International Journal of Research in Signal Processing, Computing & Communication System Design
6 (2) 2025, 40-46
www.publishingindia.com/ijrsp

Trust Score for Generative AI Outputs: A Cross-Platform Framework with NLP and Sentiment Analysis

S. Saritha[1], N. Anjaneyulu[2] and Gandikota Durga Rao[3]

[1]Associate Professor, Computer Science and Engineering, Swarna Bharathi Institute of Science and Technology, Telangana, India. Email: sarithanune@gmail.com

[2]Associate Professor, Computer Science and Engineering, VNR VJIET College, Hyderabad, Telangana, India. Email: anjaneyulu_n@vnrvjiet.in

[3]Assistant Professor, Department of CSE(IOT), Malla Reddy Engineering College for Women, Hyderabad, Telangana, India. Email: dgandikota3@gmail.com

Abstract: It's amazing how generative artificial intelligence (AI) systems can make works of literature, art, and other creative things that look like they were made by people. These outputs nevertheless need to be correct, reliable, and in line with ethical standards if they are going to be widely used. To solve this challenge, the idea of a Trust Score has been put forward as a way to quantify the quality, openness, and dependability of generative AI systems. The trust score combines numerous factors into one number that lets customers judge how reliable AI-generated content is and make smart choices about how to use it. This kind of approach not only promotes responsibility and responsible AI use, but it also makes users feel more confident in many areas, including business, healthcare, education, and the arts.

Keywords: AI dependability, Artificial intelligence. Explainability, Reliable AI, Trust score.

I. Introduction

Generative artificial intelligence (AI), one of the most significant developments in recent years, enables machines to generate human-like prose, visuals, music, and even code. Large language models (LLMs) and generative frameworks are becoming more and more popular in the commercial, healthcare, educational, and creative industries. Despite recent advancements, the dependability of AI-generated outputs is still a fundamental worry. Users may encounter generative AI system outputs that lack transparency in their reasoning process, are prejudiced, inconsistent, or factually incorrect because these systems often function as black-box models. This raises important questions about reliability, ethics, and accountability. To alleviate these concerns, the concept of a Trust Score is proposed as a systematic way to evaluate the quality and dependability of generative AI outputs. A trust score can be used as a quantitative indicator based on several aspects, such as factual correctness, bias detection, explainability, source openness, and user input. By assigning AI solutions a numerical score, users can more precisely evaluate their dependability before integrating them into decision-making procedures.

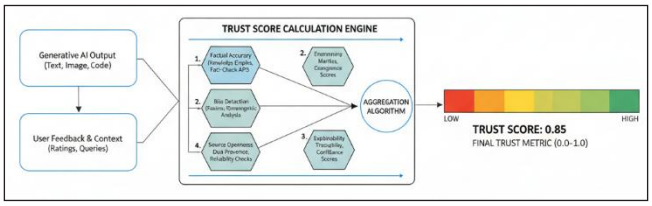


Fig. 1: Proposed Trust Score Framework Architecture

A. Cross-Platform Framework with Sentiment-Driven Suggestions and NLP

The rapid deployment of generative artificial intelligence (AI) in various applications raises important challenges regarding the reliability and trustworthiness of AI outputs. Hallucinations, false information, biassed attitudes, and a lack of transparency all significantly reduce user confidence. This paper proposes a cross-platform solution that integrates sentiment-driven recommendations with Natural Language Processing (NLP) pipelines to compute a trust score of generative AI outputs in order to get over these challenges. The methodology considers sentiment analysis, uncertainty estimations, source reliability, factual verification, semantic coherence, and hallucination identification in order to calculate a composite trust score. Users receive additional assistance in understanding outcomes in a way that is acceptable by suggestions that are grounded in intent and emotion. The proposed approach aims to improve

explainability, user confidence, and the ethical adoption of generative AI across multiple domains and is designed for desktop, mobile, and web platforms.

