CONDUCTSCIENCE
ACADEMIC PUBLISHING

Authors: Mr Tyler J Smith, Amogh Karnik, MD, Jonathan Hourmozdi, MD, James D. Thomas, MD, Adrienne S. Kline, MD, PhD

Type: Original Research

Keywords: Artificial Intelligence, Large Language Models, Retrieval Augmented Generation, Healthcare Disparity, Cardiovascular Health, Patient Education, Health Literacy, Mobile Health (mHealth)

Edited By:

Received: 2025-02-21

Published: 2025-07-23

ARTICLE IN PRESS
CONDUCTSCIENCE

LipidLlama: A Multilingual AI Chatbot for Personalized Cardiovascular Risk Assessment

Abstract

Background

Cardiovascular disease is the leading cause of death worldwide. Many patients with cardiovascular disease struggle to understand their risk factors and medical test results due to health literacy limitations and/or inadequate resources for personalized education. Existing digital solutions often provide generic, one-size-fits-all information with varying degrees of medical accuracy.

Methods

LipidLlama addresses this gap by integrating a rule-based AI system—adapted from a validated cardiovascular risk assessment tool—and a chatbot powered by a Large Language Model (LLM) enhanced with retrieval augmented generation (RAG). To improve document retrieval, a custom query-only adapter was trained on synthetic query corpus document pairs. Response quality was independently evaluated by three board-certified physicians (two cardiologists and one internist) who rated 30 responses to synthetically generated, realistic patient queries using a 5-point Likert scale across four dimensions: correctness, conciseness, comprehensiveness, and comprehensibility. Rater reliability was assessed using generalizability theory.

Results

The trained adapter improved top-k document retrieval accuracy from 67% to 80%. Responses received consistently high clinical ratings across all dimensions (mean composite score = 18.21, 95% CI: 17.80-18.62) and demonstrated strong reliability metrics across raters (generalizability coefficient $E_\rho^2 = 0.88$; dependability coefficient $\Phi = 0.84$).

Conclusion

LipidLlama provides clinically grounded, personalized explanations in response to cardiovascular health questions. With further clinical validation, this mobile health application has the potential to enhance health literacy and minimize provider burden, significantly improving access to preventive cardiovascular care, particularly in underserved communities.

Keywords: Artificial Intelligence, Large Language Models, Retrieval Augmented Generation, Healthcare Disparity, Cardiovascular Health

Background

Cardiovascular disease remains the leading cause of mortality globally. While its determinants are complex and multifactorial, modifiable lifestyle factors such as smoking status, diet, and physical activity significantly influence disease progression [1]. Despite available interventions for risk reduction, patients often struggle to implement these changes effectively due to limitations in health literacy and current patient education approaches.

Personalized patient education demonstrably improves health outcomes, particularly in chronic conditions like cardiovascular disease [2]. However, delivering individualized health literacy content creates substantial burden for already time-constrained physicians [3]. This personalization requires not only extended patient interactions but also additional time to compile patient-specific educational materials matching disease status, literacy level and language. Furthermore, physicians spend approximately one hour daily responding to patient queries through electronic health record messaging [4]. As the population ages, the burden of chronic cardiovascular disease on health systems will only increase [5], making the current paradigm of physician driven personalized health education untenable due to a scaling issue.

As a result, many patients are instead directed to general-purpose health resources, such as Northwestern Medicine's Health Library. While these repositories provide comprehensive medical information, they often fall short of meeting individual patient needs. Dense content, complex medical terminology, language barriers, and lack of personalization create significant obstacles, making it harder for patients to navigate these resources and take meaningful steps toward managing their condition. Large Language Models (LLMs) offer a potential solution for scaling

ARTICLE IN PRESS
JOURNAL SCIENCE

personalized education without adding to physicians' workload [6]. However, current implementations primarily rely on general-purpose models that lack integrated medical expertise [7]. Without access to expert-vetted clinical knowledge and patient-specific health data, these models struggle to provide accurate, tailored guidance. These barriers underscore the need for an advanced AI-solution capable of delivering truly personalized patient education.

LipidLlama meets this need through a novel approach combining validated cardiac risk assessment with AI-powered education tailored to individual patient health, knowledge, and language preferences. The mobile application guides patients through a brief survey about their lipid panel results and general health status, generating personalized risk assessments and actionable recommendations. Patients can then engage with an AI-powered chatbot, grounded in expert-curated health education materials, to ask questions in their preferred language at an appropriate comprehension level. This paper details the development and initial clinical validation of LipidLlama, a mobile health solution that delivers personalized cardiovascular education.

