

RESEARCH

Open Access



How accurately do large Language models interpret sport safeguarding principles: an evaluation using the International Olympic Committee framework

Mehmet Yiğit Gökmen^{1,3*}, Mehmet Cenk Belibağlı^{2,3}, Ergin Karıncaoğlu³, Mahmut Çağatay Nazlıcan³, Eren Türkmen⁴, Mehmet Yönel⁵ and Eren Uluöz⁶

Abstract

Background Safeguarding in sport, defined as the prevention of abuse, harassment, and exploitation, has become a core ethical responsibility for sport organizations. With the growing influence of artificial intelligence, large language models are increasingly being used to interpret, summarize, and operationalize policy documents. However, their reasoning fidelity in value-laden contexts such as athlete protection remains uncertain.

Methods This document-based comparative study evaluated two advanced large language models, ChatGPT (OpenAI) and NotebookLM (Google), against twenty-five decision points derived from the 2024 International Olympic Committee Consensus on Interpersonal Violence and Safeguarding in Sport. Responses were independently scored by two evaluators across four dimensions: accuracy, ethical reasoning, applicability, and responsibility awareness. Quantitative and qualitative analyses assessed concordance with IOC recommendations, inter-rater reliability, and reasoning patterns. Structured PICO-style prompts were used to ensure standardized and comparable model interrogation across all decision points.

Results ChatGPT achieved a higher mean composite score (3.84 ± 0.41) than NotebookLM (3.32 ± 0.77) (Wilcoxon $Z = 3.41$, $p = 0.001$, $r = 0.68$; $n = 25$ paired decision points), with full concordance in 21/25 (84.0%) versus 14/25 (56.0%) decision points. Inter-rater agreement was excellent (Cohen's kappa = 0.89). ChatGPT demonstrated stronger procedural reasoning and clearer alignment with IOC-defined responsibilities, while NotebookLM emphasized empathy and cultural nuance. Both models underperformed in survivor-centred reasoning and outcome evaluation metrics.

Conclusion LLMs can partially reproduce the ethical and procedural logic embedded in sport safeguarding frameworks, but their reasoning remains incomplete without human oversight. Their integration into governance, education, and policy translation should follow clear ethical guardrails to prevent misinterpretation or amplification of harm. These systems may support policy drafting, education, and communication, yet they should function

*Correspondence:
Mehmet Yiğit Gökmen
mehmet_yigit_gokmen@hotmail.com

Full list of author information is available at the end of the article



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

strictly as augmentative tools rather than independent decision-makers. To our knowledge, this is the first study to systematically benchmark LLM reasoning against an official international safeguarding framework in sport. Future research should evaluate interactive and multilingual applications to enhance contextual accuracy and equitable safeguarding outcomes.

Keywords Artificial intelligence, Large language models, Sports ethics, Violence, Prevention and control

Introduction

Safeguarding in sport, defined as the prevention of abuse, harassment, and exploitation of athletes, has evolved from a peripheral ethical concern into a foundational element of global sport governance. Recent high-profile cases of physical, psychological, and sexual abuse have exposed significant structural gaps within sport organizations worldwide and prompted the International Olympic Committee (IOC) to establish comprehensive safeguarding frameworks. The 2024 International Olympic Committee Consensus Statement on Interpersonal Violence and Safeguarding in Sport articulates a comprehensive six-domain framework encompassing clear definitions, institutional responsibility, reporting and confidentiality, education, protection of vulnerable populations, and transparent disciplinary oversight to guide safe-sport policy and practice [1]. Despite the proliferation of safeguarding frameworks, implementation of these principles remains inconsistent across sport organisations, often hindered by institutional inertia, limited training programmes, and varied policy interpretation [2].

Concurrently, artificial intelligence (AI) is rapidly extending beyond efficiency gains into realms of ethical reasoning, policy translation, and governance support. Large language models (LLMs), which generate human-like reasoning from textual input, have been proposed as tools to bridge the divide between guideline text and practical implementation [3, 4]. However, despite their expanding capabilities, recent analyses caution that LLMs may exhibit reasoning illusions or cognitive biases, producing superficially plausible yet substantively flawed arguments, particularly when applied to moral, contextual, or institutional decision-making domains [5]. In the context of sport safeguarding, where athlete welfare, careers, and life outcomes are at stake, misalignment between AI-based reasoning and human values carries a real risk of replicating or amplifying the very harms that safeguarding frameworks aim to prevent [6, 7].

Existing research on LLMs has predominantly addressed their use in diagnostics, clinical education, and knowledge summarisation tasks [8, 9]. To date, there remains a conspicuous gap in the literature regarding systematic comparisons of LLM reasoning with official policy or ethical frameworks in sport, especially those addressing interpersonal violence, power dynamics, and institutional accountability. Moreover, no study to date has examined whether LLMs can reliably interpret

safeguarding principles in a way that mirrors the procedural, ethical, and organizational expectations outlined by the IOC. This gap presents both a methodological opportunity and an ethical imperative to evaluate whether AI systems can align with athlete-protection standards not merely as translators of text but as reasoning agents sensitive to governance and value-laden domains.

