IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 August 2025, accepted 13 September 2025,
date of publication 22 September 2025, date of current version 30 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3612568

RESEARCH ARTICLE

Trustworthiness Evaluation of Large Language Models Using Multi-Criteria Decision Making

MELTEM AKSOY[1], AYLIN ADEM[2], AND METIN DAĞDEVIREN[2,3]
[1]Department of Computer Science, Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, Technical University Dortmund, 44227 Dortmund, Germany
[2]Department of Industrial Engineering, Gazi University, 06570 Ankara, Türkiye
[3]Council of Higher Education, 06539 Ankara, Türkiye

Corresponding author: Meltem Aksoy (meltem.aksoy@tu-dortmund.de)

ABSTRACT As Large language models (LLMs) become increasingly integrated into high-stakes applications, ensuring their trustworthiness has emerged as a critical research concern. This study proposes a novel evaluation framework that applies a multi-criteria decision making (MCDM) methodology, specifically the hesitant fuzzy analytic hierarchy process (AHP), to assess and rank LLMs based on five key trust dimensions: fairness, robustness, integrity, explainability, and safety. Drawing from expert evaluations, the framework systematically determines the relative importance of each criterion and applies a weighted scoring approach to compare seven leading LLMs, including both proprietary models such as GPT-3.5, GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 and open-source models such as Llama 3.1, Mistral Large 2 and DeepSeek V3. Results reveal GPT-4o as the most trustworthy model, significantly outperforming its peers, particularly in robustness and fairness. Open-source models showed lower scores, especially in safety and explainability, highlighting persistent gaps in their alignment with trust expectations. The findings demonstrate the effectiveness of MCDM in capturing expert uncertainty and prioritizing trust criteria, offering a robust and adaptable framework for evaluating LLMs in dynamic and sensitive domains.

INDEX TERMS Fuzzy AHP, expert evaluation, multi-criteria decision making, large language models, trustworthiness.

I. INTRODUCTION

Large language models (LLMs) have revolutionized a wide range of fields by redefining how we solve complex problems and interact with technology. These models, with their ability to generate human-like text and perform highly sophisticated tasks, excel in areas such as content creation, summarization, language translation, coding, task planning, and decision-making. Platforms like Hugging Face demonstrate the rapid expansion of this ecosystem, offering access to a wide range of models for diverse applications. These models drive innovation across fields such as healthcare [1], education [2], law [3], software engineering [4], and creative industries [5], making them invaluable tools for addressing both routine challenges and critical decision-making processes [6]. Recent research further highlights the rapid increase in LLM-related research, reflecting the growing academic interest and diverse application domains of these models [7], [8].

The remarkable capabilities of LLMs stem from their training on vast and diverse datasets, sourced from real-world text and internet-based content, combined with the use of advanced transformer-based architectures. Innovations such as low-rank adaptation (LoRA) [9] and reinforcement learning from human feedback (RLHF) [10] have further enhanced their alignment with human values, enabling these models to generate high-quality outputs. These innovations have also enabled LLMs to scale effectively, producing outputs that are accurate, contextually relevant, and coherent.

Despite their remarkable capabilities, concerns about the trustworthiness of LLMs have emerged, particularly as they are increasingly deployed in high-stakes domains like healthcare and legal decision-making. Biases in training datasets can lead to unfair or inequitable outcomes, while the models' reliability is often challenged under diverse or

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

unforeseen conditions [11]. The lack of transparency in their decision-making processes makes it difficult for users to interpret or validate outputs, undermining trust [12]. Privacy concerns also persist, as training data frequently includes sensitive information, raising risks of misuse or breaches [13]. Furthermore, ensuring that LLMs generate outputs that are safe, non-harmful, and aligned with ethical standards remains a persistent challenge [14]. These concerns highlight the need for comprehensive frameworks to ensure LLMs meet the trustworthiness expectations of users.

This study is inspired by the FRIES trust score framework proposed by [15], which evaluates the trustworthiness of machine learning models and datasets. The framework defines five trust dimensions (fairness, robustness, integrity, explainability, and safety) based on an extensive literature review. These dimensions are also highly relevant and applicable for assessing the trustworthiness of LLMs. However, while the FRIES framework provides a solid foundation, its methodology, inspired by failure mode and effects analysis (FMEA), has notable limitations. It relies on averaging individual scores, which fails to capture the nuanced perspectives of experts or account for the relative importance of different criteria. Additionally, FMEA's purpose is centered around risk assessment and failure prevention, making it less suitable for trustworthiness evaluations that require balancing multiple conflicting criteria for ranking multiple alternatives.