Methods

Technology

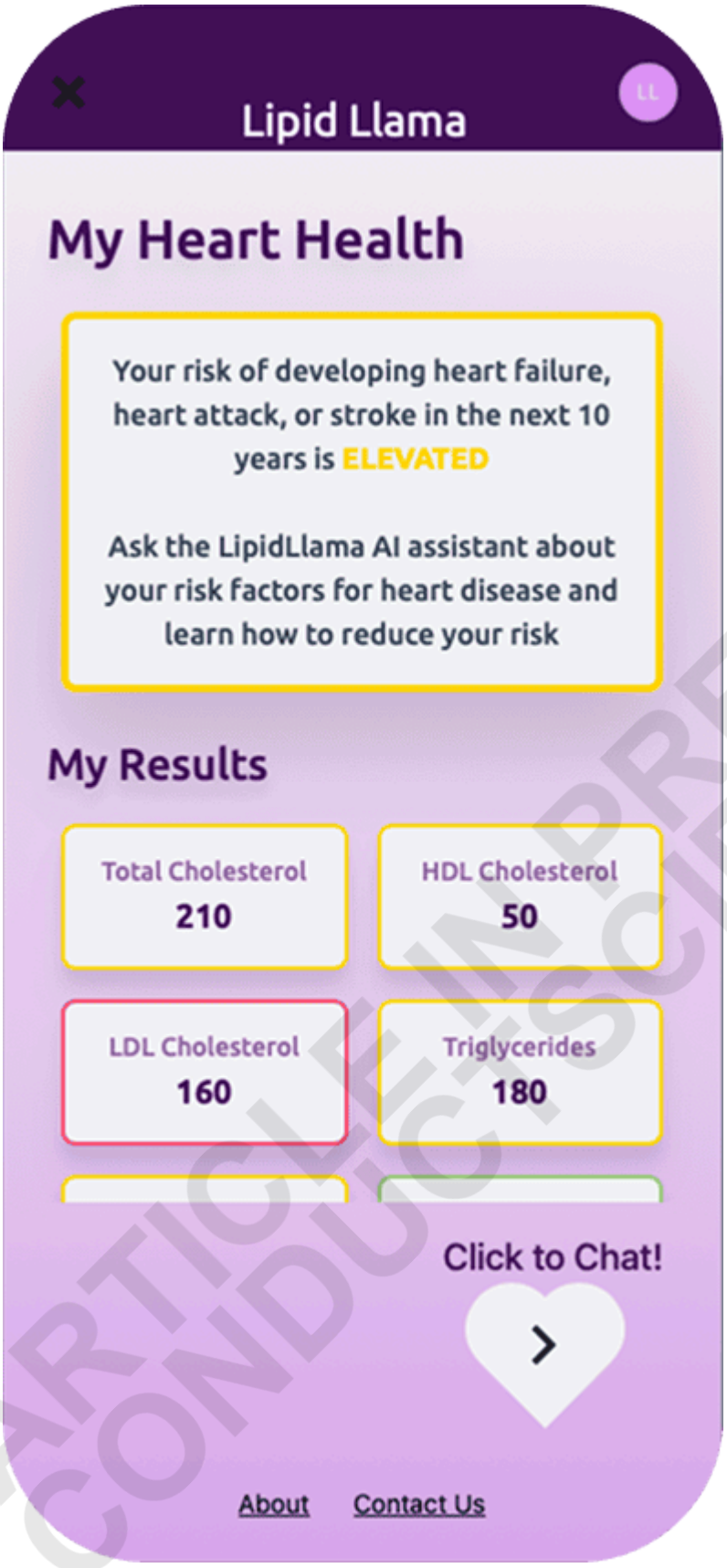
LipidLlama is a mobile application designed to deliver personalized cardiovascular risk assessment and education. It combines rule-based and generative AI to convert patient data into actionable feedback. The frontend uses modern web frameworks [8][9][10] to facilitate seamless, multilingual interaction, while the backend AI services [11][12][13] are optimized for deployment on consumer-grade infrastructure, such as an NVIDIA RTX 4090 (24GB) graphics card.

Rules-Based AI

LipidLlama’s rules-based AI system processes patient data to generate cardiac risk scores and lifestyle recommendations. This system is built upon the simplified PREVENT equations for 10-year Atherosclerotic Cardiovascular Disease (ASCVD) risk estimation [14][15], which are a set of previously validated gender specific regression models. Patients are guided through an educational health survey that collects key variables required by the PREVENT equations: total, LDL and HDL cholesterol, systolic blood pressure, estimated glomerular filtration rate (eGFR), diabetes status, smoking status, hypertension medication, and statin medication. Each variable is explained in clear patient-facing language.

Health survey inputs are encoded into numerical inputs (i.e. values for cholesterol, eGFR, and blood pressure results and 1 or 0 for binaries such as diabetes status) and processed by the appropriate PREVENT regression equation based on gender. The model applies the published regression coefficients without modification to calculate the individual’s 10-year ASCVD risk percentage. The resulting risk percentage is categorized in accordance with consensus guidelines for primary prevention of atherosclerotic heart disease into three tiers: low (<5%), elevated (5-7.5%), and high risk (>7.5%) [16]. Individual survey parameters are also classified according to standardized clinical ranges (normal, elevated, high). Actionable insights are generated from these results (e.g., “quit smoking”). Figures 1 and 2 illustrate a representative risk assessment for a synthetic female patient and the rules-based AI workflow including the PREVENT equation used for ASCVD risk calculation.

ARTICLE IN PRESS
CONDUCTSCIENCE



ARTICLE IN PRESS
CONDUCT SCIENCE

Figure 1 LipidLlama Home Screen.

LipidLlama home screen displaying the results of the rules-based AI system adapted from the 10-year ASCVD risk score for a synthetic patient."

