the consistency of safety-related content inclusion across repeated outputs, not a clinical validation of prescription safety. Full prompts and classification criteria are provided in Supplementary Material S2. As an output-length-adjusted supplementary analysis, the four category counts and their total were expressed per 1000 words of each raw output.

Table 2. Intensity Classification Criteria for Aerobic and Resistance Exercise.

Level	Aerobic Exercise	Resistance Exercise
Low	RPE < 12/%HRmax < 64%/HRR < 40%/< 3 METs	<50% 1RM
Moderate	RPE 12–13/%HRmax 64–76%/HRR 40–59%/3–5.9 METs	50–69% 1RM
High	RPE ≥ 14/%HRmax ≥ 77%/HRR ≥ 60%/≥ 6 METs	≥70% 1RM
Unclassifiable	Expression present but cannot be mapped to above criteria	Expression present but %1RM not stated or estimable

RPE = rating of perceived exertion (Borg 6–20 scale); HRmax = maximum heart rate; HRR = heart rate reserve; MET = metabolic equivalent of task; 1RM = one-repetition maximum.

2.7. Quality Verification

For a supplementary author–AI agreement assessment, 24 outputs (20%; four per scenario) were purposively selected, including five cases labeled unclassifiable for resistance intensity by the AI evaluator. The author independently reviewed the original outputs without access to Claude's labels, using the same aerobic and resistance intensity criteria and four binary safety-content categories; percentage agreement and Cohen's κ were calculated for each category. Some author judgments of safety content also considered the specificity of the provided guidance, rather than only the presence of related statements.

2.8. Statistical Analysis

Descriptive statistics are presented as mean ± standard deviation where applicable. For semantic similarity, statistical inference used the 20 generated outputs per scenario rather than the 190 dependent similarity pairs. Delete-one-output jackknife standard errors and marginal 95% t confidence intervals were calculated for the scenario-specific mean pairwise similarity (an order-two U-statistic). Between-scenario differences were evaluated with a Welch approximate omnibus F test on jackknife pseudovalues, followed by Welch–Satterthwaite pairwise contrasts with Holm correction across 15 comparisons. These procedures provide approximate rather than exact small-sample inference. A fixed-variance Rademacher multiplier omnibus test (50,000 replicates) was performed as a supplementary sensitivity check. Nonparametric Kruskal–Wallis and Dunn tests with Bonferroni correction were applied to safety-related sentence counts. Structural consistency was summarized descriptively using FITT classification frequencies. The significance level was $p < 0.05$. Analyses were performed in Python (version 3.12.13; Google Colaboratory) using sentence-transformers, scipy, numpy, pandas, and related scientific libraries. The six scenarios were additionally compared using Kruskal–Wallis tests on output-level length-adjusted safety-content rates.

2.9. Generative AI Use in Manuscript Preparation

Generative AI tools (Claude Sonnet 4.6, Anthropic; GPT-4o, OpenAI) were used to assist with translation, language refinement, and manuscript review; all AI-assisted content was reviewed and edited by the author, who takes full responsibility for the final content.

3. Results

3.1. Semantic Consistency Analysis

In the token-weighted full-prescription SBERT analysis, mean pairwise cosine similarities across 20 outputs per scenario ranged from 0.939 to 0.967 (Table 3; Figure 2). The highest descriptive mean was observed in S6 (0.967; jackknife 95% CI: 0.963–0.971), followed by S5 (0.962), while S1 was lowest (0.939; 95% CI: 0.919–0.959). The output-level jackknife Welch omnibus comparison was not statistically significant ($F\,[5, 48.71] = 2.25$, $p = 0.064$); none of the 15 Holm-adjusted pairwise comparisons were significant (Figure 3). The equal-chunk-weight sensitivity analysis yielded a similar mean range (0.939–0.967) and the same descriptive ordering but a borderline omnibus result ($p = 0.048$); again, no Holm-adjusted pairwise comparison was significant. A supplementary fixed-variance multiplier test yielded $p = 0.012$ for the token-weighted analysis, indicating sensitivity of the omnibus inference to the test calibration.

Table 3. Full-Prescription SBERT-Based Cosine Similarity by Scenario (Token-Weighted Chunk Pooling).

