diagnostics

MDPI

Article

CARDIA-TR: Faithful, Guideline-Grounded Explanation of Machine Learning Decisions for Coronary Heart Disease Using Retrieval-Augmented Large Language Models in Turkish

Keremcan Karakaya [1,*], Aytuğ Boyacı [2], Mustafa Ulaş [1,3] and Eyüp Dikmen [4]

[1] Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Fırat University, 23000 Elazig, Türkiye; mustafaulas@firat.edu.tr
[2] Department of Computer Engineering, Air Force Academy, National Defence University, 34149 Istanbul, Türkiye; aytug.boyaci@msu.edu.tr
[3] Prodrom Information and Communication Technologies, Fırat Technopolis Technology Development Zone, 23000 Elazig, Türkiye
[4] Elazığ Provincial Health Directorate, 23050 Elazig, Türkiye; eyup.dikmen@saglik.gov.tr
* Correspondence: k.karakaya@firat.edu.tr; Tel.: +90-536-924-3351

Abstract

Background/Objectives: Machine Learning (ML) models can estimate coronary heart disease risk from structured patient data, but they offer little guideline-based justification for their predictions. Explanations produced by large language models (LLMs) may also contain post hoc rationales when they are not tied to the model's actual reasoning. We developed CARDIA-TR, a framework in which an ML model makes the diagnostic decision for suspected stable coronary heart disease and an LLM explains it using Turkish and international cardiology guidelines. **Methods:** Six classifiers were evaluated on the UCI Cleveland dataset ($n = 303$) with 10×5 repeated nested cross-validation, together with a leakage audit of a widely shared version of the dataset. Eight open-source LLMs answered 150 clinical questions using retrieval-augmented generation over 52 guideline documents. Their answers were assessed with automatic metrics, blinded physician ratings, and two LLM judges whose agreement with physician ratings was measured. For 303 patients, the LLMs explained ML decisions based on exact feature contributions and could flag guideline-discordant feature use. A physician scored 80 of these explanations. **Results:** Logistic regression performed best (ROC-AUC 0.911). Confirming earlier reports, 723 duplicate rows in a widely shared 1025-row version of the dataset explained reported accuracies of up to 100%. Lexical overlap with the sources did not reflect clinical accuracy. Qwen3.6-27B and Gemma-4-26B-A4B received the highest judge and physician scores (physician 1.68 and 1.70 out of 2), but in the physician evaluation they did not differ significantly from Turkish-Gemma-9B. The physician rated all 80 faithful explanations as fully correct, including those for wrong ML predictions. The newer models detected guideline-discordant use of chest pain type caused by referral bias without false alarms (F1 0.986 and 0.935), also when the referral information was removed from the prompt. **Conclusions:** In this internally validated, referral-based cohort, an LLM explanation layer anchored to the ML model's reasoning could justify decisions and audit their consistency with clinical guidelines. External validation is needed.

Keywords: coronary heart disease; large language models; retrieval-augmented generation; clinical decision support system; explainable AI; explanation fidelity; guideline-consistency audit; data leakage; Turkish clinical natural language processing

Check for updates

Academic Editor: Kaan Orhan

Received: 9 October 2026
Revised: 7 November 2026
Accepted: 10 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Coronary Heart Disease (CHD) is a chronic disease that develops as a result of the atherosclerotic process in the coronary arteries and remains a leading cause of death worldwide [1]. Ischemic heart disease alone accounts for approximately nine million deaths annually, and a significant portion of this burden is concentrated in low- and middle-income countries [1,2]. Acute Coronary Syndrome (ACS) is the acute form of coronary artery disease. It includes unstable angina, Non-ST-Elevation Myocardial Infarction (NSTEMI) and ST-Elevation Myocardial Infarction (STEMI), and it develops when an atherosclerotic plaque in a coronary artery ruptures or erodes [3]. ACS begins suddenly, progresses quickly and carries a high risk of death, so decisions must be made early and in line with guidelines [3]. Percutaneous coronary intervention and modern drug therapy have improved care, yet many patients still have major adverse cardiovascular events during follow-up [3]. Fast, consistent and evidence-based decision support is therefore needed in coronary heart disease, both in stable patients and in ACS.

Care for stable coronary artery disease and for acute coronary syndromes is guided by clinical guidelines that change often and grow more complex over time [3,4]. Physicians must decide on risk stratification, antithrombotic therapy, the timing of revascularization and the care of special patient groups. They often make these decisions under time pressure and without easy access to the full guideline text. Decision support systems based on artificial intelligence could help in two ways, by estimating risk and by giving quick access to guideline information. Machine Learning (ML) models that predict coronary heart disease risk from structured patient data have been used widely and with good results [5,6]. However, they rarely explain their predictions with reasons that a clinician can relate to the guidelines [7,8]. An ML model can estimate how likely it is that a patient has the disease, but it cannot say which guideline-based reasons support that estimate. This is a serious obstacle in medicine, where clinical decisions have to be justified.

Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) may help to close this gap. LLMs can produce fluent text that fits the context [9]. RAG links this text to documents retrieved from an external knowledge base, which aims to make the output more factual and less prone to hallucination [10]. In medicine, recent studies have shown that RAG systems built on clinical guidelines can clearly improve the clinical accuracy of language models [11,12]. For instance, in an assessment of surgical suitability, a RAG system supported by international guidelines achieved 96.4% accuracy—significantly outperforming the 86.6% success rate of human evaluators—and produced no hallucinations [11]. Similarly, while a RAG-enabled system achieved 95.5% accuracy in generating guideline-compliant recommendations regarding prostate-specific antigen testing, the accuracy rate of junior doctors in the same scenarios remained at 62.3% under closed-book conditions [12]. However, the vast majority of existing studies have focused on English clinical resources and English question-answering datasets [13]. Research on guideline-based, RAG-supported clinical decision support systems remains limited for languages—such as Turkish—that are morphologically rich and have relatively scarce clinical natural language processing resources [13,14]. Furthermore, a significant portion of existing systems evaluates language model outputs solely using automated text similarity metrics, failing to assess deeper attributes—such as clinical accuracy, guideline adherence, and the quality of reasoning—through physician evaluation. Integrating these two fields poses unique methodological problems. First, when a language model is asked to explain a ML prediction, the model may produce convincing but contrived justifications without knowing the real reason for the prediction. In the explainable artificial intelligence literature, this situation is discussed as the problem of fidelity of the explanation to the model. Secondly, ML models learn the characteristics of the cohort on which they are trained. Relationships that

contradict clinical knowledge—such as those arising from referral bias—can also become embedded in the model. In systems that provide only a risk score or a feature contribution plot, such relationships may go unnoticed by the physician. To the best of our knowledge, there is no study in which guideline information from language models is used to audit the relationships learned by an ML model. Thirdly, high performance figures reported on widely used open datasets cease to be comparable when data versions and evaluation protocols are not clearly documented, making it difficult to evaluate the ML component of hybrid systems against a realistic benchmark. Duplicate records in a widely shared version of the Cleveland dataset have already been reported as one source of such inflated results [15,16].

This study aims to address the aforementioned gaps collectively. First, it combines ML-based risk prediction with the guideline-based explanatory capabilities of LLMs; in doing so, it ensures that the explanation remains faithful to the actual rationale behind the ML model's decision. Second, the explanation layer is used to audit whether the relationships learned by the ML model align with clinical guidelines. Third, the ML component is evaluated using a method protected against data leakage, and the results are compared with evaluation methods found in the literature. Finally, this approach has been developed within a Turkish clinical context and evaluated through physician assessment. To this end, a two-component clinical decision support framework named CARDIA-TR (CARDIac decision support in Turkish) is proposed within the context of coronary heart disease. The question-answering component covers guidelines on both stable coronary disease and acute coronary syndrome, while the patient-level component addresses risk assessment for suspected stable coronary heart disease. In this framework, the ML component is positioned as the element that estimates the risk of coronary heart disease from structured patient data and is probabilistically calibrated, while the RAG component is positioned as the element that explains this risk—based on the ML's actual decision rationale—using Turkish and international cardiology guidelines and monitors the model's consistency with those guidelines. Thus, rather than merely providing a risk score, the system supplements it with a guideline-based explanatory layer that renders the risk clinically interpretable and reveals to the physician which of the relationships learned by the ML model from the data contradict clinical knowledge. The main contributions of this study are as follows:

- A decision support framework (CARDIA-TR) tailored to the Turkish clinical context has been proposed; it integrates a calibrated ML-based risk prediction for coronary heart disease with an explanation of that prediction via RAG grounded in cardiology guidelines. In this model, the ML component makes the diagnostic decision, while the language model explains it.
- The ML component was developed using six classifiers within a repeated nested cross-validation scheme designed to eliminate data leakage. Furthermore, a leakage analysis confirmed earlier reports that the high accuracy values reported in the literature for the Cleveland dataset largely stem from data versions containing duplicate rows and the use of specific, isolated data splits, and it quantified the overlap between training and test data under each evaluation protocol.
- A hybrid (BM25 + dense + RRF) RAG architecture was designed based on a knowledge base compiled from Turkish and international cardiology guidelines, and the contribution of the retrieval strategy was examined through a controlled ablation study involving four conditions.
- Eight open-source LLMs of varying generations and scales were compared using 150 physician-reviewed clinical questions. Evaluation methods included automated metrics, a two-stage blinded physician assessment, and two LLM judges whose agree-

ment with physician ratings was measured. A consistent discrepancy between source overlap and clinical accuracy was demonstrated across the different evaluators.

- A faithful explanation method was developed—wherein the ML model's actual variable contributions are conveyed to the language model, allowing it to flag guideline contradictions without altering the decision itself—and this method was evaluated across eight models using data from 303 patients, with automatic fidelity measures and a physician rating of a stratified sample of 80 explanations.
- It has been demonstrated that the explanation layer can detect relationships learned in violation of guidelines due to cohort-specific referral bias associated with the model, and this function has been characterized as a guideline-consistency audit. The audit was evaluated across all clinical variables and also without referral information in the prompt.

In the remainder of the article, related studies are summarized in Section 2. The materials and methods of the proposed CARDIA-TR framework are presented in Section 3. Experimental findings are presented in Section 4. The findings are discussed in Section 5, and the study's conclusions and suggestions for future research are included in Section 6.

2. Related Work

This section addresses three research areas within which the proposed CARDIA-TR framework is situated. The first concerns the prediction of coronary heart disease risk from structured clinical data using ML and the explainability of these predictions. The second focuses on the performance of clinical guideline-based question-answering systems and RAG approaches. The third involves solutions developed for languages—such as Turkish—where clinical natural language processing resources are relatively limited. Finally, the proposed approach is compared with existing studies.

2.1. Prediction of Coronary Heart Disease Risk Using ML and Explainability

Estimating the risk of coronary heart disease from structured clinical features is a problem that has long been studied in the medical informatics literature. In this field, the UCI Cleveland Heart Disease dataset serves as the most widely used benchmark due to its standardized feature structure and the comparability it offers, despite its limited sample size [5,17,18]. However, recent studies have noted that the heavy use of this dataset limits the generalizability of the findings [5,8]. More recent studies have also focused more on explainability rather than classification accuracy alone. Zhao et al. used a Generative Adversarial Imputation Network (GAIN) to handle missing data in a hospital-based cohort. Their XGBoost model achieved an AUC of 0.9053, and SHAP values were used to interpret the model at both global and local levels [6]. Similarly, the XAI-HD framework addressed class imbalance and missing data issues by combining ML, deep learning, and explainable AI techniques [7]. Meanwhile, the SmartHeart framework combined explainable cardiovascular risk prediction with a conceptual architecture proposing a real-time data collection pipeline [19]. In a recent study, a pipeline designed to prevent data leakage was established using the Framingham and Cleveland datasets; SHAP-based interpretability was achieved by employing a stratified train–test split and applying scaling solely within the training set [20]. However, the authors noted that the results should be interpreted with caution, given that the test set for the Cleveland subset consisted of only 60 patients [20].

While these studies demonstrate the importance of explainability for clinical adoption, there is a qualitative difference between the type of explanation they generate and clinical justification. Feature contribution methods, such as SHAP, quantitatively indicate which variables contribute to a prediction, as well as the direction and magnitude of that contribution [21]. However, they do not explain which guideline recommendation the

contribution relates to, the level of evidence upon which it is based, or how it should be interpreted in clinical practice. A recent assessment identifies the lack of validation for SHAP-based explanations in real-world clinical settings and scalability limitations due to computational costs as fundamental shortcomings of current approaches [8]. These methods can show which variables influenced a prediction, but they do not explain why that risk is important in terms of clinical guidelines. The proposed study addresses this gap by generating guideline-based natural language explanations instead of relying only on feature contributions.

2.2. Clinical Guideline-Based Question Answer and RAG

The performance of LLMs in answering questions based on clinical guidelines has become a rapidly expanding area of research in recent years. The common approach in these studies involves generating question sets from guideline texts and scoring model responses using expert evaluations or official answer keys. Within the field of cardiology, the study most directly related to this work was conducted by Alexandrou et al. [22]. The authors compared the accuracy of ChatGPT-4o, DeepSeek R1, and the medical-domain-specific Med-PaLM 2 models on 38 open-ended questions based on acute coronary syndrome guidelines. Furthermore, by evaluating ChatGPT-4o and DeepSeek R1 both with and without RAG, they examined the impact of access to guideline content on reducing hallucinations [22]. This study clearly demonstrates the value of guideline-based access within the context of cardiology. Furthermore, it focuses exclusively on closed-source and cloud-based models. It treats access as a binary variable (present or absent) and is limited to English-language sources. Similar study designs have been implemented in other clinical fields as well. Jin et al. had three cardiologists evaluate—using a six-point scale—the responses provided by four LLMs to eight clinical questions based on the 2025 guidelines for left atrial appendage closure [23]. Pan et al. posed 52 questions derived from NCCN bone cancer guidelines to the DeepSeek V3.1, ChatGPT 5, and Grok 4 models; the responses were independently scored by two experts, and the correct response rate for all three models was found to exceed 90% [24]. Wang et al. evaluated three RAG-augmented models using 43 diabetes-related questions and reported that retrieval integration increased the correct answer rate by an average of 12%, while also improving answer comprehensiveness and clarity [25]. In a larger-scale benchmark, Wang and Chen tested state-of-the-art models on over 20,000 question–answer pairs derived from systematic reviews and clinical guidelines [26]. They demonstrated that accuracy levels reached 90% for structured guideline sources, 70% for narrative sources, and 50–60% for systematic reviews [26]. This finding indicates that guideline texts constitute a relatively favorable information source for RAG.

Common patterns across these studies point to three trends. First, the models evaluated are predominantly closed-source systems or those accessed via cloud-based service interfaces [22–26]. Even in studies involving open-weight models, access was typically gained through external interfaces; the performance of locally deployed configurations within the context of clinical guidelines has received relatively little scrutiny. This creates an applicability gap for clinical settings where the transfer of patient data outside the institution is restricted. Second, the question sets were generally limited in size; in the studies summarized above, the number of questions ranged from 8 to 52 [22–25]. Third, the contribution of the retrieval component is usually treated as a binary factor—"RAG present" versus "RAG absent"—without a controlled disentanglement of the relative contributions of lexical, semantic, and hybrid retrieval strategies [22,25].

