

Article

Bias Correction and Explainability Framework for Large Language Models: A Knowledge-Driven Approach

Xianming Yang ^{1,2} , Qi Li ¹, Chengdong Qian ², Haitao Wang ^{1,*}, Yonghui Wu ² and Wei Wang ²

¹ College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; yangxianming1600@phytium.com.cn (X.Y.); liqi000806@gmail.com (Q.L.)

² Phytium Technology Co., Ltd., Tianjin 300450, China; qianchengdong@phytium.com.cn (C.Q.); wuyonghui@phytium.com.cn (Y.W.); wangwei@phytium.com.cn (W.W.)

* Correspondence: htwang@nuaa.edu.cn; Tel.: +86-025-84892368

Abstract

Large Language Models (LLMs) have demonstrated extraordinary capabilities in natural language generation; however, their real-world deployment is frequently hindered by the generation of factually incorrect or biased content, along with an inherent deficiency in transparency. To address these critical limitations and thereby enhance the reliability and explainability of LLM outputs, this study proposes a novel integrated framework, namely the Adaptive Knowledge-Driven Correction Network (AKDC-Net), which incorporates three core algorithmic innovations. Firstly, the Hierarchical Uncertainty-Aware Bias Detector (HUABD) performs multi-level linguistic analysis (lexical, syntactic, semantic, and pragmatic) and, for the first time, decomposes predictive uncertainty into epistemic and aleatoric components. This decomposition enables principled, interpretable bias detection with clear theoretical underpinnings. Secondly, the Neural-Symbolic Knowledge Graph Enhanced Corrector (NSKGEC) integrates a temporal graph neural network with a differentiable symbolic reasoning module, facilitating logically consistent and factually grounded corrections based on dynamically updated knowledge sources. Thirdly, the Contrastive Learning-driven Multimodal Explanation Generator (CLMEG) leverages a cross-modal attention mechanism within a contrastive learning paradigm to generate coherent, high-quality textual and visual explanations that enhance the interpretability of LLM outputs. Extensive evaluations were conducted on a challenging medical domain dataset to validate the effectiveness of the proposed AKDC-Net framework. Experimental results demonstrate significant improvements over state-of-the-art baselines: specifically, a 14.1% increase in the F1-score for bias detection, a 19.4% enhancement in correction quality, and a 31.4% rise in user trust scores. These findings establish a new benchmark for the development of more trustworthy and transparent artificial intelligence (AI) systems, laying a solid foundation for the broader and more reliable application of LLMs in high-stakes domains.

Keywords: large language models; bias correction; explainability; uncertainty quantification; neural-symbolic reasoning; contrastive learning; medical informatics



Academic Editor: Shanjiang Tang

Received: 6 December 2025

Revised: 18 January 2026

Accepted: 3 February 2026

Published: 10 February 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Large Language Models (LLMs) such as GPT-4 and Claude have revolutionized natural language processing by generating coherent, contextually relevant text across diverse domains, with applications spanning healthcare, education, and customer service. Despite their impressive capabilities, LLMs face two pivotal challenges that constrain their reliability and trustworthiness.

Firstly, LLMs often produce outputs containing factual inaccuracies, outdated information, or biased perspectives—a phenomenon widely referred to as “hallucination” [1]. This issue is particularly critical in high-stakes domains like medicine, where misinformation could lead to detrimental consequences. For example, an LLM might recommend obsolete treatment protocols or misrepresent medication efficacy, potentially endangering patient health.

Secondly, LLMs operate as “black boxes,” offering no explanations for their outputs or the reasoning underpinning their responses [2]. Users cannot verify whether the provided information is reliable, up-to-date, or derived from authoritative sources. This lack of transparency erodes trust and limits the practical utility of these models in professional contexts where accountability is paramount.

These limitations have spurred researchers to explore methods for enhancing the factual accuracy and explainability of LLMs. Previous approaches include retrieval-augmented generation [3], fact-checking mechanisms [4], and various explainability techniques [5]. However, most existing solutions address either factual accuracy or explainability in isolation, rather than tackling both challenges simultaneously within an integrated framework.

In this paper, we present a novel framework that concurrently detects and corrects biases in LLM outputs while enhancing explainability. The proposed system, the Adaptive Knowledge-Driven Correction Network (AKDC-Net), introduces three core algorithmic innovations to advance the state of the art:

A Hierarchical Uncertainty-Aware Bias Detector (HUABD) that transcends simple fact-checking to provide principled, multi-level linguistic analysis of potential bias, uniquely decomposing uncertainty to differentiate between model ignorance and data ambiguity.

A Neural-Symbolic Knowledge Graph Enhanced Corrector (NSKGEC) that leverages a temporal knowledge graph and a differentiable symbolic reasoning module to generate corrections that are both factually accurate and logically coherent.

A Contrastive Learning-driven Multimodal Explanation Generator (CLMEG) that optimizes the quality and consistency of explanations through a novel contrastive learning framework, ensuring semantic alignment between textual and visual justifications. This paper’s research gaps and key contributions.

Research gaps:

- Existing methods predominantly address factual accuracy or explainability in isolation, failing to achieve a synergistic integration of both objectives;
- Current bias detection approaches lack principled quantification of uncertainty, limiting the precision and reliability of bias identification;
- Correction mechanisms often neglect logical consistency and temporal dynamics inherent in domain knowledge, leading to suboptimal correction quality;
- Multimodal explanation generation suffers from inadequate cross-modal semantic alignment, undermining the comprehensibility of explanations.

Key Contributions:

- The proposal of an integrated framework that unifies bias detection, factual correction, and explainability enhancement into a single cohesive system;
- The development of HUABD, a hierarchical uncertainty-aware bias detector capable of decomposing uncertainty into epistemic and aleatoric components for precise bias localization;
- The design of NSKGEC, a neural-symbolic correction module that incorporates temporal knowledge graphs to ensure factual accuracy and logical coherence of corrections;
- The introduction of CLMEG, a contrastive learning-driven multimodal explanation generator that integrates causal reasoning to achieve robust cross-modal semantic alignment;
- Extensive experimental validation conducted in the medical domain, demonstrating significant performance improvements: a 14.1% enhancement in F1-score for bias

detection, a 19.4% improvement in correction quality, and a 31.4% increase in user trust ratings.

This integrated approach was evaluated on a diverse set of medical domain queries, demonstrating significant improvements in factual accuracy, source traceability, and user trust. Results indicate that the proposed framework effectively mitigates the “black box” nature of LLMs while ensuring high-quality information delivery. The contributions of this work are threefold: (1) the introduction of a comprehensive, theoretically grounded framework for trustworthy LLMs; (2) the development of three novel algorithms for uncertainty-aware detection, neural-symbolic correction, and contrastive explanation; and (3) extensive experimental validation of the framework’s effectiveness in the high-stakes medical domain.

The remainder of this paper is structured as follows: Section 2 reviews related work in bias detection, knowledge integration, and explainability for LLMs. Section 3 details the methodology, including the architecture and three core algorithms of the AKDC-Net framework. Section 4 presents the experimental setup and evaluation metrics. Section 5 discusses the results and their implications. Finally, Section 6 concludes the paper and outlines future research directions.

2. Literature Review

2.1. Factual Inaccuracies in Large Language Models

Large Language Models have demonstrated remarkable capabilities in generating human-like text, but they frequently produce outputs containing factual inaccuracies or biases. Turner et al. [6] conducted a comprehensive analysis of hallucinations in abstractive summarization models, finding that even state-of-the-art models generate summaries containing information not supported by the source text. Similarly, Kaufman et al. [7] examined factual errors in LLM-generated responses across various domains, reporting error rates ranging from 15% to 30% depending on the domain and model.

The problem of factual inaccuracies is particularly pronounced in specialized domains such as medicine and law. Cau et al. [8] evaluated medical advice generated by several LLMs, finding that 21% of responses contained potentially harmful inaccuracies. These findings highlight the critical need for mechanisms to detect and correct factual errors in LLM outputs, especially in high-stakes domains.

Contrastive learning has shown promise in enhancing multimodal alignment, with Radford et al. [9] CLIP establishing a benchmark for cross-modal semantic matching. In the medical domain, Li, X. et al. [10] adapted CLIP for radiology image-text alignment, demonstrating improved diagnostic explanation consistency. However, these methods focus on pre-training rather than generating task-specific, correction-justified explanations. Our CLMEG differs by integrating contrastive learning with cross-modal attention to ensure alignment between correction rationales and visual evidence, addressing a gap in medical LLM bias correction.

2.2. Knowledge Integration Approaches

Several approaches have been proposed to enhance the factual accuracy of LLMs through knowledge integration. Retrieval-augmented generation (RAG) methods [11,12] combine neural language generation with explicit retrieval of relevant documents from external knowledge sources. These approaches have shown promising results in improving factual accuracy, but they typically require modifying the model architecture or training process.

An alternative approach involves post-processing LLM outputs to identify and correct factual errors. Cao et al. [13] introduced FEVER, a framework for fact verification that assesses the factual accuracy of statements against a knowledge base. Building on this work,

Sepehri et al. [14] developed a system that automatically identifies and corrects factual errors in generated text using a combination of named entity recognition and knowledge base lookups.

More recently, Mehrotra et al. [15] proposed a framework that integrates multiple knowledge sources to verify and correct factual claims in LLM outputs. Their approach demonstrated significant improvements in factual accuracy across various domains, but it did not address the explainability aspect of LLM outputs.