Scenario	Case Description	N outputs	Mean	Pair SD	Jackknife SE	95% CI
S1	T2DM + Obesity	20	0.939	0.042	0.010	0.919–0.959
S2	Knee OA + Fall Risk	20	0.954	0.029	0.008	0.937–0.971
S3	Post-colon CA Recovery	20	0.957	0.026	0.007	0.941–0.972
S4	HTN + T2DM + Obesity	20	0.961	0.026	0.008	0.945–0.978
S5	Fat Loss + Endurance	20	0.962	0.016	0.004	0.953–0.971
S6	Hypertrophy + Strength	20	0.967	0.008	0.002	0.963–0.971

T2DM = type 2 diabetes mellitus; OA = osteoarthritis; CA = cancer; HTN = hypertension. Cases 1–3 were adopted from Choi et al. [8].

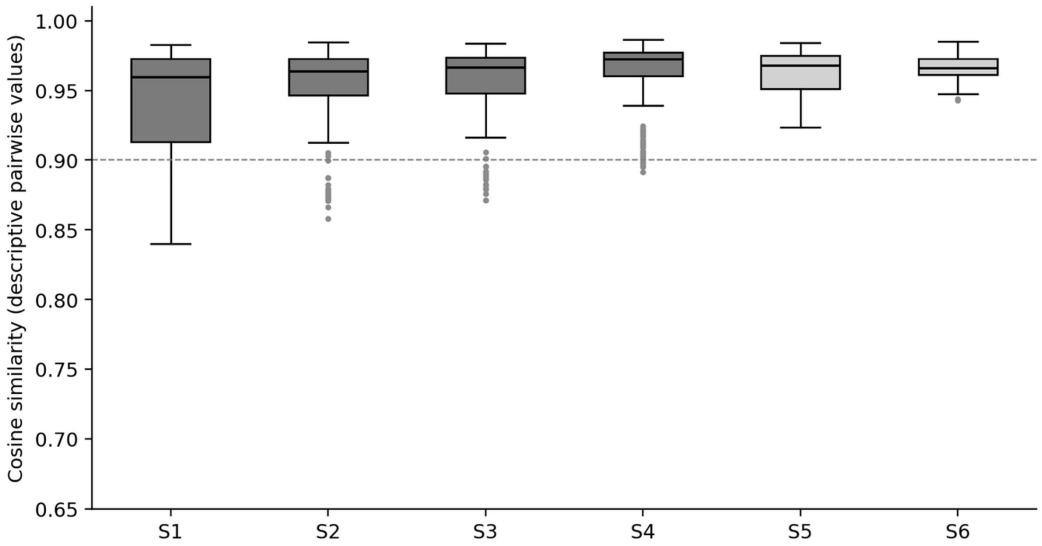


Figure 2. Descriptive distribution of full-prescription SBERT-based cosine similarities across scenarios. Each box summarizes 190 dependent pairwise values computed from 20 repeated outputs; the pairs are not independent observations. Dark gray = clinical cases (S1–S4); light gray = healthy adult cases (S5–S6). Dashed line indicates cosine similarity of 0.90. Statistical inference was performed at the output level (see Table 3).