2.3. Clinical Language Models for Turkish and Low-Resource Languages

The majority of the literature on clinical natural language processing relies on English-language resources. Although medical-domain-specific models such as Med-PaLM and BioMistral demonstrate strong performance, their direct applicability to Turkish clinical contexts is limited because they were primarily designed for English-language environments [27,28]. Research conducted for the Turkish language is relatively recent and focuses on two main areas. In the first area, a Turkish-specific medical language model was developed using real patient–physician question-and-answer data, demonstrating the potential contributions of domain-specific Turkish models within the healthcare context [13]. In the second part, Koçak et al. administered 240 questions from the Medical Specialty Exam in Turkish to four LLMs [14]. While ChatGPT-4 achieved the highest performance with 88.75% accuracy, Llama 3 70B, Gemini 1.5 Pro, and Command R+ remained at levels of 79.17%, 78.13%, and 50%, respectively [14]. Although these studies demonstrate the applicability of language models to Turkish clinical texts, they focus on patient consultation data and multiple-choice exam questions, respectively. Furthermore, they do not address the aspects of guideline-based retrieval and reasoning. The importance of the retrieval strategy for low-resource languages has been concretely demonstrated in a study conducted on Korean diabetes guidelines. Lee et al. developed a bidirectional RAG system utilizing both Korean and American diabetes guidelines and compared dense retrieval—using eleven embedding models—with BM25-based lexical retrieval employing language-specific tokenizers [29]. The study reported that multilingual embedding models outperformed language-specific models, whereas, for lexical retrieval, the language-specific tokenizer yielded significantly better results for Korean texts. It has been shown that retrieval components combining optimal dense and lexical representations enhance coverage [29]. This finding indicates that the choice of retrieval strategy directly impacts performance in morphologically rich languages. However, no study has been found that performs RAG based on a knowledge base derived from Turkish clinical guidelines, generates guideline-based rationales, and evaluates these outputs through physician assessment. The fact that Turkish is a morphologically rich and agglutinative language can directly influence the performance of methods based on word segmentation and lexical matching [30]. Therefore, the relative contributions of lexical and semantic retrieval strategies in this language require further investigation.

2.4. Comparison with Existing Studies and the Position of This Study

The literature summarized above proceeds along three distinct lines. The first comprises ML studies that perform risk prediction using structured data and address explainability through feature contribution methods [5–8,19,20]. The second consists of studies on LLMs that evaluate performance in question-answering based on clinical guidelines [22–26]. The third involves clinical language model studies focusing on Turkish and low-resource languages [13,14,29]. A comparison of the proposed approach with existing studies is presented in Table 1.

In light of this comparison, the contribution of the proposed study to the literature can be summarized in four points. First, ML-based risk prediction and guideline-based retrieval-augmented explanation are integrated within a single framework with a clearly defined division of labor. The explanation layer generates rationales by linking the variable contributions of the ML model to clinical guideline evidence and verifies the consistency of these contributions with the guidelines. Second, the contribution of the retrieval strategy is disentangled through a controlled ablation design involving four conditions, moving beyond the binary "RAG present/absent" distinction. Third, the evaluation was conducted using open-source models capable of local execution; this approach examines a configuration viable for clinical settings where the transfer of patient data outside the institution is

restricted. Fourth, the study was conducted within a Turkish clinical context, and the outputs were validated not only through automated text metrics but also via a blinded physician assessment.

Table 1. Comparison of the proposed CARDIA-TR framework with related studies.

Study Type	Risk Prediction	Explanation Format	Access Strategy	Model Type	Language	Expert Assessment
SHAP- based CHD prediction (XAI-HD, SmartHeart, etc.) [7,19]	Yes	Feature contribution	No	-	English	Usually not available
AKS guidelines + RAG (JACC, 2025) [22]	No	Guide-based text	RAG present/absent (binary)	Closed source	English	Available
Guideline-based clinical question answering (LAAO, NCCN, etc.) [23,24]	No	Guide-based text	No	Closed source	English	Available (2–3 specialists)
Large-scale clinical question-answering benchmarks [26]	No	-	RAG (PubMed)	Closed source	English	Answer key
Turkish medical language model/TUS evaluation [13,14]	No	-	No	Opened + Closed	Turkish	Answer key
Recommended: CARDIA-TR	Yes (calibrated, leak-proof)	Guideline-based rationale based on ML variable contributions + guideline-consistency audit	Four-condition controlled ablation (no-retrieval/BM25/ dense/hybrid)	Open source, local (8 models)	Turkish	Available (blinded; 480 responses + 30 pilot explanations + 80 faithful explanations across two stages; LLM judges with measured agreement with physician ratings)

3. Materials and Methods

This section presents in detail the data sources, methods, and evaluation protocol used to develop the proposed CARDIA-TR framework. First, the overall study design and its two-stage structure are described. Subsequently, the guideline corpus, patient dataset, and clinical question set utilized in the study are introduced. The following subsections address the design of the LLM component, the RAG architecture, and the hybrid explanation mechanism that integrates these two components. Finally, the evaluation protocol and statistical analysis methods used to measure the system's performance are explained. The overall workflow of the study is summarized in Figure 1.

3.1. Study Design

In this study, an artificial intelligence-based clinical decision support framework (CARDIA-TR) for coronary heart disease was developed and evaluated through two complementary components. In this study, coronary heart disease risk prediction refers to the probability of angiographically significant coronary artery disease estimated by the ML model, and the diagnostic decision refers to the binary output derived from this probability. Acute coronary syndrome is addressed only in the question-answering component, and guideline-based clinical decision support refers to the framework as a whole. The study was motivated by the fact that, ML models can generate numerical risk estimates from structured patient data but fail to explain these estimates based on clinical guidelines, Large Language Model (LLM)-based RAG systems can generate guideline-based explanations

but are limited in producing calibrated numerical decisions from structured data. To combine the strengths of these two paradigms, the CARDIA-TR framework was proposed, integrating ML-based risk estimation with RAG-based, guideline-driven explanation. The study consists of two stages. In the first stage, the RAG performance of eight open-source LLM—operating on Turkish and international cardiology guidelines—was evaluated using a physician-reviewed set of 150 clinical questions. In the second stage, patient-level risk predictions generated by the ML model are explained by language models based on clinical guidelines. This stage comprises an experiment involving all 303 patients and eight models to evaluate faithful explanations based on the ML model's feature contributions and to audit their consistency with guidelines. A physician scored a stratified random sample of 80 of these explanations. In addition, a physician scored 30 pilot explanations produced with an earlier design in which the language model received only the case vignette. In this approach, the ML component predicts the patient's risk. The RAG component explains this prediction using clinical guidelines and checks whether the ML model is consistent with those guidelines. The study is retrospective and methodological. It uses only publicly available datasets and published clinical guidelines. No prospective intervention was performed on patients. The proposed system is intended to support clinical decisions and provide explanations, not to replace clinical judgment. Since the study uses only anonymized public data and published guidelines, it does not include identifiable patient information and does not require ethics committee approval.

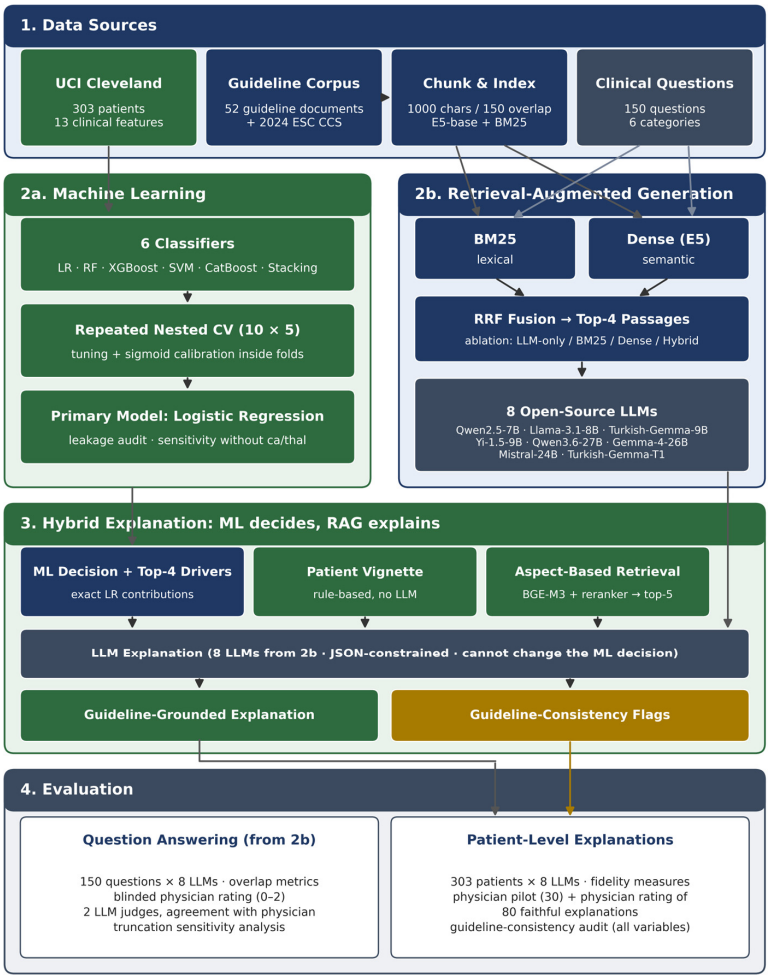


Figure 1. General workflow of the proposed CARDIA-TR framework. Dark gray arrows show the main data flow, light gray arrows show the clinical questions used as retrieval queries, and the gold arrow shows the guideline-consistency flags.

The design of CARDIA-TR consists of four components. The first is the separation of the decision from the explanation. The diagnostic decision is made by the validated and calibrated clinical decision rule ML component; the language model cannot alter this decision. The second is the explanation's fidelity to the ML. The language model is provided with the variable contributions of the ML, calculated precisely for each patient. Outputs are constrained by a structured schema, and citations are limited strictly to the accessed passages. This prevents the explanation from deviating from the model's actual rationale or relying on non-existent sources. The third component is guideline-consistency auditing. The language model is not compelled to validate the ML; instead, it can specifically report instances where the ML utilizes a variable contrary to guideline evidence or where the decision contradicts the findings as a whole. The accuracy of these reports is measured against a reference derived from the guideline, without utilizing label information. The fourth component is the validity of the evaluation. The ML component was tested using a protocol free from data leakage and through comparisons with protocols found in the literature. Language model outputs were evaluated using automated metrics, a two-stage blinded physician assessment, and two language model judges whose agreement with physician scores was measured.

3.2. Data Sources

The study utilized three primary data sources: the guideline corpus forming the RAG system's knowledge base, the patient dataset underpinning the ML component, and the clinical question set prepared to evaluate the performance of the language models. The knowledge base of the retrieval system was constructed from current guidelines and consensus documents regarding the diagnosis and management of cardiovascular diseases, with a primary focus on acute coronary syndrome. This corpus integrates international sources [3,4,31] with Turkish-language guidelines, consensus statements, and review articles published by the Turkish Society of Cardiology. A total of 52 documents were processed by converting them from PDF format to plain text. The corpus was not restricted solely to acute coronary syndrome; it was deliberately expanded to encompass related areas such as heart failure, dyslipidemia, and cardiovascular prevention. This approach also enabled the evaluation of the retrieval system's ability to identify question-specific, relevant passages within a heterogeneous knowledge base.

The UCI Cleveland Heart Disease dataset, considered a common benchmark in cardiovascular ML studies, was used for training the ML component and for patient-level risk prediction [17,18]. The data were collected in the 1980s from patients referred for coronary angiography because of suspected coronary artery disease. This introduces referral bias and limits generalizability to current or unselected populations. The dataset comprises 13 clinical attributes from 303 patients and a target variable indicating the presence of coronary heart disease. The attributes include age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic findings, maximum heart rate achieved, exercise-induced angina, exercise-induced ST depression, ST-segment slope, number of major vessels colored by fluoroscopy, and thallium scintigraphy findings. The target variable is binary-coded to indicate the presence or absence of the disease. The binary target was defined by treating an angiographic finding greater than 0 on the original 0–4 scale (indicating ≥50% stenosis in at least one major coronary artery) as the presence of CHD; 139 patients (45.9%) in the cohort tested positive for CHD. Categorical attributes were utilized using UCI coding. For instance, regarding chest pain type: 1 denotes typical angina, 2 atypical angina, 3 non-anginal chest pain, and 4 asymptomatic; for thallium scintigraphy, 3 indicates normal, 6 a fixed defect, and 7 a reversible defect. Additionally, to specifically examine the impact of the evaluation protocol on reported performance, a 1025-row version

widely distributed on an open-access platform was used [32]. The Cleveland dataset was employed to predict the risk of coronary heart disease based on structured clinical attributes. The risk predictions generated in this study were supported by explanations grounded in cardiology guidelines. To assess the guideline-based question-answering performance of language models, a set of 150 clinical questions focusing on acute coronary syndrome was compiled. The questions were categorized under six clinical headings: medication, antithrombotic therapy and secondary prevention, intervention and revascularization, diagnosis and symptomatology, special populations, complications and emergency management, and electrocardiography and biomarker interpretation. The questions were formulated based on relevant guideline content and finalized following a review by a physician for clinical relevance and adherence to guidelines. The distribution of questions by category is shown in Table 2. Here, n represents the number of questions.

Table 2. Distribution of the clinical question set by category and sample questions.

Category	*n*	Sample Question
Medication, antithrombotic therapy, and secondary prevention	32	How can the duration of dual antiplatelet therapy be shortened in a patient with NSTEMI who is at high risk of bleeding?
Intervention and revascularization	27	What is the target door-to-balloon time for primary percutaneous intervention in STEMI?
Diagnosis and symptomatology	25	What are the atypical symptoms of acute coronary syndrome in women?
Special populations	23	How is the reperfusion strategy selected when STEMI develops during pregnancy?
Complication and emergency management	22	What are the indications for mechanical circulatory support in cardiogenic shock?
ECG and biomarker interpretation	21	What is the clinical significance of the de Winter pattern, and which vessel does it implicate?
Total	150	

3.3. ML Component

Current reporting recommendations for regression and ML-based clinical prediction models were followed during the development and reporting of the ML component [33]. Six supervised classification algorithms were comparatively evaluated to predict the risk of coronary heart disease at the patient level. These included logistic regression, random forest, extreme gradient boosting (XGBoost), support vector machine, CatBoost (a gradient boosting method capable of handling categorical variables directly), and a stacking ensemble that combines these base learners with a logistic meta-model [34–39]. These algorithms were chosen because they are widely used in coronary heart disease prediction and represent different types of learning methods. Before training the models, the data were preprocessed. Missing values in continuous variables were replaced with the median, while missing categorical values were replaced with the most frequent value. Continuous variables were standardized using z-scores, and categorical variables were converted using one-hot encoding. To avoid data leakage, all preprocessing steps were included in the modeling pipeline and applied only to the training folds during cross-validation [40]. Model performance was evaluated using repeated nested cross-validation. In the outer loop, stratified five-fold cross-validation was repeated 10 times using different random splits. For each outer training fold, hyperparameters were selected using an inner five-fold cross-validation, with ROC-AUC as the selection metric. The inner cross-validation used

only the training data from that outer fold. In each outer fold, the model was trained without using the corresponding test data. Prediction probabilities were therefore obtained out of fold, meaning that each patient's prediction came from a model that had not been trained on that patient's data. This approach eliminates the optimistic bias resulting from using the same data for both training and evaluation [41]. Model performance is reported as the mean ± standard deviation across 10 repetitions, using the following metrics: area under the ROC curve (ROC-AUC), accuracy, recall, specificity, F1 score, and Brier score. Probability calibration was applied to enhance the probabilistic reliability of the risk predictions. Calibration was performed using the sigmoid (Platt) method within a nested cross-validation framework [42,43]. To prevent data leakage, the calibrator was fitted only to the training folds, while evaluation was conducted on the untouched test folds. Calibration performance was assessed using the Brier score [44]. Because the number of vessels colored by fluoroscopy (ca) and the thallium scintigraphy result (thal) come from investigations closely related to the angiographic outcome, the primary model was also evaluated without these two variables using the same protocol.