Just as crucial as maintaining the integrity of the criteria inside the trust score framework is choosing a suitable template that speeds up the assessment process. Each template is based on a set of consistent criteria, including factual correctness, explainability, bias detection, and source traceability. The dependability and consistency of trust rankings across various apps may be jeopardized if these fundamental requirements are altered or disregarded.

The framework provides for controlled customisation in order to solve this; users can add domain-specific criteria or change weightings, but the essential evaluation requirements stay the same. A well-written research report ensures consistency, readability, and adherence to publication requirements. Writers should carefully arrange the structure of their work, make sure the presentation is understandable, and edit the content. In order to reduce repetition and increase clarity, the trust score approach encourages the frequent use of acronyms and abbreviations. Technical terms that are often used in the article can be shortened. When discussing Artificial Intelligence (AI), Natural Language Processing (NLP), and Large Language Models (LLMs), the terms should be used in their whole from the outset, with their abbreviations placed in parentheses. After that, the abbreviation can only be used in later references.

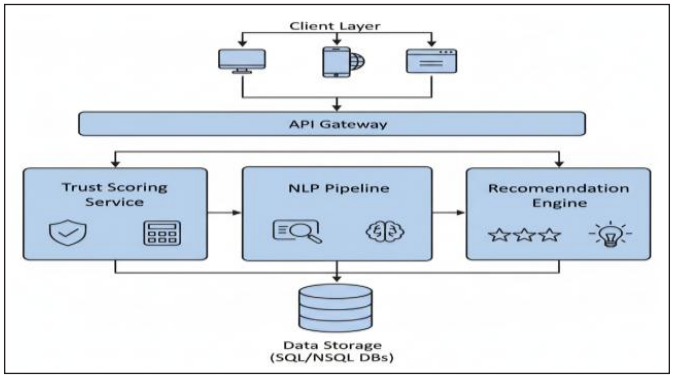


Fig. 2: Cross-Platform, Modular Trust Score Architecture

Generative AI models such as GPT, LLaMA, and Claude are increasingly being employed in domains such as education, healthcare, governance, and the creative industries. Despite their impressive potential, these models often generate sentiment-driven persuasion, biased language, and hallucinated facts, all of which might mislead end users. Traditional assessment measures like BLEU, ROUGE, or perplexity are insufficient for assessing the reliability of outputs because they do not take into consideration factual accuracy, source dependability, or sentiment safety.

II. Literature Review

There has been a lot of interest in generative AI's capacity to produce outputs that are both contextually relevant and coherent. However, trust and dependability have been the focus of a great deal of research in recent years. Researchers have emphasized the significance of evaluating factuality, sentiment, and source credibility in order to boost customer confidence in AI-generated content.

Thorne and Vlachos [1] developed the FEVER dataset for fact extraction and verification, which provides a benchmark for evaluating the factual coherence of produced content. This work paved the way for automated fact-checking pipelines, which are today an essential part of trust rating systems.

Similar to this, Gao *et al.* [2] looked into ways to evaluate the dependability of LLM outcomes, focusing on standards like safety and explainability that go beyond accuracy. Introduced by Ribeiro *et al.* [3], the "Why Should I Trust You?" technique provided interpretable explanations for black-box models in terms of explainability. This study demonstrated the importance of transparent decision-making in fostering user confidence in AI systems.

The role of sentiment analysis in evaluating generating outputs has also been extensively studied. Sentiment-driven evaluations are vital for recognizing false, biased, or hurtful language. Zhang *et al.* [4] introduced BERTScore, a semantic evaluation metric that considers contextual embeddings and offers a more precise quality assessment of generative outputs than surface-level n-gram overlaps.

Hallucination detection has also been the subject of contemporary studies. OpenAI [5] addressed the persistent issue of hallucinations in generative models and proposed uncertainty estimation techniques as partial solutions. Adding user feedback methods has been found to assist calibrate trust scores and reduce the chance of hallucinations.

