Rehabilitation & Recreation
Volume 20 No.2 (2026)

https://doi.org/10.67034/2786-8354.2026.20.2.21

COMPARATIVE EVALUATION OF AI-GENERATED AND COACH-DESIGNED TRAINING PROGRAMS IN YOUTH KICKBOXING

ПОРІВНЯЛЬНА ОЦІНКА ТРЕНУВАЛЬНИХ ПРОГРАМ У ЮНАЦЬКОМУ КІКБОКСИНГУ, ЗГЕНЕРОВАНИХ ШТУЧНИМ ІНТЕЛЕКТОМ ТА РОЗРОБЛЕНИХ ТРЕНЕРАМИ

Volsky D.[1], Pavlenko O.[1], Biletska V.[2]
[1]National Federation of Kickboxing of Ukraine WAKO, Kyiv, Ukraine
[2]Borys Grinchenko Kyiv Metropolitan University, Ukraine
ORCID: 0000-0003-2731-5611
ORCID: 0009-0008-1886-9069
ORCID: 0000-0002-8813-1747

Abstracts

Introduction. The aim of this study was to conduct an expert evaluation and comparative analysis of kickboxing training programmes developed by coaches of different qualification levels and by an artificial intelligence language model (ChatGPT).

Materials and Methods. A blinded repeated-measures design was applied. Sixteen independent experts in striking combat sports evaluated four 8-week training programmes for beginner kickboxers aged 10–11 years: three coach-designed programmes (TP-L1 – expert, TP-L2 – intermediate, TP-L3 – novice) and one AI-generated programme (TP-AI). Programme quality was assessed using a 12-item questionnaire covering three domains: structure, combat sports specificity, and practicality. All items were rated on a 5-point Likert scale. Data were analysed using non-parametric statistics (Friedman test, Wilcoxon signed-rank test with Bonferroni correction, Kendall's W).

Results. Significant differences were found between programmes ($\chi^2(3) = 32.30$, $p < 0.01$) with a strong level of inter-expert agreement ($W = 0.67$). The highest overall ratings were observed for TP-L1 (4.08 ± 0.32), followed by TP-L2 (3.50 ± 0.42) and TP-AI (3.32 ± 0.48), while TP-L3 received the lowest scores (3.01 ± 0.48). Pairwise comparisons revealed significant differences between TP-L1 and all other programmes ($p < 0.05$–0.01), as well as between TP-L2 and TP-L3 ($p < 0.05$). No significant differences were found between TP-L2 and TP-AI ($p = 0.18$) or TP-L3 and TP-AI ($p = 0.63$). Effect sizes were large for all significant comparisons ($r = 0.67$–0.99). Domain-specific analysis showed that TP-L1 achieved the highest scores across all domains, while TP-AI demonstrated relatively high structural quality (3.90 ± 0.57) but lower combat sports specificity (2.65 ± 0.64). Item-level analysis confirmed that the AI-generated programme performed well in general structure (Q1–Q4) but showed weaker results in sport-specific components (Q6–Q8).

Conclusions. AI-generated training programmes demonstrate acceptable methodological quality and may reach a level comparable to those developed by less experienced coaches. However, expert-designed programmes remain superior, particularly in sport-specific integration. Artificial intelligence can be considered a supportive tool in training design, but professional coaching expertise remains essential in combat sports.

Keywords: artificial intelligence, kickboxing, combat sports, coaching expertise, training programme, expert evaluation.

Мета. Метою дослідження було проведення експертної оцінки та порівняльного аналізу тренувальних програм з кікбоксингу, розроблених тренерами різного рівня кваліфікації та згенерованих моделлю штучного інтелекту ChatGPT.

Матеріали та методи. Дослідження виконано за сліпим дизайном з повторними вимірюваннями. У ньому взяли участь 16 незалежних експертів у галузі ударних єдиноборств. Оцінювалися чотири 8-тижневі програми для початківців 10–11 років: три тренерські (TP-L1 – експерт, TP-L2 – середній рівень, TP-L3 – початковий рівень) та одна згенерована ШІ (TP-AI). Якість програм оцінювали за 12-критеріальною

Received: 11 April 2026. Accepted: 29 June 2026. Available online: 30 July 2026

© 2026. Volsky D., Pavlenko O., Biletska V.
Rehabilitation and Recreation. ISSN 2786-8354 (Online) | ISSN 2786-8346 (Print)
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

анкетою, що охоплювала три блоки: структура, специфічність виду спорту та практичність. Оцінювання здійснювалось за 5-бальною шкалою. Для аналізу застосовано непараметричні методи (критерій Фрідмана, критерій Вілкоксона з поправкою Бонферроні, коефіцієнт конкордації Кендалла).