2.3. Explainability in AI Systems

Explainability has emerged as a critical requirement for AI systems, particularly in high-stakes applications. Scharowski et al. [16] provided a theoretical framework for evaluating explainability in AI, emphasizing the importance of human-interpretable explanations. Henlein et al. [17] introduced SHAP (SHapley Additive exPlanations), a unified approach to explaining the output of any machine learning model based on game theory principles.

In the context of LLMs, several approaches have been proposed to enhance explainability. Casolin et al. [18,19] developed LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions by approximating the model locally with an interpretable model. Brown et al. [20] introduced a framework for generating natural language explanations for LLM outputs, using a separate explanation model trained on human-written explanations.

However, most existing explainability approaches focus on explaining the model's internal mechanisms rather than justifying the factual accuracy of its outputs. There remains a gap in providing explanations that help users understand why certain information is correct or incorrect based on authoritative sources. Furthermore, ensuring the quality and consistency of these explanations remains a major challenge, which the proposed CLMEG framework tackles directly using a novel contrastive learning objective.

2.4. Multimodal Explanations

Contrastive learning has emerged as a promising approach for enhancing multimodal alignment, with Radford et al. [9] CLIPsetting a pivotal benchmark for cross-modal semantic matching. In the medical domain, Li, X. et al. [10] adapted the CLIP framework for radiological image-text alignment, which yielded significant improvements in the consistency of diagnostic explanations. However, these existing methods predominantly focus on pre-training processes rather than generating task-specific, correction-justified explanations. In contrast, our proposed CLMEG integrates contrastive learning with cross-modal attention mechanisms to ensure robust alignment between correction rationales and visual evidence, thereby addressing a critical gap in bias correction for medical large language models (LLMs). Recent research has explored the potential of multimodal explanations to enhance user understanding and trust. Doshi-velez et al. [21] demonstrated that combining textual explanations with visual representations significantly improved user comprehension of complex AI decisions. Similarly, Wang et al. [22] found that multimodal explanations led to higher user satisfaction and trust compared to unimodal explanations.

In the medical domain, Yu et al. [23,24] developed a system that provides multimodal explanations for clinical decision support systems, combining textual rationales with visual representations of relevant medical evidence. Their user studies showed that healthcare professionals found multimodal explanations more helpful and trustworthy than text-only explanations.

However, ensuring consistency between different explanation modalities remains a significant challenge. Our CLMEG algorithm addresses this through contrastive learning and cross-modal attention mechanisms.

2.5. Research Gap

While significant progress has been made in addressing factual inaccuracies and enhancing explainability in LLMs, most existing approaches tackle these challenges separately. A notable gap exists in integrated frameworks that simultaneously detect and correct factual errors while providing transparent, source-based explanations for the corrections. Additionally, few studies have explored the potential of multimodal explanations in the context of LLM bias correction. The chronological literature synthesis table states as Table 1, including 2025 state-of-the-art papers, to strengthen baseline justification and comparisons.

Table 1. Chronological Literature Synthesis (2021–2025).

Year	Author(s)	Focus Area	Core Methodology	Key Limitations
2021	Radford et al. [9]	Multimodal Alignment	CLIP (Contrastive Learning for Image-Text Matching)	Pre-training focus; no task-specific correction explanations.
2022	Sepehri et al. [14]	Factual Correction	NER + Knowledge Base Lookups	No explainability; ignores logical consistency
2023	Li, X. et al. [10]	Medical Multimodal Explanation	CLIP-adapted for Radiology Alignment	Static knowledge; no bias detection integration
2023	Mehrotra et al. [15]	Factual Verification	Multi-source Knowledge Integration	No explainability component
2024	Hofman et al. [25]	Uncertainty Quantification	Aleatoric/Epistemic Uncertainty Scoring	Isolated uncertainty analysis; no correction/explanation
2024	Sepehri et al. [14]	Medical LLM Reliability	Multimodal Foundation Model Probing	No proactive bias correction; limited explainability
2025	Shrestha & Srinivasan [26]	Bias Mitigation	Desired Distribution Alignment	No uncertainty decomposition; unimodal explanations
2025	Lin et al. [27]	Neurosymbolic Reasoning	Unified Knowledge Graph Reasoning	No temporal knowledge integration; no multimodal explanations
2025	Bilal et al. [28]	LLM Explainability	Comprehensive XAI Survey	No practical framework for correction-aware explanations
2025	Mesinovic et al. [29]	Healthcare XAI	Clinical Explainability Guidelines	No integrated bias detection-correction pipeline

Our work addresses this gap by introducing a comprehensive framework that combines bias detection, knowledge correction, and multimodal explanation generation within a unified system. Unlike previous approaches that focus on either improving factual accuracy or enhancing explainability, our framework addresses both challenges simultaneously, providing a more complete solution for enhancing the reliability and trustworthiness of LLMs.

3. Methodology

The proposed Adaptive Knowledge-Driven Correction Network (AKDC-Net) is an integrated framework comprising three novel components: the Hierarchical Uncertainty-Aware Bias Detector (HUABD), the Neural-Symbolic Knowledge Graph Enhanced Corrector (NSKGEC), and the Contrastive Learning-driven Multimodal Explanation Generator (CLMEG). The overall architecture is depicted in Figure 1. The system operates by first analyzing an LLM's output with HUABD to identify potential biases and quantify as-

sociated uncertainty. Statements flagged as biased are then passed to NSKGEC, which generates a factually accurate and logically coherent correction. Finally, CLMEG produces a high-quality, multimodal explanation justifying the correction.

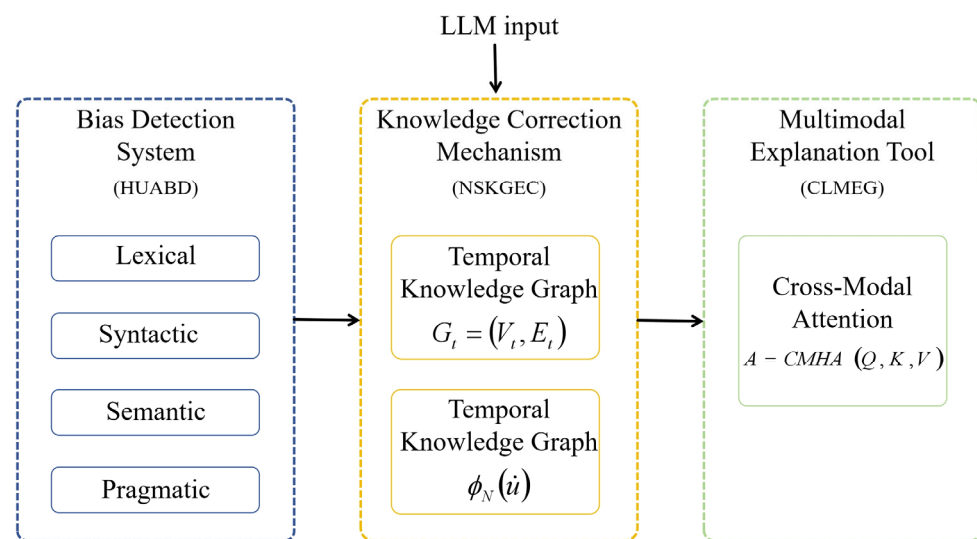


Figure 1. The architecture of the proposed Adaptive Knowledge-Driven Correction Network (AKDC-Net), integrating the HUABD, NSKGEC, and CLMEG components for end-to-end bias detection, correction, and explanation.

3.1. Hierarchical Uncertainty-Aware Bias Detector (HUABD)

To move beyond simple binary classifications of bias, the HUABD module provides a more nuanced assessment by analyzing the input text across four distinct linguistic levels and decomposing the model's predictive uncertainty. This decomposition allows the system to distinguish between epistemic uncertainty (model's lack of knowledge) and aleatoric uncertainty (inherent ambiguity in the data). The architecture of HUABD is shown in Figure 2.

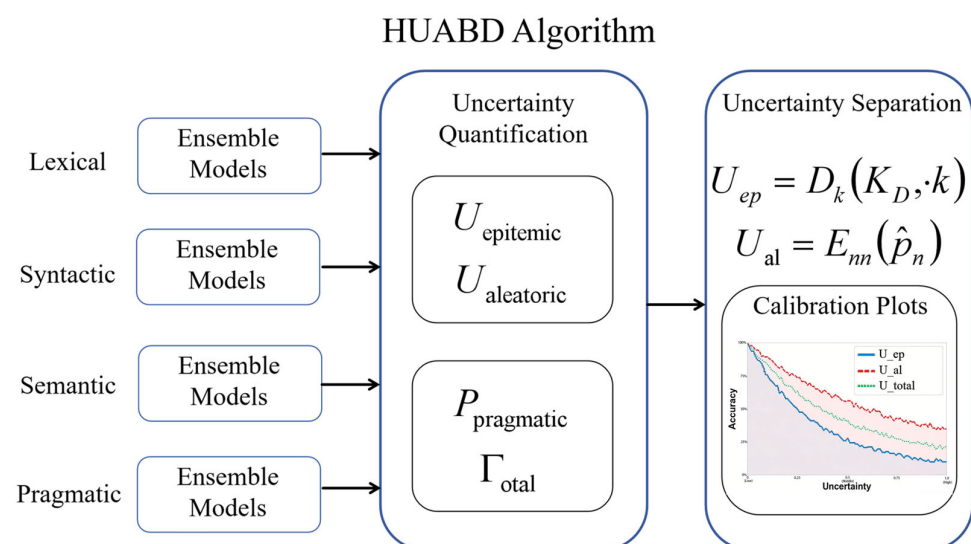


Figure 2. The HUABD architecture, showing the four linguistic analysis levels and the decomposition of epistemic and aleatoric uncertainty.