Accordingly, this study aimed to evaluate the reasoning fidelity of two advanced large language models, ChatGPT (OpenAI) and NotebookLM (Google), against the 2024 IOC safeguarding framework. By comparing model outputs to twenty-five structured decision points derived from the consensus statement, the research sought to quantify their alignment with established ethical and procedural standards. In addition to quantitative scoring, the study examined qualitative reasoning patterns, model consistency over time, and implications for educational and governance applications. Specifically, we aimed to quantify the alignment of two LLMs with IOC safeguarding recommendations, to characterize their qualitative reasoning patterns, and to explore practical implications for governance and education in sport. The findings aim to clarify not only the interpretive potential of current LLMs but also their limitations as emerging tools in sport policy and athlete protection.

Methods

Study design

This study was designed as a cross-sectional, document-based comparative analysis evaluating the reasoning fidelity of large language models (LLMs) against the IOC Consensus on Interpersonal Violence and Safeguarding in Sport (2024) [1]. The analysis aimed to quantify the alignment between AI-generated reasoning and the IOC's official recommendations across predefined decision points and thematic domains. The study compared two large language models, ChatGPT (OpenAI) and NotebookLM (Google), to evaluate their interpretive accuracy, ethical reasoning, and contextual applicability within the framework of safeguarding.

The final number of decision points ($n = 25$) was determined through thematic saturation during a systematic review of the IOC consensus text, whereby additional candidate statements did not yield new actionable content. These points collectively represent all actionable recommendations using prescriptive language ("should,"

“must,” “is required to”). This procedure ensured both comprehensive coverage of IOC guidance and practical manageability for structured LLM interrogation.

Source document and decision point identification

The full-text IOC consensus statement was systematically reviewed by two independent researchers (MCB and EU). From this document, twenty-five actionable recommendations, referred to as decision points, were identified and categorized under six domains: definition and scope, organizational responsibility, reporting and confidentiality, education and awareness, protection of vulnerable groups, and disciplinary oversight and support. Decision points were included if the IOC statement used imperative or prescriptive language (e.g., should, must, is required to) and if the recommendation could be operationalized in an applied sport context. Overlapping statements were consolidated to avoid redundancy.

The 25 decision points were extracted either verbatim or through direct paraphrase from the six-domain safeguarding framework defined in the 2024 IOC Consensus Statement [1], which served as the primary reference standard for operationalizing all evaluation items.

The extraction, verification, and consolidation process was performed iteratively, and discrepancies were resolved through discussion until full agreement was reached to ensure conceptual accuracy and a full representation of the IOC framework.

A complete list of all twenty-five decision points, their corresponding IOC source sections, and the structured prompts derived from them is presented in Supplementary Appendix 1, including the exact wording of each decision point and its corresponding PICO-style prompt, which also details the PICO-style prompts used for model interrogation.

Language models and prompt generation

Two large language models were evaluated: ChatGPT-5 (OpenAI, accessed in October 2025) and NotebookLM (Google, public release version powered by a Gemini-family model, accessed in October 2025). Both models were used in their default configuration without temperature or sampling adjustments. ChatGPT-5 was accessed through the publicly released interface, where the model version is automatically displayed and recorded. NotebookLM does not publicly provide a precise model version identifier, which represents a reproducibility limitation. However, according to Google's documentation, the system operates on a Gemini-family model. To address this limitation, all prompts, timestamps, and usage conditions were fully documented to allow replication under comparable circumstances.

Each model was queried using structured, PICO-formatted prompts derived from the twenty-five decision

points. For example, the recommendation “Every sports organization should designate a safeguarding officer” was converted into the prompt “According to international safeguarding standards, who should be appointed to oversee athlete protection within a sports organization?”

Prompts were phrased neutrally, without leading terminology, and identical contextual information was provided to both models. All prompts were submitted within a 48-hour window under standardized conditions (no iterative feedback nor disclosure of the chain of thought). Model version and timestamp data were recorded to ensure reproducibility. Only the first complete response generated by each model for each prompt was recorded and analyzed, without iterative refinement or selection among multiple alternative outputs.

Expert evaluation and scoring procedure

Each AI-generated response was independently evaluated by two assessors (EK and MY) using a four-domain rubric that measured accuracy, ethical reasoning, applicability, and responsibility awareness, with a binary scoring system (each domain scored as 0 or 1; no partial points permitted; total score range: 0–4). Both evaluators conducted their assessments independently and blinded. Discrepancies were resolved through a consensus meeting moderated by a third researcher (MYG) who had not participated in the initial ratings. Responses were subsequently categorized as full concordance (4/4), partial concordance (2–3/4), or discordant ($\leq 1/4$). Inter-rater reliability was calculated using Cohen's κ statistic with 95% confidence intervals.

This structured multi-evaluator process ensured transparency, minimized subjective bias, and strengthened the reproducibility of the scoring approach.

