MedBlock-Bot: A Blockchain-Enabled RAG System for Providing Feedback to Large Language Models Accessing Pediatric Clinical Guidelines

Mohamed Yaseen Jabarulla, Steffen Oeltze-Jafra,
Philipp Beerbaum and Theodor Uden

August 18, 2025

MedBlock-Bot: A Blockchain-Enabled RAG System for Providing Feedback to Large Language Models Accessing Pediatric Clinical Guidelines

Mohamed Yaseen Jabarulla*, Steffen Oeltze-Jafra*, Philipp Beerbaum†, Theodor Uden†
*Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School,
†Department of Pediatric Cardiology and Pediatric Intensive Care Medicine, Hannover Medical School,
Hannover, Germany.
Email: Jabarulla.Mohamed@mh-hannover.de, Uden.Theodor@mh-hannover.de

***Abstract*— Accessing reliable clinical knowledge quickly is an everyday challenge for clinicians. Large Language Models (LLMs) can assist healthcare professionals by providing this knowledge, but their responses often deviate from expert consensus or are not up to date necessitating reliable validation and possible correction. To address this, we introduce MedBlock-Bot, an interactive Streamlit-based system integrating a blockchain-enabled Retrieval-Augmented Generation (RAG) framework for expert-driven assessment and immutable feedback storage within a permissioned consortium network. Unlike traditional feedback mechanisms that may be altered or lost, MedBlock-Bot employs smart contracts to securely store and verify any feedback, ensuring transparency and auditability. We evaluated the system using three open-source LLMs—BioMistral, HippoMistral, and LLaMa 3.1—on clinical guideline interpretation for neonates with hypoplastic left heart syndrome. Human experts assessed model responses based on accuracy and relevance, revealing variations in adherence to the guideline knowledge. Additionally, deploying the blockchain component in a local permissioned environment (Ganache) ensured efficient transaction processing and tamper-proof feedback retrieval without gas cost concerns. Our results demonstrate the integration of blockchain for LLM feedback review enhancing trust, accountability, and structured knowledge retention. Clinicians can access past expert assessments for validation, while developers can leverage this feedback for potential model refinement. Taking the long-term impact into account this approach targets towards a reliable and dynamic representation of clinical knowledge and consensus. Open-Source Code: https://github.com/yaseen28/MedBlock-Bot**

Keywords*— *Medical language models, Blockchain, RAG, Pediatric cardiology, Clinical care.

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated significant potential in assisting healthcare professionals by interpreting medical documents, answering clinical queries, and by that providing decision support [1], [2]. These models, trained on vast corpora of medical and general knowledge, can streamline information retrieval and improve workflow efficiency. However, a critical challenge remains—LLMs often generate responses that do not strictly adhere to the established clinical consensus for example provided by guidelines, posing risks to patient safety [3]. Ensuring that these models align with current medical knowledge is essential for their reliable deployment in clinical practice. Furthermore, current changes in clinical knowledge or experience-based knowledge can hardly be provided in a structured way at all.

The integration of LLMs in healthcare has been widely explored, particularly in medical text summarization, and question-answering. Research has shown that models such as ChatGPT, BioBERT, and Med-PaLM can generate informative medical responses, but their accuracy and reliability remain a concern, especially in critical healthcare applications [3], [4], [5]. Several studies have highlighted the limitations of LLMs in following strict medical protocols, often producing hallucinated, incomplete, or misleading responses [3]. To mitigate these issues, Human-in-the-Loop systems, such as PubMedQA and MedQA [6], [7] have been developed, incorporating expert feedback to refine model outputs. However, these approaches rely on static, pre-annotated datasets, which limit adaptability to evolving medical guidelines. To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as a technique that enables LLMs to dynamically retrieve relevant information from external sources before generating responses [8]. Unlike standard LLMs that rely solely on pre-trained knowledge, RAG improves contextual accuracy and reduces hallucinations by incorporating real-time domain-specific documents. This approach has shown promise in clinical question-answering and medical document interpretation [9]. However, existing RAG-based systems lack mechanisms for tracking expert feedback, ensuring auditability, and maintaining accountability in clinical environments. Addressing these gaps is essential for deploying LLMs in real-world clinical applications.