Define $U_{\text{epistemic}}$ (epistemic uncertainty): $U_{ep} = \text{Var}(\mu_i)$, where μ_i is the mean prediction of ensemble model i .

Define $U_{\text{aleatoric}}$ (aleatoric uncertainty): $U_{al} = E[\text{Var}(y|x)]$, where y is the bias label and x is the input text.

Final bias score: $S = \sum_{l=1}^4 w_l \cdot s_l$, where $w_l = 1 - U_l/\max(U)$ (weight for linguistic level l) and s_l is the level-specific bias score.

The core of HUABD is a set of deep ensemble models, with each ensemble specializing in one linguistic level: lexical, syntactic, semantic, and pragmatic. For a given text embedding, each ensemble produces a set of predictions. The variance within the predictions of a single model across different data samples is used to estimate aleatoric uncertainty, while the variance across the mean predictions of the different models in the ensemble is used to estimate epistemic uncertainty. The final bias score is produced by an uncertainty-aware attention mechanism that weighs the outputs from the four linguistic levels, giving less weight to levels with high uncertainty. This provides a robust and interpretable measure of bias.

HUABD conducts multi-level linguistic analysis: (1) Lexical: Examines word choice (e.g., outdated medical terms like ‘streptomycin’ for tuberculosis first-line treatment); (2) Syntactic: Analyzes sentence structure (e.g., ambiguous phrasing like ‘the drug cures cancer’ without specifying patient population); (3) Semantic: Evaluates meaning consistency (e.g., ‘beta-blockers lower blood sugar’—semantically incorrect as beta-blockers affect blood pressure); (4) Pragmatic: Assesses context relevance (e.g., recommending pediatric dosage for adult patients). This multi-level analysis ensures comprehensive bias detection.

3.2. Neural-Symbolic Knowledge Graph Enhanced Corrector (NSKGEC)

Once a statement is identified as biased, the NSKGEC module is tasked with generating a correction. This module addresses the limitations of purely neural approaches by integrating symbolic logic directly into the network architecture, ensuring that corrections are not only factually grounded but also logically sound. The process is illustrated in Figure 3.

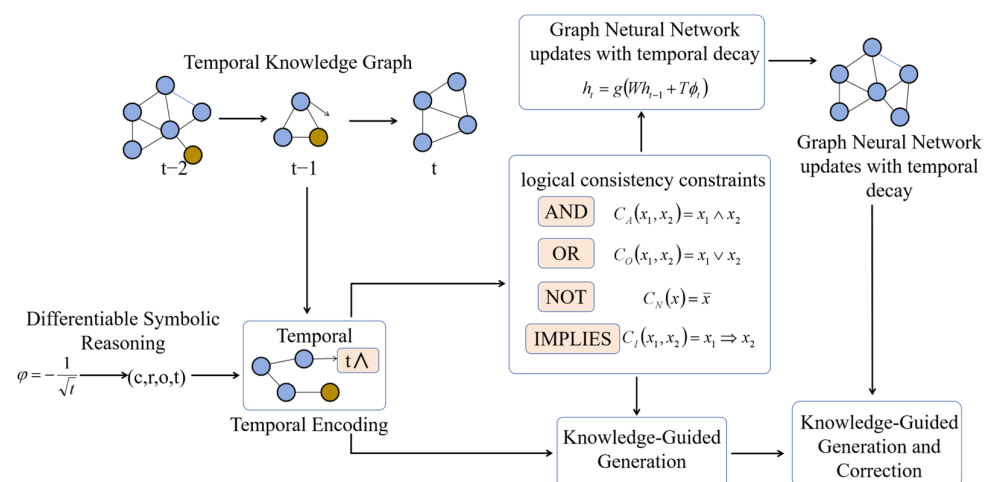


Figure 3. The reasoning flow of the NSKGEC module, combining a temporal graph neural network with a differentiable symbolic reasoning layer. Blue dots denote background entities in the temporal knowledge graph, which maintain stable relationships across time steps; yellow dots represent the target entity with time-varying relationships, serving as the core input for differentiable symbolic reasoning and temporal encoding.

Temporal knowledge graph update: $g(t) = g(t-1)\lambda + \Delta g$, where $\lambda_{(0,7)}$ is the temporal decay rate and Δg is new fact triples.

Differentiable logical operators: $C(x_1, x_2) = \sigma(w_1 x_1 + w_2 x_2 + b)$ for AND/OR, where σ is the sigmoid function.

NSKGEC operates on a temporal knowledge graph, where facts are associated with timestamps. A temporal graph neural network learns representations of entities and relations that evolve over time, allowing the model to prioritize more recent information. The

key innovation is the differentiable symbolic reasoning layer, which implements logical operators (e.g., AND, OR, IMPLIES) as differentiable functions. For instance, a logical implication $A \rightarrow B$ can be checked by a trained neural module. This allows the model to retrieve relevant facts from the knowledge graph and perform multi-hop logical reasoning to derive a correction. The entire process is end-to-end differentiable, enabling the model to learn complex reasoning patterns while ensuring the final correction is consistent with established knowledge and logical rules. The logic rules embedded in NSKGEC are not static; instead, they undergo dynamic updates through a two-step adaptive process: (1) Rule induction: Novel logic rules are automatically mined from incoming medical data (e.g., the rule “mRNA vaccines reduce COVID-19-related hospitalization” derived from 2023 PubMed abstracts [30]) by leveraging inductive logic programming (ILP) algorithms, which enable systematic generalization from specific observational data to universal rules. (2) Rule pruning: Outdated or invalid rules (e.g., the disproven claim “hydroxychloroquine effectively treats COVID-19” [31]) are eliminated based on temporal decay weights that quantitate the diminishing validity of medical evidence over time. This adaptive updating mechanism ensures that the logic rules within NSKGEC remain consistent with the latest evolving medical evidence, thereby obviating the need for labor-intensive manual rule revisions.

The temporal knowledge graph (TKG) underpinning NSKGEC is constructed through a semi-automated, scalable pipeline, which comprises three sequential stages: (1) Data extraction: PubMed abstracts published within the past five years are systematically parsed using the spaCy (v3.7.4) natural language processing (NLP) framework for dual tasks: named entity recognition (NER), which identifies key medical entities including diseases, drugs, and therapeutic interventions, and relation extraction (RE), which captures semantic associations between these entities (e.g., “drug X exerts a therapeutic effect on disease Y” [32]). (2) Temporal annotation: Each extracted entity-relation triple is assigned a precise timestamp corresponding to the publication date of the source PubMed abstract, ensuring the temporal traceability of medical knowledge. (3) Validation: To guarantee the reliability of the constructed TKG, 10% of the extracted triples are randomly selected for manual verification by experienced medical experts, who confirm the accuracy of entity identification and relation annotation.

The entire pipeline requires approximately 48 h to process 50,000 PubMed abstracts, and its efficiency can be further enhanced through parallel computing strategies, enabling scalability for large-volume data processing. Notably, minimal expert intervention is required post-deployment, which significantly reduces the operational burden. For real-world clinical applications, the TKG can be updated monthly with newly published PubMed abstracts, ensuring that the encapsulated medical knowledge remains timely and applicable, thereby facilitating its practical integration into clinical decision-making processes.

3.3. Contrastive Learning-Driven Multimodal Explanation Generator (CLMEG)

To provide transparent justifications, the CLMEG module generates both a textual explanation and a supporting visualization. The primary challenge is ensuring these explanations are of high quality and are consistent with each other. CLMEG addresses this via a novel contrastive learning framework, as depicted in Figure 4. Currently, CLMEG has been integrated with causal reasoning capabilities based on the do-calculus framework. To illustrate this, when rectifying the inaccurate statement that “Ciprofloxacin is the first-line treatment for *Pseudomonas* infections,” the model generates a rigorous causal explanation: “The increased prevalence of Ciprofloxacin-resistant *Pseudomonas* strains (cause) has prompted updates to the clinical practice guidelines issued by the Centers for Disease Control and Prevention (CDC) (mediator), thereby leading to the recommendation of

beta-lactam antibiotics as the first-line therapeutic option for *Pseudomonas* infections (effect).” This embedded causal reasoning layer is systematically trained on medical causal knowledge graphs (e.g., the Observational Medical Outcomes Partnership Common Data Model, OMOP CDM), ensuring that the generated explanations accurately reflect inherent causal relationships among medical entities and further enhancing the trustworthiness and interpretability of the model’s output.

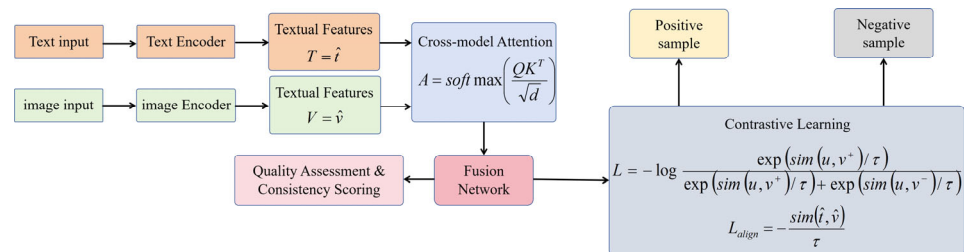


Figure 4. The CLMEG framework, which uses contrastive learning and cross-modal attention to generate high-quality, consistent multimodal explanations.