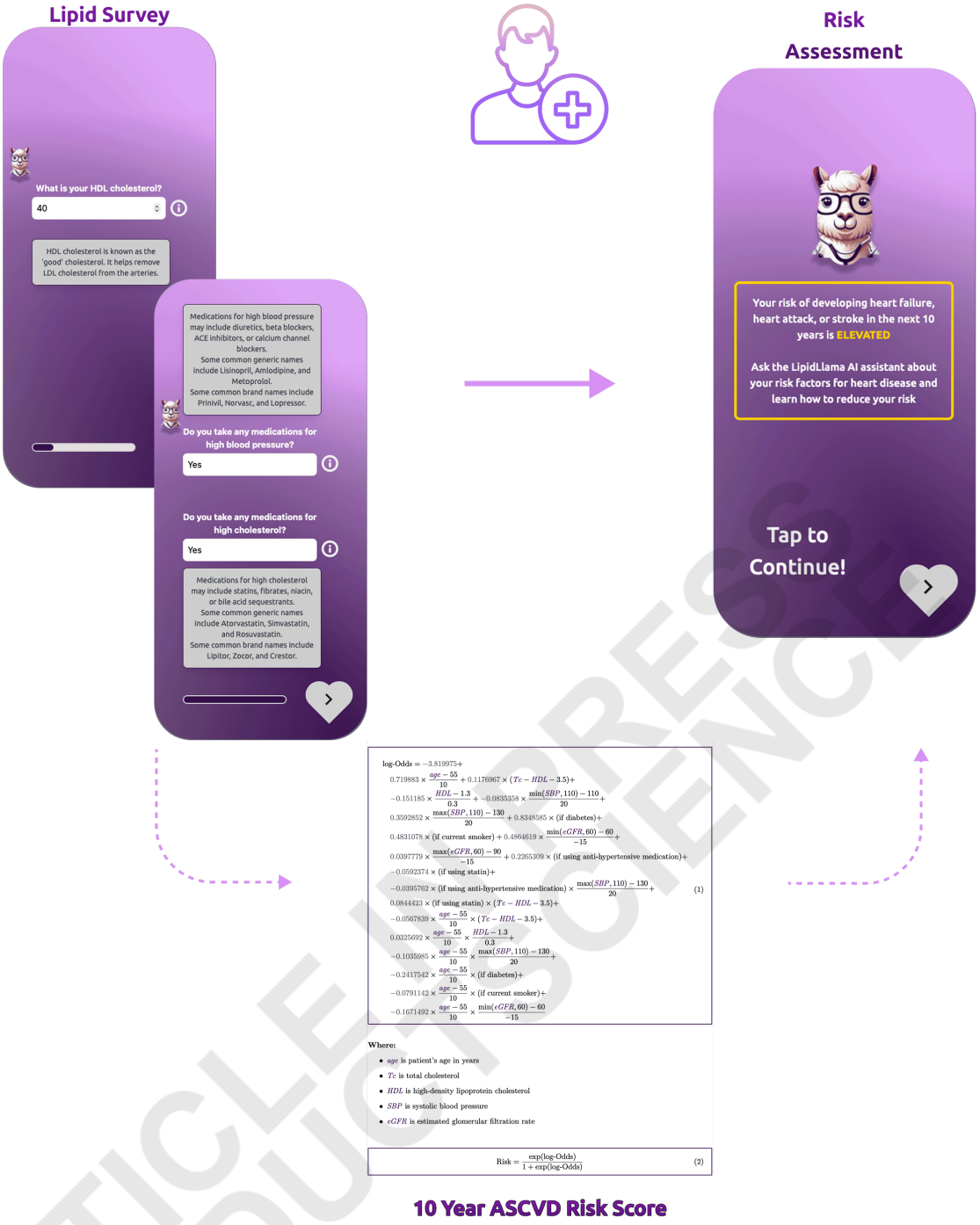


Figure 2 Rules-based AI system.

Patients respond to a lipid survey adapted from the 10-year ASCVD risk score. The system evaluates responses using predefined rules derived from the ASCVD regression model and normal value ranges, generating a personalized risk assessment.

Chatbot

Rules-based AI extracts specific information but can only give generalized feedback as possible rule enumerations grow exponentially. Large Language Models (LLMs), like ChatGPT [17], offer greater flexibility, allowing users to ask any question and receive a personalized

response. However, despite their impressive capabilities, LLMs can sometimes generate incorrect or misleading answers—a phenomenon known as hallucination [18]. This occurs because LLMs operate probabilistically: they predict words based on patterns learned from vast amounts of

internet text rather than retrieving concrete facts. While LLMs excel at mastering grammar, structure, and compositionality, their knowledge of specific facts is inherently approximate [19][20][21]. For critical uses, such as a medical chatbot like LipidLlama, hallucinations pose a serious challenge.

However, hallucination can be mitigated by providing the model with expert documents to ground its responses. Rather than relying on an LLM's internal knowledge to generate an answer, this approach ensures that the necessary information is explicitly supplied. The role of AI shifts to composing a well-formed response based on verified reference text—a task LLMs excel at. This method, known as Retrieval-Augmented-Generation (RAG) [22], enhances the accuracy of LLM-generated answers by anchoring them in verifiable, expert content.

RAG Pipeline

A RAG pipeline consists of a corpus of expert documents, an embedding model, and an LLM. The document corpus encompasses information relevant to any potential query. For LipidLlama, this corpus was curated from approximately 100 patient education articles in the Northwestern Medicine Health Library [23], covering lipid testing, cardiovascular health, medications, and lifestyle recommendations. Every document was written and reviewed by medical experts with independent review by the authors to ensure that the corpus contained only accurate and up-to-date information.

Document relevance to a given query is computed through semantic vector representations. The BGE embedding model (another type of AI language model) transforms both documents and patient queries into normalized vectors in a shared semantic space [24][25]. Relevance is quantified by calculating the Euclidean distance between vectors, which allows for fast document search [26] while maintaining angular distance properties (i.e. how aligned are documents and queries in the semantic space). This approach ensures that conceptually similar query-document pairs have similarity scores closer to 0 (i.e. are parallel in sematic space) than unrelated content.

Previous work suggests that query and document embeddings may reside in misaligned semantic spaces, limiting retrieval accuracy. However, applying a linear transformation to query embeddings, such as a query-only adapter, can enhance retrieval [27]. To improve query alignment to LipidLlama's document corpus, a lightweight query-only adapter was trained using contrastive loss on 2,000 synthetic question-document pairs. Evaluation was conducted on a held-out test set of 500 question-document pairs. For each query, the top 100 documents were returned via the similarity search method described previously. Similarity scores were standardized via z-score normalization, and documents exceeding 2 standard deviations below the mean (scores closest to 0 before z-score normalization) were considered retrieved. Threshold decision is hyperparameter chosen to balance accuracy against the number of documents returned, an inherent computational limitation for the LLM. Retrieval accuracy was measured as the proportion of queries for which the correct document appeared among the z-score normalized set post-thresholding. Comparisons were made to baseline and to BGE-Reranker-v2-M3 [28][29], a reranking model designed to improve document retrieval without the need for additional training.