Blockchain technology has recently emerged as a reliable solution for ensuring data integrity in healthcare applications, including medical records, clinical trials, and secure data sharing [10], [11]. Traditional feedback mechanisms, such as centralized databases or manual submission via email, lack transparency, are prone to tampering, and do not provide an auditable record of expert assessments [12], [13]. In a permissioned blockchain environment [14], access is limited to authorized participants, such as clinicians and researchers, ensuring secure submission, review, and validation of expert feedback while maintaining confidentiality, transparency, and auditability. Public blockchains allow open participation, whereas permissioned blockchains enforce access control through identity verification and role-based permissions [13].

Additionally, smart contracts [15], [16] autonomously execute predefined rules, regulating data access and submission processes while ensuring that expert feedback on LLM outputs remains verifiable, tamper-proof, and immutable. Despite these advantages, the integration of blockchain for managing expert feedback on LLM-generated clinical responses remains underexplored. To bridge this gap, we formulate the following research question:

"How can we design a system that securely manages expert feedback on LLM-generated responses to clinical guidelines, ensuring transparency within a healthcare consortium and providing an auditable feedback loop for future model refinement?"

In response, we propose MedBlock-Bot, an interactive, blockchain-enabled RAG system designed to evaluate the performance of LLMs in interpreting medical documents, particularly clinical guideline. As a case study, we evaluate our system using European clinical guidelines for hypoplastic left heart syndrome (HLHS), where LLM-generated responses are assessed by human experts for adherence to clinical guidelines [17]. This evaluation ensures the system's effectiveness in real-world applicability. The key contributions of this research are as follows:

1. We designed and implemented a RAG-driven clinical query processing system that enhances contextual accuracy in LLM-generated responses by retrieving relevant information from clinical guidelines before generating outputs.
2. We evaluated the performance of open-source LLMs—BioMistral, HippoMistral, and LLaMa 3.1—by assessing their adherence to clinical guidelines using expert validation.
3. We developed a blockchain-based feedback storage framework that utilizes a permissioned blockchain to securely store, audit, and ensure the tamper-proof integrity of clinician feedback.
4. We simulated clinician feedback by submitting corrected responses with ratings and identifiers while analyzing gas usage, transaction efficiency, and smart contract execution in a local Ethereum test environment.
5. We implemented an interactive dual-mode dashboard using Streamlit, enabling clinicians to review generated responses and developers to leverage structured feedback for potential model refinement.

This research demonstrates the integration of blockchain into the expert feedback loop of LLMs potentially enhancing trust, accountability, and the auditability of LLM-generated responses. This approach lays the foundation for the continuous improvement of AI-driven clinical decision support systems. Ultimately, it contributes to the long-term goal of developing a reliable, evolving clinical guideline-based LLM for medical applications.

II. MATERIALS AND METHODS

A. Ethereum Blockhain Testnet

The healthcare industry demands secure, transparent, and tamper-proof systems for managing sensitive data [18]. Traditional data storage methods often face challenges such as security breaches, lack of transparency, and difficulties in ensuring data integrity. Blockchain technology addresses these issues by maintaining a distributed, immutable ledger where recorded transactions cannot be altered or deleted without consensus from network participants [12]. This makes blockchain a promising solution for securely managing expert feedback on AI-generated medical responses, ensuring accountability and auditability.

For our implementation, we selected the Ethereum blockchain [19], [20] due to its strong developer ecosystem, well-established security mechanisms, and support for smart contracts—self-executing agreements with conditions written directly into code. Smart contracts autonomously enforce predefined rules, ensuring that expert feedback is securely collected, stored, and auditable at any time [15]. In our case, we deployed and tested these contracts using Remix IDE, a widely used tool for Ethereum-based smart contract development [21]. This ensures that all feedback transactions are recorded accurately and can be audited at any time.

Given the sensitive nature of medical data, we opted for a permissioned blockchain rather than a public network. This ensures that only authorized participants, such as medical experts and hospital personnel, can access and submit feedback, enhancing security and compliance with healthcare regulations. For development and testing, we used Ganache [21], a personal Ethereum blockchain that simulates the Ethereum network in a controlled environment. Ganache enables rapid deployment and testing of smart contracts without incurring transaction fees.