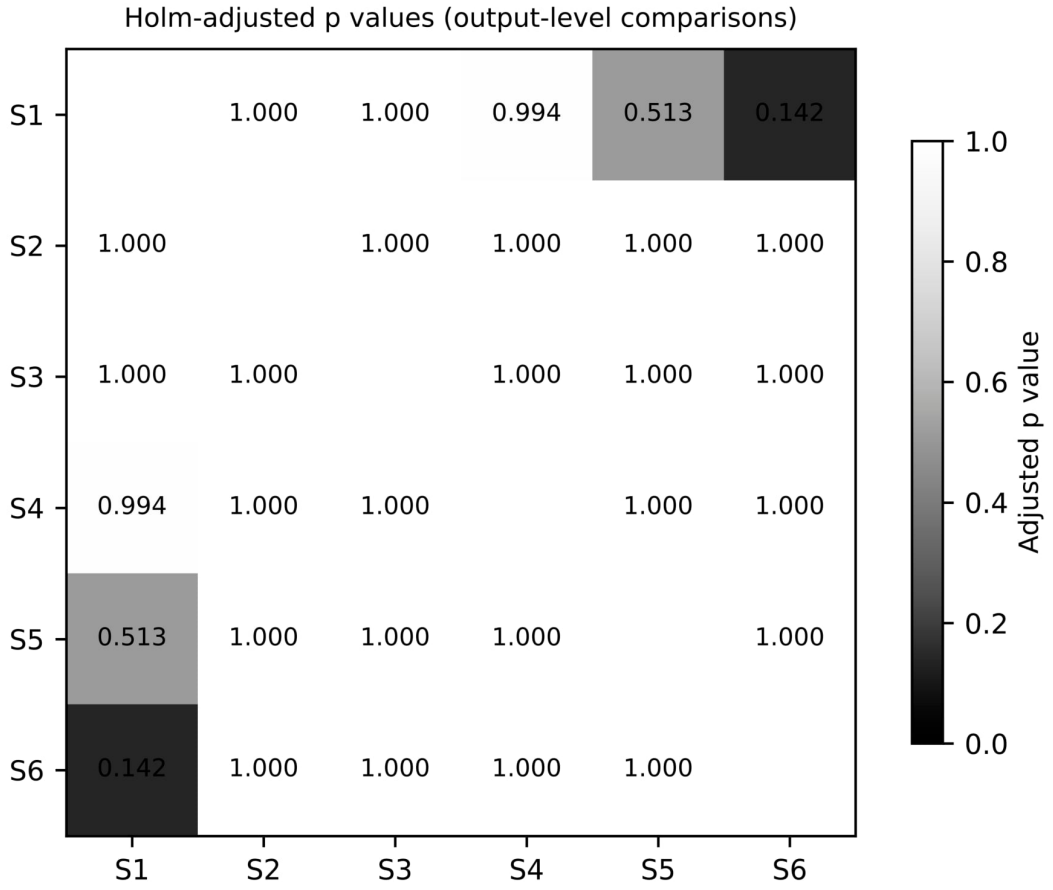


Figure 3. Holm-adjusted p values for 15 output-level jackknife Welch–Satterthwaite pairwise contrasts of full-prescription token-weighted cosine similarity. No contrast reached $p < 0.05$. Displayed p values are approximations based on 20 outputs per scenario.

3.2. Structural Consistency: FITT Classification

FITT components were classified based on the initial prescription phase (Weeks 1–4) using Claude Sonnet 4.6 as an AI judge (Table 4, Figure 4). In clinical cases (S1–S4), aerobic exercise three to four times per week combined with resistance exercise twice per week was the most common frequency pattern (45–50%), while S5 showed an aerobic-dominant pattern (4–5×/week; 65%) and S6 a resistance-centered pattern (4×/week; 35%). Aerobic intensity in clinical cases was predominantly low (90–100%), whereas healthy adult cases were classified as moderate or high (100%). Resistance exercise intensity was predominantly low in clinical cases (75–85%), with unclassifiable outputs observed in 10–25% of cases. Exercise duration ranged from 10–30 min in clinical cases, 30–40 min in S5, and 60–90 min in S6. Across all scenarios, aerobic, resistance, balance, and flexibility exercises were prescribed in combination, with walking and balance training predominant in clinical cases and resistance training predominant in healthy adult cases. Across the four clinical scenarios, 14 of 80 outputs (17.5%) lacked a classifiable initial-phase resistance %1RM intensity (S1: 2/20; S2: 3/20; S3: 4/20; S4: 5/20). Among the remaining 66 classifiable clinical outputs, 62 were rated low and four moderates. Unclassifiable indicates insufficient %1RM specification under the rubric, not necessarily absence of resistance-exercise instructions; differences among classifiable outputs represent a separate form of across-repetition variability.

Table 4. FITT Intensity Classification by Scenario (Weeks 1–4).

Scenario	Frequency	Aerobic Intensity *n* (%)	Resistance Intensity *n* (%)	Time
S1	Aerobic 3–4x/Resistance 2x per week (50%)	Low 19 (95%), Moderate 1 (5%)	Low 15 (75%), Moderate 3 (15%), Unclass. 2 (10%)	20–30 min
S2	Aerobic 3–4x/Resistance 2x per week (45%)	Low 18 (90%), Moderate 2 (10%)	Low 17 (85%), Unclass. 3 (15%)	10–25 min
S3	Aerobic 3–4x/Resistance 2x per week (50%)	Low 20 (100%)	Low 15 (75%), Unclass. 4 (20%), Moderate 1 (5%)	15–30 min
S4	Aerobic 3–4x/Resistance 2x per week (50%)	Low 20 (100%)	Low 15 (75%), Unclass. 5 (25%)	10–30 min
S5	Aerobic 4–5x/Resistance 3x per week (65%)	Moderate 20 (100%)	Moderate 19 (95%), High 1 (5%)	30–40 min
S6	Resistance 4x/Aerobic 2–3x per week (35%)	Moderate 17 (85%), High 3 (15%)	Moderate 14 (70%), High 6 (30%)	60–90 min

Unclass = unclassifiable.