Leakage Audit

To examine the extent to which reported performance depends on the evaluation method, logistic regression and random forest models were re-evaluated under six different protocols using fixed hyperparameters [45]: In the first protocol (P1), repeated stratified cross-validation was performed on the UCI data, with preprocessing steps applied within the pipeline. In the second protocol (P2), the data was split once using an 80/20 ratio with 100 different random seeds; the average accuracy across these splits and the accuracy of the best-performing split were reported separately. In the third protocol (P3), scaling and the selection of the top eight features were applied to the entire dataset prior to cross-validation. The 1025-row version of the dataset was used in the final three protocols. This version was evaluated using a single 80/20 split in P4, 10-fold cross-validation in P5, and 10-fold cross-validation in P6 after removing exact duplicate rows. Duplicates were defined as rows in which all 13 predictors and the target were identical. For each protocol, the proportion of test rows with an exact copy in the training set was calculated, using 100 random 80/20 splits for P4 and all folds for P5 and P6. To establish the provenance of this version, its unique rows were matched with the UCI Cleveland records on the eight variables whose coding is identical in both versions, and the coding of the target was compared. The purpose of this analysis was to establish a realistic performance baseline for the ML component and to demonstrate under what conditions values reported in the literature for the same dataset are comparable. Duplicate records in this version have been reported previously [15,16], and this analysis is intended as a confirmation and extension of those reports.

3.4. RAG Architecture

A RAG-based system was developed to generate clinical answers based on guidelines [46]. The system has three main steps: document indexing, retrieval of relevant passages, and answer generation with a large language model. First, the guideline documents were converted into plain text and divided into overlapping text chunks. Each chunk contained 1000 characters, with a 150-character overlap between neighboring chunks. This overlap was used to reduce the loss of context between chunks. The text chunks were then converted into dense vector representations using the multilingual-E5-base embedding model [47]. This model uses different prefixes for queries and documents ("query:" and "passage:") and works with multilingual text, including Turkish. The resulting vectors were stored in a Chroma vector database for similarity-based retrieval [48]. During retrieval, both lexical and semantic search were used. BM25 was used to find passages with exact

term matches, while dense retrieval used cosine similarity with E5 embeddings to find passages with similar meaning [49]. Lexical search is useful for terms that need exact matching, such as drug names, clinical scores, and abbreviations. Semantic search, on the other hand, can find related content even when different words are used. The ranked results obtained from these two retrieval methods are combined using the Reciprocal Rank Fusion (RRF) method [50]. By scoring each document proportionally to the inverse of their order in the two lists, RRF eliminates the problem of direct comparison of scores on different scales and balances the contribution of the two methods. From the combined sequence, the four most relevant passages were selected as context after elimination of duplicate sources. During the response generation phase, eight open-source LLMs with different architectural and educational backgrounds were comparatively evaluated. These are Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Turkish-Gemma-9B and Yi-1.5-9B-Chat, Qwen3.6-27B, Gemma-4-26B-A4B, Mistral-Small-24B-Instruct and the reasoning-oriented Turkish-Gemma-9B-T1 [51–58]. The model set was selected to encompass locally runnable open-source models in the 7–27 billion parameter range, enable comparisons between different generations of the same model families, and allow for the evaluation of Turkish-specific models (Turkish-Gemma-9B and Turkish-Gemma-9B-T1) against general-purpose models. Since the current open-weights model from the Llama family did not fit into the memory of a single 40 GB GPU, the similarly sized Mistral-Small-24B was used; for the Yi family, the Gemma-4 family was utilized because there has been no open-weights release following Yi-1.5. The prompt provided to the models was designed to ensure that responses rely solely on the provided reference passages, that no information absent from these passages is generated, and that the response is in Turkish, clinically accurate, and concise. To ensure deterministic output across all models, the temperature parameter was set to zero, and the "thinking" mode was disabled in models equipped with a reasoning capability. The retrieval strategy ablation and initial physician assessment were conducted using the first four models in quantized formats (GGUF, Q5_K_M) on the llama.cpp framework [59]. For the question-answering and explanation experiments comparing the eight models, next-generation large models were run using versions pre-quantized to FP8 precision, while the other models utilized bf16 precision; all were executed on a single NVIDIA A100 (40 GB) GPU (NVIDIA Corporation, Santa Clara, CA, USA) using the vLLM framework [60]. In the question-answering experiment comparing eight models, the knowledge base was reconstructed from the same 52 documents—using identical chunking, embedding, and retrieval configurations—because the plain text files used in the initial phase were inaccessible. Text extraction from PDFs was re-performed using PyMuPDF, and exact cosine similarity was employed instead of approximate search for retrieval. Since the retrieval results of the reconstructed knowledge base did not perfectly match those of the initial phase, all eight models—including the four from the initial phase—were re-run on this knowledge base using the same prompt template and decoding settings (temperature 0, maximum 512 tokens) to ensure the comparison was conducted under identical conditions. The ablation study in Section 4.2 is based on the responses from the initial phase, whereas the eight-model comparison in Sections 4.3 and 4.5 relies on the re-run responses. The physician evaluation in Section 4.4 covers responses from both phases, with the phase associated with each result indicated in Section 4.4. Model identifiers, quantization, and software versions are listed in Supplementary Table S1, and the settings of the two phases are compared in Supplementary Table S2. All comparisons between model generations were made within the second phase, where all models used the same knowledge base, inference framework, and settings.

Retrieval Strategy Ablation Design

A controlled ablation study was designed to isolate the contribution of hybrid retrieval to the system's retrieval-augmented response generation. In this design, the context fueling response generation was constructed under four distinct retrieval conditions, while all other components—specifically the 150 questions, four language models, prompt template, inference parameters, and the top-k = 4 passages—remained constant. This ensured a fair comparison, with the retrieval strategy serving as the sole variable across conditions. The four conditions evaluated are as follows:

- LLM-only (no retrieval): The model was tasked with generating a response using only its parametric knowledge, without being provided any reference passages. This condition represents the baseline scenario where no retrieval takes place.
- BM25-only (lexical-only): The context was constructed solely through lexical retrieval based on the BM25 ranking function.
- Dense-only (semantic-only): The context was constructed solely through dense retrieval based on cosine similarity using E5 embedding representations.
- Hybrid (BM25 + Dense + RRF): This is the proposed approach, created by combining lexical and semantic retrieval via Reciprocal Rank Fusion (RRF).

This design resulted in a total of 2400 responses (4 conditions × 4 models × 150 questions). For each retrieval condition, source groundedness was assessed using word overlap, bigram overlap, TF-IDF similarity, and rule-based evidence verification. Because no reference passages were provided in the LLM-only condition, source-groundedness metrics were not applicable. The LLM-only condition therefore served as the no-retrieval baseline, whereas physician evaluation was performed on responses generated using the hybrid retrieval configuration selected for the final system. The ablation therefore measures source grounding only and not clinical accuracy.

3.5. Hybrid Explanation Mechanism

The second component of the proposed CARDIA-TR approach is a hybrid explanation mechanism that integrates patient-level risk predictions generated by the ML model with natural language explanations based on clinical guidelines. In this mechanism, the ML component is positioned to predict risk, while the RAG component explains that risk using guideline-based rationales. The objective is to support the numerical risk prediction with a justification layer that renders the prediction clinically interpretable, rather than presenting the figure in isolation.

The mechanism comprises three stages. In the first stage, the ML model estimates the probability of coronary heart disease based on a patient's structured clinical attributes. In the second stage, the same patient's attributes are converted into a natural language clinical case summary (vignette) via a predefined, deterministic template. This conversion is executed using a rule-based method that maps numerical and coded attributes (e.g., gender, type of chest pain, ST-segment depression value, number of vessels visualized on fluoroscopy) to human-readable clinical statements; this process is carried out without the use of any language models and in a manner that leaves no room for subjective interpretation. Since the attributes in the Cleveland dataset comprise key variables that are clinically interpretable for coronary heart disease risk assessment, the generated case summaries provide natural language input suitable for the RAG system to produce guideline-based explanations. In the third stage, the generated case summary is fed into the RAG system, which retrieves relevant guideline passages and produces a text explaining the patient's risk status based on clinical reasoning derived from those passages. In the pilot, the language model received only the case summary and the retrieved passages. It did not receive the ML prediction or the feature contributions. The pilot explanations for 30 patients—which were evaluated by

a physician—were generated using a single language model (Turkish-Gemma-9B) under a hybrid retrieval setting. This specific model was selected because it achieved the highest clinical accuracy during the initial phase of the pilot study, specifically in the physician evaluation conducted at the question-answering level. Using a single model for the pilot ensured that any variability in explanation quality stemmed from case characteristics rather than differences between models. The explanation experiment, which covers the full cohort of 303 patients and compares eight different models, is described in Section 3.5.2. The process of converting attributes into a case summary is illustrated in Table 3 using a sample patient. As shown in this example, coded attributes (such as "cp = 3" indicating the type of chest pain, or "thal = 6" indicating a perfusion finding) are mapped deterministically to their corresponding clinical terms ("non-anginal chest pain," "fixed perfusion defect") and transformed into a coherent clinical narrative.

Table 3. Example conversion from structured attributes to a clinical case summary (vignette).

Attribute	Encoded Value	Clinical Correlate
Age	64	64 years old
Gender	1	male
Chest Pain Type (CP)	3	Non-anginal chest pain
Resting Blood Pressure	140	140 mmHg
Cholesterol	335	335 mg/dL
Maximum heart rate (thalach)	158	158/min
ST Depression (oldpeak)	0	0 mm
Number of Stained Vessels (ca)	0	0
Thallium (thal)	6	Fixed perfusion defect
Generated Text	64-year-old male patient. Chest pain type: non-anginal chest pain. Resting blood pressure 140 mmHg, cholesterol 335 mg/dL. Maximum heart rate achieved during exercise: 158 bpm. ST depression during exercise: 0 mm. Number of vessels visualized on fluoroscopy: 0. Thallium scintigraphy: fixed perfusion defect.	

Thirty patients were selected by the authors from the Cleveland dataset to evaluate the explanation mechanism. The selection was stratified by the outcome of the earlier ML pipeline used in the first phase to ensure a balanced representation of various clinical scenarios and model behaviors. Selection within each group was not random. With the final logistic regression model, 28 of the 30 ML decisions are identical. Accordingly, the evaluation set included both patients correctly classified and those incorrectly classified by the ML model. The selected cohort comprised 9 true positive, 9 true negative, 6 false positive, and 6 false negative cases. This design achieved a balance in which 15 patients were classified as positive and 15 as negative by the model, while also ensuring that 15 patients had a positive and 15 a negative ground-truth label. Such a setup enables an examination of the clinical coherence and guideline compliance of the rationales generated by the explanation layer—not only when the model makes correct decisions but also when it makes erroneous ones.

3.5.1. Aspect-Based Guideline Retrieval

Regarding patient-level assessments, a criterion-based retrieval process was applied to an expanded collection—comprising 52 guideline documents and a supplementary guideline on the management of chronic coronary syndromes—to access guideline-based evidence specific to the findings of each of the 303 patients. English and Turkish template queries were generated based on aspects such as pre-test probability, exercise electrocardiography,

perfusion scintigraphy, coronary calcification, risk factors, and resting electrocardiography derived from each patient's findings; no language model was used in the query generation process. The corpus was segmented into 220-word chunks with a 40-word overlap, while bibliography and table-of-contents sections were removed using rule-based methods. For each aspect, results from BM25 and multilingual BGE-M3 dense retrieval were combined using RRF, and candidate passages were re-ranked using the BGE-reranker-v2-m3 cross-encoder [61,62]. A total of five passages were selected by including the most relevant passage for each specific aspect first, followed by the most relevant passages for the general query to fill the remaining slots. The retrieval process was performed once, independently of the model, ensuring that all language models processed the same set of passages.

3.5.2. ML-Faithful Explanation Method and Guideline-Consistency Auditlanations and Guideline-CoML-Faithful Explanations and Guideline-Consistency Audit

In this study, the fidelity of the explanation is defined as the explanation being based on the actual rationale used by the ML model for that specific patient and conveying that rationale without alteration. To ensure the explanation was grounded in the ML model's actual rationale—and given that the primary model was logistic regression—the contribution of each variable to the decision for every patient was precisely calculated on the logit scale as the product of the coefficient and the transformed value; one-hot encoded categories were aggregated under their respective variables. Continuous variables were standardized within each training fold, so their contribution is measured relative to the fold mean. Categorical variables were one-hot encoded without a reference category, so the contribution of a categorical variable equals the coefficient of the patient's category and is measured relative to the model intercept. The contributions of all variables and the intercept add up to the log-odds of the ML prediction. Each patient was explained using the model that included them in the test fold during the first iteration of repeated cross-validation; the probabilities from the retrained models were matched one-to-one with the out-of-fold probabilities. The language model was provided with the patient vignette, guideline excerpts, the ML decision and probability, and the four variables contributing most to the decision—along with their directions of influence. It was asked to explain the clinical significance of each variable by referencing the excerpts. The language model cannot alter the ML decision; however, to limit the generation of post hoc rationalizations, it can report two types of contradictions in separate fields: variables with an effect direction contrary to the guiding evidence, and instances where the body of findings contradicts the ML decision. The prompt states that the patients originated from a cohort referred for coronary angiography due to suspected coronary artery disease; this information is derived from the cohort's original definition and does not include prevalence data. To test whether this information influenced the detection of guideline discordance, the experiment was repeated for the two best models with this sentence replaced by a neutral statement that the patients were evaluated for coronary artery disease. Outputs were constrained by a JSON schema and generated using grammar-based constrained decoding [63]. By restricting the reported variable names and cited passage IDs to permitted values, references to non-existent sources were prevented, although this does not verify that a cited passage supports the specific claim made in the explanation. For two models (Llama-3.1-8B, Yi-1.5-9B) that entered an infinite loop of meaningless whitespace in the JSON output, free whitespace characters were disabled at the decoding level. In cases where the context window was exceeded, the lowest-ranked passage was removed, and the process was retried.