To address these shortcomings, we propose a methodology based on the multi-criteria decision making (MCDM) approach. The decision-making process is a continuous activity that individuals engage in, either consciously or unconsciously, encompassing both simple and complex problems. In cases where multiple alternatives exist and the criteria influencing the decision conflict, MCDM methods are employed [16]. While traditional MCDM techniques predominantly utilize crisp numbers, uncertainties, and data deficiencies necessitate the use of fuzzy MCDM approaches [17], [18]. MCDM allows for the systematic aggregation of expert opinions through pairwise comparisons and assigns weights to criteria, reflecting their relative importance. Experts systematically compare preferences, which are then converted into numerical weights through MCDM. This formal process reduces arbitrary judgments, ensures logical coherence through consistency checks, and allows for a balanced consideration of all trust dimensions. Unlike FRIES, MCDM generates comparative rankings of alternatives, enabling clearer and more actionable evaluations of trustworthiness. Among MCDM techniques, the analytical hierarchy process (AHP) is one of the most widely applied methods for ranking criteria and addressing complex decision-making problems [19], [20], [21]. However, this study employs an extended version of AHP with hesitant fuzzy numbers, rather than the conventional AHP method. This approach offers greater flexibility by accommodating expert hesitation in comparative evaluations and utilizing linguistic expressions instead of rigid numerical scales [17], [22], [23], [24]. In this study, we utilized the hesitant fuzzy AHP method to determine the criteria weights for evaluating LLMs based on trust. In the second phase, we assessed alternatives using a scoring-based approach to identify the most prominent option.

The evaluation of trustworthiness in LLMs necessitates a multifaceted approach that encompasses various dimensions of performance. Initially, accuracy serves as a fundamental metric, reflecting the model's ability to generate responses that align with factual information. However, this is insufficient; trustworthiness also involves the assessment of consistency and coherence in the model outputs, which are critical for establishing reliability. Furthermore, transparency is a pivotal element, as users must understand the decision-making processes behind the LLMs' responses. Incorporating user feedback mechanisms allows for iterative improvements and enhances the model's accountability. Beyond technical metrics, ethical considerations, such as bias and fairness, must be scrutinized to ensure that the model does not perpetuate harmful stereotypes or misinformation. Together, these factors create a holistic framework for measuring trustworthiness, ultimately enabling users to have greater confidence in the outputs produced by LLMs.

Compared with existing LLM evaluation frameworks, the novelty of our approach lies in explicitly integrating expert uncertainty and relative weighting of trust dimensions into the evaluation process. Unlike FRIES [15], which relies on averaging individual scores, our hesitant fuzzy AHP approach systematically incorporates expert hesitation through linguistic term sets and provides consistency checks in pairwise comparisons. This enables more robust and interpretable rankings of multiple LLMs across several trust dimensions. Moreover, while large-scale benchmarks such as TRUSTLLM [26], HELM [27], DecodingTrust [28], and XTRUST [30] mainly adopt task-based performance metrics, our framework introduces an expert-driven, multi-criteria methodology that captures both technical and ethical considerations. This contribution is particularly valuable in high-stakes domains, where subjective expertise and contextual weighting of criteria are crucial for a comprehensive and actionable assessment of trustworthiness.

II. RELATED WORKS

A. TRUSTWORTHINESS EVALUATION OF LLMS

The evaluation of trustworthiness in LLMs has become a significant focus of research, with numerous frameworks and methodologies proposed to address the complexities of this task [25]. Current literature primarily relies on task-based assessments that use objective metrics to evaluate various dimensions of trustworthiness, such as accuracy, fairness, robustness, and safety. Reference [26] stands out as a comprehensive taxonomy that identifies eight key dimensions: truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability. TRUSTLLM [26] evaluates 16 LLMs across 30 datasets and tasks, highlighting disparities in performance between proprietary and

open-source models. HELM [27] adopts a multi-metric evaluation approach, emphasizing dimensions beyond accuracy, such as bias and toxicity. Reference [28] evaluates GPT-4 and GPT-3.5 models through dimensions such as bias, toxicity, and adversarial robustness, providing insights into their strengths and weaknesses. Reference [29] introduces a prompting strategy that uses malicious demonstrations to test the trustworthiness of open-source LLMs. XTRUST [30] benchmark extends these evaluations by incorporating multilingual capabilities, assessing models in ten languages, and expanding trustworthiness assessments beyond English-centric datasets. Furthermore, Reference [31] introduced TRUST-SCORE, a holistic task-based metric designed to assess LLM trustworthiness in Retrieval-Augmented Generation (RAG) contexts. Their approach extends the evaluation scope by incorporating dimensions such as answer correctness, citation groundedness, and refusal capability.