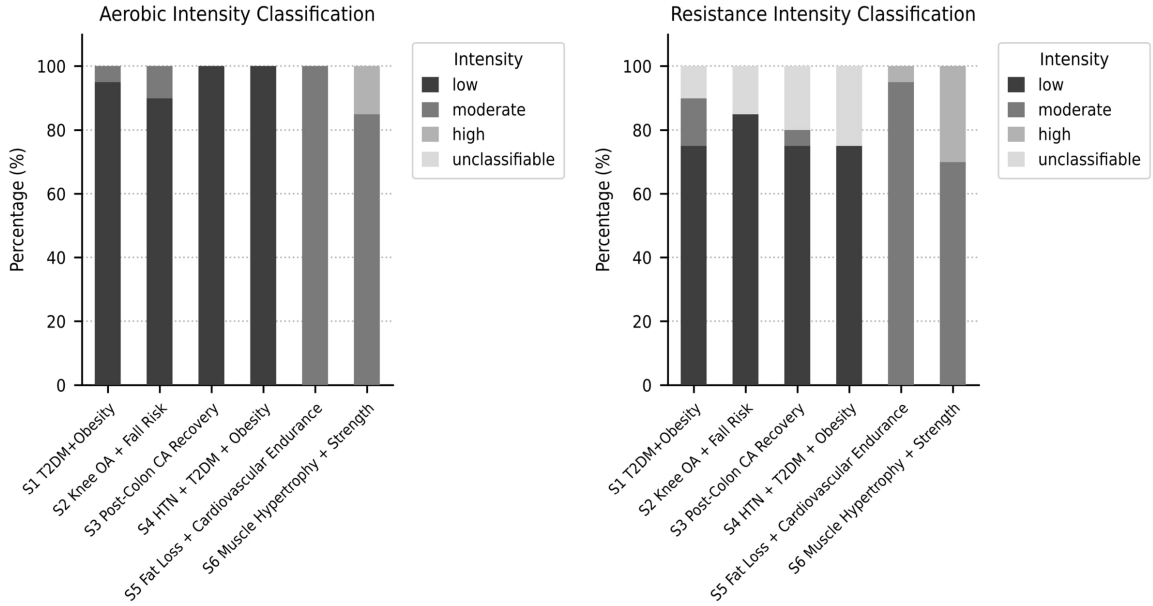


Figure 4. Distribution of FITT intensity classifications across scenarios as assessed by Claude Sonnet 4.6 (AI-as-a-Judge). **Left panel** shows aerobic intensity; **right panel** shows resistance intensity. Classifications are based on initial prescription phase (Weeks 1–4). The lightest shade indicates unclassifiable outputs.

3.3. Safety-Content Consistency

Evaluation of the presence or absence of safety-related expressions across four categories, including contraindications, precautions, symptom monitoring, and risk warnings, showed that the AI evaluator identified all four categories in 100% of the 120 outputs.

Beyond binary inclusion, additional analysis of sentence counts by category revealed significant differences in the total number of safety-related sentences across scenarios (Kruskal–Wallis: $H = 86.18$, $p < 0.001$). Clinical cases (S1–S4) generated more safety-related expressions than healthy adult cases (S5–S6), with the highest value observed in S4 (Mean = 61.40, SD = 6.62). Significant differences were also identified across all four individual categories (all $p < 0.01$, Table 5).

Table 5. Descriptive Statistics and Kruskal–Wallis Test Results of Safety-Related Sentence Counts by Scenario and Category.