3.6. Evaluation Protocol

The system's performance was evaluated at two levels, corresponding to the study's two core components: the guideline-based question-answering outputs of the language models and the patient-level clinical explanations. The question-answering outputs were primarily assessed using automated and descriptive metrics. These metrics aim to quantify the lexical and semantic overlap between the generated responses and the retrieved guideline passages. Specifically, the unigram overlap rate, bigram overlap rate, and TF-IDF weighted cosine similarity between the response and the source passages were calculated [64]. Additionally, descriptive statistics—such as response length and the results of a rule-based evidence check verifying the adequacy of the context provided by the retrieval system—were reported. It is acknowledged that while these automated metrics reflect the superficial and semantic alignment of the responses with the source material, they do not—on their own—measure deeper attributes such as clinical accuracy or the quality of the reasoning. Given this limitation, the question-answering outputs were also subjected to physician evaluation. The physician evaluation was conducted on responses generated under the hybrid retrieval condition—used in the final system—rather than on all four retrieval conditions. A blinded setup was adopted to eliminate bias; specifically, model identities were masked, and responses were presented in a randomized order. Sixty questions, balanced across clinical categories, were selected, and a total of 240 responses—four per question—were scored by a physician based on clinical accuracy. Clinical accuracy was assessed to determine whether the response contained medically correct and reliable information, using a three-point scale (0 = insufficient, 1 = partial, 2 = sufficient). The physician used intermediate values for a limited number of responses where they were uncertain. In the second stage, the same physician scored the responses from Gemma-4-26B-A4B, Qwen3.6-27B, Mistral-Small-24B, and Turkish-Gemma-9B to the same 60 questions (240 responses)—using the same rubric and blinded setup—based on the re-established knowledge base. Turkish-Gemma-9B was scored in both stages to enable a comparison between them. The adherence of responses to the retrieved guideline passages was quantified using the automated evidence-faithfulness metric defined in Section Retrieval Strategy Ablation Design, rather than through a separate physician evaluation. Thus, the models were compared not only in terms of textual overlap but also regarding clinical usability. The clinical accuracy of all eight models was also evaluated by two LLM judges (Qwen3.6-27B and Gemma-4-26B-A4B) using the same three-point scale and definitions applied by the physician, based on a 60-question subset [65]. Since the passages from the reconstructed knowledge base differed from those viewed by the physician in the initial phase, the judges were shown only the questions and answers; the answers were presented in a randomized order with model identities masked. The agreement of the judges with the physician's scores was measured. To this end, the 240 original responses from the four models in the first phase and the 240 responses evaluated by the physician in the second phase were used. To mitigate potential self-preference bias arising from the fact that the judges were themselves among the models being evaluated, judges from two distinct model families were employed, and the average of the two judges was reported. Agreement with physician ratings does not make the judges a substitute for physician evaluation, and their scores are used only for the relative ranking of models. Because the 512-token limit truncated some responses, a sensitivity analysis was performed using only questions on which the compared models produced complete responses.

Clinical explanations generated at the patient level were evaluated through a separate process. Pilot explanations generated by Turkish-Gemma-9B for 30 patients selected from the Cleveland dataset were scored by a physician using the same three-point scale. The physician, a general practitioner with two years of clinical experience, saw only the case vignette and the explanation, without the ML prediction or the true outcome. The physician

assessed clinical accuracy based on the medical validity of the cited risk rationales and their alignment with criteria in clinical guidelines. Prior to the evaluation, scoring criteria were defined in a written guideline, and pre-calibration was conducted on a limited number of samples to ensure consistency. The evaluation results were analyzed based on both question categories and patient subgroups (correctly and incorrectly classified cases), thereby allowing for an examination of performance patterns regarding both clinical subject areas and model decision accuracy. Explanations generated by eight models for 303 patients were evaluated using three automated metrics. "Coverage" indicates how many of the four variables most influential to the ML decision were explained. "Citation rate" measures the proportion of variable explanations grounded in at least one guideline passage. "Topic match" indicates whether a cited passage—when a relevant guideline passage is present in the context—actually pertains to the specific variable being explained. Fidelity to the ML model was measured separately from clinical correctness. For each variable explanation, it was checked whether the patient value was stated correctly and whether the stated direction of effect matched the sign of the ML contribution. For each general explanation, it was checked whether the ML decision and probability were stated correctly. These checks used rule-based text matching. Clinical correctness was assessed by the physician on a sample of 80 explanations from the two best models (Qwen3.6-27B and Gemma-4-26B-A4B). Forty patients were drawn at random with a fixed seed, ten from each ML outcome group (true positive, true negative, false positive and false negative), and half of each group were cases in which the ML model used chest pain type against the guideline direction. Both models explained the same patients. Model identity and the true outcome were hidden, while the ML decision and its main drivers were shown. The physician used the same three-point scale and also marked whether an explanation contained a potentially harmful statement. The physician was asked to judge whether each explanation described the clinical meaning of the ML drivers correctly and in line with the guidelines, whether it contained misleading statements, and whether any reported contradictions were clinically reasonable. The physician was explicitly instructed not to judge whether the ML decision itself was correct. This evaluation therefore assesses the clinical correctness of the explanation and not its fidelity to the ML model, which was measured separately with the automatic measures. It was also examined whether the decision inconsistencies reported in the explanations could serve as a "review" alert for the physician. To this end, the ML error rate for patients with reported inconsistencies was compared to that of patients without reported inconsistencies. Furthermore, the ML errors detected by this alert were compared against those identified by a simple uncertainty alert, which selected an equal number of patients based on ML probabilities closest to 0.5. Guideline-consistency auditing was first evaluated based on the type of chest pain. Guidelines rank the probability of coronary artery disease—from highest to lowest—as typical angina, atypical angina, non-anginal chest pain, and asymptomatic status. Patients for whom the ML model's contribution to this variable ran counter to this ranking were identified as representing "true discrepancies," independent of the ground-truth labels. The variable inconsistencies reported by the language models were compared against this reference, and precision, recall, and F1 scores were calculated for each model. The reference was then extended to all clinical variables. For each variable, the direction of risk stated in the guidelines was defined. For example, older age, male sex, exercise-induced angina, higher ST depression, lower maximum heart rate, and abnormal thallium findings increase the risk. A patient–variable pair was marked as discordant when the ML contribution had the opposite direction. For continuous variables, the direction was read relative to the cohort mean, so a variable was discordant when, for example, an above-average value lowered the log-odds although the guidelines associate it with higher risk. Because categorical contributions are measured relative to the model intercept and not

to a reference category, their sign is relative. Discordance was therefore assessed only for categories with a clear guideline direction, namely typical angina and asymptomatic presentation for chest pain type, male and female sex, presence and absence of exercise-induced angina, upsloping versus flat or downsloping ST segment, normal versus abnormal thallium findings, and fasting blood sugar above 120 mg/dL and abnormal resting ECG. Other categories were not assessed. The flags are interpreted as signals that an association learned from this referral cohort may not transfer to other populations, not as errors of the model within this cohort.

3.7. Statistical Analysis

ML model performance metrics were calculated based on out-of-fold predictions. ML performance metrics are reported as mean ± standard deviation across 10 repetitions. 95% confidence intervals for physician and judge scores were calculated using a bootstrapping method with 5000 iterations [66]. Automated metrics at the question-answering level and physician evaluation scores at the patient level were summarized using descriptive statistics (mean, standard deviation, and distributions). In model- and category-based comparisons, group means and distributions were presented via tables and graphs. Since multiple models were evaluated on the same set of questions, physician and judge clinical accuracy scores were treated as paired measurements. The Friedman test was used to assess model-based differences; in cases of statistical significance, pairwise comparisons were conducted using the Wilcoxon signed-rank test with Holm correction [67–69]. Effect sizes were reported as Kendall's W for Friedman tests and as $r = Z/\sqrt{N}$ for pairwise tests, together with bootstrap 95% confidence intervals for paired differences. In the pilot, correctly and incorrectly predicted cases were compared using the Mann–Whitney U test with the rank-biserial correlation. Exact (Clopper–Pearson) confidence intervals were used for proportions. In the alert analysis, error rates for flagged versus non-flagged patients were compared using Fisher's exact test [70]. Agreement between the LLM judges and the physician, as well as between the judges themselves, was measured using quadratically weighted Cohen's kappa and Spearman's rank correlation [71]. All analyses and visualizations were performed using the Python programming language (version 3.12) and relevant scientific computing libraries (scikit-learn 1.4, pandas 2.1, NumPy 1.26, SciPy 1.11) [40,72–74].

4. Results

This section presents findings regarding the ML component, retrieval architecture, and descriptive analysis of language model outputs for the proposed CARDIA-TR framework. Subsequently, results based on physician evaluation, a comparison of eight models in terms of clinical accuracy, and guideline-consistency audits using explanations faithful to the ML decisions are presented.

4.1. Performance of the ML Model

Six classification models were evaluated to predict the presence of coronary heart disease using a repeated nested cross-validation scheme (10 repetitions × 5 outer folds; hyperparameter selection and calibration performed in inner folds). All metrics were calculated based on out-of-fold predictions and are presented in Table 4 as the mean ± standard deviation across the ten repetitions. Logistic regression achieved the highest discrimination performance (ROC-AUC 0.911 ± 0.006). This model was followed by random forest and a stacking ensemble (both 0.908), support vector machine (0.907), XGBoost (0.900), and CatBoost (0.889). Logistic regression also yielded the best values for accuracy (0.851), recall (0.806), specificity (0.888), F1 score (0.832), and Brier score (0.117). The fact that standard

deviations across repetitions remained below 0.02 for all metrics indicates that the results are stable with respect to data splitting.

Table 4. Out-of-fold performance metrics of ML models (Cleveland dataset [17], $n = 303$; 10 repeats × 5-fold nested cross-validation, calibrated probabilities, mean ± SD).

Model	ROC-AUC	Accuracy	Recall	Specificity	F1	Brier
Logistic Regression	0.911 ± 0.006	0.851 ± 0.006	0.806 ± 0.008	0.888 ± 0.012	0.832 ± 0.006	0.117 ± 0.004
Stacking	0.908 ± 0.007	0.833 ± 0.009	0.796 ± 0.014	0.865 ± 0.012	0.814 ± 0.011	0.120 ± 0.004
Random Forest	0.908 ± 0.007	0.829 ± 0.009	0.788 ± 0.012	0.863 ± 0.014	0.808 ± 0.010	0.121 ± 0.004
Support Vector Machine	0.907 ± 0.005	0.841 ± 0.007	0.801 ± 0.014	0.874 ± 0.011	0.822 ± 0.008	0.121 ± 0.004
XGBoost	0.900 ± 0.009	0.833 ± 0.010	0.786 ± 0.012	0.873 ± 0.015	0.812 ± 0.011	0.125 ± 0.005
CatBoost	0.889 ± 0.011	0.823 ± 0.011	0.782 ± 0.016	0.858 ± 0.013	0.802 ± 0.013	0.133 ± 0.006
LR without ca and thal	0.875 ± 0.004	0.793 ± 0.005	0.741 ± 0.008	0.837 ± 0.011	0.767 ± 0.005	0.143 ± 0.002

With only 303 cases, the more complex models, including random forest, gradient boosting, and stacking, did not perform better than logistic regression. Logistic regression was therefore selected as the main ML model based on the predefined selection criteria. The model with the highest mean ROC-AUC for calibrated probabilities was preferred, with the Brier score used in case of a tie. Sigmoid (Platt) calibration slightly reduced the Brier score from 0.118 to 0.117 without changing the model's discrimination performance. At a decision threshold of 0.5, the model produced an average of 26.9 false negatives and 18.3 false positives in the out-of-fold predictions. When the number of vessels colored by fluoroscopy (ca) and the thallium scintigraphy result (thal) were removed, logistic regression reached a ROC-AUC of 0.875 ± 0.004, an accuracy of 0.793, a recall of 0.741, a specificity of 0.837, and a Brier score of 0.143 (Table 4). The model therefore keeps reasonable discrimination without these two variables, but part of its performance depends on investigations that are close to the angiographic outcome.

Impact of the Evaluation Protocol on Reported Performance

Accuracy values in the 95–100% range are frequently reported in the literature for the Cleveland dataset. To investigate the source of these values, logistic regression and random forest models were retrained under six different evaluation protocols, as shown in Table 5. Confirming earlier reports [15,16], the 1025-row version widely distributed on open-access platforms contains only 302 unique patients, with the remaining 723 rows being exact duplicates. Of the unique profiles, 187 appear three times, 114 appear four times, and one appears eight times. All 302 unique profiles match a record of the UCI Cleveland dataset on the eight variables whose coding is identical in both versions, and in all matched records the target is inversely coded, with 1 indicating no disease. When this version is split at an 80/20 ratio, on average 96.9% of the test rows (range 90.7% to 100% over 100 splits) have an exact copy in the training set. In 10-fold cross-validation, this is the case for 100% of the test rows in every fold, and after duplicates are removed the overlap is zero. Under these conditions, the random forest model achieved 100% accuracy and an ROC-AUC of 1.000 in both a single split and 10-fold cross-validation. When duplicates were removed, the same model's accuracy dropped to 0.838 and its ROC-AUC to 0.910—returning to the level (0.830/0.903) obtained with the correct protocol. Logistic regression, which has a limited capacity to memorize data patterns, was not significantly affected by the duplicates (0.810–0.840). A second major source of bias is the reporting of a single training/test split. While the average accuracy across one hundred random splits was 0.839 for logistic regression, selecting the best split among them raised the accuracy

to 0.918 and the ROC-AUC to 0.973. The impact of applying preprocessing steps (scaling and feature selection) to the entire dataset prior to cross-validation was limited (±0.01) in this dataset. These findings indicate that the values reported in this study (accuracy 0.851, ROC-AUC 0.911) are close to the upper limit of performance achievable through leakage-free protocols, whereas higher values found in the literature largely stem from dataset versions containing duplicates and from selected single splits. These conclusions refer only to this dataset version and the protocols examined here.

Table 5. Impact of the evaluation protocol on reported performance (accuracy/ROC-AUC).

Protocol	Explanation	Logistic Regression	Random Forest
P1	Reference (leakage-free) protocol: UCI ($n = 303$), repeated stratified 5-fold CV, preprocessing inside the pipeline	0.835/0.903	0.830/0.903
P2a	UCI, single 80/20 split, mean over 100 seeds	0.839/0.906	0.827/0.903
P2b	UCI, single 80/20 split, best seed	0.918/0.973	0.918/0.930
P3	UCI, scaling and feature selection on the full dataset before CV	0.837/0.903	0.823/0.894
P4	1025-row version (723 duplicates), single 80/20 split	0.810/0.930	1.000/1.000
P5	1025-row version, 10-fold CV	0.840/0.919	1.000/1.000
P6	1025-row version, duplicates removed ($n = 302$), 10-fold CV	0.828/0.907	0.838/0.910

In Table 5, CV stands for cross-validation. Since the models in P1 were trained with fixed hyperparameters, the values may differ slightly from those in Table 4.

4.2. Impact of the Retrieval Strategy

To examine the contribution of the proposed hybrid retrieval approach to guideline-based explanation generation, four retrieval conditions (LLM-only, BM25-only, Dense-only, and Hybrid) were comparatively evaluated using the same 150 questions and four models. The responses for each condition—averaged across the four models—are presented in Table 6 in terms of source faithfulness metrics.

Table 6. Comparison of retrieval conditions in terms of source-grounding metrics (average of four models, 95% confidence interval).