While these approaches uncover critical weaknesses, they lack a holistic perspective and fail to integrate multi-dimensional assessments that reflect the broader trustworthiness landscape. A recurring limitation in these frameworks is their reliance on task-based, objective evaluations, which often fail to capture the subjective, context-dependent aspects of trustworthiness. Many frameworks do not incorporate expert opinions or allow for the dynamic weighting of trust dimensions, which is crucial for reflecting their varying importance across domains. Furthermore, there is no unified framework or consensus on standardized criteria for trustworthiness evaluations, complicating cross-framework comparisons. These limitations underscore the need for methodologies that can address the inherent subjectivity and contextual nuances of trustworthiness while also enabling systematic, transparent evaluations.

B. APPLICATIONS OF MCDM IN LLM EVALUATION

MCDM techniques have been applied in various AI-driven evaluations, including LLM and chatbot assessments. Reference [32] used fuzzy-weighted zero-inconsistency (FWZIC) and MAIRCA to rank medical LLMs based on clinical concept extraction. Reference [33] developed a Fuzzy AHP (FAHP)-based evaluation framework tailored for the healthcare domain, aiming to assist providers in selecting suitable LLMs. Reference [34] integrated AHP with GPT-4 for structured decision-making in cybersecurity.

In chatbot evaluations, [35] introduced an AHP-CoCoSo framework to assess customer service chatbots, incorporating single-valued neutrosophic sets for uncertainty management. Additionally, [36] proposed LLM-as-a-Fuzzy-Judge, a hybrid framework combining supervised fine-tuning and prompt engineering to align LLM outputs with human expert evaluations in medical education settings. Their approach leverages fuzzy logic to model subjective criteria such as professionalism, medical relevance, ethical behavior, and contextual distraction when evaluating medical students' clinical communication skills. Reference [37] applied AHP to evaluate clinical health chatbots, while Reference [38] used DEMATEL for mental health chatbot selection. Although these studies provide structured evaluation frameworks, they focus on performance-based assessments rather than trustworthiness. Reference [39] also employed DEMATEL but focused on identifying criteria for user trust in LLMs, analyzing factors like credibility, data security, and user experience. However, this study does not evaluate the trustworthiness of LLMs as models but rather explores user perceptions.

C. RESEARCH GAP

Recent advances in benchmarking and evaluation frameworks have provided valuable insights into the performance of LLMs, yet critical challenges remain in capturing their overall trustworthiness in a systematic and interpretable way. Large-scale benchmarks such as TRUSTLLM [26], HELM [27], DecodingTrust [28], and XTRUST [30] have systematically assessed models across multiple tasks and dimensions. However, these frameworks primarily rely on performance-based task evaluations, focusing on metrics such as fairness, robustness, or toxicity in isolation. Their results are typically reported dimension-by-dimension, without providing a systematic aggregation into a single interpretable trust score. Moreover, these task-based approaches do not incorporate expert judgments, thereby overlooking the subjective and context-dependent aspects of trust that are crucial in high-stakes domains. While the FRIES framework [15] was not designed specifically for LLMs, it introduced a widely recognized set of trust dimensions for machine learning models and datasets. However, its evaluation methodology—based on simple averaging—does not capture expert hesitation or the relative importance of different criteria. On the other hand, MCDM-based applications [32], [33], [34], [35], [36], [37], [38], [39] demonstrate that these methods can effectively support structured evaluations and even guide LLM selection in specific contexts. However, they have not been applied with a primary focus on systematically assessing the trustworthiness of LLMs. Our study addresses this gap by applying a hesitant fuzzy AHP approach, which explicitly models expert hesitation, incorporates the relative importance of trust dimensions, and provides interpretable comparative rankings of leading LLMs, thereby combining technical and ethical considerations into a structured evaluation process.