Category	S1	S2	S3	S4	S5	S6	H	*p*
Contraindication	11.90 ± 3.99	10.00 ± 3.01	14.35 ± 2.82	13.95 ± 2.86	11.25 ± 6.23	10.55 ± 6.01	20.55	0.001
Precaution	19.75 ± 4.33	15.25 ± 3.74	16.70 ± 3.36	27.35 ± 3.22	10.00 ± 2.13	11.90 ± 2.38	84.57	<0.001
Monitoring	10.70 ± 2.66	6.75 ± 1.71	6.70 ± 1.46	13.50 ± 2.24	3.70 ± 1.08	3.35 ± 1.35	97.76	<0.001
Warning	4.35 ± 1.09	4.20 ± 1.40	4.80 ± 2.29	6.60 ± 1.23	1.90 ± 0.72	2.45 ± 0.76	77.11	<0.001
Total	46.70 ± 9.95 [ab]	36.20 ± 5.63 [ac]	42.55 ± 5.55 [a]	61.40 ± 6.62 [b]	26.85 ± 7.35 [c]	28.25 ± 7.30 [c]	86.18	<0.001

Values are presented as mean ± SD ($n = 20$ per scenario). H = Kruskal–Wallis test statistic. Superscript letters indicate Dunn's post hoc comparisons with Bonferroni correction. Means sharing at least one letter are not significantly different, whereas means with no letters in common differ significantly (adjusted $p < 0.05$).

Dunn's post hoc test with Bonferroni correction revealed that S4 differed significantly from S2, S3, S5, and S6 (all $p < 0.05$). S1 and S3 also showed significantly higher total safety sentence counts compared to S5 and S6 (all $p < 0.05$). No significant differences were observed between S2 and the healthy adult cases (S5, S6), nor between S5 and S6 (Table 5). After normalization by output length, the mean total safety-related expression rate ranged from 11.82 to 30.26 per 1000 words across scenarios (Supplementary Table S2); differences remained evident (Kruskal–Wallis H = 93.79, $p < 0.001$). S4 retained the highest mean rate (30.26 per 1000 words), whereas S6 had the lowest (11.82).

In the supplementary 24-output author–AI assessment, agreement was 95.8% for aerobic intensity ($\kappa = 0.903$) and 79.2% for resistance intensity ($\kappa = 0.713$). Agreement was 100% for contraindication and precaution presence (κ undefined because both raters assigned the same category to all outputs), 79.2% for symptom monitoring ($\kappa = 0$), and 87.5% for risk warnings ($\kappa = 0$). Claude classified all 24 outputs as positive for each safety category. In contrast, the author rated symptom monitoring as absent in five outputs and risk warnings as absent in three; these discrepancies should be interpreted cautiously because some author judgments reflected the specificity of the safety advice rather than the mere presence of related statements. Separately, κ for monitoring and warnings was constrained by the absence of variation in AI ratings. No conflicting or unsupported contraindication was flagged in the author's limited 24-output review; this was not a comprehensive clinical safety evaluation. In the five selected outputs initially labeled unclassifiable by Claude, the author independently assigned the same resistance-intensity category (5/5). The author additionally classified four outputs as unclassifiable that Claude had classified into intensity categories.

4. Discussion

This study conducted a pre-deployment reliability assessment of generative clinical decision-support outputs, using LLM-based exercise prescriptions produced by Gemini 2.5 Flash as a concrete test case, evaluated across three complementary dimensions: semantic consistency, FITT structural classification, and safety-content consistency. The results confirmed that full-prescription semantic similarity was high, while variability was observed in certain structural components, indicating that output stability may vary across scenario and prescription components even under identical prompting conditions. This finding has practical implications: the same patient profile may yield structurally different prescriptions across repeated generations, potentially resulting in non-equivalent exercise recommendations in real-world deployment.

Research on repeated generation consistency of LLMs has been conducted primarily in clinical decision support and medical question answering. Carandang et al. [20] reported high semantic consistency in repeatedly generated clinical notes, aligning with the cosine

similarity values observed in the present study (0.939–0.967), and Hanss et al. [21] similarly found high consistency in GPT-4 responses, suggesting that response consistency may serve as a useful reliability indicator. More broadly, recent multi-axial evaluation work has shown that generator models can score well on reference-oriented semantic metrics while still exhibiting dissociations on other reliability axes that those metrics do not capture [1], underscoring that no single metric is sufficient to establish generative reliability. In the exercise prescription domain, Lai et al. [6] called for standardized evaluation frameworks to ensure consistency and reliability across AI models. The present study directly addresses this gap by proposing an evaluation pipeline combining SBERT-based semantic similarity with AI-as-a-Judge FITT classification, extending the methodological scope of LLM-based exercise prescription research by introducing repeated generation consistency as a reliability dimension.

