

Transforming sports knowledge: Chatbot AI-based systems for improved training and education in sports

TRI LATHIF MARDI SURYANTO^{1*}, EKA DYAR WAHYUNI², ILMATUS SA'DIYAH³, AJI PRASETYA WIBAWA⁴, DANANG ARBIAN SULISTYO⁵, ARIDHOTUL HAQIYAH⁶

^{1,2},Department of Information Systems, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur, INDONESIA

³Department of Indonesian Linguistics Study Program, Faculty of Social, Culture, and Political Sciences, Universitas Pembangunan Nasional Veteran Jawa Timur, INDONESIA

⁴Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, INDONESIA

⁵Department of Information Technology, Institut Teknologi dan Bisnis Asia Malang, INDONESIA

⁶Department of Physical Education Health and Recreation Department, Universitas Islam, INDONESIA.

Published online: January 31, 2026

Accepted for publication: January 15, 2026

DOI:10.7752/jpes.2026.01015

Abstract

This study introduces a Multiple Closed-Domain (MCD) Question Answering (QA) system utilizing Transformer-based models to optimize access to sports knowledge, improving training and management in the sports sector. Traditional QA systems, whether closed-domain or open-domain, each come with inherent limitations: closed-domain systems excel in accuracy within specific topics but lack scalability, while open-domain systems offer broader coverage but often generate irrelevant or incorrect responses. To overcome these challenges, this research proposes an optimized MCD QA framework using Transformer models, specifically BART-base and BART-large, designed to address the needs of sports communities by providing accurate, context-aware answers across multiple closed domains. A dataset of 10,000 question-answer pairs was curated from five major Indonesian sports federations—PSSI (football), PBESI (e-sports), PBVSI (volleyball), PBSI (badminton), and PERBASI (basketball). The data was cleaned, tokenized, and validated by experts to ensure its relevance to sports-specific queries. Hyperparameter optimization, performed using Optuna, significantly improved the performance of the models. Evaluation metrics, including ROUGE, BLEU, and METEOR, assessed the quality of the generated responses in terms of fluency, accuracy, and semantic relevance. The results show that the optimized models achieved superior performance, particularly the BART-large model, which demonstrated the highest ROUGE-L score, indicating its ability to capture essential sports knowledge. The BART-base model excelled in BLEU and METEOR, highlighting its fluency and semantic coherence. This system provides a scalable, efficient solution for accessing sports regulations, training techniques, and health policies, essential for athletes, coaches, and sports managers. AI-driven system has significant potential in sports education and management, enabling quicker access to accurate information, improving decision-making processes, and enhancing the overall quality of sports training and performance.

Keywords: NLP, transformer, multiple closed domain, chatbot, sports informatics.

Introduction

Chatbots are sophisticated software agents engineered to replicate human-like dialogues, offering consumers automated replies to questions across multiple fields. Utilizing advancements in natural language processing (NLP) and machine learning, contemporary chatbots can comprehend context (Suryanto et al., 2025), discern user intent, and produce coherent and pertinent responses instantaneously (Suryanto et al., 2023). They have become increasingly prevalent in sports (Aminuddin et al., 2021; Keiper et al., 2023; Pre, 2023; Priya et al., 2024), education (Dhivvya & Karnati, 2024; Kesarwani et al., 2023; Pratita et al., 2025), and healthcare (Babu & Boddu, 2024; R. Li et al., 2025; Maini et al., 2023), and sport rehabilitation (Mcbee et al., 2024; McBee et al., 2023; Pre, 2023). Where they improve efficiency, reduce human workload, and enhance user engagement. The escalating demand for efficient and accurate Chatbot Question Answering (QA) (Y. Li et al., 2023; Shao et al., 2019) Where they augment efficiency, diminish human workload, and elevate user engagement. The increasing demand for efficient and precise Chatbot QA systems has emerged as a hallmark of the digital era. These advanced tools, adept at swiftly delivering accurate information, greatly improve user experience across numerous digital applications. As the quantity of online information expands, the capacity of QA systems to efficiently explore and extract pertinent data is increasingly essential.

In the last ten years, QA systems have transitioned from rule-based and pattern-matching methodologies to data-driven models utilizing neural networks, which enhance semantic comprehension and generalization

across various query structures (Shao et al., 2019). The advancement was further expedited by Transformer-based designs, which utilize self-attention methods to record long-range contextual dependencies and have established the basis for contemporary generative QA systems (Vaswani et al., 2017). Sequence-to-sequence models like BART exhibit exceptional efficacy in language production and question answering by acquiring resilient encoder-decoder representations (P. Lewis et al., 2020). Recent studies indicate that high linguistic fluency does not inherently ensure factual accuracy, especially in specialized fields, as generative QA systems may yield fluent but inaccurate responses, underscoring the necessity for domain-aware architectures and task-specific optimization.

Notwithstanding their increasing significance, existing QA systems encounter fundamental constraints, especially when classified into closed-domain (Cavalcante et al., 2021; Lam et al., 2020; Roy & Manik, 2024; Saad & Qawaqneh, 2024) and open-domain contexts (Karpukhin et al., 2020; Soni et al., 2024). Closed-domain systems, though praised for their precision in particular subjects, are inherently limited by their restricted range and require substantial, continuous upkeep to integrate new data. In contrast, open-domain systems, intended for extensive coverage, sometimes encounter the significant problem of hallucination, producing inaccurate or irrelevant responses, and require considerable computational resources for their functionality.