III. METHODOLOGY

In this paper, we proposed an MCDM-based hybrid calculation methodology to prioritize LLM alternatives concerning a trust-based criteria set. The general overview of the calculation process is shown in Fig. 1.

In Step 1, we determined the trust-based criteria set by deriving the criteria from the literature, using [15] as the primary reference. Although [15] addressed a similar problem, it applied a different evaluation approach for assessing alternative LLMs. In this study, we adopted the criteria set proposed by [15] and adapted it to our specific context.

In step 2, i.e., determine the alternatives, we selected seven state-of-the-art LLMs commonly used in academic and industrial settings. The selection was based on their technical attributes, frequency of use, and popularity within the AI community.

In step 3, i.e., expert selection and evaluation, experts were chosen based on their domain knowledge and familiarity with trustworthiness evaluation in AI systems. To avoid groupthink and ensure independent judgments, no formal group discussion or consensus-building method (e.g., Delphi technique) was employed. Instead, each expert independently evaluated the alternatives and criteria using the predetermined hesitant fuzzy linguistic term set.

In step 4, i.e., calculation of the criteria weights, the hesitant fuzzy AHP method was applied to aggregate the individual expert inputs and derive the final weights.

Finally, in step 5, the linguistic scoring system is used to evaluate the alternatives with respect to the criteria.

Before proceeding with the details about the calculation process, the general information about the MCDM process, hesitant fuzzy sets, and the extension of the hesitant fuzzy AHP were given to clarify why the related techniques were preferred in the analysis stage of the study.

The act of decision-making is a continuous action that people perform, whether they are aware of it or not. Decision-making processes can involve simple topics and concepts, as well as highly comprehensive and complex processes. In both simple and comprehensive decision-making processes, there are several factors that decision-makers consider, along with multiple alternative situations that need to be evaluated. The decision-maker selects the most suitable alternative based on the factors (criteria) influencing the decision, using their current knowledge and experience. This basic decision-making mechanism can be described in this way. When multiple alternative situations are subject to selection and the criteria influencing the selection conflict with each other, the phenomenon of MCDM arises [16], [40]. In the literature, there are too many techniques that have been developed to solve the MCDM problems [41]. Generally, in these techniques, crisp numbers are utilized. However, due to the nature of the decision-making process, data deficiencies and uncertainties may sometimes be encountered. As a solution to this issue, MCDM techniques in fuzzy environments have been developed [17], [18], [42].

On the other hand, nearly all MCDM techniques have their own specific usage purposes, calculation methods, and processes. AHP is one of the most widely used techniques for ranking criteria in decision-making processes, specifically for calculating criteria weights. It has been successfully applied to MCDM problems across various fields to date [43], [44], [45]. This technique, developed by Saaty, is a linear algebra-based method that relies on binary comparison logic and aims to solve complex MCDM problems [46], [47].

However, in this article, the AHP technique is not used in its traditional form but in its extended version with hesitant fuzzy numbers. This technique, which belongs to the group of methods extended with fuzzy numbers, has been frequently used in the literature and has proven to be as valid as the traditional AHP technique [17], [22], [23], [24], [48].

A key distinguishing feature of hesitant fuzzy numbers compared to other scales based on fuzzy numbers is that they provide an effective solution in cases where experts are unable to determine the superiority of one option over another [22], [49]. Another notable feature is that the evaluation criteria in the scale utilize flexible, human-like linguistic expressions. For instance, while the traditional AHP requires selecting one of the exact numbers from the 1-9 scale to express the superiority relationship between criteria x and y, the hesitant fuzzy AHP allows this relationship to be determined more flexibly [22], [50].

Due to these two features, we employed the hesitant fuzzy AHP technique to determine the weights of the criteria for evaluating the trustworthiness of LLMs. In the second stage, each alternative was assessed according to the established criteria, and a basic scoring approach was applied to identify the most prominent model.

Calculation of Alternative Weights Using the Hesitant Fuzzy Extension of AHP: The Hesitant Fuzzy Linguistic Term Set (HFLTS) is a relatively recent concept in fuzzy logic [50]. These sets consist of linguistic terms that accommodate decision-makers' hesitation or indecision, enabling them to articulate their preferences more effectively during the decision-making process [22], [51], [52]. In essence, HFLTS facilitates the incorporation of subjective expert perspectives into a structured decision-making framework.