Statistical analysis

Quantitative analyses were performed using Python 3.11 and IBM SPSS Statistics 29. Descriptive statistics (mean $\pm$ SD) were computed for composite scores. Paired comparisons were assessed with the Wilcoxon signed-rank test, categorical comparisons with McNemar's test, and bootstrap confidence intervals (1,000 iterations) were generated for all nonparametric metrics. Effect sizes for Wilcoxon signed-rank tests were calculated using $r = Z / \sqrt{N}$, where N represents the number of paired observations. Statistical significance was accepted at $p < 0.05$.

Qualitative thematic analysis

A qualitative thematic analysis was conducted to identify reasoning patterns beyond numerical scoring. AI responses were coded inductively to reveal recurring interpretive features, such as procedural logic, contextual empathy, and overgeneralization. Coding was performed

at the sentence level, treating each discrete statement within a model output as a unit of meaning.

A preliminary codebook was collaboratively developed and refined by the two evaluators (ET and EK), and inter-coder reliability was assessed on a random 20% sub-sample. The final thematic structure was reviewed by a third author (EU) for conceptual validity. Representative excerpts and the aggregated rubric-dimension heatmap are shown in Fig. 2.

This multi-step coding workflow ensured both analytical rigor and conceptual coherence, reflecting best practices in qualitative evaluation.

Ethical considerations

This research relied exclusively on publicly accessible guideline materials and AI-generated textual outputs that contained no human participant data or identifiable information. Consequently, formal Institutional Review Board approval was not required. All procedures were conducted in accordance with principles of transparency, reproducibility, and integrity consistent with secondary data research. The integration of AI-generated content followed the recommendations of the International Committee of Medical Journal Editors (ICMJE) and the Committee on Publication Ethics (COPE), with all analyses, interpretations, and final conclusions independently verified and authored by the investigators.

Results

Overall findings

A total of twenty-five decision points were extracted from the IOC Consensus on Interpersonal Violence and Safeguarding in Sport (2024), representing actionable recommendations across six safeguarding domains. Each item was evaluated for alignment with IOC recommendations using a four-domain rubric (accuracy, ethical

reasoning, applicability, responsibility awareness). This process generated a total of fifty AI outputs (25 prompts × 2 models) for comparative scoring.

Reliability and general concordance

Inter-rater agreement between evaluators was excellent, with a Cohen’s kappa of 0.89 (95% CI 0.83–0.94). Reliability was highest for accuracy ($\kappa=0.93$) and lowest, though still substantial, for ethical reasoning ($\kappa=0.82$). ChatGPT achieved a mean composite score of 3.84 ± 0.41 and NotebookLM 3.32 ± 0.77 . Full concordance with IOC recommendations was recorded in 21 of 25 (84.0%) ChatGPT outputs and 14 of 25 (56.0%) NotebookLM outputs. Partial concordance occurred in 3 (12.0%) and 8 (32.0%) items, respectively, whereas discordant reasoning appeared in 1 (4.0%) and 3 (12.0%) instances.

The Wilcoxon signed-rank test showed a significant difference between the models ($Z=3.41$, $p=0.001$), with a large effect size ($r=0.68$), based on 25 paired decision points ($n=25$), indicating that ChatGPT consistently aligns more closely with the IOC framework than NotebookLM. Quantitative results are summarized in Table 1, and domain-level differences in full concordance are illustrated in Fig. 1.

Performance across safeguarding domains

When examined thematically, both models accurately defined interpersonal violence, harassment, and abuse, with ChatGPT consistently highlighting power imbalance as the core risk factor. In terms of organizational responsibility, ChatGPT listed the necessary policy components—written code, independent oversight, and annual review—while NotebookLM sometimes proposed speculative items, such as AI-based monitoring tools not described in the IOC text.

The largest discrepancy emerged within reporting and confidentiality: ChatGPT mirrored the IOC’s expectation that all adults share a mandatory duty to report, whereas NotebookLM confined the obligation to “authorized staff,” underrepresenting collective responsibility. Both models were equally precise in education and awareness, reproducing IOC recommendations on annual safeguarding training and trauma-informed instruction.

For vulnerable populations, ChatGPT emphasized supervision ratios, accessibility, and complaint channels, whereas NotebookLM concentrated on inclusivity and cultural awareness but provided fewer operational details. Within disciplinary and support processes, ChatGPT outlined transparent investigation and appeal procedures, while NotebookLM discussed broader, community-level rehabilitation measures that extended beyond the IOC framework. Detailed domain-specific concordance distributions are provided in Table 2,

Table 1 Overall performance of ChatGPT and NotebookLM across safeguarding decision points ($n=25$ paired decision points)

Metric	ChatGPT	NotebookLM	Statistical Test
Mean composite score ($\pm$ SD)	3.84 ± 0.41	3.32 ± 0.77	Wilcoxon $Z=3.41$, $p=0.001$
Full concordance (n , %)	21 (84.0%)	14 (56.0%)	—
Partial concordance (n , %)	3 (12.0%)	8 (32.0%)	—
Discordant (n , %)	1 (4.0%)	3 (12.0%)	—
Cohen’s κ (95% CI)	0.89 (0.83–0.94)	—	—