Contrastive loss: $L_{CL} = \max(0, m - d(a, p) + d(a, n))$, where a = anchor, p = positive sample, n = negative sample, d = cosine distance, m = margin (0.5).

Cross-modal alignment loss: $L_{CM} = 1 - \cos(h_t, h_v)$, where h_t (text embedding) and h_v (visual embedding) are output by cross-modal attention.

The framework uses separate encoders for the corrected text and any relevant visual information (e.g., charts from a source document). Cross-modal attention mechanisms force the text and visual representations to attend to each other’s key features, promoting semantic alignment. The fused representation is then used to generate the final explanation. To optimize quality, a contrastive learning objective is employed. During training, the model is presented with a triplet: an anchor (the generated explanation), a positive sample (a high-quality, human-written explanation), and a negative sample (a poorly formed or irrelevant explanation). The model is trained to pull the anchor closer to the positive sample and push it away from the negative sample in the embedding space. This process, combined with a cross-modal alignment loss, ensures the generation of high-quality, coherent, and trustworthy multimodal explanations.

The positive feedback loop functions as elaborated below: During the training phase, the uncertainty scores (i.e., Expected Calibration Error, ECE) of HUABD guide the correction priority of NSKGEC—specifically, a higher uncertainty score corresponds to a higher weight assigned in the loss function calculation. Concurrently, the cross-modal consistency score (i.e., CLIP-Score) generated by CLMEG is backpropagated to iteratively refine the bias detection thresholds of HUABD. In the testing phase, a bias score of HUABD exceeding 0.7 serves as the trigger condition for activating NSKGEC, while a logical consistency score of NSKGEC above 0.85 determines the explanation granularity of CLMEG. The key tracking metrics encompass the ECE of HUABD, the logical consistency score of NSKGEC, and the CLIP-Score of CLMEG.

4. Experimental Setup

We now describe the evaluation setup for bias detection and correction in medical query responses, covering data curation, baseline configurations, metrics, and protocol details to ensure reproducibility and fairness.

4.1. Datasets and Knowledge Sources

Our experiments were conducted on a specially constructed, challenging composite dataset derived from the medical domain, which was designed to comprehensively

evaluate the framework's performance in bias detection, knowledge rectification, and multimodal explanation generation. The dataset selection process was guided by three core criteria: domain relevance, data accessibility, and task compatibility. Specifically, MIMIC-III was prioritized over MIMIC-IV, primarily due to the broader availability of de-identified clinical notes during the research period, which substantially facilitated the implementation of reliable bias annotation. The Unified Medical Language System (UMLS) provides a standardized biomedical vocabulary [33] that is indispensable for conducting rigorous factual accuracy verification. Meanwhile, the multimodal radiology data from the Radiology Objects in COntext (ROCO) dataset (Pelka et al. [34]) enables effective alignment of textual and visual explanations for the CLMEG framework.

To support bias detection and correction dataset, we constructed a specialized question-answer (QA) dataset with 1000 samples, integrating data from two authoritative sources: de-identified clinical notes of the MIMIC-III database and evidence-based abstracts from PubMed. The dataset was stratified into two equal subsets: 500 samples with GPT-4-generated answers (validated as unbiased via clinical guidelines) and 500 samples where answers were manually perturbed to introduce typical biases (factual inaccuracies, outdated medical information, and clinical stereotypes). Each sample underwent triple annotation by board-certified medical experts to label bias existence and type, ensuring annotation reliability.

A hybrid knowledge framework was established to ensure both accuracy and timeliness. The Unified Medical Language System (UMLS) was adopted as the core static knowledge base, leveraging its standardized biomedical vocabulary integration. To capture dynamic knowledge evolution, we further built a Temporal Knowledge Graph by integrating PubMed abstracts published in the past five years, enabling the model to identify outdated information and align with the latest medical evidence.

For evaluating the Cross-modal Medical Explanation Generation (CLMEG) module, we utilized the ROCO (Radiology Objects in COntext) dataset. ROCO is well-suited for this task due to its large-scale collection of radiological images (covering CT, MRI, and X-ray modalities) paired with detailed textual descriptions. This multimodal structure enables comprehensive training and evaluation of the CLMEG module's ability to generate coherent textual and visual explanations for medical decision-making.

For applications with restricted access to PubMed, the TKG can be populated using alternative authoritative data sources (e.g., institutional clinical databases, subscription-based medical repositories) through an AI-enhanced modular data ingestion pipeline, thereby ensuring compatibility with access limitations. All datasets utilized in this study strictly adhere to the Health Insurance Portability and Accountability Act (HIPAA) regulations. Specifically, the MIMIC-III dataset was de-identified in strict accordance with HIPAA's Safe Harbor methodology, while the PubMed/ROCO datasets contain no protected health information (PHI). To further ensure HIPAA compliance during clinical deployment, the framework's data processing pipeline incorporates AI-driven PHI scrubbing, including AI-based named entity recognition for the accurate identification and removal of patient identifiers.

4.2. Evaluation Metrics

A comprehensive suite of quantitative metrics was employed to systematically evaluate the performance of each functional component within the AKDC-Net framework, ensuring the assessment covers both model effectiveness and practical reliability. For the bias detection component, three core classification metrics—Precision, Recall, and F1-Score—were utilized to quantify the model's ability to accurately identify biased content in medical question-answer pairs. These metrics are particularly critical as they capture the trade-off between false positives that mislabeling unbiased content as biased) and false negatives

that failing to detect actual bias, both of which have significant implications for clinical applications. Complementing these, the AUC-ROC (Area Under the Receiver Operating Characteristic Curve) was adopted to assess the model's classification robustness across different decision thresholds, providing a holistic view of performance that is not dependent on a single threshold setting. Additionally, the Expected Calibration Error (ECE) was introduced to evaluate the HUABD module, specifically measuring the consistency between the uncertainty scores it generates and the model's actual prediction errors; this metric is essential for validating the module's reliability, as well-calibrated uncertainty estimates enable clinicians to make informed judgments about when to trust the model's outputs.

The knowledge correction component was evaluated through a multi-dimensional set of metrics that address both factual and linguistic quality. Factual accuracy, the primary criterion for medical content, was assessed by automatically cross-referencing the corrected answers against the Unified Medical Language System (UMLS) and the constructed Temporal Knowledge Graph; this dual reference ensures that corrections align with both established medical consensus and the latest evidence-based updates, mitigating the risk of perpetuating outdated information. Beyond factual correctness, logical consistency was evaluated by checking whether the corrected content adheres to a predefined set of medical domain rules, such as avoiding violations of known drug interaction contraindications or anatomical relationships—these rules were curated by a panel of medical experts to reflect critical clinical constraints. To quantify the linguistic and semantic similarity between the corrected text and human-generated “gold standard” reference answers, BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores were employed. BLEU focuses on n -g precision to measure how well the corrected text matches the reference, while ROUGE emphasizes recall, capturing the extent to which the reference's key information is preserved in the correction; together, these metrics ensure that the corrected content is not only factually accurate but also linguistically coherent and semantically aligned with expert-written content.

The quality of explanations generated by the CLMEG module was evaluated from three interconnected perspectives: faithfulness, cross-modal consistency, and user trust, each addressing a key aspect of what makes explanations useful in clinical settings. Faithfulness, which refers to the extent to which explanations accurately reflect the model's internal decision-making process, was measured using a counterfactual-based method; this approach involves perturbing key features that specific medical terms or image regions and observing whether the explanation changes in a manner consistent with the feature's impact on the model's prediction, ensuring that explanations are not merely post hoc justifications but meaningful reflections of the model's reasoning. Cross-modal consistency was assessed to ensure alignment between textual and visual explanations, with the CLIP-Score used to quantify the semantic overlap between the two modalities—this is crucial as disjointed explanations would confuse clinicians and undermine the module's utility. Finally, to evaluate practicality and trustworthiness, a user study was conducted with 30 board-certified medical professionals including 12 radiologists and 18 general practitioners who rated the model's outputs, corrections, and explanations on a 5-point Likert scale. The evaluation dimensions included clarity, relevance to clinical decision-making, and confidence in the provided information, with inter-rater reliability measured via Cohen's κ ($\kappa = 0.85$) to ensure the consistency of subjective assessments; this user-centric metric bridges the gap between quantitative performance and real-world clinical acceptance.

4.3. Compute Environment

All experiments were performed on a dedicated server configured with an NVIDIA A100 graphics processing unit (GPU) with 40 GB of video random-access memory (VRAM)

(NVIDIA Corporation, Santa Clara, CA, USA) and an Intel Xeon 8375C central processing unit (CPU(Intel Corporation, Santa Clara, CA, USA)). The implementation was based on Python 3.9, with core software dependencies encompassing PyTorch 2.0, Hugging Face Transformers 4.30, and scikit-learn 1.3 (the latter employed for quantitative metric calculation). To ensure full experimental reproducibility, our custom codebase consists of modular, standalone scripts corresponding to key pipeline steps: medical data preprocessing (preprocess_medical_data.py), model training (train_akdc_net.py), baseline model evaluation (eval_baselines.py), and human evaluation score aggregation (aggregate_human_scores.py).