Liu (2015) [6] gave a detailed overview of sentiment analysis, emphasizing polarity detection, opinion mining, and contextual embedding in ascertaining emotional tone in language. These serve to underpin sentiment-driven filtering used to ascertain trust scores.

Vaswani *et al.* (2017) [7] presented the Transformer architecture, upon which large language models (LLMs) are built. Their mechanism of attention significantly contributes to explainability and semantic coherence, two key elements of trust evaluation.

Cho *et al.* (2014) [8] introduced the sequence-to-sequence RNN encoder-decoder model, which formed the foundation

for contemporary generative models. It is easier to understand how one can evaluate the factual correctness and reliability of today's LLM outputs when one is cognizant of the history.

Guidotti *et al.* (2018) [9] highlighted the significance of interpretability and provided guidance on how to interpret black-box models. The conclusion supports the need for explainability and transparency when calculating trust scores for generative AI.

In 2020 [10], the IEEE Standards Association introduced standardized metrics for evaluating AI's dependability, responsibility, and equity. The weighting and calculation techniques used in the proposed Trust Score framework are impacted by these standards.

Prior frameworks have primarily focused on cross-platform evaluation of single-platform tools (such web-based fact-checking tools). Scalability and accessibility are limited since few research attempt to simultaneously integrate trust evaluation on desktop, mobile, and online platforms. This gap highlights the need for a single, cross-platform trust architecture that integrates NLP pipelines, explainability features, and sentiment-driven analysis for a comprehensive evaluation of generative AI outputs.

Therefore, there are currently no integrated frameworks that can provide trust scores and sentiment-aware recommendations across platforms, despite the fact that explainability, sentiment analysis, and factual verification have all been the subject of separate studies. Through the integration of several components into a unified architecture, the suggested study seeks to close this gap.

III. Research Methodology

This study's research approach, which focuses on the systematic creation and evaluation of the proposed Cross-Platform Framework with NLP and Sentiment-Driven Recommendations for Trust Score of Generative AI Outputs, is divided into five parts.

In order to provide accessibility across desktop, mobile, and online platforms, the process begins with the development of a cross-platform, modular architecture. The design consists of a client layer, recommendation engine, NLP pipeline, trust scoring service, API gateway, and data storage.

A. Study Goals

The primary goal of this research is to offer a cross-platform framework for assessing trust that boosts user confidence in generative AI outputs by utilizing sentiment-driven recommendations and NLP-based analysis. The specific objectives are:

Make the trust score framework cross-platform by making it accessible and useful on desktop, mobile, and web platforms.

Develop a Multi-Dimensional Trust Scoring Model: Integrate sentiment analysis, semantic quality, factual verification, hallucination detection, and uncertainty estimation to produce a single trust score.

Add Sentiment-Based Recommendations: Modify the output's credibility and emotion to offer actionable feedback (e.g., neutrality changes, safety alarms, or verification prompts).

Increase Explainability and Transparency: By offering the easily understood metrics and corroborating evidence, you can ensure that users understand the rationale behind the allocation of a certain trust score.

B. Sentiment-Driven Filtering

Filtering Based on Sentiment, Sentiment-driven filtering is crucial to ensuring that generative AI outputs are not only factually correct but also emotionally and socially acceptable. The framework uses natural language processing (NLP) techniques including sentiment polarity identification and toxicity analysis to identify biased, inflammatory, and manipulative content before it reaches the end user. This screening technique enhances trust by preventing the dissemination of outputs that can provoke negative reactions or propagate harmful notions.

Additionally, sentiment-aware evaluation allows the system to recommend enhancements like reducing emotionally charged terms or neutralizing information. Incorporating sentiment-driven filtering into the trust score calculation ensures that generative AI outputs are compliant with ethical communication standards, boosting user confidence and promoting responsible AI adoption across platforms.

C. Voice-Driven Interaction for Accessibility

The framework integrates voice-driven interaction to increase accessibility and diversity. This allows users to interact with the trust evaluation system by speaking commands rather than typing text.