Chatbot Response Generation and Translation

The technical details of the LipidLlama chatbot including the RAG system are available in figure 3. An instruction-finetuned version of Llama-3.1-8B serves as the LLM backbone [30]. Patient-specific cardiovascular health information, retrieved documents, and queries are compiled into a prompt and passed to the LLM for response generation. To improve accessibility, a multilingual translation pipeline to integrates the Opus-MT family of translation models before document

AHEAD OF PRESS
CONDITIONAL ACCEPTANCE

retrieval and after LLM response generation [31][32]. Spanish, Mandarin, Arabic, and Polish were chosen based upon Northwestern Medicine's patient demographics and available computational resources.

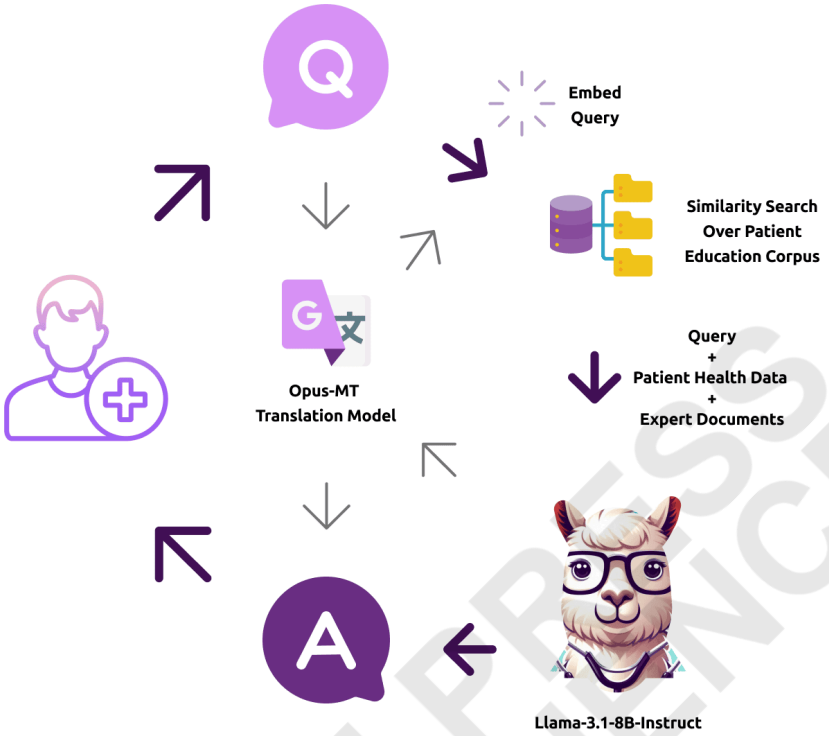


Figure 3 LipidLlama Chatbot Technical Design Diagram

Pipelines for retrieval-augmented generation (RAG) and translation. The outer RAG pipeline is highlighted in bolded purple arrows while the inner translation loop is indicated by gray arrows. A patient submits a query, which is embedded and used for similarity search over a corpus of patient education documents. The query, retrieved information, and relevant patient health data are passed to the Llama-3.1-8B-Instruct model for response generation. If necessary, queries and responses are translated using a language-specific Opus-MT model (gray arrows)

Physician Evaluation Procedure

To assess clinical accuracy, LipidLlama responses were evaluated by three board-certified physicians (2 cardiologists: JDT and JH; 1 internist: AK), each independently reviewing chatbot-generated answers to a set of synthetic patient-query pairs. All generated responses were generated using the full RAG pipeline, incorporating the trained query-only adapter described above. Synthetic patient profiles were created by sampling lipid values and disease prevalences from established population ranges. Corresponding queries were drawn from an author-reviewed, held-out test set of 238 synthetically generated questions. From this pool, 30 patient-question-response triplets were randomly sampled for formal evaluation.

Each response was rated on a 5-point Likert scale across four dimensions: "correct", "concise", "comprehensive", and "comprehensible". This rating scheme was adapted from a previously validated framework for evaluating ChatGPT responses on similar patient education topics [33]. The fully crossed design (3 raters x 30 questions x 4 dimensions) enabled decomposition of

ARTICLE IN PRESS
CONDUCT SCIENCE

variance across multiple sources. Generalizability theory (G-Theory) was applied to estimate the reliability of ratings. Unlike classical test theory, G-Theory can distinguish multiple sources of variance simultaneously, raters, questions, dimensions, and their interactions, enabling robust analysis of the generalizability of cardiologists' ratings [34][35][36].

Results

Document Retrieval Performance

The query-only adapter improved retrieval performance over the baseline embedding model. Both approaches returned a similar number of documents below z-score threshold (Baseline Mean = 4.85, 95% CI: 4.70-5.00; Query-Adapter: Mean = 5.02, 95% CI: 4.82-5.21), while the adapter achieved a substantial gain in accuracy. The proportion of queries for which the correct document was retrieved from the top-k documents increased from 67% (baseline) to 80% (query-only adapter), a 13 percentage point improvement. The query-only adapter still demonstrated superior performance when compared to the reranking model, which achieved 70% accuracy, yielding a 10 percentage point improvement.

Physician Evaluation

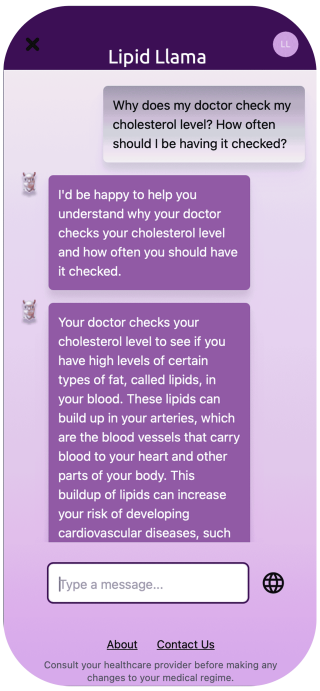
LipidLlama successfully generated completed responses for 28 of the 30 test questions. The remaining two responses requested clarifying information (e.g. "What cholesterol medication are you taking?"), which the physician raters deemed appropriate. However, these two cases were excluded from the final quantitative analysis.