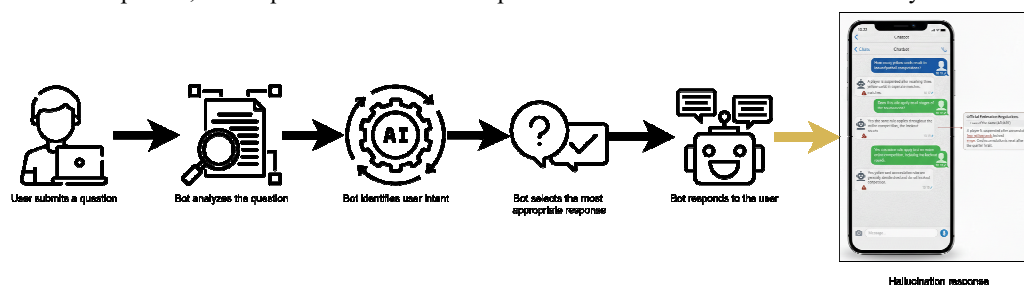


Figure 1. Illustration of hallucinated responses chatbot in Football regulated.

Figure 1 illustrates that the occurrence of hallucination in generative Question Answering systems presents a significant challenge in the context of sports-related knowledge, where precision and factual correctness are paramount. Open-domain QA models can produce articulate yet inaccurate or unsubstantiated answers since they depend on probabilistic language creation instead of verified domain expertise, especially when faced with specialized legislation or procedural inquiries. In the realm of sports, such mistakes may result in misreading of competition regulations, misguided training suggestions, or erroneous administrative choices, thereby eroding trust and practical applicability. Research on QA evaluation has underscored that measures focused solely on fluency are inadequate for assessing reliability, indicating the necessity for domain-specific systems that value factual consistency over verbal eloquence (Durmus et al., 2020; Farea et al., 2025). This highlights the necessity of creating QA systems that clearly regulate knowledge scope to enhance and guarantee reliable information dissemination in critical sports contexts.

These deficiencies highlight a considerable obstacle in creating genuinely adaptable and dependable QA solutions. In contrast to numerous general-purpose domains, sports knowledge exhibits a significant level of structural formality, regulatory authority, and contextual dependence. A substantial amount of sports-related information is regulated by official papers, including statutes, regulations, and laws of the game (AD/ART), which delineate responsibilities, competition rules, disciplinary procedures, and organizational hierarchies within sports federations. This type of information is both domain-specific and normative, indicating that precision and uniformity are crucial for ensuring compliance and equitable implementation in practice.

Moreover, sports knowledge frequently encompasses multifaceted settings, including technical laws, training methodology, health and rehabilitation protocols, and managerial policies, all of which necessitate interpretation within specific situational frameworks. Inaccuracies or ambiguities in automated responses may result in misreading of regulations, misguided training decisions, or suboptimal management practices. Previous research on chatbots in sports rehabilitation and management education has recognized the need of dependable and contextually relevant information dissemination to assist practitioners and learners (Keiper et al., 2023; McBee et al., 2024; McBee et al., 2023). The attributes of the sports domain render it more vulnerable to the constraints of open-domain QA systems, underscoring the necessity for regulated, domain-specific methodologies that can deliver precise and credible responses while reducing the potential for risk and misinformation.

In light of these restrictions, there is an evident and urgent necessity for a more sophisticated methodology: Multiple Closed-Domain (MCD) QA for the MultiSportQA document. This framework aims to reconcile the precision of closed-domain systems with the comprehensiveness of open-domain systems. The focus is on systems capable of efficiently and accurately managing numerous specialized subjects concurrently, avoiding the drawbacks of open-domain delusion and surpassing the strict topical limitations of individual closed-domain systems. This presents MCD QA as a vital advancement in the domain.

Despite progress in Transformer-based Question Answering systems, current research mostly emphasizes on single closed-domain or open-domain architectures, resulting in a significant lack of exploration into the integration of many regulated domains within a cohesive and optimal QA framework. Research on sports chatbots has primarily focused on conversational functionality or discrete applications, neglecting a systematic approach to cross-domain information integration, mitigation, and computing efficiency in regulation-driven sports environments. Despite the robust generating capabilities of models like BART, their effective use for Multiple Closed-Domain Question Answering with organized federation-level sports knowledge is constrained, prompting the creation of the suggested MultiSportQA framework.

Therefore, there is an urgent requirement for a domain-specific Question Answering system capable of delivering accurate and regulation-compliant responses in diverse sports situations, where misinformation could negatively impact training, competition, and governance. This research addresses the core challenge of developing a robust Question Answering (QA) system capable of continuously operating across many closed domains, particularly within regulation-driven sports knowledge environments. The primary objective is to achieve high response accuracy while significantly improving and maintaining computing efficiency, employing contemporary natural language processing methods based on Transformer architecture.

This study seeks to enhance Transformer-based models for question-answering tasks in specialized domains, exploring techniques for integrating varied domain knowledge into a unified Multiple Closed-Domain QA framework, and systematically evaluating model performance in terms of accuracy, semantic relevance, and efficiency. This research employs a Transformer-based architecture for Multiple Closed-Domain Question Answering, trained on curated sports regulatory documents from diverse federated domains. Model performance is evaluated by stringent experimentation and quantitative metrics to assess accuracy, semantic relevance, hallucination reduction, and computational economy.