Retrieval Condition	Word Overlap	Bigram Overlap	TF-IDF Similarity	Evidence-Commitment Ratio
LLM-only (without retrieval)	-	-	-	-
BM25-only	0.604 [0.589, 0.619]	0.328 [0.309, 0.347]	0.336 [0.325, 0.348]	0.933 [0.922, 0.944]
Dense-only	0.655 [0.641, 0.669]	0.383 [0.363, 0.403]	0.372 [0.361, 0.384]	0.961 [0.952, 0.969]
Hybrid (BM25 + Dense + RRF)	0.641 [0.627, 0.654]	0.361 [0.343, 0.380]	0.346 [0.336, 0.356]	0.949 [0.938, 0.959]

The findings indicate that all three conditions involving retrieval achieved high levels of source faithfulness. The source faithfulness score exceeded 0.93 across all conditions, demonstrating that the generated responses were largely based on the retrieved guideline passages. Dense and hybrid retrieval yielded slightly higher faithfulness values across all metrics compared to lexical (BM25) retrieval. The confidence intervals for dense and hybrid retrieval overlap significantly, indicating limited differences between these two conditions. This result suggests that, within this evaluation set of structured clinical questions, semantic

retrieval alone is sufficient to ensure high source faithfulness, and the addition of a lexical component provides no significant additional benefit in terms of source faithfulness.

Within CARDIA-TR, the retrieval component is intended to identify relevant guideline passages that can support faithful explanations of ML-generated risk predictions. The consistently high evidence-faithfulness scores across all retrieval methods indicate that guideline grounding was relatively robust to the choice of retrieval strategy. Hybrid retrieval was retained for the final system because of its potential advantage for queries involving exact lexical matches, such as drug names, clinical scores, and abbreviations, although its performance was comparable to dense retrieval in the present evaluation. Because the ablation was not evaluated by a physician, these results show source grounding only and do not show whether retrieval improves clinical accuracy compared with the LLM-only condition.

4.3. Descriptive Analysis of Language Model Outputs

Eight open-source LLMs were compared in the question-answering evaluation: Llama-3.1-8B, Qwen2.5-7B, Yi-1.5-9B, Turkish-Gemma-9B, Qwen3.6-27B, Gemma-4-26B-A4B, Mistral-Small-24B, and Turkish-Gemma-9B-T1. All models were run using the same knowledge base, the same 150 questions, the same prompt template, and the same decoding settings, resulting in a total of 1200 generated responses. The responses from the eight models are summarized in Table 7 in terms of overlap with source passages and response length. Gemma-4-26B-A4B and Llama-3.1-8B produced the highest word overlap, while Llama-3.1-8B and Gemma-4-26B-A4B achieved the highest evidence ratios. Turkish-Gemma-9B exhibited the lowest overlap values, indicating a tendency to paraphrase source content rather than reproducing it verbatim. Average response lengths ranged from 659 to 1051 characters. The rate at which the 512-token response limit was reached varied significantly among the models; this rate remained below 1% for Gemma-4-26B-A4B and Turkish-Gemma-9B, whereas it reached 29% for Qwen2.5-7B and 68% for Yi-1.5-9B. In the case of the reasoning-focused Turkish-Gemma-9B-T1, 98% of the responses were truncated at this limit.

Table 7. Descriptive analysis of responses from eight LLMs in terms of source-overlap and length.

Model	Word Overlap	Bigram Overlap	TF-IDF Similarity	Evidence Rate	Avg. Response Length	Truncated (%)
Llama-3.1-8B	0.661	0.366	0.319	0.945	1051	13.3
Qwen2.5-7B	0.550	0.235	0.325	0.924	1044	28.7
Yi-1.5-9B	0.587	0.287	0.328	0.933	929	68.0
Turkish-Gemma-9B	0.432	0.160	0.228	0.729	677	0.7
Qwen3.6-27B	0.554	0.256	0.282	0.884	721	2.7
Gemma-4-26B-A4B	0.666	0.408	0.322	0.937	674	0.7
Mistral-Small-24B	0.623	0.322	0.324	0.901	901	18.7
Turkish-Gemma-9B-T1	0.444	0.125	0.186	0.816	659	98.0

Truncated refers to the proportion of responses that hit the 512-token response limit. It should be emphasized that these descriptive metrics reflect the extent to which responses superficially adhere to the source material, but high overlap does not necessarily equate to higher clinical quality. To illustrate taken from the 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias [75]. In this example, the next-generation Qwen3.6-27B and Gemma-4-26B-A4B models accurately and concisely conveyed the guideline recommendation. Conversely, Qwen2.5-7B incorrectly linked the recommendation to LDL-C levels

and pivoted to combination therapies that were outside the scope of the question. Yi-1.5-9B failed to address the timing of statin initiation altogether, instead conveying a different recommendation from the same context. Although all four models utilized the same source passages, the responses differed markedly in terms of their focus on the question and clinical coherence (Figure 2).



Figure 2. Example comparison of guideline-grounded RAG answers from newer-generation (green) and earlier (blue) LLM for the same clinical question. All models received the same four retrieved passages. Responses were generated in Turkish and are shown in the authors' English translation; the original responses are available in the study's data The source excerpt is taken from the 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias [75]. repository. Responses are bridged for visual clarity; [...] indicates omitted text.

4.4. Physician Evaluation

Given that automated text metrics alone are limited in reflecting clinical quality, the system outputs were clinically evaluated by a physician in a blinded manner. The evaluation was conducted in two parts: first, the guideline-based question-answering outputs of the language models; and second, the patient-level clinical explanations generated from ML predictions.

4.4.1. Clinical Evaluation of Question-Answering Outputs

For 60 questions distributed evenly across clinical categories, the responses were scored by a physician for clinical accuracy on a three-point scale (0 = inadequate, 1 = partial, 2 = adequate), with model identities masked and responses presented in randomized order. The evaluation was conducted in two stages. In the first stage, the 240 responses generated by the four initial models using their original knowledge bases (i.e., the "A" rows in Table 8) were scored; in the second stage, 240 responses generated by Gemma-4-26B-A4B, Qwen3.6-27B, Mistral-Small-24B, and Turkish-Gemma-9B using the reconstructed knowledge base (i.e., the "B" rows in Table 8) were scored. Turkish-Gemma-9B was evaluated in both stages to allow for a comparison between the two. Clinical accuracy results by model are presented in Table 8. Since the "A" and "B" runs involved responses generated under different contexts, they should not be directly compared.

Table 8. Clinical accuracy results of LLM based on physician assessment ($n = 60$ for each model; 95% confidence intervals calculated using the bootstrapping method).

Model	Run	Mean	95% CI	Median	Fully Correct (%)	Insufficient (%)
Gemma-4-26B-A4B	B	1.70	[1.53, 1.85]	2.0	80.0	10.0
Qwen3.6-27B	B	1.68	[1.53, 1.82]	2.0	75.0	6.7
Mistral-Small-24B	B	1.62	[1.43, 1.78]	2.0	75.0	13.3
Turkish-Gemma-9B	B	1.55	[1.38, 1.72]	2.0	63.3	8.3
Turkish-Gemma-9B	A	1.38	[1.18, 1.57]	2.0	55	17
Qwen-2.5-7B	A	1.10	[0.91, 1.30]	1.0	37	27
Llama-3.1-8B	A	1.08	[0.87, 1.32]	1.0	42	33
Yi-1.5-9B	A	0.85	[0.65, 1.07]	1.0	28	43

The initial physician assessment revealed differences in clinical accuracy among the models. The Turkish-Gemma-9B model, trained specifically for Turkish, achieved the highest clinical accuracy, with a mean score of 1.38 and a median of 2.0; 55% of its responses were rated as completely accurate by the physician. This model outperformed the others in all comparisons, with confidence intervals showing only limited overlap with those of the other models. The general-purpose Qwen-2.5-7B and Llama-3.1-8B models demonstrated moderate performance. Conversely, the Yi-1.5-9B model exhibited the lowest clinical accuracy, with a mean score of 0.85 and 43% of its responses deemed inadequate. This result shows the opposite pattern from the automated overlap metrics reported in Section 4.3. Llama-3.1-8B had the highest word overlap with the source text, but its clinical accuracy was only in the middle range. In contrast, Turkish-Gemma-9B had the lowest overlap but achieved the highest clinical accuracy. This shows that greater textual overlap does not necessarily mean better clinical quality. A model may copy more from the source and still provide a less useful answer. Another model may paraphrase the same information and give a more clinically accurate response. For this reason, automated text similarity metrics should not be used alone to evaluate clinical performance. They should be considered together with clinical assessment.

The adherence of responses to the retrieved guideline passages was also analyzed on a per-model basis using an automated evidence-faithfulness metric. In the hybrid retrieval condition, the evidence-faithfulness rate exceeded 0.91 across all models, indicating that the source adherence provided by the RAG architecture is largely independent of model selection. Notably, the ranking of evidence-faithfulness is inversely related to the ranking of clinical accuracy. While the Yi-1.5-9B model achieved the highest evidence-faithfulness but demonstrated the lowest performance in clinical accuracy, the Turkish-Gemma-9B model—which yielded the lowest evidence-faithfulness value—ranked first in clinical accuracy. This pattern demonstrates that strictly adhering to the source text—that is, largely reproducing passages verbatim—does not, in itself, guarantee the generation of a clinically adequate response. Differences in clinical accuracy among the four models in the initial phase were evaluated using the Friedman test on paired measurements from the same question set and were found to be statistically significant ($\chi^2 = 13.40$, $p = 0.004$, Kendall's W = 0.074). Pairwise comparisons using the Wilcoxon signed-rank test with Holm correction revealed that the superiority of Turkish-Gemma-9B over Yi-1.5-9B was significant (p_Holm = 0.005), whereas the differences between Turkish-Gemma-9B and both Qwen-2.5-7B (p_Holm = 0.110) and Llama-3.1-8B (p_Holm = 0.110) did not reach statistical significance after correction. These results indicate that the Turkish-Gemma-9B

model achieved the highest average clinical accuracy across all comparisons; however, this superiority was statistically confirmed only against the lowest-performing model. Among the responses scored in the second stage, Gemma-4-26B-A4B and Qwen3.6-27B achieved the highest average clinical accuracy, while Qwen3.6-27B exhibited the lowest rate of inadequate responses. The next-generation models scored numerically higher than Turkish-Gemma-9B, but the difference among the four models was not found to be statistically significant (Friedman $\chi^2 = 3.00$, $p = 0.39$, Kendall's W = 0.017). The average scores for Turkish-Gemma-9B were 1.38 in the first stage and 1.55 in the second stage (difference 0.17, 95% CI −0.03 to 0.37, Wilcoxon $p = 0.105$). Although the difference is not significant, comparisons between rows A and B should be interpreted with caution, as the responses in the two stages were generated under different contexts. A clinical comparison covering all eight models is presented in Section 4.5.

4.4.2. Clinical Evaluation of Patient-Level Explanations

To evaluate the generation of patient-level, guideline-based explanations—the core function of the CARDIA-TR framework—a physician scored explanations generated for 30 patients selected from the Cleveland dataset, balanced according to ML prediction accuracy. In this pilot, the language model received only the case vignette and the retrieved passages and did not receive the ML prediction. The cases were selected by the authors and not at random. These 30 patients were selected to include both cases correctly and incorrectly classified by the ML model, comprising 9 true positives, 9 true negatives, 6 false positives, and 6 false negatives. The mean overall clinical accuracy score for the explanations was 1.57 (median 2.0), with 70% of the explanations rated as completely accurate by the physician. This figure is higher than the overall average observed in the initial question-answering outputs, although the two evaluations used different questions and patients and are therefore not directly comparable. However, the most notable finding is that explanation quality is strongly linked to the accuracy of the underlying ML prediction; the mean clinical accuracy score was 1.94 for cases correctly predicted by the ML model, whereas it dropped to 1.00 for cases incorrectly predicted (difference 0.94, 95% CI 0.44 to 1.42, Mann–Whitney U = 176.0, $p < 0.001$, rank-biserial r = 0.63). Inadequate explanations occurred in 0 of 18 versus 4 of 12 cases (Fisher exact test $p = 0.018$). Because the sample is small and was not selected at random, this result is exploratory. Detailed results by prediction group are presented in Table 9.

Table 9. Clinical accuracy results for patient-level explanations by ML prediction group (all descriptions were generated using Turkish-Gemma-9B and a hybrid retrieval setup).

Prediction Group	*n*	Mean	Median	Fully Correct
True Negative	9	2.00	2.0	9/9
True Positive	9	1.89	2.0	8/9
False Positive	6	1.17	1.5	3/6
False Negative	6	0.83	1.0	1/6
ML True (total)	18	1.94	2.0	17/18
ML False (total)	12	1.00	1.0	4/12

In 17 of the 18 cases correctly predicted by the ML model, the explanation was found to be entirely accurate; none of the explanations in these cases were deemed inadequate. Conversely, all four explanations rated as inadequate (i.e., scored as 0) by the physician stemmed from cases where the ML model made incorrect predictions—specifically, two false negatives and two false positives. There was not a single instance where the ML

model made a correct prediction, yet the explanation was found to be inadequate. When interpreting this finding, it is important to consider that the two components of the proposed framework operate independently of each other. The clinical case summaries used for explanation generation were constructed to include the patient's raw clinical attributes. The RAG component generated a patient-specific explanation by evaluating these data against relevant clinical characteristics and retrieved guideline-based evidence. For instance, the errors made by the ML model might be attributed to the fact that these cases involved signals that were clinically contradictory, atypical, or otherwise difficult to explain; effectively, two independent components were simultaneously challenged by the same set of cases. In this sample, the decline in explanation quality was observed to be concentrated largely in cases where the ML prediction was incorrect. This also indicates that the system's overall clinical reliability may be limited in atypical cases where both components are challenged. Given that the model produced erroneous predictions with low probability values—particularly within the false-negative group, which exhibited the lowest explanation performance—this finding underscores the importance of ML calibration and uncertainty quantification in the context of clinical decision support. The distribution of clinical accuracy for patient explanations across cases where the ML prediction was correct versus incorrect is shown in Figure 3.

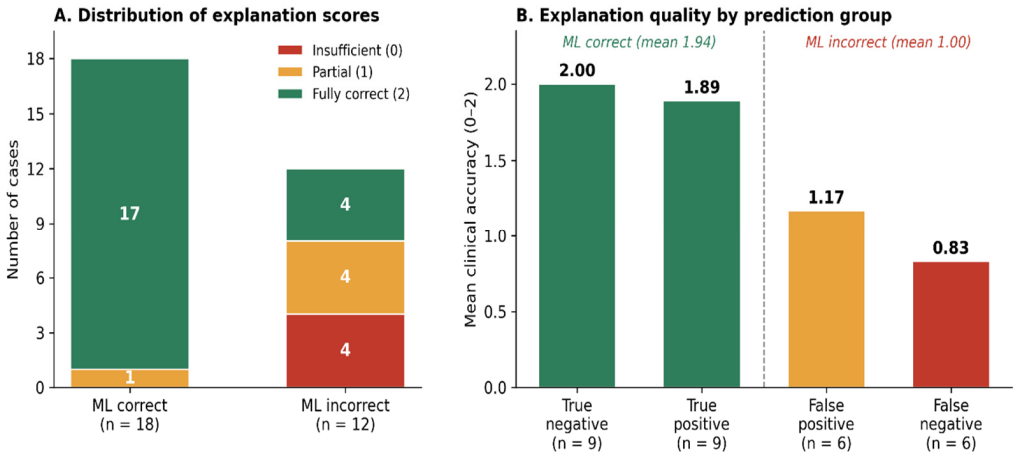


Figure 3. Distribution of clinical accuracy of patient-level explanations in cases where the ML prediction is correct and incorrect.

As shown in Figure 3, when the model makes a correct prediction, almost all explanations are evaluated as completely accurate. However, in cases of incorrect prediction, the quality of the explanation drops significantly, and all evaluations of inadequacy stem from this group.