Chatbot outputs demonstrated high clinical quality. Mean ratings (with 95% confidence intervals) were as follows: correct mean = 4.76 (95% CI: 4.66-4.88), conciseness mean = 4.46 (95% CI: 4.31-4.61), comprehensive mean = 4.37 (95% CI: 4.22-4.52), and comprehensible mean = 4.62 (95% CI: 4.47-4.77). Combination into single composite score (maximum 20) was equally strong with mean = 18.21 (95% CI: 17.80-18.62). Rater reliability of the composite score was strong with generalizability coefficient ($E\rho^2$) of 0.88 and a dependability coefficient (Φ) of 0.84, where both coefficients range from 0 to 1, with scores above 0.8 indicating very strong reliability.

Error categorization was conducted qualitatively. Of 84 individual ratings across 28 answered queries, 19 received correctness scores below 5, with only one instance rated 3 or lower. Review of these cases revealed that deductions reflected minor omissions or imprecise ordering of information rather than factual inaccuracies.

Figure 4 highlights representative results from LipidLlama by displaying two patient queries for a synthetic patient with hypertension, smoking history, and a high ASCVD risk score of 8.5%. The responses, including a Spanish-language translation, demonstrate LipidLlama's ability to provide culturally responsive, patient-specific recommendations.

ARTICLE IN PRESS
UNDER PEER REVIEW



I'd be happy to help you understand why your doctor checks your cholesterol level and how often you should have it checked.

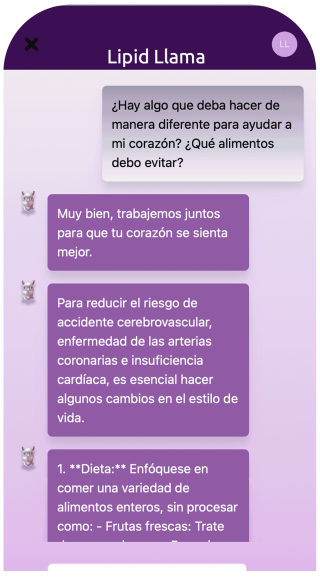
Your doctor checks your cholesterol level to see if you have high levels of certain types of fat, called lipids, in your blood. These lipids can build up in your arteries, which are the blood vessels that carry blood to your heart and other parts of your body. This buildup of lipids can increase your risk of developing cardiovascular diseases, such as heart disease, stroke, and heart failure.

Your doctor checks your cholesterol level regularly to:
1. See if your levels are within a healthy range.
2. Monitor any changes in your levels over time.
3. Make informed decisions about your treatment plan, if necessary.

The American Heart Association recommends that adults have their cholesterol level checked at least once every five years, starting at age 20. However, if you have a history of heart disease, diabetes, or high blood pressure, your doctor may recommend more frequent checks.

In your case, since your 10-year risk of developing stroke, coronary artery disease, and heart failure is 8.5%, which is too high, your doctor is likely to want to check your cholesterol level more frequently to monitor any changes and make adjustments to your treatment plan as needed.

It's great that you're taking proactive steps to manage your risk factors, including working with your healthcare provider to manage your blood pressure and quitting smoking. Remember to discuss your cholesterol level with your doctor and ask any questions you may have about your results.
Read More at Northwestern Health Library: Cholesterol in the Blood



ARTICLE IN PRESS
CONFIDENTIAL

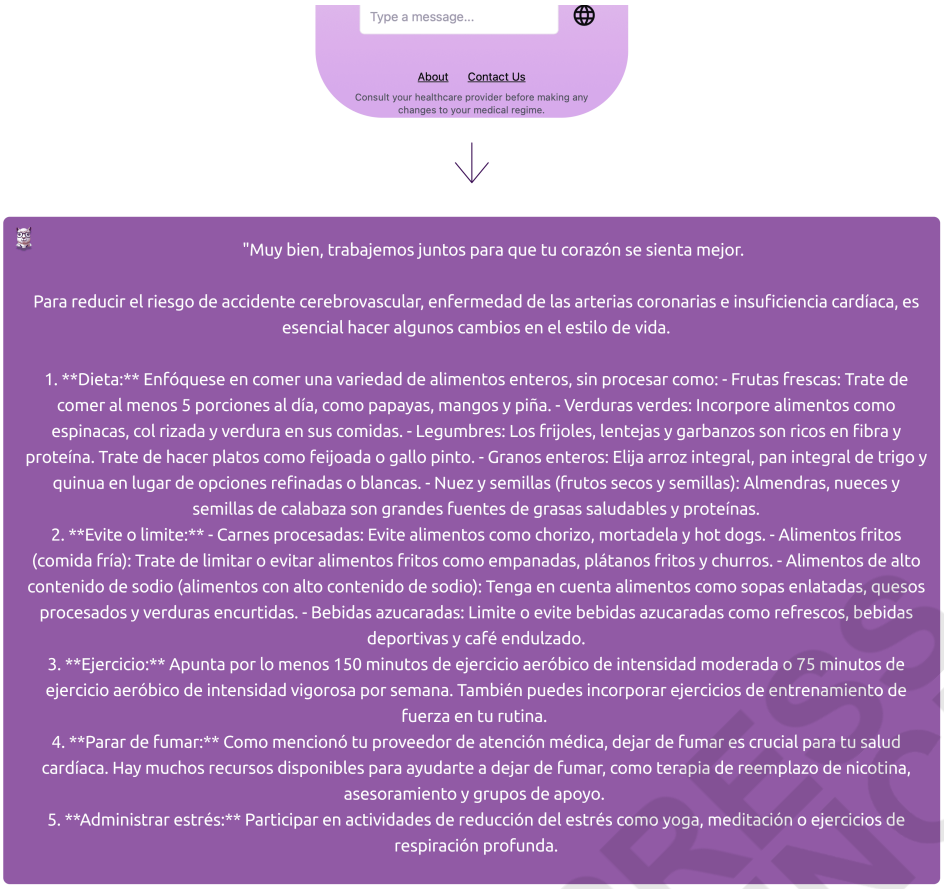


Figure 4 Example LipidLlama Patient Interactions and Culturally Responsive Translations