Material & methods

The efficacy of any effective Question Answering (QA) system is contingent upon the quality and organization of its training data. This section delineates the rigorous method employed to prepare the dataset, assuring its appropriateness for optimizing Transformer-based models, similar to the study that acquired raw data from documents in Indonesia (Sulistyo, 2025), Figure 2 illustrates that the data preparation process comprised several distinct stages: Data Cleaning, Expert Validation, Tokenization, and Dataset Splitting, resulting in a dataset prepared for training.

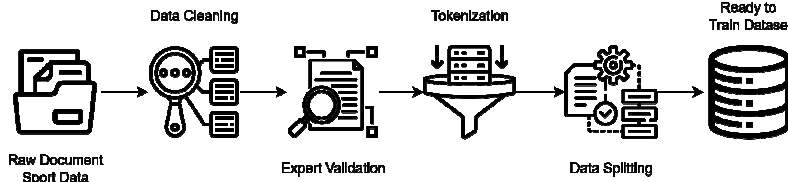


Figure 2. Dataset Preparation Process

Measure

Multiple evaluation metrics were employed to comprehensively assess the quality of the generated answers in the multiple closed-domain Question Answering (QA) system. The evaluation employed six widely used metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, BLEU, and METEOR. ROUGE-1 measures unigram overlap, reflecting the model's ability to reproduce essential keywords while ROUGE-2 captures bigram matches to assess short phrase similarity and fluency (J. Zhang et al., 2020). ROUGE-L evaluates the longest common subsequence to gauge structural similarity (Ruma et al., 2023), and ROUGE-Lsum adapts this approach for summarization by aligning sentence-level matches (Tay et al., 2019). BLEU calculates n-gram precision with a brevity penalty to assess how closely the generated text matches (Ji et al., 2022; M. Lewis et al., 2019; Liang et al., 2023; Wibawa et al., 2025) the reference in both phrasing and length. METEOR goes further by accounting for stemming (Jamroziński & Markowska-Kaczmar, 2022), synonyms, and paraphrases (Liang et al., 2023), providing a semantically richer evaluation beyond simple word matching.

- ROUGE-N:

$$ROUGE - N = \frac{\sum_{s \in Ref} \sum_{n-gram \in s} \min(Count_{gen}(n-gram), Count_{ref}(n-gram))}{\sum_{s \in Ref} \sum_{n-gram \in s} Count_{ref}(n-gram)}$$

Where:

$Count_{gen}$ = frequency of n-gram in generated text

$Count_{ref}$ = frequency of n-gram in reference text

- ROUGE-L (LCS-based):

$$ROUGE - L = \frac{(1 + \beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

Where:

$$R_{LCS} = \frac{LCS(gen, ref)}{len(ref)} = \text{recall based on LCS length}$$

$$P_{LCS} = \frac{LCS(gen, ref)}{len(gen)} = \text{precision based on LCS length}$$

β = weighting factor (commonly set to 1.2)

- BLEU:

$$BLEU = BP \cdot e^{(\sum_{n=1}^N \log p_n)}$$

Where:

p_n = precision for n-grams of size n

w_n = weight assigned to each n-gram size (uniform by default)

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r)/c} & \text{if } c \leq r \end{cases} = \text{brevity penalty, with } c \text{ as candidate length and } r \text{ as reference length}$$

- METEOR:

$$Meteor = F_{mean} \cdot (1 - Penalty)$$

$$F_{mean} = \frac{10 \cdot P \cdot R}{R + 9P} = \text{balance precision } P \text{ and recall } R$$

$$Penalty = 0.5 \cdot \left(\frac{chunks}{matches} \right)^3 \text{ penalizes fragmented matches}$$

Data collection and analysis

The dataset was obtained from five prominent Indonesian sports federations—PSSI (football), PBESI (e-sports), PBVSI (volleyball), PBSI (badminton), and PERBASI (basketball)—yielding 10,000 meticulously curated question-answer pairings (2,000 for each discipline). Sources comprised governmental rules, statutes (AD/ART), and articles, methodically transformed into plain text with distinct frameworks for Pertanyaan, Jawaban, and pertinent citations. Preprocessing encompassed column standardization, whitespace elimination, Unicode-to-ASCII conversion, lowercasing, and removal of NaN values, as illustrated in Figure 3..

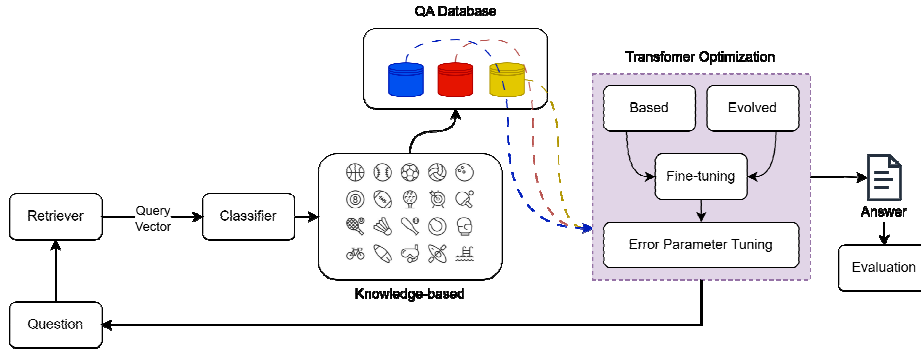


Figure 3. MCD-QA Model for MultiSportQA

A crucial expert validation phase engaged coaches, teachers, and officials to ensure domain accuracy and contextual reliability. The validated dataset was tokenized using a Transformer tokenizer, combining question, domain, and context as inputs and answers as targets, with fixed sequence lengths for training efficiency (512 tokens for input, 128 for output). Each database has its own sports group, for example, the blue database is for football, red is for basketball, yellow is for swimming, and so on. Finally, an 80/20 train-test split produced a high-quality, ready-to-train dataset for subsequent model development.