4.5. Comparison of Eight Models in Terms of Clinical Accuracy

In this section, the responses of eight models are compared in terms of clinical accuracy. Since it was not feasible in practice for a physician to score all eight models, two LLMs were used as judges for this comparison. The agreement of the judges with the physician's ratings was first examined, followed by a comparison of the judge scores for the eight models. The results for the four models scored by the physician in the second stage were also used to check this comparison.

4.5.1. Agreement of the LLM Judges with Physician Ratings

The responses of the eight models described in Section 4.3 were scored on the 60-question subset rated by the physician by two independent LLM judges (Qwen3.6-27B

and Gemma-4-26B-A4B), using the same three-point scale and definitions as the physician evaluation. The judges' alignment with the physician's assessment was tested against the 240 original responses (previously scored by the physician), and the results are presented in Table 10. The two judges demonstrated a high level of agreement with each other; however, the agreement between the average judge score and the physician's score at the response level was only slight to fair. The judges applied stricter scoring criteria than the physician. Nevertheless, the ranking of the four models observed in the physician's assessment was largely preserved in the judges' scores. The Gemma-4 judge produced a ranking identical to that of the physician. Consequently, the judge scores were interpreted not as an absolute measure of clinical accuracy, but rather as an indicator for ranking the models. The judges were not calibrated to the physician ratings, and only their agreement with these ratings was measured. In the second phase, the agreement between the judge average and the physician was similar ($\kappa = 0.279$).

Table 10. Agreement of LLM judges with physician assessment (original hybrid responses from the previous four models, $n = 240$).

Judge	Quadratic Weighted κ	Spearman ρ	Exact Fit	Average Score (Judge/Physician)
Gemma -4-26B-A4B	0.262	0.319	0.412	0.64/1.10
Qwen3.6-27B	0.180	0.261	0.379	0.37/1.10
The average of the two judges	0.278	0.322	0.429	0.50/1.10
Among judges	0.634	-	-	-

4.5.2. LLM-Judge Scores of the Eight Models

The judge scores for the eight models are presented in Table 11. A significant difference was found among the models (Friedman $\chi^2 = 111.73$, $p < 0.001$, Kendall's W = 0.266). Qwen3.6-27B received the highest score, followed by Gemma-4-26B-A4B and Turkish-Gemma-9B. The difference between Qwen3.6-27B and Gemma-4-26B-A4B is not significant (difference 0.14, 95% CI 0.01 to 0.28, r = 0.24, Holm-adjusted $p = 0.059$). This difference was sensitive to the handling of three responses with a missing judge score (adjusted $p = 0.022$ when these were excluded). Test statistics and effect sizes for all comparisons are given in Supplementary Table S3. Qwen3.6-27B scored significantly higher than the other six models. Self-preference bias cannot be fully excluded, but both judges awarded first place to Qwen3.6-27B, while the Gemma-4 judge placed its own model in second place. In the physician evaluation, Gemma-4-26B-A4B and Qwen3.6-27B were also the two models that received the highest scores. However, the difference among the four models scored by the physician is not statistically significant. The agreement between physician and judge averages regarding these responses is at a similar level to the agreement observed in the initial stage. The divergence between automated metrics and clinical accuracy was also replicated in this comparison. Llama-3.1-8B, which exhibited the highest evidence rate in Table 7, ranked fifth among the eight models in the judge evaluation. Conversely, Qwen3.6-27B, which received the highest judge score, showed a moderate level of word overlap. This result supports the finding established in Section 4.4.1 via physician evaluation—that "source overlap does not solely represent clinical quality"—using a different evaluator and eight models. The results for Turkish-Gemma-9B-T1 should be interpreted with caution. Despite the reasoning mode being disabled, this model reached the 512-token limit in 98% of its responses, and internal reasoning text was intermingled with the final answer in 40 instances. Consequently, the model's low score reflects an incompatibility with the fixed response-budget evaluation protocol rather than a deficiency in clinical knowledge.

The model was excluded from the ranking analysis. In a sensitivity analysis restricted to complete responses, the differences between models became smaller. On the 18 questions where all seven non-reasoning models produced complete answers, the differences were not significant (Friedman $\chi^2 = 12.12$, $p = 0.059$). On the 36 questions where all models except Yi-1.5-9B produced complete answers, Qwen3.6-27B remained first (1.38), followed by Turkish-Gemma-9B (1.22) and Gemma-4-26B-A4B (1.15), and the differences remained significant ($\chi^2 = 18.92$, $p = 0.002$) (Supplementary Table S4). Truncation therefore affected the ranking of the most truncated models, while Qwen3.6-27B remained first in the subset that excluded Yi-1.5-9B. Given these limitations, the eight-model ranking should be regarded as preliminary.

Table 11. Clinical accuracy results for eight LLMs based on LLM judges (60 physician-scored questions; average of two judges, 0–2 scale; 95% confidence intervals calculated using the bootstrapping method).

Model	Mean	95% CI	Median	Fully Correct (%)	Inadequate (%)
Qwen3.6-27B	1.24	[1.07, 1.42]	1.25	33.3	11.7
Gemma-4-26B-A4B	1.10	[0.93, 1.27]	1.0	26.7	13.3
Turkish-Gemma-9B	1.04	[0.85, 1.23]	1.0	25.4	22.0
Mistral-Small-24B	0.83	[0.64, 1.02]	0.5	18.6	30.5
Llama-3.1-8B	0.69	[0.50, 0.88]	0.5	16.7	43.3
Qwen2.5-7B	0.66	[0.48, 0.85]	0.5	15.3	44.1
Turkish-Gemma-9B-T1 [a]	0.61	[0.44, 0.78]	0.5	8.3	45.0
Yi-1.5-9B	0.32	[0.17, 0.48]	0.0	8.3	73.3

Responses awarded 2 points by both judges were considered fully correct. Responses awarded 0 points by both judges were deemed inadequate. Since the judge scores are stricter than the physician's scores, the values should not be directly compared with those in Table 8. For three models, one judge did not return a valid score for one response ($n = 59$). [a] Reasoning-oriented model, excluded from the ranking interpretation because of its incompatibility with the fixed response budget.

4.6. ML-Faithful Explanations and Guideline-Consistency Audit

In CARDIA-TR, the diagnostic decision is made by an ML component, while a language model explains this decision using evidence from clinical guidelines. To ensure the explanation is grounded in the ML model's actual rationale, the four variables contributing most significantly to the decision were calculated for each patient via logistic regression; these were then fed into the language model alongside the ML decision, the associated probability, and the direction of each variable's influence. The language model was tasked with explaining the clinical significance of each variable by citing relevant passages from the guidelines. While the model cannot alter the ML decision, it can separately report two types of contradictions: a "variable contradiction" occurs when the ML model utilizes a variable in a direction contrary to the guidelines, whereas a "decision contradiction" arises when the patient's findings, viewed holistically, do not align with the ML model's decision. All eight models generated explanations for 303 patients.

4.6.1. Scope and Evidence Base of the Explanations

The explanation metrics are presented in Table 12. All models individually explained nearly all of the most influential ML variables. The rate of citation to the guideline passage within the variable explanations ranged from 0.733 to 1.000. In instances where the relevant

guideline passage was present in the context, the rate at which the cited passage's topic matched the explained variable ranged from 0.562 to 0.681. Qwen3.6-27B and Turkish-Gemma-9B were among the models with the highest values. A high citation rate does not, by itself, imply an accurate citation. Although Llama-3.1-8B provided citations for all explanations, it ranked sixth among the eight models in terms of topic matching. The frequency of reporting decision conflicts varied significantly across models; Mistral-Small-24B reported a decision conflict in only 1.3% of explanations, whereas Qwen2.5-7B did so in 47.5%. Among patients for whom a decision conflict was reported, the ML error rate was found to be significantly higher only for Gemma-4-26B-A4B. Even for this model, flagging the same number of patients based on ML uncertainty captured slightly more errors. This result indicates that the language model's decision-level objection does not provide additional warning value beyond the ML's own uncertainty. Fidelity to the ML model is reported in Table 13. The patient value was stated correctly in 84% to 100% of the variable explanations, depending on the model. A direction of effect was stated explicitly in 51% to 86% of the variable explanations, and among these the stated direction matched the sign of the ML contribution in 92% to 99%. The ML decision was stated correctly in 47% to 98% of the general explanations. Because these measures rely on rule-based text matching, they should be regarded as approximate.

Table 12. Evaluation criteria for explanations consistent with the ML decision (n = 303 patients).

Model	Valid Response	Scope	Citation Rate	Topic Matching	Decision Inconsistency (%)	Variable Discrepancy (%)
Qwen3.6-27B	300	0.998	0.838	0.667	12.0	45.7
Gemma-4-26B-A4B	303	1.000	0.733	0.634	21.8	57.8
Turkish-Gemma-9B	303	1.000	0.767	0.661	8.9	33.0
Turkish-Gemma-9B-T1	302	1.000	0.998	0.633	6.3	18.9
Mistral-Small-24B	297	1.000	0.985	0.569	1.3	6.4
Llama-3.1-8B	301	1.000	1.000	0.589	36.2	98.3
Qwen2.5-7B	301	0.998	0.914	0.562	47.5	47.2
Yi-1.5-9B	280	1.000	1.000	0.681	37.1	86.1

Table 13. Fidelity of the explanations to the ML model (n = 303 patients). Values are proportions. The direction agreement is calculated among explanations that state a direction explicitly. All measures are based on rule-based text matching.

Model	Value Stated Correctly	Direction Stated Explicitly	Direction Matches ML	Decision Stated Correctly	Decision Stated Incorrectly
Qwen3.6-27B	1.000	0.707	0.953	0.907	0.003
Gemma-4-26B-A4B	0.965	0.764	0.981	0.983	0.000
Turkish-Gemma-9B	0.939	0.594	0.918	0.818	0.007
Turkish-Gemma-9B-T1	1.000	0.737	0.935	0.983	0.000
Mistral-Small-24B	0.999	0.856	0.985	0.909	0.003
Llama-3.1-8B	0.997	0.836	0.988	0.811	0.010
Qwen2.5-7B	0.840	0.672	0.949	0.870	0.010
Yi-1.5-9B	0.835	0.511	0.958	0.468	0.021

4.6.2. Guideline-Consistency Audit

An analysis of the variable discrepancies revealed that "chest pain type" was the most frequently reported variable across seven of the eight models. This finding aligns with the internal structure of the ML model: chest pain type ranks among the four most influential variables for 218 patients, yet in 150 of these cases, it is utilized in a manner contrary to clinical guidelines. While guidelines associate typical angina with the highest pre-test probability and the absence of symptoms with the lowest, the prevalence of CHD among asymptomatic patients in the Cleveland cohort is 73%. This reflects a referral bias stemming from the fact that the cohort consists of patients referred for angiography. The ML model learned this cohort-specific relationship from the data and, consequently, treated the absence of symptoms as a factor increasing the likelihood of CHD. The extent to which language models could detect this discrepancy was measured against a reference derived from the symptom hierarchy in the guidelines—without utilizing label information—and the results are presented in Table 14. Gemma-4-26B-A4B identified 146 of the 150 discrepant cases without a single false alarm. Qwen3.6-27B also produced no false alarms and generated the fewest unnecessary alerts per patient regarding other variables. Two distinct patterns of failure were observed among the other models: models such as Yi-1.5-9B and Llama-3.1-8B generated excessive and indiscriminate alerts, whereas Mistral-Small-24B and Turkish-Gemma-9B overlooked the majority of the discrepancies. In particular, the fact that Mistral-Small-24B hardly ever questions the ML decision points to a risky tendency toward affirmation in the context of clinical decision support. These findings indicate that the explanation layer can serve not only to justify the ML decision but also to function as a guideline-consistency audit, making visible to the physician any relationships the ML has learned from the data that contradict clinical guidelines. The reliability of this function, however, depends significantly on the model's output. When the reference was extended to all clinical variables, 153 of the 1137 patient–variable pairs that could be assessed were used by the ML model against the guideline direction. These involved chest pain type (149), a fixed thallium defect (3), and age (1). Qwen3.6-27B detected these with a precision of 0.903, a recall of 0.867, and an F1 score of 0.884 (Table 15). Gemma-4-26B-A4B reached the highest recall (0.967) but produced more false alerts, mainly for maximum heart rate and ST slope (precision 0.725, F1 0.829). The other six models reached F1 scores between 0.18 and 0.44. When the referral information was removed from the prompt, the detection of discordant chest pain use changed very little. The F1 score was 0.924 compared with 0.935 for Qwen3.6-27B and 0.990 compared with 0.986 for Gemma-4-26B-A4B, and precision remained 1.000 for both models (Supplementary Table S5). This suggests that the detection had limited sensitivity to the referral statement. Within this referral cohort, the association learned by the ML model is consistent with the data. The flags therefore indicate an association that conflicts with guideline knowledge and may not transfer to other populations, rather than an error of the model within this cohort.

Table 14. Performance of language models in detecting guideline-discordant use of chest pain type by the ML model.

Model	Patient Evaluated	True Anomaly	Marked	Precision	Recall	F1	Alert on Other Variables (Per Patient)
Gemma-4-26B-A4B	218	150	146	1.000	0.973	0.986	0.191
Qwen3.6-27B	215	147	129	1.000	0.878	0.935	0.057
Llama-3.1-8B	216	148	120	0.908	0.736	0.813	0.797
Yi-1.5-9B	197	133	147	0.694	0.767	0.729	2.411

Table 14. *Cont.*

Model	Patient Evaluated	True Anomaly	Marked	Precision	Recall	F1	Alert on Other Variables (Per Patient)
Qwen2.5-7B	216	149	79	0.759	0.403	0.526	0.575
Turkish-Gemma-9B-T1	218	150	31	0.839	0.173	0.287	0.215
Turkish-Gemma-9B	218	150	30	0.767	0.153	0.256	0.300
Mistral-Small-24B	215	148	18	0.944	0.115	0.205	0.020

Table 15. Guideline-consistency audit extended to all clinical variables (1137 assessable patient–variable pairs, of which 153 were used by the ML model against the guideline direction).

Model	Correctly Flagged	False Alerts	Missed	Precision	Recall	F1
Qwen3.6-27B	130	14	20	0.903	0.867	0.884
Gemma-4-26B-A4B	148	56	5	0.725	0.967	0.829
Turkish-Gemma-9B	24	90	129	0.211	0.157	0.180
Turkish-Gemma-9B-T1	26	65	127	0.286	0.170	0.213
Mistral-Small-24B	17	6	133	0.739	0.113	0.197
Llama-3.1-8B	110	237	41	0.317	0.728	0.442
Qwen2.5-7B	61	171	91	0.263	0.401	0.318
Yi-1.5-9B	103	669	33	0.133	0.757	0.227

4.6.3. Physician Evaluation of Faithful Explanations

The physician scored 80 faithful explanations from Qwen3.6-27B and Gemma-4-26B-A4B for 40 randomly selected patients, stratified by ML outcome. All 80 explanations were rated fully correct (95% CI 95.5% to 100%), and none was marked as containing a potentially harmful statement (0 of 80, upper 95% confidence limit 4.5%). This also held for the 40 explanations of patients in whom the ML prediction was wrong (95% CI 91.2% to 100%). In the pilot, where the language model did not receive the ML prediction, explanation quality dropped when the ML prediction was wrong. In the faithful design, the explanation conveys the actual reasoning of the ML model, and the physician judged this reasoning to be explained correctly even when the final decision was incorrect. Because all scores were identical, the two models could not be compared and no test between subgroups was possible. This result shows that the physician judged the sampled explanations to be clinically sound. It does not show that every explanation was fully faithful to the ML model, or that the ML decisions themselves were clinically justified. The automatic measures in Table 13 were computed on all 303 patients and capture a different property. For example, Qwen3.6-27B stated the ML decision correctly in 90.7% of all general explanations, while all sampled explanations were rated fully correct.