Результати. Виявлено статистично значущі відмінності між програмами ($p < 0.01$) при високому рівні узгодженості експертів ($W = 0.67$). Найвищі оцінки отримала програма TP-L1 (4.08 ± 0.32), далі TP-L2 (3.50 ± 0.42) та TP-AI (3.32 ± 0.48), найнижчі – TP-L3 (3.01 ± 0.48). Парні порівняння показали значущі відмінності між TP-L1 і всіма іншими програмами ($p < 0.05$–0.01), а також між TP-L2 і TP-L3 ($p < 0.05$). Водночас між TP-L2 і TP-AI ($p = 0.18$) та TP-L3 і TP-AI ($p = 0.63$) статистично значущих відмінностей не виявлено. Розміри ефекту для значущих порівнянь були великими ($r = 0.67$–0.99). Аналіз за доменами показав, що TP-L1 має найвищі значення у всіх блоках, тоді як TP-AI демонструє відносно високі показники структури (3.90 ± 0.57), але нижчі – специфічні відповідно до виду спорту (2.65 ± 0.64). Порівняльний аналіз підтвердив сильні сторони ШІ у структурі та слабші – у специфічних компонентах.

Висновки. Програми, згенеровані штучним інтелектом, демонструють прийнятну методичну якість і можуть відповідати рівню програм менш досвідчених тренерів. Водночас програми, розроблені експертами, залишаються кращими, особливо щодо спортивної специфіки. Штучний інтелект доцільно розглядати як допоміжний інструмент, але не як заміна тренера.

Ключові слова: штучний інтелект, кікбоксинг, єдиноборства, тренерська діяльність, тренувальні програми, експертна оцінка.

1. Introduction

The rapid development of artificial intelligence (AI) technologies has led to their widespread implementation their widespread implementation across various domains of sport, including performance analysis, monitoring of athletes' physical condition, optimisation of the training process, and decision-making support during both preparation and competition. AI enables the analysis of large volumes of data for injury prediction, evaluation of training effectiveness, and personalisation of training loads, which may enhance the quality of athletic preparation and the effectiveness of training planning [14, 18, 31]. However, alongside the numerous advantages associated with the implementation of AI technologies in sport, there are also challenges related to the diversity of approaches, issues of model standardisation, and ethical concerns regarding the processing of athletes' sensitive data [16].

In sports training, the design of training programmes is based on the principles of periodisation, progressive overload, individualisation, and sport specificity, which have been established through decades of research in training theory. Contemporary technologies, including digital platforms, allow for more precise planning of training volume and intensity, monitoring of recovery, and programme adjustments [4]. At the same time, the quality of a training programme is determined not only by the use of technology, but primarily by its methodological coherence: a clear structure, logical progression, alignment with age-related and individual characteristics, and a balanced integration of physical, technical, tactical, and psychological preparation [28]. These requirements are particularly significant in combat sports, where competitive performance is highly variable, technical and tactical decision-making plays a substantial role, and direct interaction with an opponent is inherent to the activity [5, 9]. In addition, combat sports require the integration of high-intensity intermittent efforts, anaerobic power, rapid cognitive responses, and precise management strategies, which together increase the complexity of programme design. Kickboxing, while characterised by its own rule-specific technical and tactical features, shares fundamental structural, physiological, and performance demands with other striking combat sports [1, 22].

One of the contemporary directions in the application of artificial intelligence involves the use of large language models (LLMs) for the generation of training programmes and recommendations within the field of sport. Such models are capable of processing natural language based on task descriptions, thereby opening new opportunities for the automated creation of diverse training plans and advisory content [11]. Nevertheless, AI-generated programmes may appear structurally coherent while lacking contextual sensitivity, microcycle-level adaptation, or deep sport-specific integration. Furthermore, concerns have been raised regarding the potential for overly generic recommendations or unverified

content when prompts are insufficiently constrained. Existing studies demonstrate a variety of domains in which large language models are already being applied. Although research directly focused on the generation of training programmes for athletes remains limited, some pilot projects have been conducted [23], and users of AI-generated plans may evaluate them as useful and convenient. Furthermore, there are studies addressing specific aspects of this direction. For example, the use of artificial intelligence for exercise prescription in personalised health contexts has been explored [10], and the LLM-based PlanFitting model [25] demonstrated that AI can generate personalised physical activity plans. In addition, there are examples comparing different levels of AI application in programme design, including comparisons of training programmes for footballers generated by various AI models [30], as well as evaluations of selected components of AI-generated training programmes [13]. Collectively, these findings indicate the technological potential of LLMs in training design, while simultaneously highlighting the need for further systematic and methodologically rigorous research in this area. However, the unresolved issue is not whether AI systems are capable of generating training plans, but whether the methodological quality, sport specificity, and structural coherence of such plans are comparable to those developed by qualified coaches [29].

The evaluation and design of training programmes have traditionally been regarded as complex professional processes combining scientific knowledge with practical coaching expertise. Research in the field of coaching science emphasises that effective planning is grounded in the integration of knowledge, skills, and an in-depth understanding of sport-specific demands [8], while coaching expertise develops through a combination of formal education, experience, and applied practice [12, 21]. A coach's qualification level determines their ability to structure a programme appropriately, construct a justified progression of training loads, account for athletes' individual characteristics, and consider the specific features of competitive performance. Despite the rapid advancement of digital technologies, empirical studies directly comparing the quality of AI-generated training programmes in combat sports with those developed by coaches of different qualification levels remain extremely scarce. Therefore, an objective comparison of the quality of AI-generated and coach-designed training programmes within a specific sport is highly relevant. Accordingly, the aim of the present study was to conduct – to conduct an expert evaluation and comparative analysis of kickboxing training programmes developed by coaches of varying qualification levels and by the ChatGPT language model..