Results

This study investigates the performance of Transformer-based models in a MultiSportQA system within a specific multiple closed-domain Law of the game Indonesia's Sports Federation. The evaluation employed standard text generation metrics, including ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Koshy et al., 2023) for measuring word overlap, BLEU (Bilingual Evaluation Understudy) for fluency, and METEOR (Metric for Evaluation of Translation with Explicit Ordering) for semantic similarity. The comparative performance of the models is presented in Table 1.

Table 1. Performance of BART Models under Different Training Configurations

Model	Training Parameter	ROUGE1	ROUGE2	ROUGEL	ROUGELsum	BLEU	METEOR
BART-base	Epoch = 5	0.470	0.225	0.440	0.440	0.158	0.452
BART-base	Epoch = 10	0.464	0.219	0.431	0.430	0.177	0.470
BART-base	Learning rate = 8.375365110990092e-05, Batch size = 16 Epochs = 7 Weight decay = 0.0813999276676497	0.471	0.220	0.441	0.440	0.161	0.454
BART-base	Learning rate = 3e-5, Batch size = 16 Epochs = 20 Weight decay = 0.3	0.493	0.263	0.451	0.452	0.195	0.487
BART-large	Learning rate = 2e-5, Batch size = 16 Epochs = 20 Weight decay = 0.3	0.502	0.259	0.472	0.471	0.173	0.481
BART-large	Learning rate = 2e-5, Batch size = 16 Epochs = 20 Weight decay = 0.3 Preprocessing adds quotes	0.481	0.233	0.445	0.445	0.182	0.489

Analysis of the BART-base model reveals an interesting pattern concerning the number of epochs. An increase from 5 to 10 epochs (BART-base - E5 to BART-base - E10) generally led to a decrease in ROUGE metrics (e.g., ROUGE-Lsum dropped from 0.4407 to 0.4309). This decline in ROUGE, which measures overlap in words or phrases, might suggest that the model began to overfit the training data or produced less precise answers in capturing key information with uncontrolled epoch increases. The phenomenon of overfitting as the number of epochs increases is a well-documented issue in transformer-based models, as more epochs can lead to memorization of the training data rather than generalization (Armand et al., 2017). A similar effect is noted in models where excessive training leads to diminishing returns on the evaluation metrics (Vaswani et al., 2017).

Conversely, both BLEU (from 0.1581 to 0.1770) and METEOR (from 0.4528 to 0.4701) scores improved during this transition. The improvement in BLEU and METEOR, despite lower ROUGE scores, indicates that the model might be generating more fluent and semantically relevant summaries overall, even if they are less exact in specific keyword capture. This is further supported by an increase in brevity_penalty and length_ratio, the increase in brevity penalty and length ratio further indicates that the model produced longer responses closer to the reference length, which is typically favored by fluency metrics like BLEU (Ruma et al., 2023). This trade-off between fluency and precision highlights the complexity of training transformer-based models, where an increase in certain metrics can come at the cost of others, especially when the model shifts towards generating more fluent, albeit less precise, answers (Farea et al., 2025). signifying that the model produced longer responses closer to the reference length, which metrics like BLEU often favor for fluency.

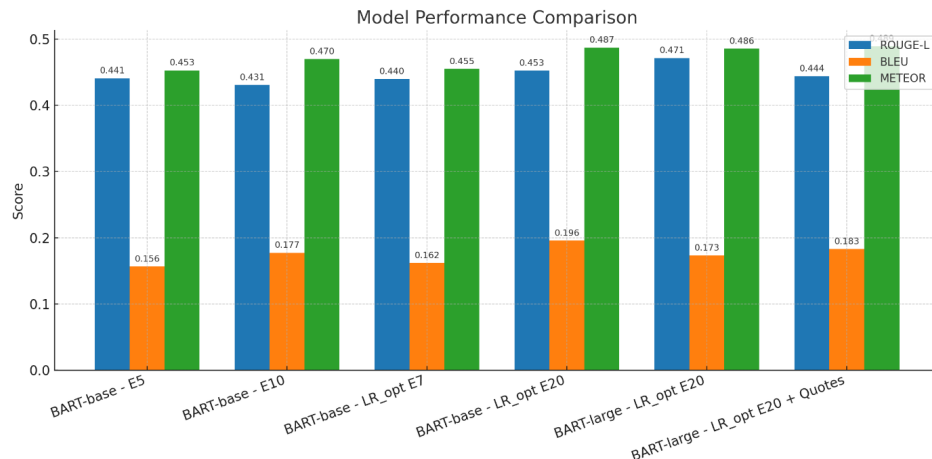


Figure 4. Comparison Transformer-Based Models for MultiSportQA

Furthermore, employing Optuna for hyperparameter optimization has demonstrated a marked superiority over conventional techniques such as grid search in transformer models (Raffel et al., 2020). The integration of reduced learning rates, increased epochs, and efficient weight decay facilitates the model's convergence to a more optimal solution, as demonstrated in this study. This aligns with findings (M. Zhang & Shang, 2022), who demonstrated that fine-tuning using optimal configurations leads to improved model performance on tasks like QA and summarization.