Table 1 presents the numerical representations of the linguistic expressions. We directly adapted this scale, including its numerical equivalents and its application in the Hesitant Fuzzy-AHP process, from [22]. The process of integrating the AHP method with the HFLTS involves the following calculation process:

Firstly,the semantics and syntax of the linguistic term set (S) is established. In this study, S consists of the following components: ''No importance (ni), very low importance (vli), low importance (li), medium importance (mi), high importance (hi), very high importance (vhi), absolute importance (ai)'' [22].

After that, the context-free grammar (GH) is defined, where GH = {VN, VT, I, P}, as shown in the equation at the bottom of the next page.

VN can be expressed as a set of variable symbols (primary term, composite term, unary relation, binary relation, and conjunction). VT contains all the expressions that will be used in generating phrases through the production rules.

For the context-free grammar, the production rules are as shown in Fig. 2.

In addition to the linguistic expressions mentioned above, conjunction words such as ''more than,'' ''at least,'' and ''at most'' are also included. Using production rules, these conjunctions enable the formation of complex phrases like

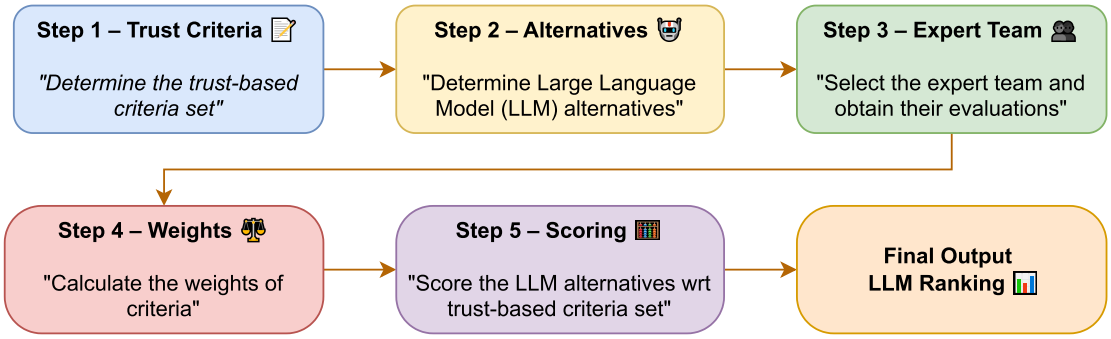


FIGURE 1. Proposed framework for evaluating LLM alternatives based on trust criteria.

TABLE 1. The scale for linguistic expressions.

ni	vli	li	mi	hi	vhi	ai
0	1	2	3	4	5	6

$$P = \begin{pmatrix} \text{I} = \langle \text{primary term} \rangle \,|\, \langle \text{composite term} \rangle, \\ \langle \text{composite term} \rangle ::= \\ \langle \text{unary relation} \rangle \langle \text{primary term} \rangle | \langle \text{binary relation} \rangle \\ \langle \text{primary term} \rangle | \langle \text{conjuction} \rangle \langle \text{primary term} \rangle, \\ \text{primary term} ::= S_0 \,|\, S_1 \,|\, ... \,|\, \text{S}_g, \langle \text{unary relation} \rangle ::= \\ \text{lower than} \,|\, \text{greater than} \,|\, \text{at least} \,|\, \text{at most}, \\ \langle \text{binary relation} \rangle ::= \text{between}, \langle \text{conjuction} \rangle ::= \text{and} \end{pmatrix}$$

FIGURE 2. Production rules [22].

"more than very high importance" or "at least medium importance", enhancing the flexibility and expressiveness of linguistic evaluations.

Then, the gathering of the preferences relations p^k provided by expert $k \in \{1, 2, \ldots, m\}$ for criteria is performed. At this stage, experts are requested to assess the criteria.

Following the gathering process, by utilizing the transformation function E_{GH}, the preference relations are transformed into HFLTS. The linguistic evaluations provided by the experts are transformed using the E_{GH} function. The operation of this function is demonstrated, as shown in the equation at the bottom of the page.

Then, envelope $\left[p_{ijk}^{-}, p_{ijk}^{+}\right]$ for each HFLTS is obtained. At this stage, a range is assigned to the results obtained in the previous step, defining both lower and upper limits to represent the uncertainty in the evaluations.