2. Materials and Methods

Study Design and Participants

The study was conducted as a blinded comparative expert evaluation study using a repeated-measures design. The aim was to compare the quality of a training programme generated using artificial intelligence with programmes developed by coaches of different qualification levels. In designing the study protocol for evaluating training plans generated by ChatGPT, we referred to and built upon previous research [2, 15, 29], as well as methodological examples from other domains, including medicine [20].

Sixteen independent experts representing striking-based combat sports (kickboxing, boxing, Muay Thai, taekwondo) and sports training participated in the study. The inclusion criteria were: higher education in physical culture and sport, at least ten years of professional experience, experience in preparing athletes at national level, and no involvement in the development of any of the evaluated programmes. Each of the sixteen experts independently assessed all four training programmes, which enabled the application of a repeated-measures design and ensured comparability of evaluations across programmes.

To assess the quality of the training programmes, a questionnaire for evaluating training plan quality was employed. The instrument was developed based on a previously validated and informative study [11] and further adapted to the context of combat sports [5, 28]. The questionnaire consisted of 12 items grouped into three conceptual blocks. The first block, Overall Structure and Methodological Principles, evaluated the overall organisation (example of question Q1: Rate the overall structure of the program), programme logic, structural coherence, and consideration of athletes' preparedness. The second block, Combat Sports Specificity, focused on the extent to which the programme reflected the

specific demands of kickboxing training. The third block, Practicality and Applicability, assessed the feasibility of implementing the programme under typical training conditions.

All items were rated using a 5-point Likert scale, where 1 corresponded to a very low level and 5 to a very high level. For each programme, the overall mean score across all items, mean scores for each block, and an integral quality index reflecting the aggregated evaluation across all criteria were calculated.

The study did not involve the direct participation of athletes, the collection of personal or medical data, or any intervention in the training process. All procedures were limited to the analysis of written training programmes and expert evaluation. The experts participated voluntarily, were informed about the aim of the study, and provided their consent. The anonymity of all participants was guaranteed, and the data were used exclusively in aggregated form for scientific analysis. The study received approval from the local ethics committee of the relevant university department in accordance with international guidelines for research not involving vulnerable populations or personal data.

Development of Training Programmes

All training programmes, both coach-designed and AI-generated, were developed based on a single standardised written prompt. Coaches were instructed to create the programme relying exclusively on the written description, without receiving any additional oral or written clarification. The prompt was as follows: «Develop an 8-week training programme for a group of beginner kickboxers aged 10–11 years. Participants have no medical contraindications for sport participation and have passed a medical examination. The group includes several weight categories. The athletes have no prior experience in kickboxing or have up to one year of training experience, but may have experience in other sports, including combat sports. The programme should last eight weeks and include general physical development, acquisition of basic technical elements (stance, footwork, straight punches and kicks), and development of coordination abilities. The first week should focus on familiarisation and adaptation to the specific demands of the sport. Training sessions are conducted three times per week and must include a warm-up, main part, and cool-down. The programme should combine general and specific physical preparation, consider safety requirements, allow the use of basic equipment (punching bag, focus mitts, basic protective gear), and clearly describe the logic of load progression throughout the entire period».

One programme was generated using the artificial intelligence language model ChatGPT (OpenAI, based on the GPT-4 architecture, November 14, 2025) based on the detailed structured prompt containing all the above information. Three additional programmes were independently prepared by kickboxing coaches who did not communicate with each other during the development process. The programmes were labelled as TP-L1 (Coach Level 1 – Expert), TP-L2 (Coach Level 2 – Intermediate), TP-L3 (Coach Level 3 – Novice), and TP-AI (AI-Generated Training Programme).

Coach classification was performed according to a Multidimensional Coach Qualification Framework grounded in contemporary concepts of coaching expertise [8, 12, 21]. The main criteria included coaching experience duration, the competitive level of athletes coached, possession of an official coaching licence, and experience in competitions of different levels. Coach Level 3 had ≤3 years of experience, worked primarily with beginners, had not prepared athletes at national level, and held basic certification. Coach Level 2 had 4–10 years of experience, worked with regional or national-level athletes, possessed an official federation licence, and had experience preparing athletes for national competitions. Coach Level 1 had ≥10 years of experience, prepared athletes at national or international level, held the highest category of coaching licence, and had experience participating in international tournaments as a coach. To minimise potential bias, all programmes were standardised into a unified structural format, stripped of authorship identifiers, and presented to experts in randomised order. Experts were blinded to the origin of each programme (coach-designed or AI-generated), ensuring unbiased evaluation.

Statistical Analysis

All data were processed and analysed using the Python programming language (version 3.12.7) with appropriate statistical and data visualisation libraries. Given the ordinal nature of Likert-scale data and the repeated-measures design, non-parametric statistical methods were applied

throughout the analysis. For each training programme, an integral quality indicator was calculated as the mean score across all 12 evaluation criteria for each expert. Descriptive statistics included mean and standard deviation (Mean ± SD), median and interquartile range (Median [Q1–Q3]), range (minimum–maximum), and 95% confidence intervals (95% CI). Inter-rater agreement among experts was assessed using Kendall's coefficient of concordance (Kendall's W).