In the full-prescription semantic analysis, mean similarity was descriptively highest in S6 (0.967) and lowest in S1 (0.939). However, the output-level Welch omnibus test for the token-weighted analysis did not reach the conventional significance threshold ($p = 0.064$), and no Holm-adjusted pairwise comparison was significant. Other pooling and inferential calibrations yielded borderline or significant omnibus results, highlighting sensitivity to analytical choices. Consequently, the descriptive ranking should not be interpreted as a statistically robust hierarchy of scenario-specific consistency.

It is also important to note that high semantic similarity does not necessarily imply clinical equivalence. SBERT-based similarity reflects overall semantic structure but may not capture clinically meaningful numerical variations, such as a shift in aerobic intensity from 50% to 70% HRmax. Semantic consistency should therefore be complemented by quantitative evaluation approaches such as the FITT classification applied in this study. The very high full-text cosine similarities may partly reflect recurring program structure and smoothing through chunk aggregation, rather than equality of exercise doses or clinical recommendations.

The FITT classification results indicate that Gemini 2.5 Flash showed prescription patterns broadly consistent with guideline-based intensity ranges according to case characteristics. Clinical cases (S1–S4) were predominantly prescribed low-intensity aerobic (90–100%) and resistance exercise (75–85%), while a clear shift toward moderate-to-high intensity was observed in healthy adult cases (S5–S6), consistent with ACSM recommendations [19].

However, variability was observed in several quantitative components. Resistance exercise intensity showed the greatest instability, with unclassifiable outputs in 10–25% of clinical cases despite explicit prompting for %1RM values, and inconsistent classifications within the same scenario in S6. Frequency variability was also notable in S6, where the most common pattern accounted for only 35% of outputs. Session duration was not consistently stated, with some outputs providing estimated rather than explicit values. These patterns of quantitative instability are consistent with findings by Qiu et al. [22] on LLM performance decline in treatment planning tasks, underscoring that clinically actionable prescription elements require verification beyond semantic-level consistency checks.

The AI-judged 100% inclusion rate suggests that the explicitly requested safety-content categories were consistently identified across repeated generations, although the supplementary author review did not fully reproduce these presence judgments for monitoring and risk warnings. However, the quantity of safety expressions varied significantly by scenario ($H = 86.18$, $p < 0.001$), with S4 generating the highest volume (Mean = 61.40), and precaution and monitoring showing the greatest variability. Dunn's post hoc analysis revealed that S2 did not differ significantly from healthy adult cases, indicating that safety expression volume may reflect clinical complexity rather than clinical status alone. As

emphasized by previous study [23] and further supported by evidence from the exercise prescription domain [8], expert validation remains essential prior to clinical deployment. This pattern persisted when safety-expression counts were normalized by output length; however, both absolute and normalized counts remain indicators of content volume rather than clinical appropriateness.

Although the LLM-as-a-Judge paradigm is emerging as a scalable and cost-effective assessment alternative, careful design and standardization are required to ensure reliability [24]. The present study constitutes an early systematic attempt to apply this methodology within a concrete clinical decision-support task, and the resulting evaluation pipeline is intended to generalize to other generative clinical decision-support systems where repeated-generation reliability, not only single-output accuracy, determines deployment readiness. Several limitations should be acknowledged. First, the present study characterized within-model repeated-generation variability using a single model (Gemini 2.5 Flash) under a stochastic generation setting (temperature = 1.0), whereas our subsequent cross-model study examined model-specific reproducibility across three LLMs under standardized temperature-0 conditions [25]. Notably, variability in resistance-exercise intensity specification remained evident in that subsequent study, suggesting that this issue may not be explained solely by decoding stochasticity. Second, safety evaluation was limited to the presence and quantity of safety-related expressions, which may reflect verbosity or repetition rather than clinical appropriateness; potentially conflicting or unsupported contraindications were not systematically assessed. Third, although a blinded single-author comparison was performed for 24 purposively selected outputs, independent external expert validation and inter-human-rater reliability were not assessed. Fourth, the six scenarios may not fully represent the diversity of clinical practice. Fifth, the choice of chunk boundaries and pooling weights in document-level embeddings may influence semantic similarity estimates; scenario-level statistical inference was sensitive to test calibration, and cosine similarity is not a validated measure of clinical equivalence.