After that, since there is more than one expert,the selection of a linguistic aggregation operator φ is required. In our study, seven experts assessed the criteria. To prevent data loss, we choose the arithmetic mean method to aggregate the experts' linguistic evaluations [53]. The arithmetic mean method uses the following steps:

$$\tilde{x} = \Delta\left(\frac{1}{n}\sum_{i=1}^{n}\Delta^{-1}(s_i, \alpha_i)\right) = \Delta\left(\frac{1}{n}\sum_{i=1}^{n}\beta_i\right) \tag{1}$$

The 2-tuple representation corresponding to S is defined as $S = S \times [0.5, 0.5]$. The function: $\Delta \rightarrow [0, g] \rightarrow S$ is given by (2).

$$\Delta(\beta) = (s_i, \alpha_i) \text{ with } \begin{pmatrix} i = \text{round}(\beta) \\ \alpha = \beta - i \end{pmatrix} \tag{2}$$

where, β is approximated to the nearest integer i, $(i \in \{0, 1, \ldots, g\})$ using a rounding function.

$$\text{VT} = \left\{\text{lower than, greater than, at least, at most, between,and } \text{S}_0, \text{S}_1 \ldots\ldots\ldots, \text{S}_\text{g}\right\}$$
$$\text{I} \in \text{VN}$$

$$\text{E}_{\text{GH}}(\text{very low importance}) = [\text{very low importance}]$$
$$\text{E}_{\text{GH}}(\text{at least very high importance} = [\text{very high importance, absolute importance}]$$
$$\text{EE}_{\text{GH}}(\text{more than medium importance}) = \left[\text{high importance, very high importance, absolute importance}\right]$$

$\Delta^{-1} : \langle S \rangle \rightarrow [0, g]$ is defined by (3).

$$\Delta^{-1}(s_i, \alpha_i) = i + \alpha \tag{3}$$

The experts' linguistic evaluations are converted into numerical values through the use of the arithmetic mean operator. After that, the determination of the pessimistic and optimistic collective preferences, the calculation of interval-based utilities, the normalization of these utilities, and the computation of the criteria weights are performed, respectively.

IV. CASE PROBLEM

In this section, the MCDM's main problem focuses on the evaluation and assessment of LLM models with respect to their trustworthiness across various critical dimensions. We examine different widely used LLMs, including GPT-4o (A1), GPT-3.5 (A2), Claude 3.5 Sonnet (A4), Gemini 1.5 (A5), representing closed-source models, along with Llama 3.1 (A3), Mistral Large 2 (A6), and DeepSeek V3 (A7) as open-source alternatives. These models were selected to provide a comprehensive evaluation across both proprietary and open-source solutions. These 7 LLMs are identified and referred to in the context of this work as alternatives (A1-A7). It should be noted that these alternatives are utilized in the context of the current research as a case study, and due to the rapid advancement of LLM technology, the results of this assessment may evolve as new models emerge with enhanced capabilities. We employed various criteria based on the FRIES Trust Score framework dimensions (fairness, robustness, integrity, explainability, and safety) to assess each alternative's trustworthiness and suitability for high-stakes applications.

A. ALTERNATIVES

We identified seven ($n = 7$) widely used LLMs with effective capabilities as alternatives. The models were chosen to ensure a balanced comparison between closed-source and open-source models while covering a range of parameter sizes and performance levels. Details of these LLMs are as follows:

GPT-4o (A1): Building upon the foundation of GPT-4, OpenAI introduced GPT-4o (o for omni) [54], a model designed for enhanced multimodal capabilities and improved efficiency. GPT-4o surpasses its predecessor GPTs not only in speed and cost-effectiveness, being 50% cheaper and faster in the API, but also in its ability to process text, images, and even audio with heightened accuracy. Moreover, it demonstrates improved performance in non-English languages, broadening accessibility and usability across diverse linguistic contexts. GPT-4o supports a 128k token context window.

GPT-3.5 (A2): OpenAI's generative pre-trained transformer (GPT) series has significantly advanced the field of natural language processing. GPT-3.5 [55] is a sophisticated language model that belongs to the GPT-3 series. It leverages the transformer model to generate human-like text based on the input it receives. It is widely used due to its cost-effectiveness compared to the GPT-4 model series. GPT-3.5 has 175B parameters and a context window of 4,096 tokens.