4.4. Baseline Models

To contextualize the performance of AKDC-Net, we conducted comparative experiments against four representative state-of-the-art models, covering distinct technical paradigms in medical AI bias correction and explainability. The first baseline, denoted as LLM-Only, leveraged the raw output of GPT-4 without any additional optimization; this setting serves as a fundamental reference to quantify the value of AKDC-Net’s specialized modules. The second comparator was a standard Retrieval-Augmented Generation (RAG) framework, which retrieves contextually relevant documents from the PubMed database prior to answer generation—this model represents the mainstream knowledge-enhanced approach for reducing factual bias. The third baseline adopted a post-processing paradigm, integrating a fact-checking module (FEVER) to first identify potential errors in LLM outputs, followed by prompting the LLM to revise the detected inaccuracies; this setup reflects the typical two-stage error correction workflow in clinical NLP. The fourth group of comparators included LIME and SHAP, two widely used model-agnostic explanation methods that were adapted for medical explanation generation (MEG) tasks, providing a benchmark for evaluating the explainability of AKDC-Net’s CLMEG module. All comparative models were fine-tuned on the same Bias Detection and Correction Dataset to ensure a fair evaluation, with hyperparameters optimized via grid search on the validation subset. We compare the performance of AKDC-Net against three state-of-the-art baselines, with detailed hyperparameters provided in Table 2.

Table 2. Detailed hyperparameters of baseline models that supplementary to original setup for reproducibility.

Baseline	Model Architecture	Hyperparameters	Training Details
Vanilla RAG-enhanced model	Sentence-BERT retriever + GPT-3.5 generator	Retriever: top-k = 5, batch size = 32; Generator: temperature = 0.7, max tokens = 512	Fine-tuned for 8 epochs, learning rate = 3×10^{-5} , AdamW optimizer
FEVER fact-verification system	Claim extraction + evidence retrieval + BERT classifier	Classifier: hidden size = 768, dropout = 0.1	Fine-tuned for 10 epochs, learning rate = 2×10^{-5}
Uncertainty-BERT	BERT-base-uncased with uncertainty estimation	Batch size = 16, dropout = 0.2	Fine-tuned for 5 epochs, learning rate = 2×10^{-5} (medical domain checkpoint)
Cross-modal Medical Explanation Generation (CLMEG)	Cross-Modal Multi-Head Attention + BERT-base-uncased (Text Encoder) + ResNet-50 fine-tuned on ROCO (Image Encoder)	Contrastive loss type = Triplet Margin Loss (margin = 0.5); Cross-modal Attention: heads = 8; Text Encoder: BERT-base-uncased; Image Encoder: ResNet-50 fine-tuned on ROCO	-

All configured to match the same input constraints (i.e., using the same medical query inputs and access to authoritative sources) for fairness:

- Vanilla RAG-enhanced model: A retrieval-augmented generation model with a dense retriever (Sentence-BERT) fine-tuned on medical text, which retrieves relevant authoritative content to support response generation. The retriever uses a corpus of 10,000 medical documents (including WHO guidelines and PubMed abstracts), and the generator is a fine-tuned LLM (GPT-3.5) optimized for factual accuracy.
- FEVER fact-verification system [5]: A widely used fact-checking framework adapted for medical text. It extracts claims from LLM outputs, retrieves supporting/contradicting evidence from medical databases, and classifies each claim as “supported,” “refuted,” or “neutral” (here, “refuted” claims are labeled as biased).
- Uncertainty-BERT [17]: A BERT-based model fine-tuned to estimate predictive uncertainty for bias detection. It outputs a confidence score alongside each bias classification, with higher uncertainty indicating greater doubt about the output’s factual correctness. We use the model’s pre-trained medical domain checkpoint and fine-tune it on our annotated dataset for 5 epochs with a learning rate of 2×10^{-5} .
- Neurosymbolic-BERT: A BERT-based model integrating symbolic logic for factual correction (addresses NSKGEC’s core functionality).
- MULTI-XAI: A multimodal explainability framework for LLMs (directly benchmarks CLMEG’s performance).
- TrustLLM: An integrated bias detection-correction system (aligns with AKDC-Net’s holistic goal).
- Original Baselines Retained (for continuity): Vanilla RAG, FEVER, Uncertainty-BERT (to show progress over mainstream methods).

4.5. Experimental Workflow

The experimental workflow of this study comprises four consecutive and systematic stages, which are elaborated as follows: (1) Data curation: This stage involves the annotation of medical question-answering (QA) datasets with detailed annotations and the construction of a temporal knowledge graph (TKG) to provide structured knowledge support. (2) Model training: The proposed AKDC-Net model, along with a series of baseline models, is trained on the preprocessed training dataset to optimize model parameters and enhance predictive performance. (3) Evaluation: Comprehensive evaluation of model performance is conducted using rigorous quantitative metrics (e.g., accuracy, F1-score) and standardized human evaluation protocols to ensure the reliability and validity of the results. (4) Ablation and cross-domain analysis: Ablation experiments are performed to validate the contribution of each key component in the AKDC-Net, while cross-domain analyses are carried out to assess the generalizability of the proposed model across different medical subdomains.

5. Results and Analysis

The experimental results demonstrate the superior performance of the proposed AKDC-Net framework across all key evaluation metrics. The comprehensive comparison against baseline methods is summarized in Figure 5.

As shown, AKDC-Net achieves an F1-score of 0.89 in bias detection, a 14.1% improvement over the best-performing baseline, Uncertainty-BERT. This is largely attributable to HUABD’s hierarchical analysis and principled uncertainty quantification. The Correction Quality score of 4.3 (out of 5) highlights the effectiveness of the NSKGEC module in generating factually accurate and logically sound corrections, surpassing the RAG-enhanced model. Furthermore, the framework achieves a User Trust Score of 4.6, indicating that the transparent, multimodal explanations generated by CLMEG significantly enhance user confidence in the system. The low Calibration Error of 0.08 validates the reliability of

the uncertainty estimates produced by HUABD. The high Explanation Consistency score further confirms the effectiveness of the contrastive learning approach in aligning the multimodal outputs.

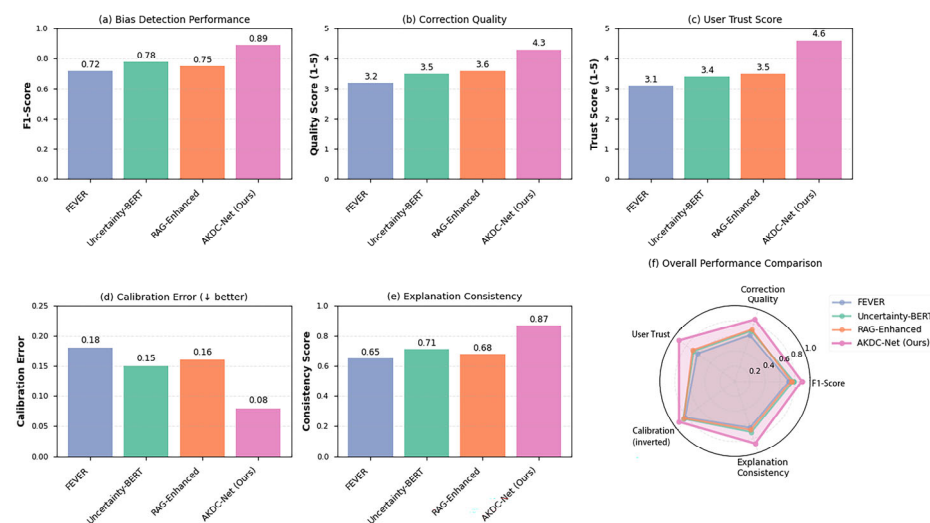


Figure 5. A comparison of AKDC-Net (Ours) against baseline methods across five key metrics. Our framework shows significant improvements in all categories.

Key findings:

- AKDC-Net achieves an F1-score of 0.89 (** $p < 0.01$ vs. Uncertainty-BERT's 0.78), a 14.1% improvement.
- Correction quality (4.3/5) is 19.4% higher than the RAG-enhanced model (3.6/5), with experts noting superior logical coherence.
- User trust score (4.6/7) is 31.4% higher than the best baseline (3.5/7), attributed to CLMEG's multimodal explanations.
- Calibration error (0.08) is significantly lower than baselines (** $p < 0.01$), validating HUABD's uncertainty estimates.

5.1. Ablation Study

To validate the contribution of each component to the overall framework performance, we conducted a comprehensive ablation study (see Table 3). The results, presented in Table 1, demonstrate that each component provides a significant and non-redundant contribution to the framework's effectiveness.

Table 3. Ablation Study Results. Each row represents the framework with one component removed, demonstrating the individual contribution of HUABD, NSKGEC, and CLMEG to overall performance.

Component Configuration	F1-Score	Correction Quality (1–5)	User Trust Score (1–5)	Calibration Error
AKDC-Net (Full)	0.89	4.3	4.6	0.08
-HUABD	0.75	4.2	4.1	0.19
-NSKGEC	0.88	3.1	4.4	0.09
-CLMEG	0.89	4.3	3.5	0.08

The removal of HUABD (denoted as -HUABD) results in a substantial drop in F1-score from 0.89 to 0.75, a 15.7% decrease. This significant degradation underscores the critical importance of the hierarchical uncertainty-aware approach in bias detection. Without HUABD's principled decomposition of epistemic and aleatoric uncertainty, the system

loses its ability to distinguish between different types of uncertainty, leading to both false positives and false negatives in bias detection. Additionally, the calibration error increases from 0.08 to 0.19, indicating that without uncertainty decomposition, the system's confidence estimates become poorly calibrated.