5. Conclusions

This study evaluated the repeated generation consistency of LLM-generated exercise prescriptions across three dimensions: semantic similarity, FITT structural classification, and safety-content consistency. Full-prescription SBERT analysis produced high mean semantic similarity across all six scenarios, but did not establish a robust ordering of scenarios or significant pairwise differences after multiplicity correction. Variability in key quantitative components, particularly exercise intensity, remained evident despite high semantic similarity. These findings indicate that use of generative models for exercise prescription requires structured decision-making processes and quantitative control mechanisms, particularly when outputs are intended to support clinical or health-related decisions.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/healthcare14234070/s1: Supplementary Material S1: Exercise Prescription Generation Prompt; Supplementary Table S1: Detailed Clinical Characteristics of the Six Scenarios Used for Repeated Exercise Prescription Generation; Supplementary Material S2: Evaluation Prompts and Classification Criteria; Supplementary Table S2: Output-length-normalized safety-related sentence/clause counts; Supplementary Data S1: Complete raw model-generated exercise prescriptions (120 outputs; CSV format).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve human participants, human data, or biological material; no animal subjects were used.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full prompt text, case scenario descriptions, and classification criteria used in this study are provided in the Supplementary Materials. The 120 raw model-generated exercise prescriptions are provided in Supplementary Data S1. Additional information is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this work, the author used Claude Sonnet 4.6 (Anthropic) to assist with translation and language refinement, and GPT-4o (OpenAI) to review the manuscript for language clarity and consistency. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Conflicts of Interest: The author is the founder and representative of KINAI Co., Ltd., a company engaged in the development of AI-based exercise and health-related technologies. The company did not provide funding or other support for this study. The author declares no other competing interests.

References

1. Byun, H.; Lee, D.; Jung, M.; Jang, B. Multi-axial analysis of clinical reasoning in large language models: Inter-verifier disagreement and its implications for automated evaluation. *J. Med. Syst.* **2026**, *50*, 117. [CrossRef] [PubMed]
2. Enichen, E.J.; Young, C.C.; Frates, E.P. The potential of AI to create personalized exercise plans. *Health Promot. Pract.* **2025**, *27*, 1063–1067. [CrossRef] [PubMed]
3. Washif, J.; Pagaduan, J.; James, C.; Dergaa, I.; Beaven, C. Artificial intelligence in sport: Exploring the potential of using ChatGPT in resistance training prescription. *Biol. Sport* **2024**, *41*, 209–220. [CrossRef] [PubMed]
4. 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.* **2024**, *23*, 56–72. [CrossRef] [PubMed]
5. Akrimi, S.; Schwensfeier, L.; Düking, P.; Kreutz, T.; Brinkmann, C. ChatGPT-4o-generated exercise plans for patients with type 2 diabetes mellitus: Assessment of their safety and other quality criteria by coaching experts. *Sports* **2025**, *13*, 92. [CrossRef] [PubMed]
6. Lai, X.; Chen, J.; Lai, Y.; Huang, S.; Cai, Y.; Sun, Z.; Wang, X.; Pan, K.; Gao, Q.; Huang, C. Using large language models to enhance exercise recommendations and physical activity in clinical and healthy populations: Scoping review. *JMIR Med. Inform.* **2025**, *13*, e59309. [CrossRef] [PubMed]
7. van der Leeden, M.; Stuiver, M.M.; Huijsmans, R.; Geleijn, E.; de Rooij, M.; Dekker, J. Structured clinical reasoning for exercise prescription in patients with comorbidity. *Disabil. Rehabil.* **2020**, *42*, 1474–1479. [CrossRef] [PubMed]
8. Choi, M.; Park, J.; Lee, M.; Beom, J.; Jung, S.Y.; Lee, K. AI-generated exercise prescriptions for at-risk populations: Safety and feasibility of a large language model assessed by expert evaluation. *J. Clin. Med.* **2026**, *15*, 2457. [CrossRef] [PubMed]
9. Shyr, C.; Ren, B.; Hsu, C.-Y.; Yan, C.; Tinker, R.J.; Cassini, T.A.; Hamid, R.; Wright, A. A statistical framework for evaluating the repeatability and reproducibility of large language models. *Journal of the American Medical Informatics Association* **2026**. [CrossRef] [PubMed]
10. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623. [CrossRef]
11. Croxford, E.; Gao, Y.; First, E.; Pellegrino, N.; Schnier, M.; Caskey, J.; Oguss, M.; Wills, G.; Chen, G.; Dligach, D.; et al. Evaluating clinical AI summaries with large language models as judges. *npj Digit. Med.* **2025**, *8*, 640. [CrossRef] [PubMed]
12. Wataoka, K.; Takahashi, T.; Ri, R. Self-preference bias in LLM-as-a-judge. *arXiv* **2024**, arXiv:2410.21819.
13. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
14. Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; Liu, Y. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv* **2024**, arXiv:2412.05579.
15. Garber, C.E.; Blissmer, B.; Deschenes, M.R.; Franklin, B.A.; Lamonte, M.J.; Lee, I.-M.; Nieman, D.C.; Swain, D.P. American College of Sports Medicine position stand: Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: Guidance for prescribing exercise. *Med. Sci. Sports Exerc.* **2011**, *43*, 1334–1359. [CrossRef] [PubMed]
16. Bishop, D.J.; Beck, B.; Biddle, S.J.H.; Denay, K.L.; Ferri, A.; Gibala, M.J.; Headley, S.; Jones, A.M.; Jung, M.; Lee, M.J.-C.; et al. Physical activity and exercise intensity terminology: A joint ACSM expert statement and ESSA consensus statement. *Med. Sci. Sports Exerc.* **2025**, *57*, 2599–2613. [CrossRef] [PubMed]