Llama 3.1 (A3): Meta AI has made significant contributions to natural language processing with its open-source Llama series. Llama 3.1 [56] extends the context length to 128k tokens and introduces eight language supports. This model is available in 8B, 70B, and 405B parameter sizes and includes both pre-trained and instruction-tuned versions optimized for multilingual dialog applications. Using an improved transformer architecture, Llama 3.1 has undergone supervised fine-tuning and RLHF to better match human preferences.

Claude 3.5 Sonnet (A4): Anthropic launched Claude 3.5 Sonnet [57] as part of their Claude 3 model family, emphasizing improvements in reasoning, accuracy, and multimodal capabilities. With a 200k token context window, Claude 3.5 is optimized for extended conversations and complex reasoning tasks. It incorporates refined training techniques to enhance safety and alignment with human values.

Gemini 1.5 (A5): Google DeepMind introduced Gemini 1.5 [58] as an evolution of their Gemini series, featuring significant advances in context length handling and multimodal processing capabilities. With a context window of up to 1 million tokens, this model is optimized for long-range dependencies and cross-modal understanding. Its advanced architecture enables efficient reasoning across diverse tasks.

Mistral Large 2 (A6): Mistral AI released Mistral Large 2 [59] as their most powerful model, featuring significant improvements in reasoning and multilingual capabilities. With 123B parameters and a 128k context window, Mistral Large 2 incorporates architectural innovations to achieve state-of-the-art results while being notably compute-efficient.

DeepSeek V3 (A7): DeepSeek-AI introduced DeepSeek-V3 [60], which is a 671B-parameter mixture-of-experts (MoE) model with 37B active parameters per token. It features multi-head latent attention (MLA) and DeepSeekMoE for efficiency and cost-effectiveness. Trained on 14.8T tokens with fine-tuning and reinforcement learning, it introduces auxiliary-loss-free load balancing and multi-token prediction. Benchmark results show it outperforms other open-source models and rivals leading closed-source models like GPT-4o and Claude-3.5 Sonnet, all while maintaining cost-effective training.

In addition to considering technical specifications, we selected the alternatives based on current usage trends, community preference rankings, and domain popularity reported in public benchmarks such as the LMSYS Chatbot Arena [61] and the Hugging Face Open LLM Leaderboard [62]. Table 2 summarizes these aspects. Proprietary models such as GPT-4o and GPT-3.5 are the most widely deployed in academia and industry and are consistently used as reference baselines in large-scale trustworthiness studies [26], [28]. Claude 3.5 Sonnet was included for its emphasis on safety and alignment, reflecting Anthropic's constitutional AI framework, which directly connects to our

evaluation criteria. Gemini 1.5 represents a frontier model with ultra-long context handling and advanced multimodal capabilities, making it particularly relevant for assessing robustness in practical scenarios. On the open-source side, Llama 3.1 was selected as the most widely adopted open-source LLM family in both academia and industry, serving as a de facto reference point in benchmarking studies. Mistral Large 2 is recognized for its strong multilingual performance and efficiency, with consistent top rankings on community leaderboards [62]. DeepSeek V3, with its novel mixture-of-experts architecture, has rapidly gained attention in the research community for achieving performance close to proprietary state-of-the-art models. Together, these selections ensure a balanced and scientifically grounded comparison between proprietary and open-source approaches, thereby enhancing the robustness and generalizability of our evaluation.

TABLE 2. Community popularity and usage domains of selected LLMs.

LLM	Community Popularity/Ranking	Typical Usage Domains
GPT-4o	Top-ranked, widely preferred	General-purpose, multimodal, enterprise applications
GPT-3.5	High usage for cost-effective tasks	Chatbots, low-cost NLP tasks
Llama 3.1	Top open-source LLM	Research, open-source applications
Claude 3.5 Sonnet	Rapidly rising in human preference votes	Enterprise solutions, long-context reasoning
Gemini 1.5	Growing adoption	Google services, multimodal applications
Mistral Large 2	Top-ranked open-source	Research, chat, and multilingual tasks
DeepSeek V3	Fast adoption, strong leaderboard entry	Research, developer-focused tools

B. CRITERIA

This research focuses on assessing the trustworthiness of LLMs using five key criteria: Fairness, Robustness, Integrity, Explainability, and Safety. We adapted the criteria proposed by [15] for evaluating the trustworthiness of machine learning datasets and algorithms. These criteria reflect trustworthiness aspects of LLMs highlighted in recent studies [26], [63], emphasizing their relevance and the growing consensus on these standards within the field.