LipidLlama's response generation for a synthetic patient with hypertension, history of smoking, and an ASCVD risk score of 8.5%. The chatbot provides personalized recommendations based on patient-specific details retrieved from its rules-based AI system. The top panel displays an English-language response explaining cholesterol management and cardiovascular risk factors. The bottom panel highlights LipidLlama's translation capabilities, demonstrating a culturally responsive Spanish-language response about dietary and lifestyle modifications.

Discussion

LipidLlama demonstrates a novel integration of rules-based and generative AI to deliver medically accurate, patient-specific cardiovascular risk assessment and education. By grounding LLM output in expert-reviewed patient education documents and structured patient health data, the system addresses known limitations of generative models, hallucination, while maintaining flexible natural language interaction. This hybrid RAG framework is well-suited to address the challenges of generative AI in patient-facing clinical applications where accuracy, interpretability, and trust are all important.

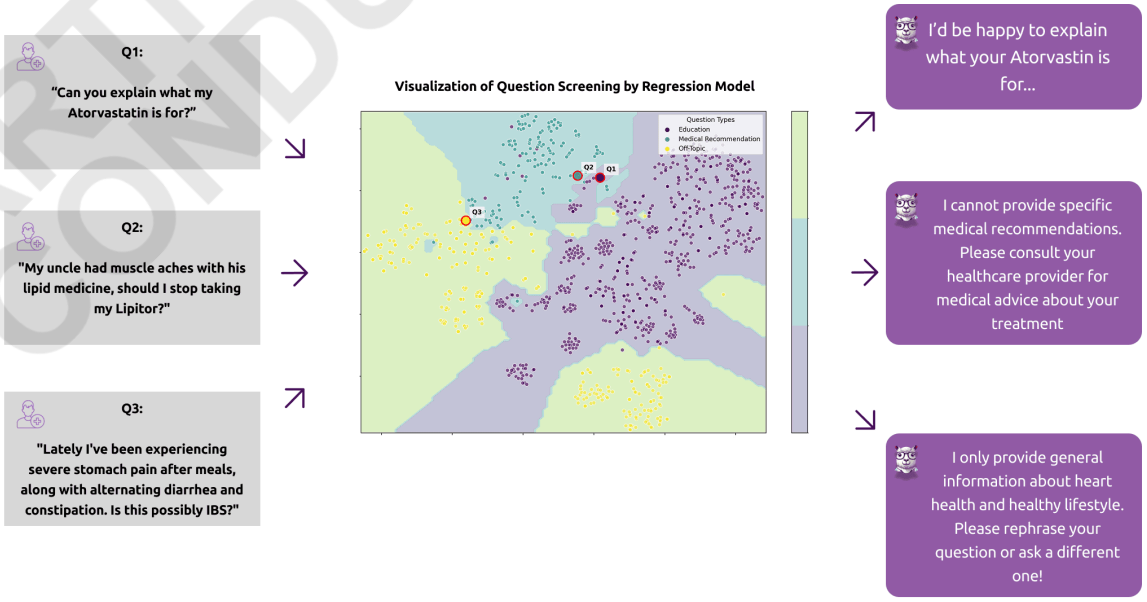
This study demonstrates the technical strength and clinical validity of LipidLlama's outputs. The query-only adapter significantly improved top-k document retrieval accuracy by 13 percentage points over baseline (80% vs. 67%), addressing a key challenge in RAG systems. Notably, this lightweight approach outperformed a reranking model that required 100× more parameters and achieved this gain without the computational burden of full parameter fine tuning, using only 2,500 query-document pairs. Importantly, improvement in retrieval accuracy enhances the grounding of generative responses in relevant source material. Retrieval is challenging in narrow corpora such as cardiovascular patient education documents, where substantial topic overlap creates many

semantically similar yet distinctive documents. In such cases, a “missed” document does not imply incorrect or conflicting output as every document was thoroughly vetted to ensure clinical accuracy. Instead, such cases involve semantic neighbors that provide clinically equivalent information. By improving alignment between patient queries and document embeddings, the adapter increases the likelihood of retrieving the most contextually appropriate information, supporting stronger, more grounded responses.

Board-certified physicians, including two cardiologists and an internist, independently evaluated LipidLlama’s clinical performance across four dimensions. Generations received high mean scores across all measures, correctness (4.76), conciseness (4.46), comprehensiveness (4.37), and comprehensibility (4.62), yielding a strong composite score (18.21/20). Reliability analysis using G-Theory revealed strong generalizability ($E\rho^2 = 0.88$) and dependability ($\Phi = 0.84$) coefficients indicating that both relative and absolute question scoring was strong across raters. These metrics account for the multiple sources of random effects variance (rater, item, and interaction), demonstrating that the observed scoring patterns are robust despite the subjectivity of rater measurement and small sample size. Qualitative review of low-scoring responses showed that deductions stemmed from minor omissions or imprecise phrasing rather than factual inaccuracies. This clinical validation, combined with the strong retrieval scores, provides an initial evaluation of LipidLlama’s ability to provide personalized cardiovascular patient education.