In the public Ethereum network, gas refers to the computational cost required to execute operations, such as transactions or executing smart contracts. Each operation has a specific gas cost, and users must pay for gas using Ether (ETH), Ethereum's native cryptocurrency [19], [20]. In our permissioned blockchain setup, there is no real ETH cost—gas usage is simulated to analyze system efficiency. This approach allows us to optimize performance and evaluate transaction costs without financial constraints, making blockchain a viable solution for secure and scalable hospital-based implementations.

B. Open-Source Large Language Models

In our study, we evaluated three open-source LLMs—BioMistral, HippoMistral, and LLaMa 3.1—for their ability to interpret medical guidelines stored in PDF format [22], [23], [24] . This selection was informed by our previous work with MedDoc-Bot, which enabled clinicians to upload medical guidelines and choose from various LLMs, including Meditron, MedAlpaca, Mistral, and LLaMA-2 [9]. Our findings indicated that LLaMA-2 and Mistral exhibited reasonable fidelity and relevance in processing clinical queries, establishing it as a promising candidate for healthcare applications. Building on these results, we selected BioMistral and HippoMistral, which are specifically designed for medical interpretation, and included LLaMA 3.1, an updated model in the LLaMA series—known for its enhanced language capabilities.

BioMistral [22] and HippoMistral [23] are fine-tuned models derived from Mistral LLM [25], optimized using domain-specific datasets that include various medical texts. BioMistral is fine-tuned specifically to generate coherent and contextually relevant responses to complex medical inquiries, while HippoMistral is tailored to handle diverse clinical scenarios, leveraging a wide range of healthcare datasets. Both models are built upon the Mistral architecture, maintaining strong performance while being specialized for medical applications. The LLaMA series model [24], LLaMA 3.1, builds on the strengths of its predecessors, integrating advanced techniques for improved language understanding. With a focus on handling large contexts and generating high-quality text, LLaMA 3.1 is a robust candidate for clinical interpretation tasks. For our experiments, we utilized the 7B-parameter versions of BioMistral and HippoMistral, as well as the 8B-parameter Instruct version of LLaMA 3.1, ensuring the models could process complex queries while maintaining efficiency in resource-constrained environments.

The three models are pre-quantized in GGUF (GPT-Generated Unified Format) using the llama.cpp Python library with 4-bit (Q4) quantization. This pre-quantization enables efficient processing across different computing environments, accommodating CPU or GPU limitations while ensuring high-performance analysis [26].

C. Dataset and Clinical Use Case

To evaluate our blockchain system, we curated 20 clinically relevant questions from the Guidelines for the Management of Neonates and Infants with HLHS [17]. These guidelines offer standardized recommendations for diagnosing, treating, and managing HLHS, a critical and complex congenital heart defect [27]. For model performance evaluation, we selected 20 questions from a specific chapter on imaging modalities, allowing a feasible, in-depth review by a pediatric specialist with more than four years of experience in pediatric cardiology.

D. Methodology

As depicted in Figure 1, the MedBlock-Bot system integrates two primary components: a RAG system and a blockchain-enabled feedback system, both designed to optimize the generation and validation of clinical responses.

1) RAG System: The RAG system begins when a user uploads clinical guideline documents in PDF format. These documents are processed through LangChain, which creates embeddings that capture their semantic meaning [28]. These embeddings are stored in a FAISS (Facebook AI Similarity Search) vector database, enabling efficient retrieval [29]. When a user submits a query, it is also converted into a numerical representation, allowing the system to conduct a semantic search within the FAISS database. The most relevant information retrieved is then fed into the selected LLM, such as Llama 3.1, BioMistral or HippoMistral, ensuring that the generated response is contextually aligned with the query and supported by authoritative clinical sources. This RAG approach minimizes the risk of hallucinations and enhances the precision of the AI-generated responses.

2) Blockchain Feedback System: In parallel, the blockchain-enabled feedback system is essential for maintaining the accuracy and integrity of the responses generated by the LLM. Medical professionals review responses, provide feedback, and assign scores ranging from 0 (poor accuracy and relevance) to 6 (high accuracy and strong clinical alignment), with intermediate scores reflecting varying degrees of correctness and coherence. These evaluations are recorded on a local test blockchain (Ganache) as tamper-proof transactions, guaranteeing accountability and transparency in AI-driven medical assessments. Users (developers or AI engineers) can query the blockchain to retrieve expert evaluations, enabling them to analyze feedback trends and refine future model iterations.