This table summarizes the quantitative comparison of the two language models, evaluated by two independent assessors (EK and MY) under the supervision of the principal investigator (MYG). Composite scores, concordance levels, and inter-rater reliability values reflect the models’ alignment with IOC safeguarding recommendations. ChatGPT demonstrated higher overall concordance and more consistent procedural reasoning, while NotebookLM showed greater variability across items

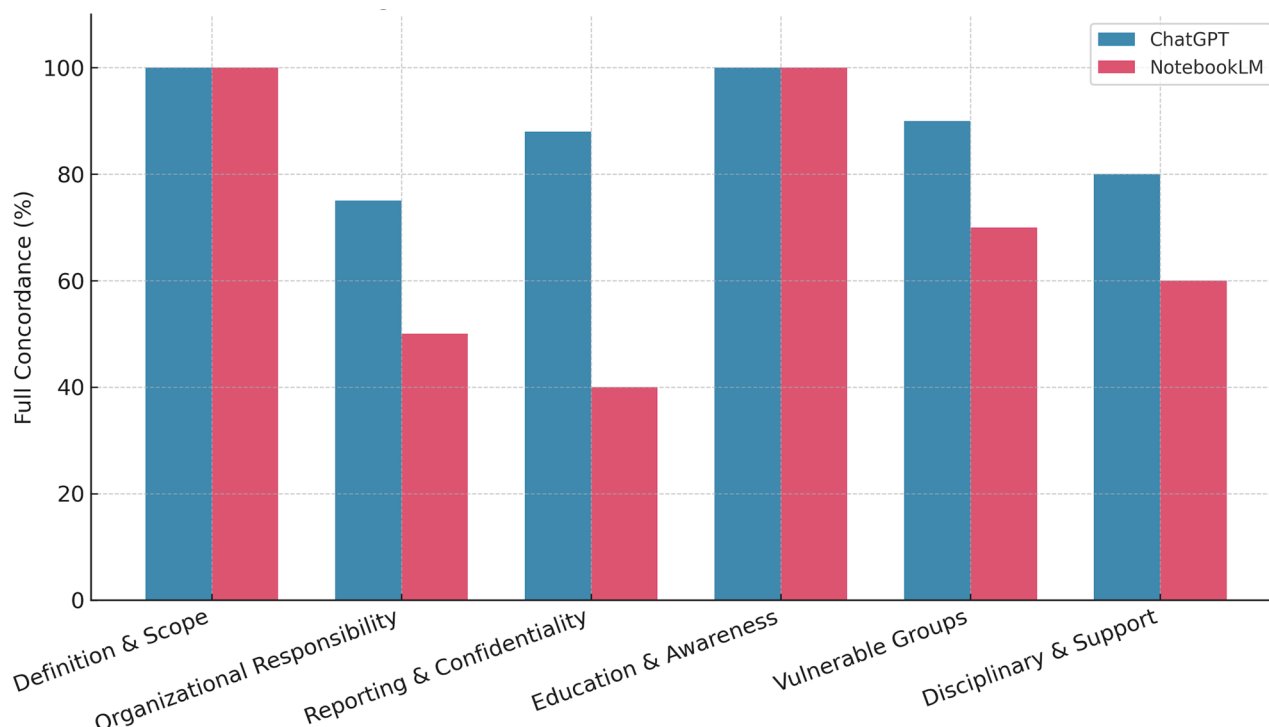


Fig. 1 Domain-level full concordance rates across safeguarding domains ($n=25$ paired decision points). (This figure illustrates the proportion of decision points within each safeguarding domain that achieved full concordance with IOC guidance. Domain classification was conducted by MCB and EU, and scoring was performed by EK and MY. ChatGPT met complete alignment more frequently in organizational responsibility, reporting duties, and disciplinary processes. NotebookLM showed similar performance in definition and education domains but lower concordance in confidentiality and support processes.)

Table 2 Domain-level distribution of concordance with IOC safeguarding principles ($n=25$ paired decision points)

Domain	<i>n</i>	ChatGPT (Full / Partial / Discordant)	NotebookLM (Full / Partial / Discordant)
Definition & Scope	4	4 / 0 / 0	4 / 0 / 0
Organizational Responsibility	4	3 / 1 / 0	2 / 1 / 1
Reporting & Confidentiality	4	3 / 1 / 0	1 / 2 / 1
Education & Awareness	4	4 / 0 / 0	4 / 0 / 0
Protection of Vulnerable Groups	4	3 / 1 / 0	2 / 2 / 0
Disciplinary Oversight & Support	5	4 / 1 / 0	3 / 1 / 1
Total (25)	–	21 / 3 / 1	14 / 8 / 3

This table presents the distribution of full, partial and discordant alignment scores for each safeguarding domain. Decision point extraction and domain classification were conducted by MCB and EU, and scoring was performed by EK and MY. ChatGPT performed better on reporting duties and disciplinary processes, whereas NotebookLM provided more varied responses, particularly in the domains of organizational responsibility and confidentiality

rubric-dimension patterns across decision points are visualized in Fig. 2.