17. Schoenfeld, B.J.; Grgic, J.; Van Every, D.W.; Plotkin, D.L. Loading recommendations for muscle strength, hypertrophy, and local endurance: A re-examination of the repetition continuum. *Sports* **2021**, *9*, 32. [CrossRef] [PubMed]
18. Currier, B.S.; D'Souza, A.C.; Fiatarone Singh, M.A.; Lowisz, C.V.; Rawson, E.S.; Schoenfeld, B.J.; Smith-Ryan, A.E.; Steen, J.P.; Thomas, G.A.; Triplett, N.T.; et al. American College of Sports Medicine position stand: Resistance training prescription for muscle function, hypertrophy, and physical performance in healthy adults: An overview of reviews. *Med. Sci. Sports Exerc.* **2026**, *58*, 851–872. [CrossRef] [PubMed]
19. American College of Sports Medicine. *ACSM's Guidelines for Exercise Testing and Prescription*, 12th ed.; Wolters Kluwer: Philadelphia, PA, USA, 2024.
20. Carandang, K.A.M.; Arana, J.M.; Casin, E.R.; Monterola, C.; Tan, D.S.; Valenzuela, J.F.B.; Alis, C. Are LLMs reliable? An exploration of the reliability of large language models in clinical note generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), Vienna, Austria, 27 July–1 August 2025; pp. 1413–1422. [CrossRef]
21. Hanss, K.; Sarma, K.V.; Glowinski, A.L.; Krystal, A.; Saunders, R.; Halls, A.; Gorrell, S.; Reilly, E. Assessing the accuracy and reliability of large language models in psychiatry using standardized multiple-choice questions: Cross-sectional study. *J. Med. Internet Res.* **2025**, *27*, e69910. [CrossRef] [PubMed]
22. Qiu, P.; Wu, C.; Liu, S.; Fan, Y.; Zhao, W.; Chen, Z.; Gu, H.; Peng, C.; Zhang, Y.; Wang, Y.; et al. Quantifying the reasoning abilities of LLMs on clinical cases. *Nat. Commun.* **2025**, *16*, 9799. [CrossRef] [PubMed]
23. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef] [PubMed]
24. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on LLM-as-a-judge. *Innovation* **2026**, *7*, 101253. [CrossRef]
25. Lee, K. Cross-model consistency of AI-generated exercise prescriptions: A repeated generation study across three large language models. *Int. J. Med. Inform.* **2026**, *220*, 106594. [CrossRef] [PubMed]

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.