To facilitate this mechanism, a Solidity-based smart contract (Refer Sub-Section E) securely stores structured feedback, including the original query, LLM-generated response, clinician corrections, reviewer names, scores, and timestamps for auditing purposes. Key functions within the contract allow for feedback submission, retrieval, counting, and management. By integrating the RAG system with blockchain, MedBlock-Bot ensures responses are accurate, reliable, and used by clinicians to review past feedback and by developers for potential future model fine-tuning, assuming expert feedback as the foundation.

Our local blockchain setup prioritizes computational efficiency. When medical experts review and score LLM-generated responses, their feedback is stored as immutable transactions. We analyze simulated gas consumption to assess its impact on system scalability and processing speed, ensuring optimal performance and resource allocation for potential large-scale deployment.

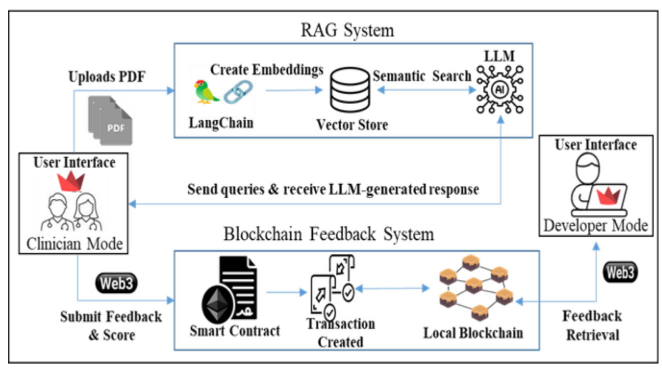

Fig. 1: Overview of a proposed Streamlit- powered Medblock-Bot system combining RAG and Blockchain-based feedback storage.

3) Medblock-Bot User Interface: The MedBlock-Bot dashboard provides an interactive interface with two distinct operational modes. The *clinician mode* enables healthcare professionals to select LLM models, configure parameters, upload relevant clinical documents, submit queries, and provide feedback on generated responses. The *developer mode* allows access to stored feedback, predefined prompt templates, and insights from clinician evaluations, supporting model fine-tuning and performance improvements. The dashboard is built using Streamlit [30], offering an intuitive and user-friendly web-based interface. Web3.py [21] is used to establish secure interactions with the blockchain, allowing clinicians to submit feedback and developers to retrieve it for further analysis. This integration ensures seamless communication between the front-end interface and the blockchain network, enhancing system efficiency and usability.

E. Smart Contract Functions

The smart contract, written in Solidity, was developed using Remix IDE and deployed via MetaMask [31] on a local Ethereum test network (Ganache). MetaMask, a browser-based Ethereum wallet, was configured to connect with Ganache, which provided test Ether for executing and simulating transactions. This setup ensured a secure and cost-free environment for testing smart contract functionalities before real-world deployment. The key functions of the smart contract include:

• *submitFeedback()* – This function stores the query, LLM response, clinician's corrected response, clinician's name, score, and timestamp on-chain.

•*getFeedbackCount()* – This function returns the total number of submitted feedback records.

•*getFeedback(index)* – This functions retrieves individual feedback records based on their index in the blockchain.

To optimize transaction processing, gas costs are minimized during feedback submission, with each transaction generating a unique Ethereum transaction hash for verification. The contract ensures data integrity by leveraging Ethereum's Proof-of-Authority (PoA) consensus mechanism [14] within the local Ganache environment.

F. Blockchain Deployement and Workflow

The MedBlock-Bot system was deployed, and simulations were performed on a local machine with the following hardware configuration: a 12th Gen Intel i9 processor, 64 GB DDR4 RAM, and an NVIDIA GeForce RTX 3090 GPU. The blockchain component is tested using Ganache, which simulates blockchain interactions without real Ether, ensuring cost-free testing and development. As illustrated in Figure 1, the blockchain feedback workflow follows these steps:

1. A clinician submits feedback via the MedBlock-Bot Web3-enabled interface after interacting with the LLM.
2. The smart contract records the feedback on-chain and generates a transaction hash as proof of submission.
3. Stored feedback can be retrieved using the *getFeedback()* function for validation and auditing.
4. The immutable nature of blockchain prevents unauthorized modifications, maintaining data integrity and accountability.