AI Safety and Ethics

Generative AI is the core engine behind the conversational capability of LipidLlama, raising critical questions about safety, bias, and scope. A key mitigating tool against potentially harmful responses is a linear regression classifier that screens user inputs before document retrieval (Figure 5). The classifier categorizes queries into three classes: patient education questions (answerable), medical recommendation requests (declined), and off-topic medical questions (declined). Trained on 900 synthetic, clinically realistic patient queries reviewed and verified by the authors, the classifier achieved a 98.47% binary F1 score (patient education vs. not), with 97.72% sensitivity and 98.92% specificity on 112 held-out test queries.



ARTICLE IN PRESS
CONDUCTSCIENCE

Figure 5 LipidLlama AI Safety

This classifier provides a robust safeguard against potentially inappropriate responses. It ensures responses to queries like "*Should I discontinue my statin due to muscle aches?"* direct users to consult their healthcare provider without offering medical advice. Similarly, off-topic queries unrelated to cardiovascular health receive clear explanation of LipidLlama's limitations, reinforcing its role as a patient education tool for cardiovascular health.

Beyond inappropriate medical responses, AI poses potential for biased responses. The system utilizes the PREVENT equations rather than the Pooled Cohort Equations for ASCVD risk assessment, acknowledging ethnicity as a social construct rather than a biological determinant [37] [38]. While the cardiologist validation study included diverse synthetic patient profiles reviewed by the authors, comprehensive bias evaluation remains an active area for future research.

Data privacy is inherent in LipidLlama's architecture. The system has been designed and optimized for consumer-grade hardware, enabling deployment completely within HIPAA-compliant health care system servers. This design ensures that patient data remains secure within the institutional network and eliminates dependence on external LLM providers that could compromise sensitive health information.

Following best practices in ethical AI, the authors have reviewed the "Do No Harm" AI safety checklist and determined that LipidLlama does not pose immediate biological-psychological, economic, or social risks [39]. However, as the tool is deployed, continuous monitoring will be necessary to detect and mitigate any potential harm or misuse.

Future Directions

LipidLlama has the potential to expand access to preventive cardiology by providing multilingual, personalized cardiovascular education in a scalable manner. It has reached a significant milestone with a fully operational minimum viable product (MVP)—a functional version of the platform that includes its core features, enabling early user feedback and validation in real-world settings. Future development efforts will focus on direct integration with electronic health records (EHRs) and patient validation.

EHR integration, still in the planning phase, faces several implementation challenges: establishing secure data protocols that comply with HIPAA and FHIR standards; building standardized interfaces to accommodate heterogenous EHR systems and their data schemas; and establishing clear patient consent workflows for automated data sharing. Moreover, healthcare system IT approval processes typically require extensive security auditing and validation testing. Despite these challenges, successful EHR integration would eliminate reliance on self-reported surveys, reduce data entry burden, improve clinical accuracy through verified medical information, and facilitate seamless data sharing with healthcare providers.

Developing a patient validation study to rigorously assess LipidLlama's impact on patient outcomes, including improvements in heart health literacy, lifestyle modifications, and medication adherence is centrally important. Given the growing evidence supporting AI's role in alleviating provider-patient communication burden [40], evaluating changes in clinician behavior and communication patterns will be essential. Equally important is the inclusion of diverse patient populations to evaluate translation pipeline accuracy and identify potential model bias. These efforts will inform iterative improvements to the system's pipeline and contribute to safe, equitable deployment at scale.

ARTICLE IN PRESS
CONDUCTSCIENCE

Limitations

This study has several important limitations. First, LipidLlama has not yet been tested with real patients or integrated into clinical workflows. Its impact on clinical outcomes, health literacy, or behavior change remains unvalidated. Second, all expert evaluation was performed at a single academic institution by physicians familiar with AI technologies, which may limit the generalizability of our population of raters and thus chatbot response scores. Third, while the system supports multilingual interaction, translations have not yet been rigorously evaluated for cultural or semantic accuracy. Finally, although synthetic patients and queries were designed to reflect real-world distributions, they may not capture the full variability of actual populations, which could affect real-world performance.

Conclusion

LipidLlama is a novel and practical solution to the challenge of personalized cardiovascular patient education. It combines a validated cardiac risk score model, expert-verified content, and AI-driven language generation to deliver clinically strong cardiovascular health education. With continued research funding, and patient validation, LipidLlama has the potential to enhance primary prevention of cardiovascular disease, especially in underserved communities, and become a cornerstone in the global effort to improve cardiovascular health outcomes.

References

1. Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The Global Burden of Cardiovascular Diseases and Risk. Journal of the American College of Cardiology. 2022 Dec;80(25):2361–71.

2. Adams RJ. Improving health outcomes with better patient understanding and education. Risk Management and Healthcare Policy. 2010 Oct 14;3:61–72.

3. Bhattad PB, Pacifico L. Empowering Patients: Promoting Patient Education and Health Literacy. Cureus. 14(7):e27336.

4. Murphy DR, Reis B, Sittig DF, Singh H. Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis. Am J Med. 2012 Feb;125(2):209.e1-7.

5. Chong B, Jayabaskaran J, Jauhari SM, Chan SP, Goh R, Kueh MTW, et al. Global burden of cardiovascular diseases: projections from 2025 to 2050. Eur J Prev Cardiol. 2024 Sep 13;zwae281.

6. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. European Heart Journal. 2024 Feb 1;45(5):332–45.