Based on Figure. 4, the model performance comparison graph, increasing the number of epochs and optimising hyperparameters has a significant effect on improving the ROUGE-L, BLEU, and METEOR metric scores. The BART-large model with the best hyperparameter configuration (LR_opt E20) shows a clear improvement in performance, especially in the ROUGE-L metric (0.471), which reflects the model's ability to capture key information and relationships between text elements. However, despite improvements in ROUGE-L and METEOR, the BLEU metric decreased slightly in the larger model (BART-large), indicating a trade-off between information accuracy and text fluency. Statistically, this shows that the BART-large model tends to be more thorough in its understanding of context, but at the same time, there is a sacrifice in terms of alignment with reference n-grams, which is related to fluency and sentence structure alignment.

The enhancement in performance demonstrates that hyperparameter optimization is essential for increasing model efficiency across multiple evaluation metrics. In the context of sports, models with this level of performance are highly useful in creating automated question-and-answer systems to support training, education, and sports management. More accurate models, such as those obtained with optimal settings, can provide more relevant and precise responses regarding the laws, regulations, and strategies needed by coaches, athletes, and sports event management. In addition, this technology can be used to improve access to information about sports, provide deeper insights on tactics, and assist in data-driven decision-making, which is vital in increasing athlete performance and the effectiveness of sports event management.

Discussion

A significant finding is the positive impact of hyperparameter optimization on the BART-base model. The configuration employing a learning rate of $3e-5$, batch size of 16, 20 epochs, and a weight decay of 0.3 (BART-base - LR_opt E20) demonstrated substantial performance gains across key metrics compared to simpler epoch settings or other learning rate configurations. This optimized setup achieved the highest ROUGE-Lsum for the base model (0.4526), the highest overall BLEU score (0.1956), and the highest METEOR for the base model (0.4871). This improvement confirms that the optimization of training parameters, such as learning rate, batch size, and weight decay, plays a crucial role in maximizing the performance of transformer-based models, as highlighted by recent studies on hyperparameter tuning (Deb et al., 2022). Hyperparameter optimization helps to balance the model's learning process, preventing overfitting and ensuring more accurate results (Liu et al., 2019).

When compared to the BART-base model, the BART-large model, fine-tuned with optimized parameters (BART-large - LR_opt E20), generally exhibited superior performance, especially in ROUGE metrics. It achieved the highest ROUGE-Lsum score (0.4715), indicating a stronger ability to capture essential information from the source text. The larger model size of BART-large enables it to learn richer, more complex language representations, leading to deeper contextual understanding, as described (Liu et al., 2019). Larger models like BART-large tend to outperform smaller variants in accuracy and relevance, particularly in tasks that require contextual depth and semantic understanding (Zhu et al., 2024). However, it is also worth noting that while BART-large achieved higher ROUGE scores, its BLEU score was slightly lower than the optimized BART-base (0.1730 vs. 0.1956). This suggests that while larger models may better capture key content, they may sometimes sacrifice fluency or match less precisely at the n-gram level.

The effect of preprocessing on model performance of "adding quotes" preprocessing to the BART-large model (BART-large - LR_opt E20 + Quotes) yielded mixed results. While there was a slight decrease in ROUGE-Lsum (from 0.4715 to 0.4459), the BLEU score increased (from 0.1730 to 0.1828), and METEOR reached its highest value (0.4891). This suggests a trade-off where the added quotes altered the structure or emphasis of the generated text, making the model less precise in matching n-grams (ROUGE), but enhancing overall fluency and semantic relevance, as indicated by the improvement in BLEU and METEOR. Preprocessing methods, such as text formatting, can have nuanced yet significant impacts on model performance, and their effects may vary depending on the evaluation metric used (Siino et al., 2024). These findings support previous work (Raffel et al., 2020), which noted that preprocessing techniques could improve semantic understanding but also affect precision in the model's output.

Beyond quantitative performance, the observed results indicate several practical implications for the deployment of Multiple Closed-Domain QA systems in sports contexts. Multiple Closed-Domain Question Answering framework offers practical benefits for sports stakeholders by enabling reliable and regulation-compliant information access. For coaches and athletes, the system supports informed decision-making in training and competition preparation by reducing the risk of misinformation derived from open-domain responses. For sports managers and federation officials, it facilitates consistent interpretation of regulatory

documents and operational guidelines, improving administrative efficiency. In sports education contexts, the framework provides a scalable AI-assisted learning tool grounded in authoritative domain knowledge, opening opportunities for standardized, context-aware digital learning and knowledge dissemination across multiple sports domains.

The strong performance within a well-defined closed domain like "AD/ART Indonesia's Sports Federation" suggests that transformer-based models hold significant potential for mitigating hallucinations by providing answers grounded in relevant and constrained knowledge sources. This supports the argument that fine-tuned models, when trained on domain-specific datasets, are less likely to generate irrelevant or incorrect answers, a common issue in open-domain QA systems. Additionally, the ability to minimize hallucinations is particularly beneficial in specialized chatbots, where the goal is to generate precise and accurate responses rather than relying on broader generalization. The findings of this study align with RoBERTa (Liu et al., 2019), who emphasized the effectiveness of fine-tuning and domain-specific optimization in reducing hallucination and improving answer accuracy.