The combination of Streamlit for interface design and Web3.py for blockchain integration ensures an efficient and secure feedback management system.

III. Experimental Results and Discussion

This section presents the analysis of the proposed system along with the evaluation of LLMs in a clinical setting: LLaMa 3.1, BioMistral, and HippoMistral. Each model was tested using 20 predefined clinical queries, generating 60 responses. One human expert assessed these responses for accuracy and relevance, providing structured evaluations. To evaluate the blockchain evaluation process, we simulated a feedback submission by adding exemplary corrected responses, ratings, and anonymized clinician identifiers (example: clinician A, B, C) for 60 model-generated outputs. These feedback entries were securely stored on a local Ethereum blockchain, ensuring the system's ability to reliably process, retrieve, and verify expert feedback while maintaining data integrity.

1) LLM Evaluation and Human Assessment

The models are accessed using the Llama.Cpp library for efficient inference, with hyperparameters adjusted as follows: Temperature: Temperature = 0.35, Max Tokens = 200, and Top-P = 0.75**.**

Each model-generated response was evaluated by human experts based on accuracy (alignment with clinical guidelines) and relevance (contextual appropriateness and completeness). Experts assigned scores from 0 to 6, which were then normalized to a 0–100% scale for consistent comparison and visualization (Figure 2).

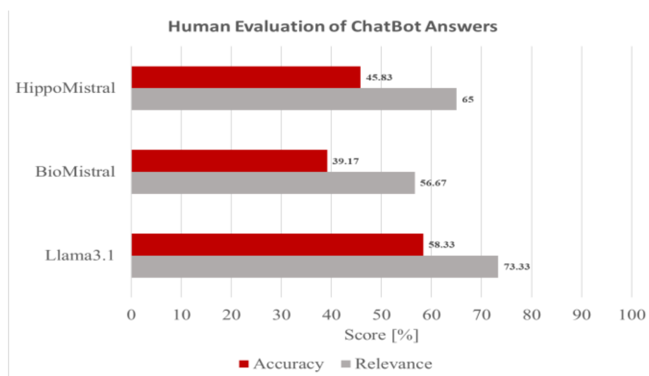

Fig. 2. Average relevance and accuracy scores for Llama 3.1, BioMistral, and HippoMistral, Models, highlighting their performance in relevance and accuracy.

The figure 2 results indicate that LLaMa 3.1 outperformed the other models, achieving higher accuracy and relevance scores across 20 queries. However, BioMistral and HippoMistral performed competitively in many cases. BioMistral was rated as good as Llama 3.1 or better regarding relevance in 9/20 questions and regarding accuracy in 8/20 questions. HippoMistral was rated as good as Llama 3.1 or better regarding relevance in 10/20 questions and regarding accuracy in 10/20 questions. Overall performance remained limited, particularly in cases requiring specialized medical knowledge, precise numerical reasoning, or nuanced interpretations of complex details from the guideline. These findings highlight a key challenge in applying LLMs to specialized medical domains—while they can capture broad medical knowledge, they struggle with fine-grained, high-stakes decision-making. Future improvements could include fine-tuning models on more pediatric guideline datasets and incorporating clinician-in-the-loop validation to iteratively refine responses.

2) Blockchain-Based Feedback Storage and Retrieval

a) Transaction Processing Time: Figure 3 compares the average submission and confirmation times across the three models. Submission time, representing the duration to send feedback to the blockchain, remained consistent across all models (0.075–0.078 seconds) due to the controlled, low-latency environment. However, confirmation time, which measures transaction validation and finalization, varied. Llama 3.1 confirmed transactions faster (~0.0236 seconds), while BioMistral and HippoMistral took slightly longer (~0.045–0.048 seconds), likely due to differences in feedback size and processing complexity.