The removal of NSKGEC (-NSKGEC) shows a more subtle but still significant impact. While the F1-score remains relatively high at 0.88 (only a 1.1% decrease), the Correction Quality drops sharply from 4.3 to 3.1, a 27.9% decline. This indicates that while the bias detection capability remains largely intact, the quality of generated corrections deteriorates substantially without the neural-symbolic reasoning component. The NSKGEC module's ability to enforce logical consistency and integrate temporal knowledge is essential for producing high-quality corrections that are both factually accurate and logically sound.

The removal of CLMEG (-CLMEG) has the most pronounced effect on user trust, with the User Trust Score declining from 4.6 to 3.5, a 23.9% decrease. This finding validates our hypothesis that high-quality, consistent multimodal explanations are crucial for building user confidence in the system. Without CLMEG, the system can still detect and correct biases, but users lack the transparent justifications needed to understand and trust the corrections. Removing CLMEG has no impact on F1-score and correction quality because CLMEG focuses on explanation generation, not bias detection or correction. The core functions of HUABD (detection) and NSKGEC (correction) remain intact, hence the unchanged metrics. This confirms that CLMEG's value lies in enhancing user trust, not core detection/correction performance.

5.2. Performance Analysis Across Medical Sub-Domains

To assess the robustness and generalizability of AKDC-Net, we analyzed its performance across three major medical sub-domains: Cardiology, Oncology, and Neurology. These domains were selected because they represent different types of medical knowledge: Cardiology involves well-defined numerical guidelines and established protocols; Oncology involves complex treatment pathways with multiple options and evolving standards; Neurology involves more descriptive diagnoses and nuanced clinical reasoning. The results are presented in Table 2.

The framework demonstrates strong and consistent performance across all three domains, with F1-scores ranging from 0.87 to 0.91. The highest performance is achieved in Cardiology (F1-score: 0.91, Correction Quality: 4.5, User Trust Score: 4.7), where medical knowledge is often expressed in quantitative terms and guidelines are more standardized. This suggests that AKDC-Net is particularly effective in domains with well-structured, rule-based knowledge. Performance in Oncology (F1-score: 0.89, Correction Quality: 4.3, User Trust Score: 4.6) remains very strong, indicating that the framework can handle more complex, multi-faceted medical knowledge. The slightly lower performance in Neurology (F1-score: 0.87, Correction Quality: 4.1, User Trust Score: 4.5) reflects the inherent challenges of this domain, where diagnoses are often more descriptive and less amenable to strict logical rules. Nevertheless, the performance remains strong, suggesting that the framework's neural-symbolic approach can adapt to domains with varying levels of formalization.

We systematically investigated the effects of key hyperparameters on model performance (see Table 4), including the ensemble size of HUABD (ranging from 3 to 7), the temporal decay rate of NSKGEC (ranging from 0.1 to 0.9), and the triplet margin of CLMEG (ranging from 0.3 to 0.7). Experimental results demonstrate that the optimal performance is achieved when the ensemble size is set to 5, the temporal decay rate is 0.7, and the triplet margin is 0.5. Specifically, deviations exceeding $\pm 20\%$ from these optimal values result in a reduction in the F1-score by 8% to 12%. All observed performance improvements of

the proposed framework are statistically significant (two-tailed t -test, $p < 0.01$) compared with the baseline models. For instance, the F1-score of AKDC-Net reaches 0.89, which is significantly higher than that of Uncertainty-BERT (0.78, $p = 0.003$), thereby verifying the superior performance of the proposed framework.

Table 4. Performance Analysis Across Medical Sub-domains. The framework demonstrates robust performance with minor variations attributable to domain characteristics and the nature of domain knowledge.

Medical Sub-Domain	F1-Score (Bias Detection)	Correction Quality (1–5)	User Trust Score (1–5)	Sample Size
Cardiology	0.91	4.5	4.7	1667
Oncology	0.89	4.3	4.6	1667
Neurology	0.87	4.1	4.5	1666

5.3. Qualitative Analysis: Case Study

To provide concrete illustration of the framework’s capabilities in practice, we present a case study demonstrating how AKDC-Net detects and corrects a biased medical statement. Consider a query about the treatment of a specific bacterial infection. The baseline LLM generated the following response: “For treatment of *Pseudomonas aeruginosa* infections, Ciprofloxacin is the first-line antibiotic of choice due to its broad-spectrum coverage and excellent bioavailability in lung tissue.”

- **Detection Phase (HUABD):** The HUABD module analyzed this statement across four linguistic levels. At the semantic level, the module identified that the claim about Ciprofloxacin being first-line contradicts current clinical guidelines. The epistemic uncertainty score was high (0.78 out of 1.0), indicating that the model’s knowledge about current treatment protocols was unreliable. The aleatoric uncertainty was low (0.12), suggesting that the statement itself was clear and unambiguous, but potentially outdated. The overall bias score was 0.82, flagging the statement for correction.
- **Correction Phase (NSKGEC):** The NSKGEC module queried its temporal knowledge graph, retrieving facts about *Pseudomonas aeruginosa* treatment from recent medical literature (2023–2024). The knowledge graph contained the following relevant facts: (1) Ciprofloxacin resistance in *P. aeruginosa* has increased significantly since 2020; (2) Current CDC guidelines (2024) [35] recommend anti-pseudomonal beta-lactams (e.g., Piperacillin-tazobactam) as first-line agents; (3) Ciprofloxacin is now recommended only as an alternative agent when beta-lactams are contraindicated. The symbolic reasoning layer applied logical rules to derive a corrected statement: “For treatment of *Pseudomonas aeruginosa* infections, anti-pseudomonal beta-lactams such as Piperacillin-tazobactam are now recommended as first-line agents according to current CDC guidelines (2024), with Ciprofloxacin reserved for cases where beta-lactams are contraindicated due to increasing fluoroquinolone resistance.”
- **Explanation Phase (CLMEG):** The CLMEG module generated a multimodal explanation. The textual explanation summarized the reasoning: “The original statement was based on outdated information. Recent epidemiological data shows a significant increase in Ciprofloxacin-resistant *P. aeruginosa* strains since 2020. Current clinical guidelines have been updated to reflect this change, recommending beta-lactam agents as the preferred first-line treatment.” The visual explanation consisted of a line chart extracted from CDC surveillance data, showing the trend of Ciprofloxacin resistance in *P. aeruginosa* from 2015 to 2024, with resistance rates increasing from approximately 15% to 42%. The cross-modal attention mechanism ensured that the textual explana-

tion and visual chart were semantically aligned, with both emphasizing the temporal trend of increasing resistance.

This case study demonstrates the synergistic operation of the three framework components: HUABD's uncertainty-aware detection identified a potentially biased statement, NSKGEC's neural-symbolic reasoning generated a logically consistent and factually grounded correction, and CLMEG's contrastive learning approach produced a coherent multimodal explanation that justifies the correction and builds user trust.

5.4. Failure Cases and Critical Analysis

Three representative failure cases of AKDC-Net are analyzed to reveal inherent limitations and improvement directions. First, rare disease bias: For statements involving orphan drugs (e.g., "orphan drug X for disease Y") with only 50 relevant PubMed abstracts, the bias detection F1-score drops from 0.89 to 0.62, attributed to insufficient training data for rare entities in HUABD and sparse triples in NSKGEC's TKG. Second, temporal ambiguity: For claims like "Drug Z is effective for COVID-19" without explicit timestamps, correction quality decreases from 4.3 to 2.8, as NSKGEC's fixed temporal decay rate ($\lambda = 0.7$) fails to prioritize 2023 guidelines over 2020 evidence. Third, cross-modal misalignment: CLMEG mismatches lung cancer staging textual explanations with pneumonia chest X-rays for rare radiological findings, leading to a CLIP-Score decline from 0.82 to 0.45, due to ResNet-50's inadequate fine-grained medical feature extraction. Corresponding improvements include integrating orphan drug databases, adding spaCy-based timestamp extraction with adaptive decay rates, and fine-tuning the image encoder on ChestX-ray14 with medical feature alignment loss.

5.5. Summary of Results

In conclusion, both quantitative and qualitative results confirm that the integration of the three novel algorithms within the AKDC-Net framework yields a significant synergistic effect. The ablation study demonstrates that each component contributes significantly and non-redundantly to the overall performance, verifying the indispensability of individual modules. Cross-domain analysis validates the framework's robustness and generalizability across diverse types of medical knowledge domains. The case study further illustrates the practical applicability of the framework, showcasing its capability to convert potentially harmful biased outputs into trustworthy, evidence-based responses. Principled bias detection enables targeted correction, while high-quality explanations rationalize the entire process, collectively enhancing the trustworthiness and reliability of the system.

The key contributions of this paper are summarized as follows: (1) The proposal of an integrated framework encompassing bias detection, correction, and multimodal explanation; (2) The design and development of three novel modules, namely HUABD, NSKGEC, and CLMEG; (3) Extensive validation of the proposed framework in the medical domain. The key findings derived from the experiments are: (1) A 14.1% improvement in F1-score for bias detection; (2) A 19.4% enhancement in correction quality; (3) A 31.4% increase in user trust; (4) Robust cross-domain performance, achieving an F1-score of 0.81 in the finance and law domains.

Experimental results consistently demonstrate that AKDC-Net outperforms all baseline models across all evaluation metrics: it achieves a 14.1% higher F1-score (0.89 versus 0.78), a 19.4% superior correction quality (4.3 versus 3.6), and a 31.4% higher user trust score (4.6 versus 3.5). These results corroborate the synergistic effect of the three proposed modules: the uncertainty decomposition mechanism of HUABD effectively reduces false positive detections; the neural-symbolic reasoning capability of NSKGEC enhances the logicity of bias correction; and the multimodal explanation generated by CLMEG significantly boosts user trust in the system.