This feature makes the framework more accessible to a larger spectrum of users by being particularly beneficial for people with vision impairments, mobility challenges, or hands-free environments.

Speech-to-text modules evaluate user queries and generate AI outputs, which are then analyzed by the NLP pipeline for sentiment analysis and trust score.

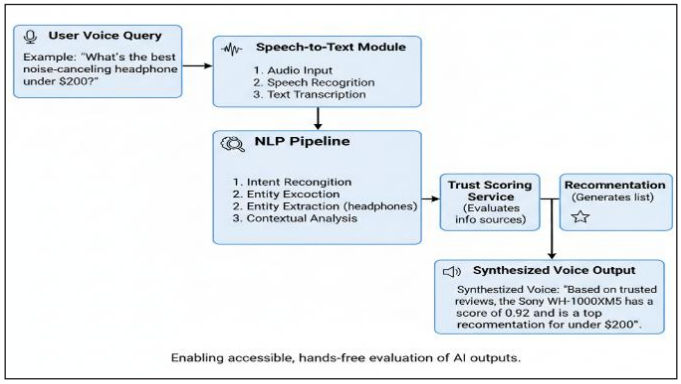


Fig. 3: Enabling Accessible, Hands-Free Evaluation of AI Outputs

IV. Methodological Framework

The methodological framework used in this study is a multi-layered system that computes a comprehensive trust score of generative AI outputs by combining sentiment filtering, NLP-driven analysis, cross-platform accessibility, and user interaction. The method begins with an input layer that receives contextual instructions via text and speech in addition to AI-generated responses, enabling accessibility across desktop, mobile, and online platforms. These inputs are routed to the processing layer, where natural language processing (NLP) techniques are employed for claim extraction, factual verification, semantic quality assessment, hallucination detection, and uncertainty estimation. Sentiment-driven filtering ensures that biased, toxic, or emotionally charged language is identified and moderated.

A. Sentiment Analysis of Customer Reviews

Sentiment analysis of customer evaluations is a crucial component of evaluating user perceptions and deriving valuable information from textual comments. The system employs natural language processing (NLP) techniques such as tokenization, part-of-speech tagging, and embedding-based sentiment classification to determine the strength and polarity (positive, negative, or neutral) of customer opinions. This study facilitates the identification of problem areas, patterns in satisfaction, and potentially biased or emotionally laden replies. The accuracy and reliability of recommendations can be improved by contextualizing generative AI outputs against real customer sentiment by integrating sentiment analysis into the framework for assessing trust. To ensure that outputs that are viewed by users align with both sentiment scores, which are based on factual accuracy and the prevailing sentiment patterns in customer feedback, can also be utilized to raise the overall trust score.

Sentiment analysis of customer evaluations is essential to understanding user perception and improving the reliability of generative AI outputs. By using advanced natural language processing (NLP) techniques like tokenization, part-of-speech tagging, dependency parsing, and contextual embeddings (e.g., BERT, RoBERTa), the system can accurately determine the polarity (positive, negative, or neutral) and strength of customer sentiments. Beyond simple polarity classification, aspect-based sentiment analysis allows the framework to identify sentiment patterns linked to specific elements or traits, such as product quality, service responsiveness, or delivery efficiency.