To compare the overall evaluations of the four training programmes, a non-parametric repeated-measures analysis of variance (Friedman test) was applied. When significant differences were identified, post hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction. Effect size for pairwise comparisons was calculated as $r = Z/\sqrt{N}$ and interpreted as small (<0.3), medium (≥0.3 and <0.5), and large (≥0.5). Statistical significance was set at $p < 0.05$ [27].

3. Results and discussion

Sixteen independent experts representing striking-based combat sports in the study. A repeated-measures design was employed, in which each expert evaluated all four training programmes across 12 criteria using a 5-point Likert scale. The distribution of overall expert evaluations for the four training programmes is presented in Table 1 and Figure 1.

Table 1

Descriptive statistics of integral quality scores for the evaluated training programmes

Program	Mean ± SD	Median [Q1–Q3]	Range (Min–Max)	95% CI
TP-L1	4.08 ± 0.32	4.04 [3.90–4.33]	3.58–4.67	3.91–4.25
TP-L2	3.50 ± 0.42	3.46 [3.23–3.75]	3.00–4.50	3.27–3.73
TP-L3	3.01 ± 0.48	3.00 [2.65–3.29]	2.25–4.00	2.75–3.26
TP-AI	3.32 ± 0.48	3.38 [2.98–3.54]	2.67–4.08	3.07–3.58

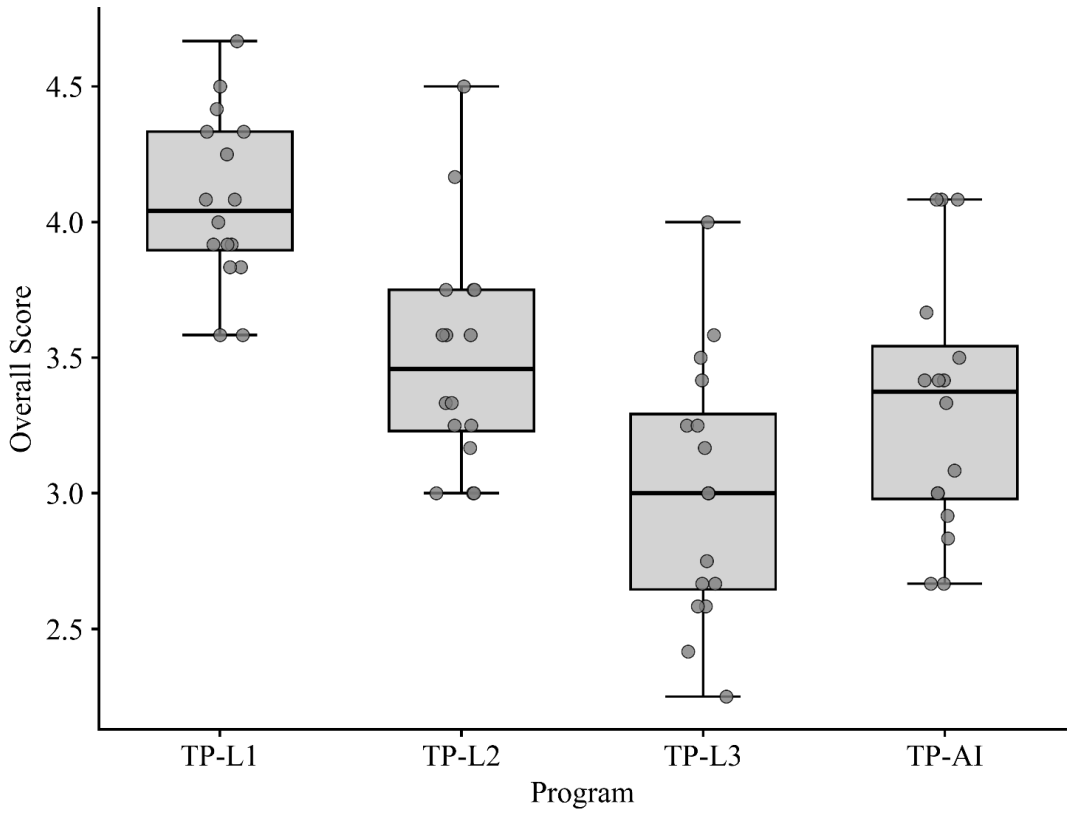

Figure 1. Distribution of overall expert evaluations of the training programmes (boxplot with individual data points)

Table 2

Pairwise comparisons of training programmes based on expert evaluations (Wilcoxon signed-rank test with Bonferroni correction)

Comparison of programmes		**Z**	**p (corrected)**	**Effect size (r)**
TP-L1	TP-L2	-3.41	<0.05*	0.85
TP-L1	TP-L3	-3.98	<0.01*	0.99
TP-L1	TP-AI	-3.77	<0.01*	0.94
TP-L2	TP-L3	-2.67	<0.05*	0.67
TP-L2	TP-AI	-2.16	0.18	0.54
TP-L3	TP-AI	-1.62	0.63	0.41

The highest values of the integral quality indicator were observed for programme TP-L1 (Mean = 4.08 ± 0.32), followed by TP-L2 and TP-AI, while the lowest scores were recorded for TP-L3 (Mean = 3.01 ± 0.48). Median values and interquartile ranges demonstrated a similar pattern, with TP-L1 showing the highest central tendency (Median = 4.04 [3.90–4.33]) and TP-L3 the lowest (Median = 3.00 [2.65–3.29]).