5. Discussion

In this study, the CARDIA-TR framework—which integrates ML-based risk prediction with RAG-based guideline explanations—was developed and evaluated within the context of guidelines for coronary heart disease, with a question-answering component that also covers acute coronary syndrome. The study's key findings can be summarized in four points. First, logistic regression achieved the highest discrimination performance under leakage-free repeated nested cross-validation, while more complex ensemble methods provided no additional benefit. Confirming earlier reports, higher values reported in the literature

largely stem from data versions containing duplicates and from selected single data splits. Second, a disconnect between automated source-overlap metrics and clinical accuracy was consistently observed across eight models by both the physician and the LLM judges. Third, next-generation models achieved the highest clinical accuracy scores in both physician and judge assessments; however, this superiority did not reach statistical significance in the physician assessment. Fourth, the explanation layer—designed to be faithful to the ML decision—not only justified the decision but also identified relationships learned by the ML model that were specific to the cohort yet inconsistent with the guidelines. The reliability of this function was markedly dependent on the model generation. The fact that logistic regression yielded the best results among the six models—in terms of both discrimination and probabilistic accuracy—demonstrates that structured clinical variables in a sample of 303 cases can be modeled using a largely linear relationship pattern. More flexible approaches, including random forest, gradient boosting, and stacking, did not outperform logistic regression. This is consistent with the idea that increasing model complexity does not necessarily improve performance when the available clinical dataset is relatively small. Logistic regression was therefore a suitable choice as the primary model for two reasons. First, it achieved the best overall performance. Second, its structure allowed the contribution of each variable to an individual prediction to be calculated directly. These patient-specific contributions could then be passed to the language model, allowing the explanation to reflect the reasoning actually used by the ML model rather than relying on a post hoc justification. Calibration also produced a small improvement in the Brier score. Although the change was limited, reliable probability estimates are important in the proposed framework because the explanation layer also interprets the predicted risk value. Part of the model's performance depends on two variables, the number of vessels colored by fluoroscopy and the thallium scintigraphy result, which come from investigations closely related to the angiographic outcome. Without these variables, the ROC-AUC decreased from 0.911 to 0.875. The model therefore remains informative without them, but its use as decision support before invasive or imaging tests would require a model built only on information available at that stage. In line with earlier reports [15,16], the leakage analysis also explained some of the unusually high results reported for the Cleveland dataset. The commonly distributed 1025-row version contained only 302 unique patients. In random 80/20 splits, on average 96.9% of the test samples had an identical counterpart in the training data, and in 10-fold cross-validation this was the case for all test samples. Under these conditions random forest reached 100% accuracy [32,45]. All unique records matched the UCI Cleveland data, and the target was inversely coded in this version, which is important for studies that use it. Once duplicate records were removed, its performance returned to a level similar to that obtained with the leakage-free evaluation protocol. Logistic regression, which is less able to memorize duplicated observations, was affected much less by this problem. Together, these results indicate that the exceptionally high performance values were mainly caused by overlap between the training and test sets rather than by true generalization. A similar issue was observed with single train/test splits. Choosing only the most favorable split increased the reported accuracy by approximately eight percentage points. For this reason, results obtained on the Cleveland dataset should be interpreted together with the exact dataset version, splitting procedure, and number of repetitions used in the evaluation. Under a leakage-free protocol, the results obtained in the present study appear to be close to the realistic upper range for this dataset. The retrieval ablation produced another relevant observation. All three conditions that included retrieval showed high evidence-faithfulness scores, indicating that the guideline corpus and retrieval pipeline were generally able to provide useful context for the clinical questions [46,49,50]. Dense retrieval produced the highest values for the automated overlap measures, suggesting that semantic similarity alone

was already effective for many of the structured questions used in the study. Hybrid retrieval gave results similar to dense retrieval. It did not clearly perform better on the automated metrics. However, it was kept in the final system because lexical search can still be useful when exact terms are important, such as drug names, clinical scores, and abbreviations. The ablation was assessed only with source-grounding metrics, so it does not show whether retrieval improves clinical accuracy compared with answering without retrieval. Another clear finding was that higher overlap with the source text did not always mean higher clinical accuracy. Llama-3.1-8B had very high overlap with the retrieved passages, but its physician score was only in the middle range. Turkish-Gemma-9B showed much lower overlap, yet it achieved the highest clinical accuracy in the initial physician evaluation. The same pattern was also seen in the eight-model comparison using the reconstructed knowledge base. Llama-3.1-8B again had one of the highest evidence-related scores but ranked fifth in the LLM-judge evaluation. In contrast, Qwen3.6-27B achieved the highest clinical score even though its source overlap was only moderate. This shows that a model can copy a large part of a guideline passage and still give a weak answer. It may miss the main point, leave out important information, or use the source in the wrong context. Figure 2 gives an example of this problem. Yi-1.5-9B received the same retrieved passages as the other models but still did not answer the question directly. For this reason, automated overlap measures are useful for showing how closely an answer follows the source, but they should not be treated as direct measures of clinical quality. Expert assessment remains necessary. The newer-generation models generally performed best in terms of clinical accuracy. Qwen3.6-27B and Gemma-4-26B-A4B occupied the top two positions in the judge-based comparison, and Qwen3.6-27B scored significantly higher than six of the other models. These two models also had the highest mean physician scores. However, the differences among the four models included in that physician comparison were not statistically significant. The physician evaluation included only 60 questions, which limits the ability to detect relatively small differences between models. The observed advantage may also reflect both model generation and parameter size, and the current experimental design cannot separate these two factors. Because truncation affected some models, the eight-model ranking should be regarded as preliminary. When only complete responses were analyzed, the differences between models became smaller, although Qwen3.6-27B remained first in the subset that excluded Yi-1.5-9B. Turkish-Gemma-9B is notable in this respect. Despite being substantially smaller than the newer models, it achieved the highest clinical accuracy in the initial phase and ranked third in the eight-model comparison. In the second physician assessment, its performance was not significantly different from that of the larger next-generation models. This suggests that language-specific training may partly compensate for model size in Turkish clinical applications, particularly because Turkish is morphologically rich and has fewer clinical NLP resources than English [13,14,30]. Another observation concerned Turkish-Gemma-9B-T1. Nearly all of its responses reached the fixed generation limit. This makes its results difficult to interpret in the same way as those of the other models. Reasoning-oriented models may require a larger response budget, and applying the same short output limit to all models may disadvantage models that use longer reasoning processes. The comparison between the LLM judges and the physician showed a similar mixture of strengths and limitations. The two judges agreed closely with each other and reproduced the overall physician-based ranking of the models reasonably well. At the individual-response level, however, agreement with the physician was only slight to fair in both evaluation stages. The judges also tended to assign lower scores than the physician. The judges were not calibrated, and only their agreement was measured. Self-preference bias cannot be fully excluded because both judges were among the evaluated models, although the Gemma-4 judge ranked its own model second and the physician placed the same two models at the

top. These findings support the use of LLM judges as a practical screening tool when many models or responses need to be compared, particularly for relative ranking [65]. They should not, however, be treated as a substitute for physician evaluation when the aim is to determine absolute clinical accuracy.

In the exploratory patient-level pilot, the quality of explanations was closely linked to the accuracy of the underlying ML prediction. While the mean clinical accuracy score for explanations was 1.94 in cases where the ML prediction was correct, it dropped to 1.00 in cases where the prediction was incorrect; furthermore, all explanations deemed inadequate by the physician occurred in cases where the ML model had erred. Since the explanation component in the pilot operated based solely on the case summary—without access to the ML prediction—this pattern suggests that both components struggled with the same cases; in other words, the cases where the ML model failed contained signals that were clinically more atypical or contradictory. This finding indicates that the explanation layer should not be viewed as a self-correcting oversight mechanism for errors in the prediction layer, and that the system's clinical reliability may be limited, particularly in atypical cases. The faithful design changes this picture. When the language model received the ML decision and its actual feature contributions, the physician rated all 80 sampled explanations as fully correct, including the 40 explanations of wrong ML predictions. In this design the explanation describes what the ML model did, so its quality no longer depends on whether the decision is correct. A faithful explanation can therefore describe a wrong decision correctly. Fidelity to the model and correctness of the decision are different properties, and the guideline-consistency audit is needed to reveal reasoning that conflicts with clinical knowledge. A "faithfulness" experiment involving eight models and 303 patients revealed an additional function for the explanation layer—representing a central finding of the study: a guideline-consistency audit. In the Cleveland cohort, the prevalence of CHD among patients referred for angiography was 73% even for those who did not report chest pain—a finding that contrasts with clinical guidelines, which associate the absence of symptoms with the lowest pre-test probability [18,31]. The ML model learned this cohort-specific association from the data and utilized the "chest pain type" variable in a manner contrary to guidelines for 150 of the 218 patients for whom this variable was influential. New-generation models have detected this anomaly with high accuracy: Gemma-4-26B-A4B flagged 97% of the anomalous cases without a single false alarm, while Qwen3.6-27B flagged 88%, also without any false alarms. When the reference was extended to all clinical variables, Qwen3.6-27B kept a high F1 score (0.884), and removing the referral information from the prompt changed the detection very little. This suggests limited sensitivity to the referral statement. This finding demonstrates that the explanation layer can make visible to the clinician relationships learned by the ML from the data—such as those stemming from cohort characteristics like referral bias—that contradict clinical knowledge. Within this cohort, the learned association is consistent with the data, so the flags do not mark errors of the model. They mark associations that may not transfer to other populations. Such scrutiny can also provide an early warning regarding which relationships might fail to generalize if the ML model were transferred to a different patient population. In contrast, some other models generated excessive and non-discriminatory alerts, while others largely overlooked anomalies. Notably, the fact that Mistral-Small-24B almost never questioned the ML decision suggests that language models in clinical decision support systems may tend to uncritically endorse model outputs, a behavior that warrants specific consideration during model selection. The value of the guideline-consistency audit emerged not at the decision level, but at the variable level. When the language model disagreed with the overall ML decision, the ML error rate was clearly higher in only one model. Even in that model, flagging the same number of cases by ML uncertainty caught more errors. The

value of the explanation layer is therefore not in predicting which cases the ML model gets wrong. Its value is that it shows the reasons behind each decision and points out which of these reasons conflict with clinical knowledge. In CARDIA-TR, we deliberately kept the diagnostic decision in the ML component and limited the language model to explanation. The language model sees the same structured variables as the ML model and has no other information about the patient. The ML model, on the other hand, is validated with out-of-fold predictions, calibrated and reproducible, and the reasons for each of its decisions can be calculated exactly. The language model is useful because it can connect guideline evidence to the patient's findings, describe this in plain language and point out reasons that go against that evidence. We based the explanation on the actual variable contributions of the ML model and did not allow the language model to change the decision. This reduces the risk of post hoc rationalization, a frequent criticism of explainable AI, while the model can still question the ML component by reporting contradictions [7,8,21]. These results have several practical implications for clinical decision support. First, risk estimates from structured data may be easier for physicians to interpret when they come with a plain-language explanation that is based on the real reasons for the decision and supported by guidelines. Second, in RAG systems it is not enough that sources are retrieved or that the answer overlaps strongly with them. Whether the model uses these sources correctly has to be checked separately. Third, in Turkish clinical decision support, models trained for the language can compete with larger general-purpose models, so model choice should take language fit into account and not only model size. Fourth, LLM judges are a practical way to screen many models, but clinical accuracy should finally be judged by physicians. Finally, open-source models that run locally performed well enough in these tasks, which supports decision support systems in which patient data stay within the institution.

The study has some limitations. First, the Cleveland dataset used for the ML component was designed for predicting coronary heart disease risk from structured data. It does not include variables that are key to the management of acute coronary syndrome, such as troponin changes, time of symptom onset, serial ECG findings and hemodynamic status [3,17,18]. Furthermore, due to its composition of patients referred for angiography, it is subject to referral bias. Consequently, patient-level results should be interpreted within the context of chronic risk assessment and should not be generalized to other populations. The data were collected in the 1980s, and the reported ML performance reflects internal validation only. External validation in independent and contemporary cohorts is required before clinical use. In addition, two predictors come from investigations closely related to the angiographic outcome. Second, the entire clinical assessment was conducted by a single physician who was not a cardiology specialist; an inter-rater reliability analysis could not be performed. The 30 pilot cases were selected by the authors and not at random. In the evaluation of 80 faithful explanations, all explanations received the highest score. This ceiling effect did not allow the two models to be compared, and the three-point scale may not distinguish between explanations of similar quality. The physician assessment did not cover all eight models but was based on responses generated in different contexts across two stages. The physician evaluation of patient explanations covered a pilot sample of 30 cases and a sample of 80 faithful explanations. The remaining explanations were evaluated only with automated metrics, and the fidelity measures rely on rule-based text matching. Topic matching checks only whether a cited passage concerns the explained variable. It is not a claim-level verification that the passage supports each statement in the explanation, so the consistency between the narrative and the guideline evidence was assessed only to a limited extent. Third, the knowledge base was reconstructed for the eight-model comparison, and the retrieved passages did not exactly match those from the initial stage. Additionally, the inference infrastructure and the level of quantization differed between the two stages. The

retrieval ablation was assessed only with source-grounding metrics. Fourth, the agreement between the LLM judges and the physician at the response level was limited, and the fact that the judges were also among the models being evaluated introduced a risk of self-preference bias. The judges were not calibrated to the physician ratings. A judge from outside the evaluated model set was not added. Fifth, the 512-token response limit resulted in the truncation of significant portions of responses for some models, which may have negatively impacted their performance. A sensitivity analysis restricted to complete responses reduced the differences between models. Sixth, the guideline-consistency audit was evaluated using rule-based references. The reference was extended to all clinical variables, but it remains based on simple risk directions and cannot capture more complex clinical reasoning. Seventh, the knowledge base relies on text extracted from PDF documents; the incomplete transfer of information from figures and tables—such as pre-test probability charts—into the text may affect the quality of retrieval and explanation. Finally, the models evaluated were limited to open-source models with up to 27 billion parameters that could be run locally on a single GPU; larger or closed-source models might yield different results. As the study examined only Turkish-language response generation and specific cardiology question categories, the generalizability of the findings to other languages and clinical domains requires further investigation.