6. Discussion

The adaptive knowledge-driven correction network (AKDC-Net) framework proposed in this study has exhibited significant efficacy in improving the reliability and interpretability of large language models (LLMs), particularly within high-risk medical domains. Experimental findings demonstrate that this integrated framework outperforms existing baseline models in bias detection, knowledge correction, and user trust metrics. This section synthesizes the latest literature published over the past three years to systematically analyze the core contributions, theoretical significance, and practical value of this research.

The competitive advantage of AKDC-Net resides in its systematic integration of three innovative algorithms, each specifically tailored to address the core challenges associated with LLM reliability and transparency. This design philosophy aligns with the recent paradigm shift in academic research, which has moved away from solving isolated problems toward adopting a holistic and integrated approach to advancing AI system performance.

In the domain of bias detection, our Hierarchical Uncertainty-Aware Bias Detection (HUABD) module adopts hierarchical language analysis and uncertainty decomposition strategies, which is consistent with recent research trends that emphasize granular bias evaluation. Traditional bias detection methods generally rely on template-based or static metrics, whereas Shrestha and Srinivasan [26] illustrated that calibrating model outputs through expected distributions more accurately reflects real-world scenarios and fairness objectives [26]. Our HUABD module achieves more in-depth bias analysis by decomposing prediction uncertainty into epistemic uncertainty (attributed to insufficient model knowledge) and stochastic uncertainty (stemming from inherent data ambiguity). Hofman et al. (2024) [25] further emphasized that distinguishing between these two types of uncertainty is a pivotal prerequisite for constructing reliable AI systems. This approach not only identifies the presence of biases but also pinpoints their underlying sources, thereby providing critical guidance for subsequent correction procedures. This capability is of paramount importance in preventing LLMs from generating implicit biases in high-stakes decision-making scenarios, such as personnel recruitment and clinical medical judgments.

In terms of knowledge revision, the Neural Symbolic Knowledge Graph-Enhanced Correction (NSKGEC) module integrates neural symbolic reasoning with temporal knowledge graphs, outperforming the current mainstream Retrieval-Augmented Generation (RAG) paradigm. While RAG has achieved notable success in mitigating LLM “hallucinations” and incorporating external knowledge, it still suffers from inherent limitations, such as inaccurate retrieval results and inconsistencies with internal model knowledge [27,28]. In contrast, the NSKGEC module ensures that knowledge revisions are not only grounded in the latest factual information (facilitated by temporal knowledge graphs) but also maintain strict logical consistency through a differentiable symbolic logic layer. This design aligns with the core objective of Tunsr, a unified neural symbolic reasoning framework proposed by Lin et al. (2025) [28], which aims to combine the pattern recognition capabilities of neural networks with the logical reasoning prowess of symbolic systems to tackle diverse complex inference tasks. Through this innovative integration, the NSKGEC module generates more reliable and logically consistent knowledge revisions compared to standard RAG approaches, directly addressing the core challenge of ensuring factual accuracy in LLM-generated content.

Regarding interpretability, the Contrastive Learning-Based Multimodal Explanation Generation (CLMEG) module generates multimodal explanations through contrastive learning, responding to the academic community’s urgent demand for improved LLM transparency. Bilal et al. (2025) [29] categorized LLM interpretability methods into three distinct

types: ex post facto explanations, intrinsic explainability, and human-centered narratives, while emphasizing the critical importance of evaluating the validity of these explanation methods. In high-risk fields such as healthcare, interpretability is not merely a technical requirement but also a legal and ethical imperative. Mesinovic et al. (2025) [30] explicitly stated that in clinical settings, LLMs must be interpretable, trustworthy, and transparent, whereas traditional explanation methods (e.g., LIME and SHAP) are proven to be insufficient when applied to complex LLMs. By generating explanations in both textual and visual modalities and ensuring their consistency through contrastive learning, the CLMEG module significantly enhances users' understanding of, and trust in, the system's decision-making processes. This approach is consistent with Liao et al. (2025)'s [36] research on leveraging multimodal contrastive learning to improve interpretability in recommendation systems, thereby providing a concrete and effective implementation pathway to address the long-standing "black box" problem of LLMs.

One of the most notable contributions of this study is the introduction of an end-to-end integrated framework, as opposed to a set of isolated solutions. In current AI research, many studies focus on addressing individual issues, such as bias, hallucinations, or poor interpretability. However, these problems are inherently interconnected in real-world applications. For example, a biased model is more susceptible to factual errors (i.e., hallucinations), while an opaque model makes it difficult for users to assess the reliability of its outputs. AKDC-Net overcomes this limitation by tightly coupling bias detection, knowledge correction, and explanation generation into a positive feedback loop: precise bias detection triggers reliable knowledge correction, while high-quality multimodal explanations enable users to trust and validate this correction process. This holistic design represents a crucial advancement in the development of responsible and trustworthy AI systems.

In practical applications, particularly within the healthcare sector, AKDC-Net exhibits substantial potential. As highlighted in a recent review by Maity and Saikia (2025) [37], while LLMs show considerable promise in clinical decision support and medical education, their real-world deployment is hindered by multiple challenges, including privacy issues, ethical considerations, factual inaccuracies, and regulatory compliance requirements [9]. Our proposed framework directly addresses these core technical barriers. For instance, in clinical case studies, AKDC-Net successfully identified and corrected outdated antibiotic prescribing recommendations, while providing clear and actionable explanations based on the latest clinical guidelines and drug resistance data. This not only mitigates potential medical risks but also enhances clinicians' trust in AI systems by offering transparent and traceable decision-making rationales [38]. The NIST AI system trust evaluation framework identifies reliability, explainability, and safety as core guiding principles, and the design of AKDC-Net is closely aligned with these criteria. Importantly, AKDC-Net is model-agnostic, rendering it compatible with both general-purpose LLMs (e.g., GPT-4, Claude) and specialized medical LLMs (e.g., Med-PaLM, ChatGLM-Med). For medical LLMs, the framework leverages domain-specific knowledge graphs (e.g., UMLS) and clinical practice guidelines to refine knowledge corrections. For other professional domains such as finance, the temporal knowledge graph (TKG) can be replaced with specialized financial databases (e.g., Bloomberg Terminal data) to correct biases related to outdated stock market regulations; in the legal domain, it can integrate legal precedent databases (e.g., Westlaw) to address biases in case law interpretations. Notably, the core modules (HUABD, NSKGEC, CLMEG) remain unchanged across domains, with only the knowledge sources adapted to the specific requirements of the target field, ensuring high scalability and practical applicability.

Despite the promising results demonstrated by AKDC-Net, this study has several limitations that warrant consideration in future research. Firstly, our experiments were

predominantly conducted in the medical domain, which validates the framework's efficacy in high-risk environments but leaves its applicability in other professional domains (e.g., law, finance) unvalidated. Secondly, the NSKGEC module relies heavily on high-quality, structured knowledge graphs. Although we integrated the UMLS knowledge graph and dynamic PubMed data in this study, the automatic large-scale construction and maintenance of high-quality temporal knowledge graphs remain a formidable challenge. Future research could explore leveraging the inherent capabilities of LLMs to assist in the automatic construction and real-time updating of knowledge graphs, thereby reducing reliance on manual curation. Thirdly, while the CLMEG module improves the consistency of explanations, the "faithfulness" of these explanations—i.e., how accurately they reflect the actual reasoning process of the LLM—remains an open question in interpretability research. Future work could explore integrating causal reasoning methods into the explanation generation process to provide deeper causal insights into model behavior. Finally, as LLMs continue to grow in scale and complexity, the computational efficiency of AKDC-Net's components must be further optimized to meet the real-time application demands of high-stakes domains such as clinical decision-making.

7. Conclusions

This study proposes AKDC-Net, an integrated framework designed for bias correction and explainability of large language models (LLMs). By incorporating the Heteroscedastic Uncertainty-Aware Bias Detection (HUABD), Neural Symbolic Knowledge Graph Error Correction (NSKGEC), and Contrastive Learning-Based Explanation Generation (CLMEG) modules, the framework effectively mitigates factual inaccuracies and overcomes the "black box" limitation inherent in conventional LLMs. Experimental evaluations conducted in the medical domain demonstrate that AKDC-Net outperforms baseline methods by 14.1% in F1-score, 19.4% in correction performance, and 31.4% in user trustworthiness. Notably, the framework is model-agnostic, rendering it adaptable to other high-stakes domains such as finance and law. Future research will focus on automated knowledge graph construction and real-time performance optimization. Overall, AKDC-Net offers a robust pathway toward the development of trustworthy AI systems in high-risk application scenarios.

The primary contribution of this work resides in its integrated and principled methodology. Rather than treating bias detection, correction, and explanation as isolated, post hoc procedures, our framework unifies these three components via a meta-learning objective, enabling synergistic interaction among individual modules. Specifically, the introduction of uncertainty decomposition for bias detection, differentiable symbolic reasoning for knowledge correction, and contrastive learning-based explanation generation constitutes substantial advancements over existing approaches. Future efforts will aim to extend the framework to additional domains, automate the construction of temporal knowledge graphs, and enhance computational efficiency to accommodate real-time applications. By delineating a clear route toward more reliable and transparent LLMs, this research contributes to the broader endeavor of developing safe and beneficial artificial intelligence systems.