These variations may stem from differences in block generation intervals, computational overhead, and storage demands, where larger feedback entries require additional processing. While minor in a permissioned blockchain setting, these differences could become more significant as the dataset scales, highlighting the need for efficient transaction handling and storage optimization.

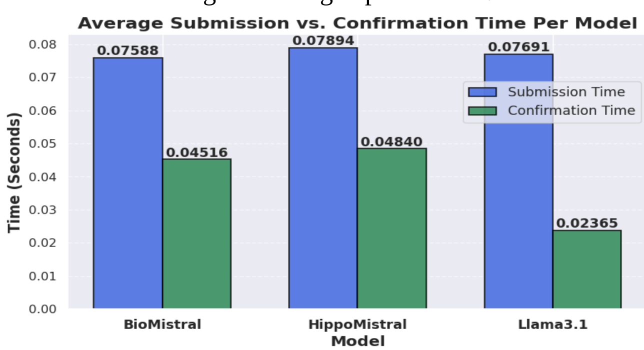

Fig. 3. Average submission and confirmation times for different models, highlighting the efficiency of feedback transaction processing within the blockchain system.

b) Gas Consumption and Storage Costs: Figure 4 compares gas consumption across different models, representing the computational effort required to store feedback on the blockchain. Llama 3.1 exhibited the highest gas usage (~1.14M units per transaction), while BioMistral and HippoMistral consumed approximately 900K units each. Although LLaMa 3.1 generally produced more accurate responses, its model-generated outputs may have been longer or required additional metadata storage, leading to higher gas costs. Furthermore, the simulated feedback process have introduced variations in correction length, affecting transaction sizes. In a private blockchain environment, gas costs are not a limiting factor as they are on Ethereum's public mainnet. However, efficient resource management remains crucial for scalability, particularly as feedback data grows over time.

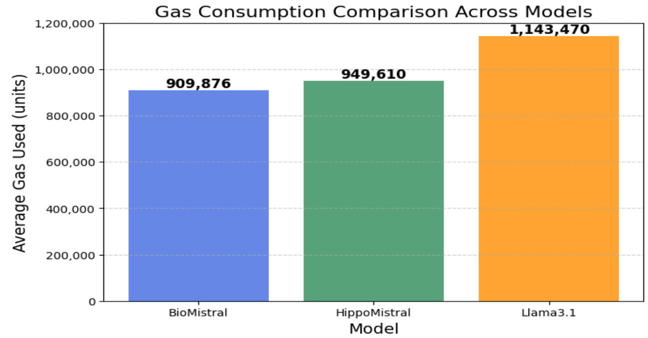

Fig. 4. Average gas consumption per feedback transaction across different models, providing insights into the computational efficiency and cost-effectiveness of the smart contract execution.

c) Feedback Retrieval Efficiency: Figure 5 illustrates the relationship between blockchain retrieval time and the number of stored feedback transactions. The retrieval time remained below 3 seconds for small datasets (≤ 5 transactions) but increased linearly as the number of stored entries grew. For instance, retrieval took 5.6 seconds for 25 transactions and 9.9 seconds for 50 transactions, eventually peaking at 11.3 seconds for 60 queries. While these times remain acceptable for real-time clinical applications, larger datasets may introduce performance bottlenecks. To mitigate this, implementing optimized indexing mechanisms (e.g., Merkle Trees), off-chain caching, and parallel processing could enhance efficiency, ensuring smooth data access for model evaluation and refinement [10].

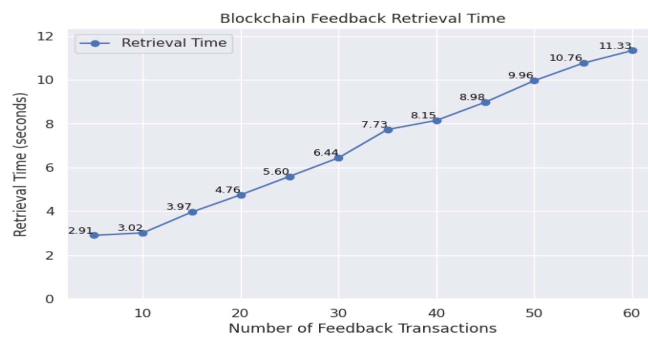

Fig. 5. The retrieval times for stored feedback transactions, demonstrating the system's ability to quickly access expert assessments while maintaining data integrity and immutability.