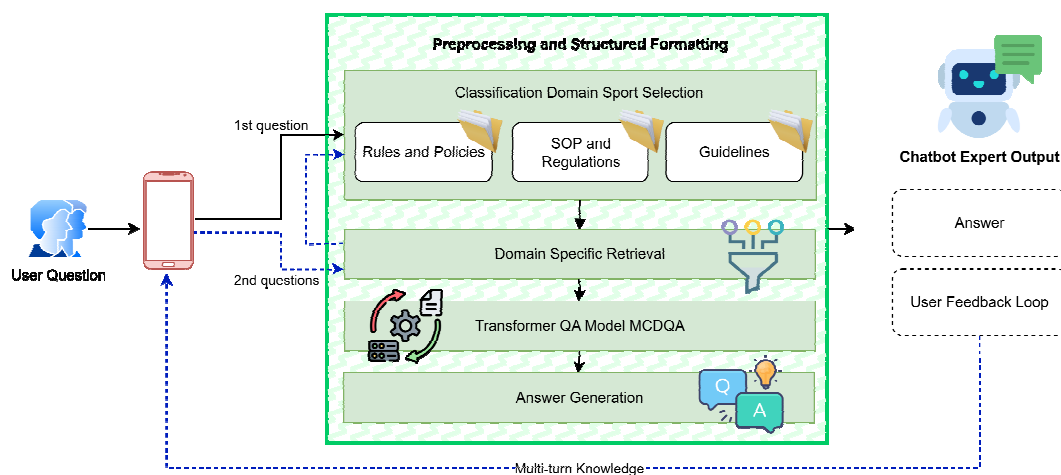


Figure 5. Chatbot Framework for Training and Education Sports

This research, as depicted in Figure 5, has substantial implications for social and sports education, particularly in advancing digital literacy among coaches, athletes, and sports administrators. In modern sporting environments, the capacity to acquire, evaluate, and utilize digital information accurately has emerged as an essential skill. The suggested AI-driven Question Answering (QA) system directly addresses this requirement by delivering structured, dependable, and readily accessible sports information, thus facilitating the cultivation of knowledgeable and digitally proficient sports professionals.

The implementation of an AI-driven QA system provides an effective remedy for persistent issues in sports education, especially in obtaining reliable information regarding laws and regulations, training methodologies, and health-related policies. These domains of knowledge are essential for guaranteeing safe, equitable, and efficient sports practices. By centralizing access to this information, the system diminishes reliance on fragmented or informal sources, which are frequently susceptible to inconsistency or inaccuracy, hence enhancing educational quality standards across sports disciplines.

The technology improves learning experiences for trainers and athletes by promoting self-directed and context-aware learning at the instructional level. Coaches can swiftly confirm regulatory specifics or training protocols, while athletes can attain a more lucid comprehension of rules and optimal practices that influence their performance and welfare. This accessibility facilitates ongoing learning and promotes reflective practice, hence enhancing performance outcomes and fostering more professional sports education environments from grassroots to top levels.

Engagement with essential stakeholders—such as sports federations, coaching groups, and educational institutions—is vital for preserving the relevance and legitimacy of the proposed system. These relationships guarantee the ongoing enhancement of the knowledge base with specialized insights, regulatory developments, and contextual expertise derived from practical experience. Engagement of stakeholders enhances trust in the system, promoting wider adoption and ensuring alignment with established sports governance frameworks.

Furthermore, stakeholder participation facilitates the iterative enhancement of the AI model to accommodate the varied requirements of distinct sports disciplines, competition tiers, and geographical settings. By integrating feedback from professionals and organizations, the system can tailor its responses to discipline-specific terminology, regulatory subtleties, and regional practices. This versatility is crucial for facilitating multi-domain sports knowledge while ensuring precision and contextual awareness across diverse applications.

Ultimately, empirical assessments of the AI-driven QA system—especially in training programs and school physical education environments—are crucial for determining its feasibility, scalability, and educational efficacy. Implementation in genuine learning environments facilitates rigorous evaluation of the system's impact on knowledge acquisition, decision-making quality, and instructional efficacy. The system's evolution, together with its incorporation into wider educational frameworks and extension into other areas, presents significant opportunities for enhancing sports culture, fostering educational innovation, and strategically applying artificial intelligence in sports and physical activity settings.

Conclusion

This study shows how Transformer-based Multiple Closed-Domain Question Answering (MCD-QA) systems help sports federations organize and regulate knowledge. The method shows that domain-specific retrieval, query categorization, targeted fine-tuning, and hyperparameter optimization reduce hallucination while maintaining high answer accuracy and semantic fluency through systematic experimentation. Optimised Transformer architectures can produce reliable and contextually relevant responses for high-stakes sports knowledge scenarios.

This research expands Question Answering literature by translating closed-domain QA models into a Multiple Closed-Domain framework that integrates multiple regulated knowledge domains and ensures factual consistency. It bridges generative capabilities and domain fidelity to provide empirical insights into Transformer optimization, filling a gap in previous studies that prioritize linguistic fluency over reliability in specialized sectors.