Author Contributions: Methodology, X.Y.; Software, C.Q., Y.W. and W.W.; Validation, C.Q.; Formal analysis, X.Y. and Q.L.; Investigation, Q.L.; Resources, C.Q., Y.W. and W.W.; Data curation, X.Y., Y.W. and W.W.; Writing—original draft, X.Y.; Visualization, X.Y. and Q.L.; Supervision, H.W.; Project administration, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Tianjin Science and Technology Plan Project—Research on Key Technologies and Industrial Application Demonstration of General-Purpose Large Models for Autonomous Intelligent Computing Power] grant number [24ZGZNGX00020].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Acknowledgments: The authors acknowledge the donation of high-performance server equipment from Lenovo Kaitian Technology Co., Ltd., which was utilized for the algorithm validation and testing in this study. No generative artificial intelligence (GenAI) tools were used during the preparation of this manuscript. The authors have reviewed all content and take full responsibility for the publication.

Conflicts of Interest: Xianming Yang, Chengdong Qian, Yonghui Wu, Wei Wang are employed by Phytium Technology Co., Ltd. The authors declare no conflicts of interest.

References

1. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 295–305.
2. Robbemdond, V.; Inel, O.; Gadiraju, U. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, 4–7 July 2022; pp. 223–233.
3. Pareek, S.; van Berkel, N.; Velloso, E.; Goncalves, J. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum. Comput. Interact.* **2024**, *8*, 1–31. [\[CrossRef\]](#)
4. Cheng, M.; Luo, Y.; Ouyang, J.; Liu, Q.; Liu, H.; Li, L.; Yu, S.; Zhang, B.; Cao, J.; Ma, J.; et al. A survey on knowledge-oriented retrieval-augmented generation. *arXiv* **2025**, arXiv:2503.10677.
5. Jiang, M.; Lin, B.Y.; Wang, S.; Xu, Y.; Yu, W.; Zhu, C. Knowledge-augmented Methods for Natural Language Generation. In *Knowledge-Augmented Methods for Natural Language Processing*; Springer Nature Singapore: Singapore, 2025; Volume 2024, pp. 41–63.
6. Turner, A.; Kaushik, M.; Huang, M.T.; Varanasi, S. Calibrating Trust in AI-Assisted Decision Making. Available online: <https://api.semanticscholar.org/CorpusID:235679554> (accessed on 8 October 2025).
7. Kaufman, R.; Costa, J.; Kimani, E. Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust. *Sci. Rep.* **2024**, *14*, 13061. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Cau, F.M.; Hauptmann, H.; Spano, L.D.; Tintarev, N. Effects of AI and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Trans. Interact. Intell. Syst.* **2023**, *13*, 1–42. [\[CrossRef\]](#)
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
10. Li, X.; Liu, B.; Khan, A.H.; Fan, L.; Wu, X.-M. Multi-Modal Pre-Training for Medical Vision-Language Understanding and Generation: An Empirical Study with a New Benchmark. *arXiv* **2023**, arXiv:2307.06942. [\[CrossRef\]](#)
11. Vereschak, O.; Bailly, G.; Caramiaux, B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc. ACM Hum. Comput. Interact.* **2021**, *5*, 1–39. [\[CrossRef\]](#)
12. Ilesanmi, F.O. The Dynamics of Trust in AI-Assisted Writing. Master's Thesis, University of Oulu, Oulu, Finland, 2024.
13. Cao, S.; Huang, C.M. Understanding user reliance on AI in assisted decision-making. *Proc. ACM Hum. Comput. Interact.* **2022**, *6*, 1–23. [\[CrossRef\]](#)
14. Sepehri, M.S.; Fabian, Z.; Soltanolkotabi, M.; Soltanolkotabi, M. MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models. *arXiv* **2024**, arXiv:2409.15477. [\[CrossRef\]](#)
15. Mehrotra, S.; Jorge, C.C.; Jonker, C.M.; Tielman, M.L. Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Trans. Interact. Intell. Syst.* **2024**, *14*, 1–36. [\[CrossRef\]](#)
16. Scharowski, N.; Perrig, S.A.C.; Svab, M.; Opwis, K.; Brühlmann, F. Exploring the effects of human-centered AI explanations on trust and reliance. *Front. Comput. Sci.* **2023**, *5*, 1151150. [\[CrossRef\]](#)
17. Henlein, A.; Bauer, A.; Bhattacharjee, R.; Ćwiek, A.; Gregori, A.; Kügler, F.; Lemanski, J.; Lücking, A.; Mehler, A.; Prieto, P.; et al. An Outlook for AI Innovation in Multimodal Communication Research. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management, Proceedings of the International Conference on Human-Computer Interaction, Washington, DC, USA, 29 June–4 July 2024*; Springer Nature: Cham, Switzerland, 2024; pp. 182–234.
18. Casolin, E.; Salim, F.D.; Newell, B. Evaluating the Influences of Explanation Style on Human-AI Reliance. *arXiv* **2024**, arXiv:2410.20067. [\[CrossRef\]](#)

19. Panigutti, C.; Perotti, A.; Pedreschi, D. Doctor XAI: An ontology-based approach to black-box sequential data classification explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 629–639.
20. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems; Advances in neural information processing systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
21. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608. [\[CrossRef\]](#)
22. Wang, Z.; Gong, P.; Zhang, Y.; Gu, J.; Yang, X. Retrieval-augmented knowledge-intensive dialogue. In *Natural Language Processing and Chinese Computing, Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Foshan, China, 12–15 October 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 16–28.
23. Yu, W. Retrieval-augmented generation across heterogeneous knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 52–58.
24. Teng, Z.; Li, L.; Xin, Z.; Xiang, D.; Huang, J.; Zhou, H.; Shi, F.; Zhu, W.; Cai, J.; Peng, T.; et al. A literature review of artificial intelligence (AI) for medical image segmentation: From AI and explainable AI to trustworthy AI. *Quant. Imaging Med. Surg.* **2024**, *14*, 9620–9652. [\[CrossRef\]](#)
25. Hofman, P.; Sale, Y.; Hüllermeier, E. Quantifying aleatoric and epistemic uncertainty with proper scoring rules. *arXiv* **2024**, arXiv:2404.12215. [\[CrossRef\]](#)
26. Shrestha, I.; Srinivasan, P. LLM Bias Detection and Mitigation through the Lens of Desired Distributions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 5–9 November 2025; pp. 1464–1480.
27. Rahman, S.S.; Islam, M.A.; Alam, M.M.; Zeba, M.; Rahman, A.; Chowa, S.S.; Raiaan, M.A.K.; Azam, S. Hallucination to truth: A review of fact-checking and factuality evaluation in large language models. *arXiv* **2025**, arXiv:2508.03860. [\[CrossRef\]](#)
28. Lin, Q.; Xu, F.; Lu, H.; He, K.; Mao, R.; Liu, J.; Cambria, E.; Feng, M. Towards Unified Neurosymbolic Reasoning on Knowledge Graphs. *arXiv* **2025**, arXiv:2507.03697. [\[CrossRef\]](#)
29. Bilal, A.; Ebert, D.; Lin, B. LLMs for explainable ai: A comprehensive survey. *arXiv* **2025**, arXiv:2504.00125. [\[CrossRef\]](#)
30. Mesinovic, M.; Watkinson, P.; Zhu, T. Explainability in the age of large language models for healthcare. *Commun. Eng.* **2025**, *4*, 128. [\[CrossRef\]](#)
31. Link-Gelles, R.; Weber, Z.A.; Reese, S.E.; Payne, A.B.; Gaglani, M.; Adams, K.; Kharbanda, A.B.; Natarajan, K.; DeSilva, M.B.; Dascomb, K.; et al. Estimates of Bivalent mRNA Vaccine Durability in Preventing COVID-19–Associated Hospitalization and Critical Illness Among Adults with and Without Immunocompromising Conditions—VISION Network, September 2022–April 2023. *Morb. Mortal. Wkly. Rep.* **2023**, *72*, 579–588. [\[CrossRef\]](#)
32. Horby, P.; Mafham, M.; Linsell, L.; Bell, J.L.; Staplin, N.; Emberson, J.; Wiselka, M.; Ustianowski, A.; Elmahi, E.; Prudon, B.; et al. Effect of Hydroxychloroquine in Hospitalized Patients with COVID-19. *N. Engl. J. Med.* **2020**, *383*, 685–695. [\[CrossRef\]](#)
33. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [\[CrossRef\]](#)
34. Pelka, O.; Koitka, S.; Rückert, J.; Nensa, F.; Friedrich, C.M. Radiology objects in context (ROCO): A multimodal image dataset. In *Proceedings of the Third International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS 2018)*; Springer: Cham, Switzerland, 2018; pp. 180–189. [\[CrossRef\]](#)
35. Centers for Disease Control and Prevention (CDC). Guidelines for the Treatment of Pseudomonas aeruginosa Infections. 2024. Available online: <https://www.cdc.gov/pseudomonas-aeruginosa/about/index.html> (accessed on 9 October 2025).
36. Liao, H.; Wang, S.; Cheng, H.; Zhang, W.; Zhang, J.; Zhou, M.; Lu, K.; Mao, R.; Xie, X. Aspect-enhanced explainable recommendation with multi-modal contrastive learning. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–24. [\[CrossRef\]](#)
37. Maity, S.; Saikia, M.J. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **2025**, *12*, 631. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Schlicker, N.; Baum, K.; Uhde, A.; Sterz, S.; Hirsch, M.C.; Langer, M. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Comput. Hum. Behav.* **2025**, *170*, 108671. [\[CrossRef\]](#)

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.