3) Limitations and Future Work

Our permissioned blockchain eliminates common issues of public blockchains, such as congestion and gas fee volatility, by operating in a closed consortium. Unlike Ethereum's Proof-of-Work (PoW), we use an efficient consensus mechanism (PoA) to improve transaction speed and scalability. However, our Ganache-based simulation does not fully account for real-world deployment challenges like regulatory compliance and integration with hospital networks. Additionally, expert engagement remains crucial for maintaining response accuracy, and optimizing smart contract execution is necessary for reducing computational overhead. A limitation of our system is the manual feedback integration, where the developer able to retrieves and formats feedback for fine-tuning. Future iterations will incorporate

Reinforcement Learning from Human Feedback (RLHF) [3] for automated feedback processing, allowing continuous model improvement. Expanding the expert pool and integrating with electronic health records (EHRs) will enhance clinical applicability. Future work will also focus on improving data privacy through federated learning and addressing regulatory compliance.

IV. CONCLUSION

This study introduced MedBlock-Bot, a blockchain-enabled RAG system designed to enhance AI-driven clinical assessments. Using European HLHS guidelines as a case study, we evaluated LLaMa 3.1, BioMistral, and HippoMistral, assessing their adherence to clinical protocols through expert validation. While LLaMa 3.1 demonstrated moderate interpretation accuracy, fine-tuning on complex cases might further improve performance. To ensure transparent and auditable expert assessments, we developed a permissioned blockchain framework that securely stores clinician feedback, reinforcing trust, accountability, and data integrity. Blockchain performance simulations confirmed efficient transaction processing without financial constraints. Additionally, an interactive dashboard allows clinicians to review responses and developers to refine models using structured feedback. This work establishes a foundation for enhancing guideline-based LLMs in clinical decision support. Future research should focus on EHR integration, retrieval optimization, and automated model refinement. Real-world deployment in hospital networks will further evaluate scalability, regulatory compliance, and clinician adoption.

REFERENCES

[1] G. Feder, R. Grol, M. Eccles, and J. Grimshaw, "Using clinical guidelines," *Bmj*, vol. 318, no. 7185, pp. 728–730, 1999, doi: 10.1136/bmj.318.7185.728.

[2] Z. Al Nazi and W. Peng, "Large Language Models in Healthcare and Medical Domain: A Review," *Informatics*, vol. 11, no. 3, p. 57, 2024, doi: 10.3390/informatics11030057.

[3] S. Tian et al., "Opportunities and challenges for ChatGPT and large language models in biomedicine and health," *Brief. Bioinform.*, vol. 25, no. 1, pp. 1–13, 2024, doi: 10.1093/bib/bbad493.

[4] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[5] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[6] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, "Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding," pp. 1–13, 2023, [Online]. Available: http://arxiv.org/abs/2305.12031.

[7] Y. Kim, J. Wu, Y. Abdulle, and H. Wu, "MedExQA: Medical Question Answering Benchmark with Multiple Explanations," pp. 167–181, 2024, [Online]. Available: http://arxiv.org/abs/2406.06331

[8] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," pp. 1–21, 2023, [Online]. Available: http://arxiv.org/abs/2312.10997.

[9] M. Y. Jabarulla, S. Oeltze-Jafra, P. Beerbaum, and T. Uden, "MedDoc-Bot: A Chat Tool for Comparative Analysis of Large Language Models in the Context of the Pediatric Hypertension Guideline," *in 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, FL USA, 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10781509.

[10] C. Agboo, Q. Mahmoud, and J. Eklund, "Blockchain technology in healthcare: a systematic review," *Healthcare*, vol. 7, no. 2, p. 56, 2019.

[11] M. Hölbl, M. Kompara, A. Kamišalić, and L. N. Zlatolas, "A systematic review of the use of blockchain in healthcare," *Symmetry (Basel)*, vol. 10, no. 10, 2018, doi: 10.3390/sym10100470.

[12] H. Li, L. Zhu, M. Shen, F. Gao, X. Tao, and S. Liu, "Blockchain-Based Data Preservation System for Medical Data," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–13, 2018, doi: 10.1007/s10916-018-0997-3.