Sports stakeholders benefit from the proposed MultiSportQA framework, which helps athletes and coaches access official rules and regulations for informed decision-making. This AI-enhanced educational resource can also be used by sports science educators, students, managers, and health professionals to support administrative tasks, rehabilitation guidance, and policy analysis, reducing the risk of misinformation from open-domain systems.

This research affects sports governance and policy formulation by promoting consistent interpretation of federation policies, rule enforcement transparency, and sports organization communication. MCD-QA systems improve data-driven governance and policy execution by reducing misinformation and enforcing official documents.

Transformer-based MCD-QA systems show promise, but more research is needed. More sports federations and related fields like sports, health and nutrition should be covered in future efforts, with practical deployment assessments and user feedback integration essential for long-term effectiveness and adoption. It lays the groundwork for trustworthy AI-driven sports knowledge systems, aiding scientific inquiry and practical applications in sports sciences and policymaking.

Acknowledgment: We would like to express our gratitude to Ministry of Higher Education, Science, and Technology of the Republic of Indonesia programme DPPM for providing funding assistance under Grant Number 098/C3/DT.05.00/PL/2025, which enabled this research to be conducted. We also thank Universitas Pembangunan Nasional Veteran Jawa Timur for their support under Contract Number 10/UN63.8/LT-Kontrak/VI/2025. Additionally, we appreciate the guidance provided by Universitas Negeri Malang in carrying out this research. The authors declare no conflict of interest in this research.

Use of Generative AI statement: During the preparation of this scholarly work, the author used ChatGPT as a writing assistance tool, specifically for language refinement, clarification of sentence structure, and improvement of academic clarity and readability. After using this tool, the author independently reviewed, edited, and verified all content, and assumed full responsibility for the accuracy, originality, scholarly integrity, and academic substance of the published article.

Conflicts of interest: No conflicts of interest

References

- Aminuddin, R., Noor, M. H. M., Ilias, N. F., & Wahab, N. I. F. A. (2021). Framework for a mobile application with a chatbot to self-report injuries and carry out contact tracing for athletes and sports trainers. *2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, 1–6. <https://doi.org/10.1109/ISIEA51897.2021.9509978>
- Armand, J., Edouard, G., Piotr, B., & Tomas, M. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2, 427–431. <https://doi.org/https://doi.org/10.48550/arXiv.1607.01759>

- Babu, A., & Boddu, S. B. (2024). BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding. *Exploratory Research in Clinical and Social Pharmacy*, 13, 100419. <https://doi.org/10.1016/j.rcsop.2024.100419>
- Cavalcante, H. G., Soares, J. N., & Maia, J. E. B. (2021). Question Expansion in a Question-Answering System in a Closed-Domain System. *International Journal of Computer Applications*, 183(23), 975–8887.
- Deb, K., Zhang, X., & Duh, K. (2022). Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines. *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 51–61. <https://doi.org/10.18653/v1/2022.blackboxnlp-1.5>
- Dhivya, J. P., & Karnati, S. B. (2024). BuddyBot: AI Powered Chatbot for Enhancing English Language Learning. *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024*. <https://doi.org/10.1109/IATMSI60426.2024.10502595>
- Durmus, E., He, H., & Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5055–5070. <https://doi.org/10.18653/v1/2020.acl-main.454>
- Farea, A., Yang, Z., Duong, K., Perera, N., & Emmert-Streib, F. (2025). Evaluation of Question Answering Systems: Complexity of Judging a Natural Language. *ACM Computing Surveys*. <https://doi.org/10.1145/3744663>
- Jamroziński, S., & Markowska-Kaczmar, U. (2022). Semi-supervised classifier guided by discriminator. *Scientific Reports 2022 12:1*, 12(1), 1–17. <https://doi.org/10.1038/s41598-022-18947-6>
- Ji, T., Lyu, C., Jones, G., Zhou, L., & Graham, Y. (2022). QAScore—an unsupervised unreference metric for the question generation evaluation. *Entropy*, 24(11), 1514. <https://doi.org/10.3390/e24111514>
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Keiper, M. C., Fried, G., Lupinek, J., & Nordstrom, H. (2023). Artificial intelligence in sport management education: Playing the AI game with ChatGPT. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 33, 100456. <https://doi.org/10.1016/j.jhlste.2023.100456>
- Kesarwani, S., Titiksha, & Juneja, S. (2023). Student Chatbot System: A Review on Educational Chatbot. *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1578–1583. <https://doi.org/10.1109/ICOEI56765.2023.10125876>
- Koshy, N. R., Dixit, A., Jadhav, S. S., Penmatsa, A. V., Samanthapudi, S. V., Kumar, M. G. A., Anuyah, S. O., Vemula, G., Herzog, P. S., & Bolchini, D. (2023). Data-To-Question Generation Using Deep Learning. *2023 4th International Conference on Big Data Analytics and Practices, IBDAP 2023*. <https://doi.org/10.1109/IBDAP58581.2023.10271940>
- Lam, K. N., Le, N. N., & Kalita, J. (2020). Building a Chatbot on a Closed Domain using RASA. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 144–148. <https://doi.org/10.1145/3443279.3443308>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 2020-Decem*.
- Li, R., Wang, Y., Wen, Z., Cui, M., & Miao, Q. (2025). Different paths to the same destination: Diversifying LLMs generation for multi-hop open-domain question answering. *Knowledge-Based Systems*, 309(November 2024), 112789. <https://doi.org/10.1016/j.knosys.2024.112789>
- Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., Wang, W., & Zhang, M. (2023). *A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering*. <https://arxiv.org/abs/2311.07536v3>
- Liang, Y., Wang, J., Zhu, H., Wang, L., Qian, W., & Lan, Y. (2023). Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 4329–4343. <https://doi.org/10.18653/v1/2023.emnlp-main.263>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach. 1*. <http://arxiv.org/abs/1907.11692>
- Maini, Y., Jha, A., Jha, P., & Sharma, D. J. (2023). NORA – HealthCare Voice Based Chatbot. *International Journal for Research in Applied Science and Engineering Technology*, 11(1), 839–847. <https://doi.org/10.22214/ijraset.2023.48660>