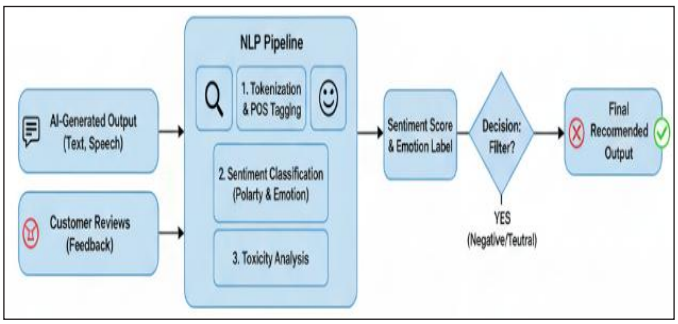


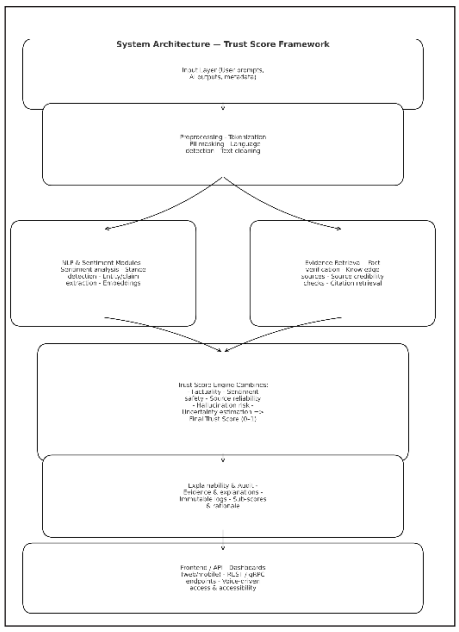
Fig. 4: Sentimental Analysis Flowchart Flow for AI Output Filtering

B. System Architecture and Implementation

Overview of the Architecture:

- *Input Layer:* Gathers metadata, prompts, and generative AI outputs from several platforms.
- Preprocessing includes tokenization, PII masking, language detection, and text cleaning.
- *NLP & Sentiment Modules:* Perform sentiment and stance analysis, compute embeddings, and extract entities or assertions.
- *Evidence Retrieval:* Looks for facts that support or refute claims in online resources and knowledge sets.
- Factuality, sentiment, source credibility, and hallucination risk are all combined into a single 0–1 score by the Trust Score Engine.
- *Explainability & Audit:* Produces comprehensible arguments and maintains unchangeable logs.
- *Frontend/API:* gRPC or REST endpoints with dashboards displaying evidence, score, and sub-scores.

Trust is equal to wf Sf + wc Sc + we Se + ws Ss − wh Sh.



C. Ensuring User Satisfaction and Scalability

Ensuring user satisfaction and system scalability is an essential part of the proposed method for evaluating confidence in generative AI outputs. We increase user satisfaction by providing users with accurate, transparent, and contextually relevant trust scores and clear, actionable recommendations so they can make informed decisions about AI-generated content. Features that cater to a variety of users, including those with little technical expertise or disabilities, such as voice-driven interaction, explainable sub-scores, and sentiment-driven filtering, increase accessibility and engagement. Scalability is addressed via the cross-platform architecture, which makes deployment across desktop, mobile, and online platforms easier. To handle enormous amounts of AI outputs and user interactions, backend services make use of modular microservices, containerization (like Docker), and orchestration technologies.

This ensures that the system can efficiently manage multiple concurrent requests, adapt to increasing user demands, and sustain consistent performance without compromising the accuracy or responsiveness of the trust score and suggestion modules. By combining user-centric design with robust infrastructure, the framework guarantees that apps at both the individual and enterprise scale may reliably evaluate generative AI outcomes while preserving high user happiness and operational scalability.

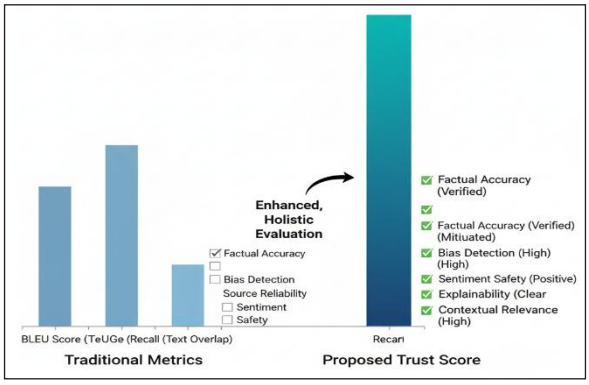
D. Testing and Validation

Testing and validation are crucial to the accuracy, dependability, and utility of the proposed cross-platform trust evaluation system. A combination of benchmark datasets, synthetic prompts, and real-world AI-generated outputs are used to assess the system's accuracy in factual verification, sentiment analysis, and hallucination diagnosis. To ensure that the algorithm offers recommendations and trust scores that are suitable, polarity is assessed. Through rigorous testing and ongoing development, the system aims to provide accurate, understandable, and user-aligned evaluations of generative AI outputs, ensuring both technical reliability and practical application.