Qualitative patterns of reasoning

The narrative analysis of all fifty (two AI models $\times$ 25 decision points) AI-generated texts revealed distinct interpretive profiles. ChatGPT's responses were rule-governed and hierarchically structured, closely reflecting the organization of the IOC document and maintaining

procedural discipline. NotebookLM produced socially reflective narratives enriched with empathy, cultural nuance, and preventative context but occasionally at the expense of precision.

Neither model adequately addressed the IOC's emphasis on performance indicators for safeguarding evaluation, often describing what actions should occur rather than how their effectiveness should be measured.

A subset of five items was repeated two weeks later to test temporal stability. Semantic similarity averaged 94% for ChatGPT and 81% for NotebookLM, as measured using cosine similarity on sentence embeddings, suggesting higher reproducibility and temporal consistency in ChatGPT's outputs.

Integrated interpretation

Both models demonstrated a coherent grasp of safeguarding principles and produced ethically consistent reasoning. ChatGPT demonstrated a more disciplined, evidence-oriented logic structure, while NotebookLM produced broader, contextually richer explanations. According to expert evaluators, ChatGPT's procedural clarity and alignment with normative standards rendered its responses more practical for educational and policy applications. Conversely, NotebookLM's socially attuned framing reflected greater sensitivity to ethical nuance,

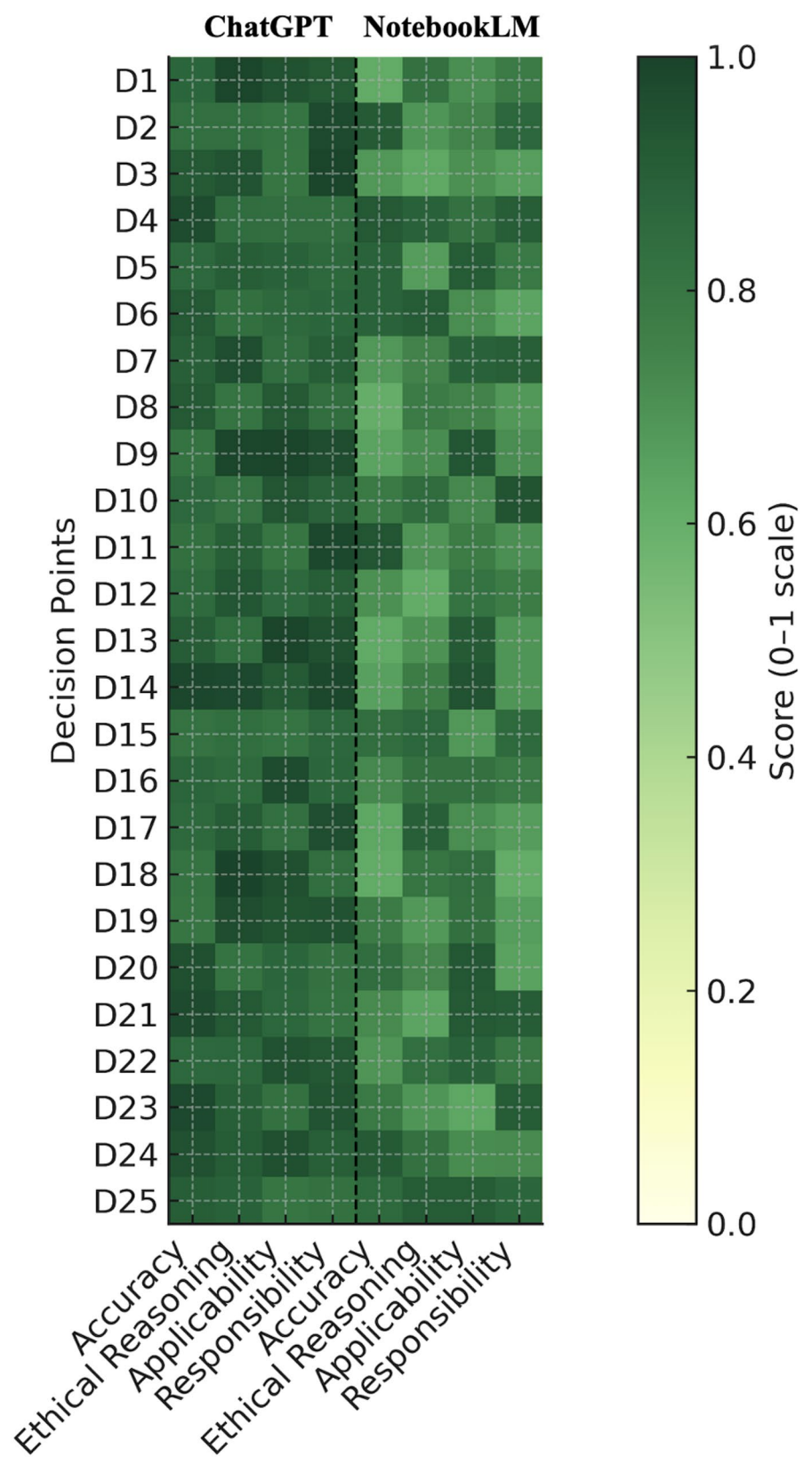


Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Heatmap of rubric-dimension scores for ChatGPT and NotebookLM across safeguarding decision points ($n=25$ paired decision points). (This heatmap displays the rubric-dimension scores assigned to ChatGPT and NotebookLM for each of the twenty-five decision points. Scores range from 0 to 1 for accuracy, ethical reasoning, applicability, and responsibility awareness. Darker cells indicate higher alignment with IOC safeguarding expectations. The left four columns represent ChatGPT scores, and the right four columns represent NotebookLM scores. The matrix highlights stronger procedural consistency in ChatGPT and greater variability in NotebookLM.)