- McBee, J. C., Han, D. Y., Liu, L., Ma, L., Adjeroh, D. A., Dong, X., & Hu, G. (2023). *Interdisciplinary Inquiry via PanelGPT: Application to Explore Chatbot Application in Sports Rehabilitation*. 1–12.
- Mcbee, J. C., Han, D. Y., Liu, L., Ma, L., Adjeroh, D. A., & Xu, D. (2024). *Assessing ChatGPT 's Competency in Addressing Interdisciplinary Inquiries on Chatbot Uses in Sports Rehabilitation: Simulation Study Corresponding Author : 10*, 1–9. <https://doi.org/10.2196/51157>
- Pratita, A., Tri Lathif Mardi, S., Arista, P., & Wibowo, A. (2025). ChatGPT in Education: Investigating Students Online Learning Behaviors. *International Journal of Information and Education Technology*, 15(3), 510–524. <https://doi.org/10.18178/ijiet.2025.15.3.2262>
- Pre, C. G. (2023). *Chatbot Generative Pre - trained Transformer and artificial intelligence in sports physical therapy and rehabilitation*. 61–62. <https://doi.org/10.4103/sjism.sjism>
- Priya, M. Y., Kamble, S. S., Shendre, S. P., Sridhar, S., & K, A. M. (2024). Megan - A Sports Chatbot using OpenAI APIs and Django Framework with Python. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–8. <https://doi.org/10.1109/I2CT61223.2024.10543499>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 1–67.
- Roy, S. C., & Manik, M. M. H. (2024). *Question-Answering System for Bangla: Fine-tuning BERT-Bangla for a Closed Domain*. <https://arxiv.org/abs/2410.03923v1>
- Ruma, J. F., Mayeesha, T. T., & Rahman, R. M. (2023). Transformer based Answer-Aware Bengali Question Generation. *International Journal of Cognitive Computing in Engineering*, 4(September), 314–326. <https://doi.org/10.1016/j.ijcce.2023.09.003>
- Saad, M., & Qawaqneh, Z. (2024). Closed Domain Question-Answering Techniques in an Institutional Chatbot. *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2024*. <https://doi.org/10.1109/ICECCME62383.2024.10796881>
- Shao, T., Guo, Y., Chen, H., & Hao, Z. (2019). Transformer-Based Neural Network for Answer Selection in Question Answering. *IEEE Access*, 7, 26146–26156. <https://doi.org/10.1109/ACCESS.2019.2900753>
- Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, 121, 102342. <https://doi.org/10.1016/j.is.2023.102342>
- Soni, A., Arora, R., Kumar, A., & Panwar, D. (2024). Evaluating Domain Coverage in Low-Resource Generative Chatbots: A Comparative Study of Open-Domain and Closed-Domain Approaches Using BLEU Scores. *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, 1–6. <https://doi.org/10.1109/ICEECT61758.2024.10738994>
- Suryanto, T. L. M., Wibawa, A. P., Hariyono, H., & Nafalski, A. (2023). Evolving Conversations: A Review of Chatbots and Implications in Natural Language Processing for Cultural Heritage Ecosystems. *International Journal of Robotics and Control Systems*, 3(4), 955–1006. <https://doi.org/10.31763/ijrcs.v3i4.1195>
- Suryanto, T. L. M., Wibawa, A. P., Hariyono, H., & Shili, H. (2025). Contextual Relevance-Driven Question Answering Generation: Experimental Insights Using Transformer-Based Models. *International Journal of Engineering, Science and Information Technology*, 5(4), 103–113. <https://doi.org/10.52088/ijesty.v5i4.989>
- Tay, W., Joshi, A., Zhang, X. (Jenny), Karimi, S., & Wan, S. (2019). *Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation* (pp. 52–60). <https://aclanthology.org/U19-1008>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-Decem(Nips)*, 5999–6009.
- Wibawa, A. P., Handayani, A. N., Dedes, K., Utama, A. B. P., Shakti, M. C., & Hernandez Collante, L. (2025). *Wulangreh and Wedhatama Translation Using Adam-NMT with Considering Unique Macapat Rules* (pp. 178–203). https://doi.org/10.1007/978-3-031-92977-9_12
- Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization. *37th International Conference on Machine Learning, ICML 2020, Part F16814*, 11265–11276.
- Zhang, M., & Shang, X. (2022). Chinese Short Text Classification by ERNIE Based on LTC_Block. *Wireless Communications and Mobile Computing*, 2022. <https://doi.org/10.1155/2022/1411744>
- Zhu, Z., Liu, L., Free, R. C., Anjum, A., & Panneerselvam, J. (2024). OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks. *Information Sciences*, 680, 121141. <https://doi.org/10.1016/j.ins.2024.121141>