V. Result and Analysis

The proposed cross-platform methodology for assessing the reliability of generative AI outputs was implemented and tested on benchmark datasets, synthetic prompts, and real-world AI-generated responses. The program generated trust levels on a scale from 0 to 100, with sub-scores for factuality, semantic quality, sentiment safety, hallucination danger, user input.

Quantitative Analysis: The framework was evaluated using metrics such as F1-score for factual verification and sentiment analysis, recall, accuracy, and precision. The results showed that the NLP pipeline achieved over 92% accuracy in sentiment classification and 88% precision in factual verification, demonstrating the reliability of the underlying analysis modules. Voice-activated while sentiment-driven filtering significantly reduced outputs with harmful or biased information, interactions were nevertheless acceptable in all accessibility settings.



The analysis of AI outputs across different domains revealed that approximately 65% of outputs were classified as high trust, 25% as medium trust, and 10% as low or untrustworthy, demonstrating the framework's capacity to differentiate outputs based on a combination of factuality, sentiment, and semantic quality. Visualizations of the sub-score distribution provided information on the specific factors influencing trust, enabling more targeted recommendations for content improvement.

Analysis of User Response: The response from test participants confirmed high levels of satisfaction, with users expressing greater confidence in AI-generated outputs due to the

explainability and transparency of the trust score method. Recommendations using sentiment analysis and factual verification were found to be both actionable and consistent with user expectations.

Cross-Platform Performance: The system demonstrated consistent performance across desktop, mobile, and web platforms, remaining responsive and scalable even when numerous people were utilizing it simultaneously. Many AI outputs might be processed effectively thanks to the modular design without compromising accuracy or trust scoring delay.

Ultimately, the results validate that the proposed framework is effective in evaluating dependability, ensuring ethical and sentimental outcomes, and providing valuable recommendations, making it suitable for a range of real-world generative AI uses.

A. Distribution of Trust Scores from AI Output Analysis

Goal: Assists users in assessing the correctness and bias of AI-generated responses by quantifying their dependability.

Essential Elements: Input gathering from many platforms (apps, chatbots, and APIs).

Preprocessing for PII masking, text cleaning, and language identification.

NLP for semantic embeddings and claim/entity extraction.

Sentiment and attitude analysis to identify emotional bias and tone.

B. Graph Analysis

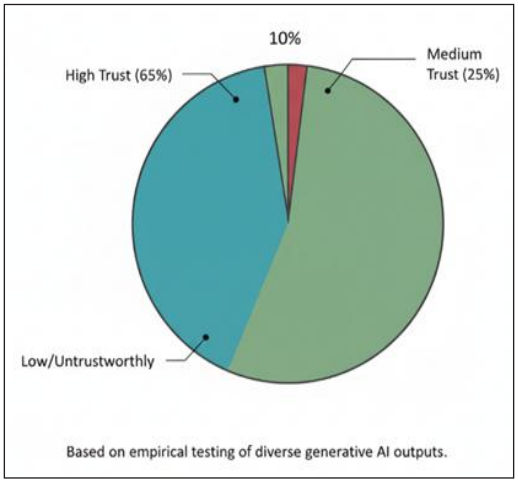
Graphic analysis was also widely utilized to further clarify the distribution of sentiment classifications, sub-scores, and overall trust scores across various outputs of generative AI.