7. Gendler M, Nadkarni GN, Sudri K, Cohen-Shelly M, Glicksberg BS, Efros O, et al. Large Language Models in Cardiology: A Systematic Review [Internet]. medRxiv; 2024 [cited 2025 Feb 19]. p. 2024.09.01.24312887. Available from: https://www.medrxiv.org/content/10.1101/2024.09.01.24312887v1

8. Vitejs. Vite [Internet]. Version 6.0.5; [cited 2025 May 13]. Available from: https://vitejs.dev/

ARTICLE IN PRESS
CONDUCT SCIENCE

9. Meta Inc. React [Internet]. Version 18.3.1. Menlo Park: Meta; [cited 2025 May 13]. Available from: https://react.dev/

10. Tailwind Labs Inc. Tailwind CSS [Internet]. Version 3.4.17. San Francisco: Tailwind Labs; [cited 2025 May 13]. Available from: https://tailwindcss.com/

11. Python Software Foundation. Python [Internet]. Version 3.11. Wilmington, DE: Python Software Foundation; [cited 2025 May 13]. Available from: https://www.python.org

12. Gerganov G. llama.cpp [Internet]. ggml-org; [cited 2025 May 13]. Available from: https://github.com/ggerganov/llama.cpp

13. Ramírez S. FastAPI [Internet]. Version 0.115.12. Tiangolo; [cited 2025 May 13]. Available from: https://fastapi.tiangolo.com/

14. Khan SS, Coresh J, Pencina MJ, Ndumele CE, Rangaswami J, Chow SL, et al. Novel Prediction Equations for Absolute Risk Assessment of Total Cardiovascular Disease Incorporating Cardiovascular-Kidney-Metabolic Health: A Scientific Statement From the American Heart Association. Circulation. 2023 Dec 12;148(24):1982–2004.

15. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and Validation of the American Heart Association's PREVENT Equations. Circulation. 2024 Feb 6;149(6):430–49.

16. Wong ND, Budoff MJ, Ferdinand K, Graham IM, Michos ED, Reddy T, et al. Atherosclerotic cardiovascular disease risk assessment: An American Society for Preventive Cardiology clinical practice statement. American Journal of Preventive Cardiology. 2022 Jun;10:100335.

17. OpenAI. Introducing ChatGPT [Internet]. OpenAI; 2024 [cited 2025 Feb 20]. Available from: https://openai.com/index/chatgpt/

18. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst. 2025 Mar 31;43(2):1–55.

19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arXiv; 2023 [cited 2025 Feb 19]. Available from: http://arxiv.org/abs/1706.03762

20. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [cited 2025 Feb 19]. Available from: http://arxiv.org/abs/2005.14165

21. Kambhampati S. Can Large Language Models Reason and Plan? Annals of the New York Academy of Sciences. 2024 Apr;1534(1):15–8.

22. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Internet]. arXiv; 2021 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2005.11401

23. Northwestern Medicine. Health Encyclopedia [Internet]. Northwestern Medicine [cited 2025 Feb 20]. Available from: http://encyclopedia.nm.org/

24. Chandrasekaran D, Mago V. Evolution of Semantic Similarity—A Survey. ACM Comput Surv. 2021 Feb 18;54(2):41:1-41:37.

ARTICLE IN PRESS
COMPUTER SCIENCE

25. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D, Nie JY. C-Pack: Packed Resources For General Chinese Embeddings [Internet]. arXiv; 2024 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2309.07597

26. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The Faiss library [Internet]. arXiv; 2025 [cited 2025 Apr 23]. Available from: http://arxiv.org/abs/2401.08281

27. Sanjeev S, Troynikov A. Embedding Adapters [Internet]. Chroma; 2024 May. Available from: https://research.trychroma.com/embedding-adapters

28. Li C, Liu Z, Xiao S, Shao Y. Making Large Language Models A Better Foundation For Dense Retrieval. 2023.

29. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. 2024.

30. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 Herd of Models [Internet]. arXiv; 2024 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2407.21783

31. Tiedemann J, Thottingal S. OPUS-MT – Building open translation services for the World. In: Martins A, Moniz H, Fumega S, Martins B, Batista F, Coheur L, et al., editors. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation [Internet]. Lisboa, Portugal: European Association for Machine Translation; 2020 [cited 2025 Feb 17]. p. 479–80. Available from: https://aclanthology.org/2020.eamt-1.61/

32. Tiedemann J, Aulamo M, Bakshandaeva D, Boggia M, Grönroos SA, Nieminen T, et al. Democratizing Neural Machine Translation with OPUS-MT [Internet]. arXiv; 2023 [cited 2025 Feb 17]. Available from: http://arxiv.org/abs/2212.01936

33. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. Open Heart. 2023 Nov;10(2):e002455.

34. Brennan RL. Generalizability Theory and Classical Test Theory. Applied Measurement in Education. 2010 Dec 30;24(1):1–21.

35. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: A practical guide to study design, implementation, and interpretation. Journal of School Psychology. 2014 Feb;52(1):13–35.

36. Smith TJ, Kline TJB, Kline A. GeneralizIT: A Python Solution for Generalizability Theory Computations.

37. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. Journal of the American College of Cardiology. 2014 Jul;63(25):2935–59.

38. Khan SS, Yancy CW. Race, Racism, and Risk—Implications of Social Determinants of Health in Cardiovascular Disease Prediction. JAMA Cardiol. 2024 Jan 1;9(1):63.

39. Khan WU, Seto E. A "Do No Harm" Novel Safety Checklist and Research Approach to Determine Whether to Launch an Artificial Intelligence–Based Medical Technology: Introducing the Biological-Psychological, Economic, and Social (BPES) Framework. Journal of Medical Internet Research. 2023 Apr 5;25(1):e43386.

ARTICLE IN PRESS
CONDUCT SCIENCE

40. Anderson BJ, Haq MZ ul, Zhu Y, Hornback A, Cowan AD, Mott M, et al. Development and Evaluation of a Model to Manage Patient Portal Messages. NEJM AI. 2025;2(3):AIoa2400354.

41. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(11).

Figures

ARTICLE IN PRESS
CONDUCT SCIENCE