[13] M. Jabarulla and H.-N. Lee, "Blockchain-Based Distributed Patient-Centric Image Management System," *Appl. Sci.*, vol. 11, no. 1, p. 196, 2020.

[14] R. Kumar, P. Kumar, G. Gupta, A. K. M. Islam, and M. Shorfuzzaman, "Permissioned Blockchain and Deep Learning for Secure and Efficient Data Sharing in Industrial Healthcare Systems," *IEEE Trans. Ind. Informatics*, vol. 18, no. 11, pp. 8065–8073.

[15] M. Sookhak, M. R. Jabbarpour, N. S. Safa, and F. R. Yu, "Blockchain and smart contract for access control in healthcare: A survey, issues and challenges, and open issues," *J. Netw. Comput. Appl.*, vol. 178, no. November 2020, p. 102950, 2021, doi: 10.1016/j.jnca.2020.102950.

[16] A. Khatoon, "A Blockchain-Based Smart Contract System for Healthcare Management," *Electronics*, vol. 9, no. 1, p. 94, 2020.

[17] N. Alphonso et al., "Guidelines for the management of neonates and infants with hypoplastic left heart syndrome: The European Association for Cardio-Thoracic Surgery (EACTS) and the Association for European Paediatric and Congenital Cardiology (AEPC) Hypoplastic Left Heart Sy," *Eur. J. Cardio-thoracic Surg.*, vol. 58, no. 3, pp. 416–499, 2020, doi: 10.1093/ejcts/ezaa188.

[18] A. AbuHalimeh and O. Ali, "Comprehensive review for healthcare data quality challenges in blockchain technology," *Front. Big Data*, vol. 6, 2023, doi: 10.3389/fdata.2023.1173620.

[19] G. Wood, "ETHEREUM: A secure decentralised generalised transaction ledger," *Tech. Rep.* Accessed: Jan. 02, 2019. [Online]. Available: https://gavwood.com/paper.pdf.

[20] V. Buterin, "Ethereum White Paper." Accessed: Jan. 02, 2019. [Online]. Available: https://github.com/ethereum/wiki/wiki/White-Paper.

[21] S.M. Jain, Introduction to Remix IDE and Truffle. In: A Brief Introduction to Web3. *Apress*, Berkeley, CA, 2023. doi: https://doi.org/10.1007/978-1-4842-8975-4.

[22] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," 2024, [Online]. Available: http://arxiv.org/abs/2402.10373.

[23] E. C. Acikgoz et al., "Hippocrates: An Open-Source Framework for Advancing Large Language Models in Healthcare," pp. 1–29, 2024, [Online]. Available: http://arxiv.org/abs/2404.16621.

[24] A. Grattafiori et al., Llama Team, "The Llama 3 Herd of Models | Research - AI at Meta," 2024, [Online]. https://arxiv.org/abs/2407.21783.

[25] A. Q. Jiang et al., "Mixtral of Experts," 2024, [Online]. Available: http://arxiv.org/abs/2401.04088.

[26] H. Shen, H. Chang, B. Dong, Y. Luo, and H. Meng, "Efficient LLM Inference on CPUs," no. *NeurIPS*, pp. 1–6, 2023, [Online]. Available: http://arxiv.org/abs/2311.00502.

[27] F. J. A et al., "Hypoplastic Left Heart Syndrome," J. Am. Coll. Cardiol., vol. 59, no. 1_Supplement, pp. S1–S42, Jan. 2012, doi: 10.1016/j.jacc.2011.09.022.

[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. *Process. 9th Int. Jt. Conf. Nat. Lang. Process.* Proc. Conf., pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[29] M. Douze et al., "The Faiss library," 2024, [Online]. Available: http://arxiv.org/abs/2401.08281.

[30] "Streamlit," Github, 2019. https://github.com/streamlit/streamlit.

[31] Lee, W. Getting Started with Smart Contract and using the MetaMask Chrome Extension. In Beginning Ethereum Smart Contracts Programming; Apress: Berkeley, CA, USA, 2019; ISBN 978-1-4842-5086-0. doi: https://doi.org/10.1007/978-1-4842-5086-0_5.