The observed ranges of scores were relatively narrow for TP-L1 (3.58–4.67), indicating a high level of consistency in expert evaluations, whereas TP-AI (2.67–4.08) and TP-L3 (2.25–4.00) demonstrated greater variability. The 95% confidence intervals further confirmed the stability of the highest-rated programme (TP-L1: 3.91–4.25). At the same time, the overlap between the confidence intervals of TP-L2, TP-AI, and TP-L3 suggests more comparable evaluations among these programmes, in contrast to the consistently higher scores TP-L1.

The results of Kendall's coefficient of concordance indicated a strong level of agreement among experts across all training programmes ($W = 0.67$). A non-parametric analysis of variance for related samples (Friedman test) revealed statistically significant differences in the overall evaluations of the four training programmes ($\chi^2(3) = 32.30$, $p < 0.01$). These findings indicate that the experts consistently differentiated between the programmes in terms of their overall quality, supporting the validity of subsequent pairwise comparisons.

Pairwise comparisons between programmes were conducted using the Wilcoxon signed-rank test with Bonferroni correction (Table 2). Statistically significant differences were identified between TP-L1 and all other programmes, including TP-L2 ($p < 0.05$), TP-L3 ($p < 0.01$), and TP-AI ($p < 0.01$). In addition, a statistically significant difference was observed between TP-L2 and TP-L3 ($p < 0.05$). No statistically significant differences were found between TP-L2 and TP-AI ($p = 0.18$) or between TP-L3 and TP-AI ($p = 0.63$).

The effect size analysis demonstrated large effects for all statistically significant comparisons ($r = 0.67–0.99$), indicating substantial differences in expert evaluations between programmes. Moderate effect sizes were observed for non-significant comparisons, suggesting the presence of practical differences.

The item-level analysis (Table 3) provided a more detailed insight into the differences between programmes across individual evaluation criteria.

TP-L1 consistently achieved the highest scores across most items, particularly those related to programme structure (Q1–Q4) and practical implementation (Q11–Q12). In contrast, TP-L3 demonstrated lower evaluations across the majority of items, confirming its overall weaker performance.

The AI-generated programme (TP-AI) showed relatively high scores in structurally oriented items (e.g., Q1, Q2, Q4), but substantially lower values in items related to combat-specific content (Q6–Q8) and tactical applicability, aligning with the domain-level findings. TP-L2 generally occupied an intermediate position across items, without pronounced strengths or weaknesses.

To summarize the domains, Figure 2 was constructed. It presents the comparison of the training programmes across the three aggregated evaluation domains: Structure, Sports Specificity and Practicality.

Table 3

Item-level comparison of training programmes based on expert evaluations (Mean ± SD)

Question	TP-L1	TP-L2	TP-L3	TP-AI
Q1	4.25 ± 1.00	3.44 ± 0.89	2.75 ± 0.77	4.12 ± 0.72
Q2	4.00 ± 0.82	3.12 ± 0.72	2.75 ± 1.06	4.12 ± 0.81
Q3	3.38 ± 0.96	3.44 ± 1.03	2.69 ± 1.25	3.69 ± 1.14
Q4	4.38 ± 0.72	3.00 ± 1.03	3.00 ± 1.21	4.12 ± 0.89
Q5	4.06 ± 1.06	3.38 ± 0.72	3.19 ± 1.38	3.44 ± 1.36
Q6	4.44 ± 0.96	3.81 ± 0.91	3.50 ± 1.03	3.00 ± 1.26
Q7	3.69 ± 1.14	3.44 ± 0.96	3.44 ± 1.15	3.06 ± 1.24
Q8	4.25 ± 0.77	3.94 ± 1.06	3.62 ± 0.96	2.06 ± 0.77
Q9	3.88 ± 0.89	3.56 ± 1.15	3.00 ± 0.89	2.44 ± 1.31
Q10	3.88 ± 1.02	3.38 ± 1.15	2.69 ± 0.87	2.69 ± 0.60
Q11	4.25 ± 0.77	3.81 ± 0.54	2.38 ± 0.62	3.56 ± 0.73
Q12	4.50 ± 0.52	3.69 ± 0.60	3.06 ± 0.68	3.56 ± 0.73