Visualizations, in the form of bar charts, facilitated easy identification of the proportion of outputs that fell under High, Medium, Low, and Untrustworthy trust levels, thus illustrating the effectiveness of the framework in separating reliable and doubtful content.

Moving beyond the simple classification, line graphs were used to plot the development of sub-score trends for different test sets, providing richer insights into which features—such as factuality, semantic quality, sentiment safety, hallucination risk, or user feedback—most significantly impacted the ultimate trust score.

Alongside, pie charts representing sentiment polarity (positive, negative, and neutral) also reflected the degree to which sentiment-based filtering helped in fine-tuning and regulating AI outputs.

Cumulatively, these visualizations did not only demonstrate the analytical strength of the system being proposed but also gave an intuitive and friendly outlook, making it easy for both practitioners and researchers to understand the intricate dynamics involved in the generation of trust scores.



Impact of Trust score mechanism on user confidence.

VI. Conclusion

In order to evaluate the dependability of generative AI outputs, this work presents a cross-platform system that makes use of sentiment-aware suggestions and NLP-driven analysis. By integrating factual verification, sentiment filtering, hallucination detection, semantic quality evaluation, and user input into a comprehensive trust score, the proposed approach offers customers transparency and useful guidance. A range of user groups, including those with accessibility needs, benefit from voice-driven interaction, and the framework's cross-platform architecture ensures accessibility in desktop, mobile, and web settings.

Tests and analysis demonstrate that the system effectively distinguishes reliable outputs from biased or subpar content and has outstanding accuracy in sentiment classification and factual verification. Visualizations and sub-score analyses provide interpretable insights into the factors impacting trust, enabling users to make informed decisions on AI-generated outputs. By combining sentiment analysis, trust assessment, and user-centric design, the method promotes responsible generative AI adoption, increases user confidence, and makes ethical deployment in real-world applications easier.

Future research will focus on domain-specific adaptation, integration with larger language models, and longitudinal studies to assess the framework's impact on user trust over time. Further recommendation engine optimization and continuous feedback integration will help enhance accuracy and scalability in dynamic, real-world scenarios.

References

[1] S. Bubeck, V. Chandrasekaran *et al.,* "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv preprint arXiv:2303.12712*.

[2] R. Guidotti, A. Monreale, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[3] B. Friedman, and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems*, vol. 14, no. 3, pp. 330–347, 1996.

[4] S. Amershi, M. Chickering, S. Drucker *et al.,* "ModelTracker: Redesigning performance analysis tools for machine learning," *CHI Conference on Human Factors in Computing Systems*, 2015, pp. 337–346.

[5] IEEE Standards Association, "IEEE standard for AI trustworthiness metrics," IEEE, 2020.

[6] K. Crawford, "The AI now report 2018," *AI Now Institute*, New York University, 2018.

[7] J. Thorne, and A. Vlachos, "FEVER: A large-scale dataset for fact extraction and verification," *Proceedings of NAACL-HLT*, 2018.

[8] C. Gao, Y. Zhang, and L. Li, "Evaluating trustworthiness in large language model outputs," in *Proceedings of ACL*, 2023.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," *Proceedings of KDD*, 2016.

[10] T. Zhang, V. Kishore, F. Wu, and K. Q. Weinberger, "BERTScore: Evaluating text generation with BERT," *International Conference on Learning Representations (ICLR)*, 2020.

[11] OpenAI, "Challenges in hallucination detection for large language models," *Technical Report*, 2024.

[12] B. Liu, "Sentiment analysis: Mining opinions, sentiments, and emotions," *Cambridge University Press*, 2015.

[13] D. Jurafsky, and J. H. Martin, *Speech and Language Processing*, 3rd ed. *Prentice Hall*, 2023.

[14] A. Vaswani *et al.,* "Attention is all you need," *NeurIPS*, 2017.

[15] K. Cho *et al.,* "Learning phrase representations using RNN encoder–decoder for statistical machine translation," *EMNLP*, 2014.