6. Conclusions

In this study, the CARDIA-TR framework—which integrates ML-based risk prediction with RAG-based guideline explanations within the context of Turkish cardiology—was proposed and evaluated. Within the framework, the diagnostic decision is generated by a calibrated and validated ML component, while an LLM explains this decision based on the model's actual variable contributions and clinical guideline evidence. Among the ML components, logistic regression achieved the highest performance (ROC-AUC 0.911, accuracy 0.851) under leakage-free repeated nested cross-validation; more complex ensemble methods did not yield additional improvements. Confirming earlier reports, the leakage analysis showed that, for the dataset version and protocols examined, accuracy rates of 95–100% such as those reported in the literature for the Cleveland dataset can be explained largely by duplicate rows and the use of selected single splits, highlighting the necessity of interpreting reported performance metrics in conjunction with the evaluation protocol. The guideline-based question-answering performance of eight open-source LLMs was compared using automated metrics, physician evaluations, and two LLM judges whose agreement with physician ratings was measured. It was consistently demonstrated across the eight models and various evaluators that high overlap with the source text does not equate to high clinical accuracy. The next-generation models Qwen3.6-27B and Gemma-4-26B-A4B achieved the highest clinical accuracy scores in both judge and physician evaluations. Meanwhile, Turkish-Gemma-9B—a model trained specifically for Turkish—demonstrated performance statistically indistinguishable from these models in physician evaluations, despite being approximately three times smaller. In the exploratory pilot, in which the language model did not receive the ML prediction, the quality of the explanation was closely related to the accuracy of the ML prediction. With the faithful design, the physician rated all 80 sampled explanations as fully correct, including those for wrong predictions. This finding indicates that the explanation layer should be positioned not as a mechanism that automatically corrects prediction errors, but as a support layer that renders the rationale behind the decision transparent. Furthermore, the explanation layer—which remained faithful to the ML decision—was able to detect relationships learned by the model that were specific to the cohort yet inconsistent with clinical guidelines. Next-generation models flagged cases where the ML utilized chest pain type in a manner contrary to guidelines

with high accuracy and without generating false alarms (F1 scores of 0.986 and 0.935), also without referral information in the prompt. This guideline-consistency audit may help make data-driven biases visible to physicians within explainable clinical decision support systems. These results come from a single, internally validated referral cohort and require external validation. The audit should be regarded as an experimental mechanism for comparing model contributions with clinical knowledge, not as a validated clinical safety system.

Future studies will involve the evaluation of explanations and question-answering outputs by multiple cardiology specialists, the testing of the system using prospective acute coronary syndrome data—incorporating variables such as troponin dynamics and serial ECG findings—and the development of guideline-consistency references that go beyond simple risk directions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics16233986/s1, Table S1: Model identifiers, quantization and inference frameworks; Table S2: Settings of the two experimental phases and of the patient-level explanation experiment; Table S3: Paired comparisons of clinical accuracy scores with effect sizes; Table S4: Sensitivity of the judge-based comparison to response truncation; Table S5: Detection of guideline-discordant use of chest pain type with and without referral information in the prompt.

Author Contributions: Conceptualization, K.K. and M.U.; methodology, K.K., A.B. and M.U.; software, K.K.; validation, E.D.; formal analysis, K.K. and A.B.; investigation, K.K. and E.D.; data curation, K.K.; writing—original draft preparation, K.K.; writing—review and editing, M.U., A.B. and E.D.; visualization, K.K.; supervision, M.U.; project administration, M.U. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Firat University Scientific Research Projects Coordination Unit under the Comprehensive Research Project program, with project number MF.26.91 titled "Systematic Analysis of Data Sets Used in Cybersecurity and Artificial Intelligence Systems". The authors also gratefully acknowledge the financial support provided by the Fırat University Scientific Research Projects Coordination Unit through Project No. FUBAP-ADEP.22.06.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it used a publicly available, fully anonymized secondary dataset, no interventions involving human participants were performed, and the clinical evaluation was conducted on language model outputs rather than on patient data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Cleveland Heart Disease dataset is publicly available from the UCI Machine Learning Repository [17]. The 150 clinical questions, all model responses, physician and LLM reviewer evaluations, patient-level explanations, out-of-fold machine learning predictions and feature contributions, prompt templates, the case scenario template, the scoring rubric, and all analysis code are openly provided on the GitHub page at https://github.com/keremcankarakaya/CARDIA-TR (accessed on 8 October 2026). Guideline documents included in the knowledge base are not redistributed due to copyright restrictions; however, a list containing the full bibliographic details of these documents is available in the same repository.

Acknowledgments: This study is derived from the master's thesis of Keremcan Karakaya, completed at the Graduate School of Natural and Applied Sciences, Fırat University, under the supervision of Mustafa Ulaş. During the preparation of this study, the authors used Claude (Anthropic, Claude Opus 5.5) for preparing figures and language editing of the manuscript. The authors reviewed and edited all outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors have declared that there is no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CHD	Coronary Heart Disease
ACS	Acute Coronary Syndrome
NSTEMI	Non-ST-Elevation Myocardial Infarction
STEMI	ST-Elevation Myocardial Infarction
ML	Machine Learning
LLM	Large Language Model
RAG	Retrieval-Augmented Generation

References

1. Liu, C.; Jin, Q.; Han, C.; Jiao, M. Global, regional, and national burden of ischaemic heart disease from 1990 to 2021: A comprehensive analysis based on the Global Burden of Disease study 2021. *J. Glob. Health* **2025**, *15*, 04291. [CrossRef]
2. Jing, S.; Chen, A.; Xia, Y.; Ma, J.; Lu, D.; Chen, Z.; Qian, J.; Ge, J. Global burden of ischemic heart disease from 1990 to 2021: Findings from the Global Burden of Disease Study 2021 and predictions to 2030. *Chin. Med. J.* **2026**, *139*, 1181–1190. [CrossRef]
3. Byrne, R.A.; Rossello, X.; Coughlan, J.J.; Barbato, E.; Berry, C.; Chieffo, A.; Claeys, M.J.; Dan, G.-A.; Dweck, M.R.; Galbraith, M.; et al. 2023 ESC Guidelines for the management of acute coronary syndromes. *Eur. Heart J.* **2023**, *44*, 3720–3826. [CrossRef]
4. Virani, S.S.; Newby, L.K.; Arnold, S.V.; Bittner, V.; Brewer, L.C.; Demeter, S.H.; Dixon, D.L.; Fearon, W.F.; Hess, B.; Johnson, H.M.; et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the Management of Patients With Chronic Coronary Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. *Circulation* **2023**, *148*, e9–e119. [CrossRef]
5. Kahraman, A. Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data. *Front. Artif. Intell.* **2025**, *8*, 1694450. [CrossRef]
6. Zhao, S.; Nan, B.; Guo, J.; Xu, W.; Li, Z. Coronary heart disease risk prediction based on GAIN imputation and interpretable machine learning. *Front. Genet.* **2026**, *16*, 1752811. [CrossRef]
7. Talukder, A., Md.; Talaat, A.S.; Kazi, M.; Khraisat, A. XAI-HD: An explainable artificial intelligence framework for heart disease detection. *Artif. Intell. Rev.* **2025**, *58*, 385. [CrossRef]
8. Paul, V.V.; Masood, J.A.I.S. An explainable machine learning framework for cardiovascular risk prediction using structured health data. *Front. Artif. Intell.* **2026**, *9*, 1812622. [CrossRef]
9. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
10. Yang, R.; Ning, Y.; Keppo, E.; Liu, M.; Hong, C.; Bitterman, D.S.; Ong, J.C.L.; Ting, D.S.W.; Liu, N. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* **2025**, *2*, 2. [CrossRef]
11. Ke, Y.H.; Jin, L.; Elangovan, K.; Abdullah, H.R.; Liu, N.; Sia, A.T.H.; Soh, C.R.; Tung, J.Y.M.; Ong, J.C.L.; Kuo, C.-F.; et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digit. Med.* **2025**, *8*, 187. [CrossRef]
12. Tung, J.Y.M.; Le, Q.; Yao, J.; Huang, Y.; Lim, D.Y.Z.; Sng, G.G.R.; Lau, R.S.E.; Tan, Y.G.; Chen, K.; Tay, K.J.; et al. Performance of Retrieval-Augmented Generation Large Language Models in Guideline-Concordant Prostate-Specific Antigen Testing: Comparative Study With Junior Clinicians. *J. Med. Internet Res.* **2025**, *27*, e78393. [CrossRef]
13. Bayram, M.A.; Diri, B.; Yildirim, S. Healthcare-Focused Turkish Medical LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–13. [CrossRef]
14. Koçak, M.; Oğuz, A.K.; Akçalı, Z. The role of artificial intelligence in medical education: An evaluation of Large Language Models (LLMs) on the Turkish Medical Specialty Training Entrance Exam. *BMC Med. Educ.* **2025**, *25*, 609. [CrossRef]
15. Kilim, O. Duplicate clinical profiles as a source of data leakage in heart disease classification: A benchmark audit of validation, calibration, and explainability. *Adv. Artif. Intell. Res.* **2026**, *6*, 52–66. [CrossRef]
16. Simmons, B. Investigating Heart Disease Datasets and Building Predictive Models. Master's Thesis, Elizabeth City State University, Elizabeth City, NC, USA, 2021.
17. Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. *Heart Disease. UCI Machine Learning Repository*; University of California, Irvine (UCI): Irvine, CA, USA, 1989. [CrossRef]
18. Detrano, R.; Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Schmid, J.-J.; Sandhu, S.; Guppy, K.H.; Lee, S.; Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Am. J. Cardiol.* **1989**, *64*, 304–310. [CrossRef]
19. Mridha, K.; Kuri, A.C.; Saha, T.; Shukla, M. SmartHeart: A conceptual framework for explainable machine learning in cardiovascular risk prediction. *Comput. Biol. Med.* **2025**, *198*, 111221. [CrossRef]

20. Raman, S.; Thakkar, D.; Calixte, J.; Kumar, R.; Sporn, K.; Marla, K.; Goel, D.; Gopali, R.; Chetla, N.; Pasha, S.; et al. Machine Learning for Coronary Heart Disease Prediction: Comparative Analysis of Framingham and Cleveland Subset of the UCI Dataset with SHAP-Based Interpretability. *Epidemiologia* **2026**, *7*, 75. [CrossRef]
21. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
22. Alexandrou, M.; Kumar, S.; Mahtani, A.U.; Strepkos, D.; Carvalho, P.E.; Mutlu, D.; Ser, O.S.; Rempakos, A.; Mastrodemos, O.C.; Rangan, B.V.; et al. Performance of Large Language Models on the Acute Coronary Syndrome Guidelines Using Retrieval-Augmented Generation. *JACC Cardiovasc. Interv.* **2025**, *18*, 2458–2467. [CrossRef]
23. Jin, Y.; Feng, C.; Zhao, J.; Lin, W.; Li, B. Accuracy of ChatGPT and DeepSeek in answering clinical questions from the 2025 Society for Cardiovascular Angiography & Interventions/Heart Rhythm Society left atrial appendage occlusion guidelines. *J. Int. Med. Res.* **2026**, *54*, 03000605261438348. [CrossRef]
24. Pan, Q.; Huang, L.; Liu, N.; Lin, F.; Ye, S. Evaluation of the accuracy of large language models in answering bone cancer-related questions. *Front. Public Health* **2025**, *13*, 1719047. [CrossRef]
25. Wang, D.; Liang, J.; Ye, J.; Li, J.; Li, J.; Zhang, Q.; Hu, Q.; Pan, C.; Wang, D.; Liu, Z.; et al. Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. *J. Med. Internet Res.* **2024**, *26*, e58041. [CrossRef]
26. Wang, C.; Chen, Y. Evaluating large language models for evidence-based clinical question answering. *Patterns* **2026**, *7*, 101519. [CrossRef]
27. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv* **2023**, arXiv:2305.09617. [CrossRef]
28. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5848–5864. [CrossRef]
29. Lee, J.; Cha, H.; Hwangbo, Y.; Cheon, W. Enhancing Large Language Model Reliability: Minimizing Hallucinations with Dual Retrieval-Augmented Generation Based on the Latest Diabetes Guidelines. *J. Pers. Med.* **2024**, *14*, 1131. [CrossRef]
30. Toraman, C.; Yilmaz, E.H.; Şahïnuç, F.; Ozcelik, O. Impact of Tokenization on Language Models: An Analysis for Turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–21. [CrossRef]
31. Gulati, M.; Levy, P.D.; Mukherjee, D.; Amsterdam, E.; Bhatt, D.L.; Birtcher, K.K.; Blankstein, R.; Boyd, J.; Bullock-Palmer, R.P.; Conejo, T.; et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **2021**, *144*, e368–e454. [CrossRef]
32. Heart Disease Dataset. Available online: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset (accessed on 6 October 2026).
33. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *Br. Med. J.* **2024**, *385*, e078378. [CrossRef]
34. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013. [CrossRef]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
37. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn* **1995**, *20*, 273–297. [CrossRef]
38. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018. [CrossRef]
39. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
42. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002*; ACM: New York, NY, USA, 2002; pp. 694–699. [CrossRef]
43. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.

44. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
45. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15:1–15:21. [CrossRef]
46. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474. Available online: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (accessed on 24 July 2026).
47. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Multilingual E5 Text Embeddings: A Technical Report. *arXiv* **2024**, arXiv:2402.05672. [CrossRef]
48. Chroma. 24 July 2026. Available online: https://github.com/chroma-core/chroma (accessed on 24 July 2026).
49. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *4*, 1–174. [CrossRef]
50. Cormack, G.V.; Clarke, C.L.A.; Buettcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, 19–23 July 2009*; ACM: New York, NY, USA, 2009; pp. 758–759. [CrossRef]
51. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. [CrossRef]
52. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
53. Ytu-Ce-Cosmos/Turkish-Gemma-9b-v0.1 · Hugging Face. Available online: https://huggingface.co/ytu-ce-cosmos/Turkish-Gemma-9b-v0.1 (accessed on 24 July 2026).
54. Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; et al. Yi: Open Foundation Models by 01.AI. *arXiv* **2024**, arXiv:2403.04652. [CrossRef]
55. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]
56. Google/Gemma-4-26B-A4B-it · Hugging Face. Available online: https://huggingface.co/google/gemma-4-26B-A4B-it (accessed on 6 October 2026).
57. Mistralai/Mistral-Small-24B-Instruct-2501 · Hugging Face. Available online: https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501 (accessed on 6 October 2026).
58. ytu-ce-Cosmos/Turkish-Gemma-9b-T1 · Hugging Face. Available online: https://huggingface.co/ytu-ce-cosmos/Turkish-Gemma-9b-T1 (accessed on 6 October 2026).
59. ggml-org/llama.cpp. (24 July 2026). C++. ggml. Available online: https://github.com/ggml-org/llama.cpp (accessed on 24 July 2026).
60. Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C.H.; Gonzalez, J.; Zhang, H.; Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles, in SOSP '23, New York, NY, USA, 23–26 October 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 611–626. [CrossRef]
61. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2318–2335. [CrossRef]
62. BAAI/bge-Reranker-v2-m3 · Hugging Face. Available online: https://huggingface.co/BAAI/bge-reranker-v2-m3 (accessed on 6 October 2026).
63. Dong, Y.; Ruan, C.F.; Cai, Y.; Lai, R.; Xu, Z.; Zhao, Y.; Chen, T. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models, May 12. *arXiv* **2025**, arXiv:2411.15100. [CrossRef]
64. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
65. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Gonzalez, J.; Stoica, I.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023. Available online: https://openreview.net/forum?id=uccHPGDlao (accessed on 6 October 2026).
66. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
67. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
68. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
69. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

70. Fisher, R.A. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J. R. Stat. Soc.* **1922**, *85*, 87–94. [CrossRef]
71. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef]
72. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
73. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
74. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
75. Catapano, A.L.; Graham, I.; De Backer, G.; Wiklund, O.; Chapman, M.J.; Drexel, H.; et al. 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias. *Eur. Heart J.* **2016**, *37*, 2999–3058. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.