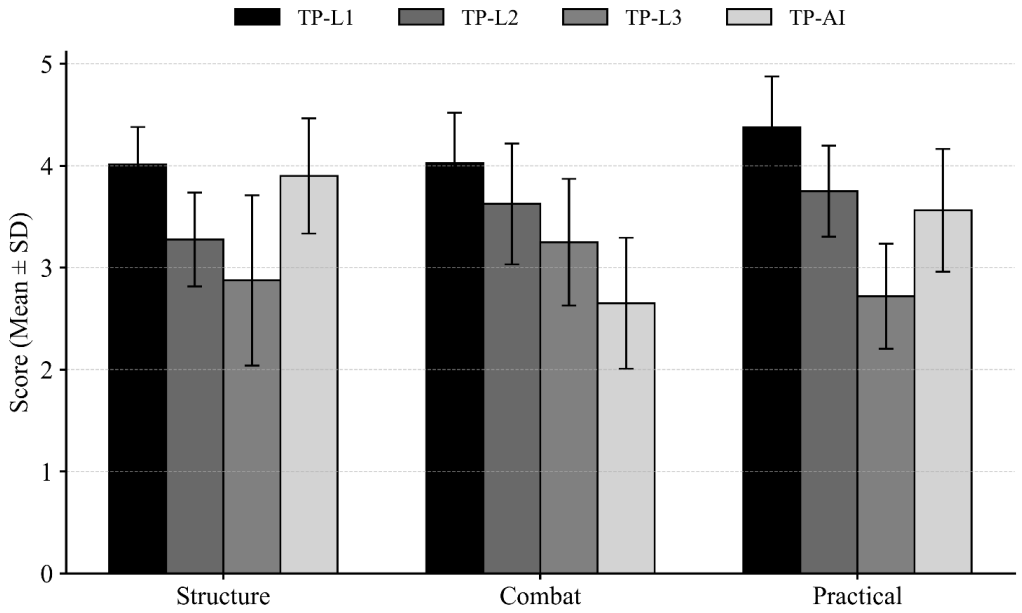

Figure 2. Domain-level comparison of training programme quality based on expert evaluations

The highest scores across all domains were consistently observed for TP-L1, confirming its leading position in overall programme quality. TP-L2 demonstrated moderate and relatively balanced evaluations across domains (3.28–3.75), whereas TP-L3 showed the lowest scores overall, particularly in Structure (2.76 ± 0.76) and Practicality (2.72 ± 0.52). The AI-generated programme (TP-AI) exhibited a differentiated profile, with a relatively high score in Structure (3.90 ± 0.57) and Practicality (3.56 ± 0.60), but substantially lower values in Combat Sports Specificity (2.65 ± 0.64).

Discussion. In contemporary sports science and practice, the generation of training programmes using artificial intelligence, particularly large language models, is becoming increasingly widespread and is attracting growing attention from both researchers and practitioners. The rapid development of AI tools enables the creation of digital training plans that may potentially support and complement traditional coaching practice, especially in the context of self-directed training among athletes and recreational participants, where direct supervision by a highly qualified coach is absent [16, 24].

In the present study, the comparison of overall evaluations of four training programmes revealed statistically significant differences, as confirmed by the Friedman test ($p<0.05$) and Kendall's

coefficient of concordance (W=0.67), indicating agreement among the sixteen experts. The mean ranks demonstrated the highest position for TP-L1, moderate values for TP-L2, lower values for TP-AI, and the lowest for TP-L3. Pairwise analysis identified statistically significant differences between TP-L1 and each of the other programmes ($p<0.05$), whereas differences between TP-L2 and TP-AI, as well as between TP-L3 and TP-AI, were not statistically significant. Notably, a significant difference was observed between TP-L2 and TP-L3. In addition, the large effect sizes observed for all statistically significant comparisons ($r = 0.67–0.99$) indicate that the detected differences are not only statistically significant but also practically meaningful. These findings suggest that more experienced coaches develop training programmes that receive higher overall expert evaluations, whereas the other programmes, including the AI-generated version, are positioned at a comparable level in terms of overall quality. Importantly, the absence of significant differences between TP-AI and both TP-L2 and TP-L3 suggests that AI-generated programmes may reach a level comparable to those developed by less experienced coaches.

The obtained results are consistent with previous research emphasising that expert coaches are characterised by a more systematic approach, a deeper understanding of sport-specific demands, and a greater ability to integrate various components of preparation into a coherent training programme [17, 21]. Studies on the development of coaching expertise also indicate that many years of experience and work with high-level athletes are associated with higher-quality decision-making in structuring the training process [19].

The comparative analysis of domains demonstrated that programmes developed by expert coaches were distinguished not only by higher overall scores but also by superior ratings in structure, combat sports specificity, and practicality. In contrast, the ChatGPT-generated programme (TP-AI) achieved a relatively high score in overall structure but lower values in the domain of combat sports specificity. This finding aligns with evidence in the literature indicating that AI-generated training plans often adhere to fundamental training principles, yet may lack individualisation or sport-specific precision without additional refinement and optimisation by domain experts. From a practical perspective, the findings suggest that artificial intelligence can be effectively used as a supportive tool in the development of training programmes, particularly for novice coaches or in situations where access to expert guidance is limited. However, the results also highlight the necessity of expert supervision to ensure adequate sport-specific adaptation and contextual relevance of training content [6, 24, 29].

Similar patterns have been reported in review studies examining other applications of AI in sport, which suggest that artificial intelligence systems demonstrate strong capabilities in structured planning in accordance with general principles of periodisation and load management. However, their specificity in relation to the tactical and technical demands of a particular sport largely depends on the quality of input parameters and the level of domain expertise embedded in the process [11, 26]. Research on AI applications in sport further emphasises that the integration of coaching oversight substantially enhances the relevance and individualisation of recommendations, particularly in complex coordination-based and contact sports [14, 30].

These conclusions are partially consistent with research in the field of generative AI, which demonstrates that experts across various professional domains are not always able to reliably distinguish AI-generated content from human-produced material, particularly when dealing with structured texts or recommendation-based documents [3]. This supports the assumption that, without additional expert adaptation, AI models may remain at the level of general methodological correctness rather than achieving full sport-specific precision [7].