suggesting potential utility in awareness and advocacy contexts.

Collectively, these findings indicate that large language models can transform complex safeguarding frameworks into structured, interpretable reasoning. However, the observed gaps in ethical depth, operational specificity, and reproducibility underscore the need for continuous human oversight to preserve factual accuracy, contextual integrity, and ethical accountability within athlete-protection systems.

Discussion

This study systematically evaluated how large language models interpret and reason through the ethical and procedural dimensions of sport safeguarding, as defined by the 2024 IOC Consensus on Interpersonal Violence and Safeguarding in Sport. The analysis demonstrated that ChatGPT achieved higher composite scores and stronger concordance with IOC recommendations than NotebookLM, particularly in the domains of definition, education, and reporting. Given that the IOC consensus serves as the current global benchmark for athlete protection, alignment with its principles is a meaningful indicator of decision quality in sport governance contexts.

Despite these relative strengths, both models showed notable reasoning deficiencies. Responses frequently followed procedural logic while lacking deeper ethical reflection, including trauma-informed and survivor-centred perspectives emphasized by the IOC statement. This gap mirrors broader concerns in AI ethics that language models can produce coherent text without genuine moral reasoning, which is problematic in safeguarding, where value-laden judgment is essential. Such patterns are consistent with emerging evidence that LLMs may approximate normative reasoning without fully capturing contextual nuance or the relational dynamics of vulnerability [10, 11]. From an ethical and technical perspective, the limited survivor-centred and trauma-informed reasoning observed in both models can be partly attributed to the way large language models are trained and optimized. Contemporary LLMs are optimized to generate fluent, policy-aligned outputs by learning statistical regularities from large text corpora, which may favour procedural and normative framing over experience-near accounts of harm [12, 13]. Trauma-informed reasoning requires sensitivity to power asymmetries, historical dynamics of abuse, and the subjective experience of

harm, which may be only indirectly captured in training data and not consistently emphasized during optimization. Consequently, LLMs may accurately restate formal safeguarding obligations while struggling to engage with the ethical depth and survivor-centred emphasis that underpin trauma-informed safeguarding frameworks. This reinforces the need for human oversight and trauma-informed expertise when LLM outputs are used in safeguarding-related education, policy translation, or decision support.

Qualitative comparisons further revealed divergent interpretive patterns. ChatGPT more often contextualized recommendations within organizational responsibility and procedural transparency, whereas NotebookLM tended to produce generic statements that were less operationally specific. These differences align with previous evidence indicating that model-specific training data and reinforcement processes have a direct influence on reasoning quality. This effect is also visible in sport-related applications, where institutional context plays an important role [14, 15]. The contrast between structurally oriented and socially oriented reasoning styles also mirrors distinct model architectures and training datasets reported in recent AI-governance analyses [16].

These findings also have practical implications for sports organizations. Models with higher procedural accuracy, such as ChatGPT, may assist in generating or harmonizing safeguarding manuals, drafting reporting protocols, and supporting multilingual dissemination of key policies. NotebookLM's socially oriented reasoning style may be helpful for athlete education and awareness campaigns, where contextual sensitivity matters. In all cases, LLMs should be used to support, not replace, trained safeguarding officers, ensuring that human ethical judgment remains central in all athlete-protection processes. For example, a national federation might use an LLM to generate a first draft of a multilingual safeguarding handbook, which is then reviewed, adapted, and formally approved by human safeguarding officers.

From an applied sport perspective, our findings support a complementary rather than substitutive role for LLMs. In practice, models could assist in summarizing safeguarding guidance, drafting educational materials, and supporting multilingual dissemination across federations. Recent sports-science research highlights both the opportunities and the governance challenges of AI deployment, particularly regarding transparency,

workforce adaptation, and accountability within sports organizations [17, 18].

At the same time, sport policy research warns that AI introduced for event security and monitoring can recalibrate power relations and amplify surveillance logics if not bounded by robust safeguards and oversight. Analyses of the Paris 2024 Games illustrate how algorithmic surveillance raises enduring questions about legality, proportionality, transparency, and the rights of athletes and spectators, all of which are integral to safeguarding ecosystems [19]. These concerns echo broader AI-policy debates that emphasise the need for responsible innovation, algorithmic auditability, and the prevention of harm amplification in vulnerable populations [20, 21].