However, it should be emphasised that our findings partially indicate that the quality of plans generated by large language models has already reached a level at which experts occasionally experience difficulty distinguishing between programmes created by humans and those generated by AI, which is consistent with a number of recent studies [29]. At the same time, according to several reviews, AI-generated training programmes may demonstrate limitations in terms of sport-specific specialisation. These conclusions are also partly aligned with research in the field of generative AI, which suggests that experts across various professional domains are not always able to reliably distinguish AI-generated content from human-produced material, particularly when

dealing with structured texts or recommendation-based outputs [6, 24].

At the same time, this study has several limitations that should be considered when interpreting the results. First, the sample included only sixteen experts, which limits the overall generalisability of the findings. Furthermore, the study relied on expert evaluation rather than objective performance outcomes, which may limit the ability to generalise the findings to real-world training effectiveness. Second, the evaluation concerned only one sport (kickboxing); therefore, the conclusions may be specific to this discipline. Third, although the evaluation criteria and questionnaire were applied in previous research and have been validated, they capture only certain aspects of programme quality and do not encompass other potentially relevant dimensions. These limitations provide a basis for further research, particularly regarding the integration of AI-generated solutions into the development of personalised training programmes, as well as hybrid approaches in which AI tools support, but do not fully replace, the coach's professional expertise.

4. Conclusions

The results of the study demonstrated statistically significant differences in the quality of training programmes developed by coaches of different qualification levels and by an AI-based model. The programme designed by the expert coach (TP-L1) consistently received the highest evaluations across all criteria and domains, confirming the decisive role of coaching experience in ensuring methodological coherence and sport-specific integration.

The AI-generated programme showed relatively high scores in structural and organisational aspects, but lower evaluations in combat sports specificity, indicating limitations in adapting content to the technical and tactical demands of kickboxing. At the same time, the absence of significant differences between the AI-generated programme and those developed by less experienced coaches (TP-L2 and TP-L3) suggests that AI can produce training plans of comparable overall quality at intermediate and novice levels. Thus, artificial intelligence can be considered an effective supportive tool in training programme design; however, the involvement of highly qualified coaches remains essential to ensure sport-specific accuracy, contextual relevance, and optimal training effectiveness.

Conflicts of interests

The authors declare no conflicts of interest related to this work.

Acknowledgment

The authors sincerely thank all the participants and their trainers for their time and effort in contributing to this study.

References

1. Ambroży, T., Bąk, R., Niewczas, M., & Rydzik, Ł. (2022). Physical and physiological characteristics of kickboxers: a systematic review. *Science of Martial Arts*, 18, 111-120.
2. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838
3. Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., ... & Liang, P. (2023). The foundation model transparency index. *arXiv preprint arXiv:2310.12941*. https://doi.org/10.48550/arXiv.2310.12941
4. Bompa, T. O., & Buzzichelli, C. (2019). *Periodization: Theory and methodology of training* (6th ed.). Human Kinetics.
5. Bujak, Z., Gierczuk, D., & Litwiniuk, S. (2013). Professional activities of a coach of martial arts and combat sports. *Journal of Combat Sports and Martial Arts*, *4*(2), 191-195. https://doi.org/10.2478/pjst-2013-0004
6. Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, *47*(1), 33. https://doi.org/10.1007/s10916-023-01925-4
7. Cavazzotto, T. G., Dantas, D. B., & Queiroga, M. R. (2024). ChatGPT and exercise prescription: Human vs. machine or human plus machine?. *Journal of Sport and Health Science*, *13*(5), 661-662. https://doi.org/10.1016/j.jshs.2023.10.008
8. Côté, J., & Gilbert, W. (2009). An integrative definition of coaching effectiveness and expertise. *International Journal of Sports Science & Coaching*, *4*(3), 307–323. https://doi.org/10.1260/174795409789623892
9. Curby, D., Dokmanac, M., Kerimov, F., Tropin, Y., Latyshev, M., Bezkorovainyi, D., & Korobeynikov, G. (2023). Performance of wrestlers at the Olympic Games: Gender aspect. *Pedagogy of Physical Culture and Sports*, 27(6), 487–493. https://doi.org/10.15561/26649837.2023.0607
10. Dergaa, I., Saad, H. B., El Omri, A., Glenn, J., Clark, C., Washif, J., ... & Chamari, K. (2024). Using artificial intelligence for exercise prescription in personalised

health promotion: A critical evaluation of OpenAI's GPT-4 model. *Biology of Sport*, *41*(2), 221-241. https://doi.org/10.5114/biolsport.2024.133661