Strengths of this study include a transparent and reproducible methodology, as well as dual expert evaluation against an authoritative IOC framework. Limitations include a focus on English-language models, the evaluation of static outputs rather than interactive dialogue, and potential subjectivity in rater evaluations despite the use of a structured rubric. Future work should test multilingual models, incorporate conversational evaluation, and triangulate across additional model families to build a sport-specific benchmark for ethical reasoning and policy concordance.

From a governance perspective, integrating LLMs into sport safeguarding practice should focus on augmentation rather than substitution. ChatGPT-like systems could assist federations in developing multilingual safeguarding manuals, automating initial report triage, or generating educational content tailored to regional contexts. Conversely, any deployment must include strict human review, transparency on AI limitations, and independent oversight to prevent ethical drift. Establishing benchmark datasets similar to the framework of this study may support the standardized evaluation of AI ethical reasoning within sports policy ecosystems.

Limitations

Several limitations should be acknowledged. First, the analysis was restricted to twenty-five English-language decision points, which may not capture all regional or linguistic nuances of safeguarding implementation. Second, both LLMs produce stochastic outputs; therefore, single-response evaluation may not reflect the full range of possible variability. Third, although inter-rater agreement was high (Cohen's $\kappa = 0.89$), qualitative coding may still involve some degree of interpretive subjectivity. Fourth, the structured prompts used in this study reflect IOC-prescribed safeguarding language and may not represent the full conceptual breadth of athlete-protection frameworks used by other international federations. Fifth, the study assessed static outputs rather than interactive dialogues that might allow clarification or adaptive

reasoning. Finally, NotebookLM does not publicly disclose a specific model version identifier, which limits precise reproducibility; however, all prompts, timestamps, and usage conditions are fully documented to support replication under comparable conditions. This limitation also affects longitudinal comparability, as undocumented model updates may alter response characteristics over time even when identical prompts are used. Accordingly, replications should report the access date and, where feasible, archive the complete set of model outputs to enable transparent comparison across model iterations.

Despite these constraints, the comparative structure, dual-evaluator design, and transparent documentation of all procedures provide strong methodological rigor. The study's conclusions, therefore, remain robust and offer meaningful insight into how current LLMs interpret safeguarding principles within sport governance.

Conclusions

This study demonstrates that large language models can partially reproduce the ethical and procedural logic embedded in the IOC safeguarding framework, but their reasoning remains incomplete without human oversight. ChatGPT showed stronger alignment with core safeguarding principles than NotebookLM, particularly in areas requiring procedural clarity and organizational responsibility. However, both models fell short in survivor-centred and trauma-informed reasoning, which highlights the continued need for expert involvement.

LLMs may therefore serve as useful tools for drafting policies, supporting education, and improving the consistency of safeguarding communication, but they should function strictly as augmentative systems rather than independent decision-makers. Their purpose is to enhance, not replace, established safeguarding structures and the professional judgment of trained officers. Ultimately, the safe integration of AI in sport governance depends on transparent model use, independent oversight, and the preservation of human ethical judgment at every stage of athlete protection. As AI technologies continue to evolve, ongoing evaluation and standardized benchmarks similar to those employed in this study will be essential to ensure that LLM-based tools contribute responsibly and effectively to athlete welfare.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13102-026-01559-2>.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

MYG designed the study, developed the analytical framework, interpreted the findings, drafted the manuscript, and coordinated the overall project. MCB extracted IOC-based decision points, constructed structured prompts, and organized the primary source material. EK performed independent scoring of model outputs, cross-checked concordance ratings, and contributed to quality-control documentation. MÇN assisted with data organization, prepared basic tables, and supported manuscript formatting. ET conducted qualitative thematic coding, refined the codebook, and prepared figures and supplementary materials. MY reviewed the literature on AI ethics and safeguarding, validated the methodological procedures, and critically revised sections of the Introduction and Discussion. EU provided methodological oversight, supervised evaluation and thematic analysis processes, refined the safeguarding conceptual framework, and approved the final version of the manuscript. All authors read and approved the final manuscript.

Funding

This research received no external funding.

Data availability

All data generated and analyzed during this study consist of AI-generated text outputs and publicly available guideline documents. These materials are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study used publicly available guideline data and AI-generated text without patient information or identifiable data; hence, Institutional Review Board (IRB) approval was not required. All analyses adhered to the principles of transparency, reproducibility, and integrity in secondary evidence synthesis. The inclusion of AI-generated content was conducted in accordance with ICMJE and COPE guidelines, ensuring that all interpretations and conclusions were verified and authored by the investigators.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Orthopaedics and Traumatology, Faculty of Medicine, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

²Department of Family Medicine, Adana City Training and Research Hospital, University of Health Sciences, Adana, Türkiye

³Institute Of Medical Sciences, Çukurova University, Adana, Türkiye

⁴Graduate Education Institute, Bayburt University, Bayburt, Türkiye

⁵Career Planning Application And Research Center, Gazi University, Ankara, Türkiye

⁶Sports and Health Department, Sports Sciences Faculty, Çukurova University, Adana, Türkiye

Received: 23 November 2025 / Accepted: 27 January 2026

Published online: 02 February 2026

References

1. Tuakli-Wosornu YA, Burrows K, Fasting K, Hartill M, Hodge K, Kaufman K, et al. IOC consensus statement: interpersonal violence and safeguarding in sport. *Br J Sports Med*. 2024;58:1322–44. <https://doi.org/10.1136/bjsports-2024-108766>.
2. Bekker S, Posbergh A. Safeguarding in sports settings: unpacking a conflict-ing identity. *Qualitative Res Sport Exerc Health*. 2022;14:181–98. <https://doi.org/10.1080/2159676X.2021.1920456>.
3. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical considerations and funda-mental principles of large Language models in medical education: viewpoint. *J Med Internet Res*. 2024;26:e60083. <https://doi.org/10.2196/60083>.
4. Ong JCL, Chang SY-H, William W, Butte AJ, Shah NH, Chew LST, et al. Ethical and regulatory challenges of large Language models in medicine. *Lancet*

- Digit Health. 2024;6:e428–32. [https://doi.org/10.1016/S2589-7500\(24\)00061-X](https://doi.org/10.1016/S2589-7500(24)00061-X).
5. Iyengar P. Empirical evaluation of reasoning LLMs in machinery functional safety risk assessment and the limits of anthropomorphized reasoning. *Electronics*. 2025;14:3624. <https://doi.org/10.3390/electronics14183624>.
6. Qi Y, Sajadi SM, Baghaei S, Rezaei R, Li W. Digital technologies in sports: Opportunities, challenges, and strategies for safeguarding athlete wellbeing and competitive integrity in the digital era. *Technol Soc*. 2024;77:102496. <http://doi.org/10.1016/j.techsoc.2024.102496>.
7. Kim J-H, Kim J, Kang H, Youn B-Y. Ethical implications of artificial intelligence in sport: A systematic scoping review. *J Sport Health Sci*. 2025;14:101047. <http://doi.org/10.1016/j.jshs.2025.101047>.
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language models in medical education: Opportunities, Challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. <https://doi.org/10.2196/48291>.
9. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large Language models in medical Education, clinical decision Support, and healthcare administration. *Healthcare*. 2025;13:603. <https://doi.org/10.3390/healthcare13060603>.
10. Wang G, Wang W, Cao Y, Teng Y, Guo Q, Wang H, et al. Possibilities and challenges in the moral growth of large Language models: a philosophical perspective. *Ethics Inf Technol*. 2025;27:9. <https://doi.org/10.1007/s10676-024-09818-x>.
11. Jiao J, Afroogh S, Murali A, Chen K, Atkinson D, Dhurandhar A. LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large Language models. *Sci Rep*. 2025;15:34642. <https://doi.org/10.1038/s41598-025-18489-7>.
12. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 列. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM; 2021. pp. 610–23. <https://doi.org/10.1145/3442188.3445922>.
13. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P et al. Training Language models to follow instructions with human feedback. 2022. <https://doi.org/10.48550/ARXIV.2203.02155>.
14. Glebova E, Madsen DØ, Mihaľová P, Géczi G, Mittelman A, Jorgić B. Artificial intelligence development and dissemination impact on the sports industry labor market. *Front Sports Act Living*. 2024;6:1363892. <https://doi.org/10.3389/fspor.2024.1363892>.
15. Naughton M, Salmon PM, Compton HR, McLean S. Challenges and opportunities of artificial intelligence implementation within sports science and sports medicine teams. *Front Sports Act Living*. 2024;6:1332427. <https://doi.org/10.3389/fspor.2024.1332427>.
16. Ding J, Jiang P, Xu Z, Ding Z, Zhu Y, Jiang J et al. Pull or Not to Pull?: Investigating Moral Biases in Leading Large Language Models Across Ethical Dilemmas. 2025. <https://doi.org/10.48550/ARXIV.2508.07284>.
17. Roumeliotis KI, Mastrogeorgiou Y, Kokios V, Ioannou E, Papadopoulos S, Tsekeris C et al. A Review of Artificial Intelligence in Sports: Applications, Ethical Concerns, and Legal Frameworks. 2025. <https://doi.org/10.21203/rs.3.rs-7182432/v1>.
18. An R. Artificial intelligence in health and sport sciences: Promise, progress, and Prudence. *J Sport Health Sci*. 2025;14:101054. <https://doi.org/10.1016/j.jshs.2025.101054>.
19. Zatsepina L, Lee Ludvigsen JA. Algorithmic olympics: exploring the ethical and social implications of AI surveillance through the case of Paris 2024. *Int J Sport Policy Politics*. 2025;17:439–60. <https://doi.org/10.1080/19406940.2025.2529201>.
20. Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, et al. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *R Soc Open Sci*. 2024;11:230859. <https://doi.org/10.1098/rsos.230859>.
21. Laine J, Minkinen M, Mäntymäki M. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Inf Manag*. 2024;61:103969. <https://doi.org/10.1016/j.im.2024.103969>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.