11. Düking, P., Sperlich, B., Voigt, L., Van Hooren, B., Zanini, M., & Zinner, C. (2024). ChatGPT generated training plans for runners are not rated optimal by coaching experts, but increase in quality with additional input information. *Journal of Sports Science & Medicine*, 23(1), 56–64. https://doi.org/10.52082/jssm.2024.56
12. Erickson, K., Côté, J., & Fraser-Thomas, J. (2007). Sport experiences, milestones, and educational activities associated with high-performance coaches' development. *The Sport Psychologist, 21*(3), 302–316.
13. Genç, A., Aydın, G. R., Kasap, M., Özkan, A., Rakhymzhanov, A., Gürcan, H. H., ... & Demirhan, B. (2026). Comparative evaluation of ChatGPT versions in training program design: scientific approach, accuracy, and practical applicability. *BMC Sports Science, Medicine and Rehabilitation*, *18*(1), 19. https://doi.org/10.1186/s13102-025-01409-7
14. Guatam, D., Saini, P., Haq, A. U., Yadav, S. K., Singh, A., Kumar, R., ... & Reddy, T. O. (2026). Artificial intelligence in sport sciences: A systematic review of models and methods. *Sports Orthopaedics and Traumatology*. https://doi.org/10.1016/j.orthtr.2025.08.001
15. Havers, T., Masur, L., Isenmann, E., Geisler, S., Zinner, C., Sperlich, B., & Düking, P. (2025). Reproducibility and quality of hypertrophy-related training plans generated by GPT-4 and Google Gemini as evaluated by coaching experts. *Biology of Sport*, 42(2), 289–329. https://doi.org/10.5114/biolsport.2025.145911
16. Kuliś, S., Zając, A., & Gołaś, A. (2024). *The role of artificial intelligence in sports analytics: A systematic review and meta-analysis of performance trends*. Applied Sciences, 15(13), 7254.
17. LaPlaca, D. A., & Schempp, P. G. (2020). The characteristics differentiating expert and competent strength and conditioning coaches. *Research Quarterly for Exercise and Sport*, *91*(3), 488-499. https://doi.org/10.1080/02701367.2019.1686451
18. Latyshev, M., Lopatenko, G., Shandryhos, V., Yarmoliuk, O., Pryimak, M., & Kvasnytsia, I. (2024). Computer vision technologies for human pose estimation in exercise: Accuracy and practicality. *In Society. Integration. Education. Proceedings of the International Scientific Conference*. Vol. 2, 626–636. https://doi.org/10.17770/sie2024vol2.7842
19. Li, L., Olson, H. O., Tereschenko, I., Wang, A., & McCleery, J. (2025). Impact of coach education on coaching effectiveness in youth sport: A systematic review and meta-analysis. *International Journal of Sports Science & Coaching*, *20*(1), 340-356. https://doi.org/10.1177/17479541241283442
20. Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., Khalili, J., Shih, Y. C. T., ... & Mafi, J. N. (2025). A randomized-clinical trial of two ambient artificial intelligence scribes: Measuring documentation efficiency and physician burnout. *medRxiv*. Preprint. https://doi.org/10.1101/2025.07.10.25331333
21. Nash, C., Ashford, M., & Collins, L. (2023). Expertise in coach development: The need for clarity. *Behavioral Sciences*, *13*(11), 924. https://doi.org/10.3390/bs13110924
22. Ouergui, I., Hssin, N., Haddad, M., Padulo, J., Franchini, E., Gmada, N., & Bouhlel, E. (2014). The effects of five weeks of kickboxing training on physical fitness. *Muscles, ligaments and tendons journal*, *4*(2), 106.
23. Philuek, P., Kusump, S., Sathianpoonsook, T., Jansupom, C., Sawanyawisuth, P., Sawanyawisuth, K., & Chainarong, A. (2025). The effects of chat GPT generated exercise program in healthy overweight young adults: A pilot study. *Journal of Human Sport and Exercise*, *20*(1), 169-179. https://doi.org/10.55860/1epqgp77
24. Puce, L., Bragazzi, N. L., Curra, A., & Trompetto, C. (2025). Harnessing Generative Artificial Intelligence for Exercise and Training Prescription: Applications and Implications in Sports and Physical Activity – A Systematic Literature Review. *Applied Sciences (2076-3417)*, *15*(7). https://doi.org/10.3390/app15073497
25. Shin, D., Hsieh, G., & Kim, Y. H. (2025, July). PlanFitting: Personalized Exercise Planning with Large Language Model-driven Conversational Agent. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. 1-19. https://doi.org/10.1145/3719160.3736607
26. Sperlich, B., Düking, P., Leppich, R., & Holmberg, H. C. (2023). Strengths, weaknesses, opportunities, and threats associated with the application of artificial intelligence in connection with sport research, coaching, and optimization of athletic performance: a brief SWOT analysis. *Frontiers in Sports and Active Living*, *5*, 1258562. https://doi.org/10.3389/fspor.2023.1258562
27. Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. TRENDS in Sport Sciences. 1(21): 19-25.
28. Toros, T. (2011). Training exercise performance questionnaire (TEPQ) – development study: A study on sportsmen from branches of judo, taekwondo, karate. *Archives of Budo*, 7(2), 81–86.
29. Wachholz, F., Manno, S., Schlachter, D., Gamper, N., & Schnitzer, M. (2025). Acceptance and trust in AI-generated exercise plans among recreational athletes and quality evaluation by experienced coaches: A pilot study. *BMC Research Notes*, 18(1), 112. https://doi.org/10.1186/s13104-025-07172-9
30. Yang, J., Qin, S., & Ren, D. (2025). Artificial intelligence coaches on the sidelines: evaluating readability and quality of soccer training plans from six generative models. *International Journal of Sports Science & Coaching*, 17479541251369593. https://doi.org/10.1177/17479541251369593
31. Zhou, D., Keogh, J. W. L., Ma, Y., Tong, R. K. Y., Khan, A. R., & Jennings, N. R. (2025). *Artificial intelligence in sport: A narrative review of applications, challenges and future trends. Journal of Sports Sciences*, https://doi.org/10.1080/02640414.2025.2518694