While other dimensions such as privacy, transparency, and accountability are also discussed in the literatures [26] and [30], they were not included as separate criteria in this study. Instead, they are conceptually reflected within the selected FRIES-based dimensions: transparency aligns closely with explainability, accountability relates to integrity and fairness, and privacy concerns are addressed under safety and integrity. Focusing on this established five-criteria set not only follows prior work [15] but also ensures that the evaluation framework remains tractable for expert judgment while still covering the essential aspects of trustworthiness.

Details of these criteria are as follows:

Fairness (C1): This criterion focuses on ensuring that the LLMs operate impartially, promoting equitable outcomes for all users across different scenarios without bias.

Robustness (C2): This criterion evaluates the LLMs' capacity to deliver consistent and reliable performance, even under diverse or adverse conditions.

Integrity (C3): This criterion aims to safeguard LLMs against unauthorized modifications, preserve their reliability, and mitigate tampering risks.

Explainability (C4): This criterion assesses whether the decision-making processes and outputs of the LLMs are transparent and interpretable, allowing users to comprehend and validate the reasoning behind their actions.

Safety (C5): This criterion examines the LLMs' ability to prevent harmful or inappropriate outputs, protect sensitive data, and comply with privacy standards and ethical guidelines.

C. EVALUATION OF LLMS BASED ON TRUSTWORTHINESS

In this study, the expert team consisted of seven professionals with backgrounds in computer science and statistics, ranging in age from their mid-20s to late 50s and representing a mix of early-career researchers and senior academics. Experts were selected based on their academic qualifications, publication records, and professional experience in areas such as LLM development, model alignment, responsible AI, and human impact assessment. To ensure diversity and minimize potential bias, the team included a balanced mix in terms of expertise domains, career stages, and gender (four male and three female experts). Each expert independently evaluated the trustworthiness criteria using a hesitant fuzzy linguistic term set. Table 3 presents their assessments.

After gathering their opinions, the evaluations need to be represented as envelopes. Table 10 presents the enveloped matrices of the expert team's assessments (see Appendix).

After representing the experts' linguistic evaluations as envelopes, the optimistic and pessimistic collective preferences were calculated and are presented in Tables 4 and 5. The calculation details of the optimistic and pessimistic collective preference derivation process are illustrated as follows (optimistic and pessimistic collective preferences for fairness with respect to robustness P^{-}_{c12} and P^{+}_{c12} (4) and (5), as shown at the bottom of the next page.

After obtaining the optimistic and pessimistic collective preferences from the expert evaluations, we calculated the interval utilities to take a step closer to determining the criteria weights. (6) illustrates the computation process for the linguistic intervals.

$$\begin{aligned}&[(((mi,-.14)+(mi,-.14)+(mi,-.29)\\&\quad+(mi,+.14))/4),(((vhi,+.00)+(hi,+.43)\\&\quad+(hi,+.43)+(vhi,+.00))/4)]\\&\quad[(\text{mi},-.11),(\text{vhi},-.29)]\end{aligned}\tag{6}$$

We calculated all linguistic intervals using the same method and presented them in Table 6.

The next step involves expressing linguistic intervals as interval utilities. Then, the midpoints of these interval utilities need to be obtained. Finally, the weights can be calculated by normalizing the midpoints. The results of the criteria weighting phase in this study show that the experts ranked fairness as the most important criterion, while they considered explainability the least important.

The second step focuses on evaluating alternatives based on the determined set of criteria. To identify the best alternative based on the trust-based criteria set, we consulted the same expert group. In this phase, the experts used a basic scoring scale to evaluate the alternatives. This scale represents a rating system used to evaluate alternatives based on specific criteria. The evaluation is conducted on a scale ranging from 1 to 10, with each value representing a certain level of a particular attribute. The scale begins with ''Extremely Low'' and ends with ''Maximum.'' This range is used to define the strength, impact, or importance of each criterion. Each level of the scale allows experts to convert their opinions into numerical values, enabling a clearer and more objective analysis when comparing alternatives. Table 11 presents the experts' evaluations of the alternatives according to the criteria set (see Appendix).

We aggregated the experts' evaluations using the arithmetic mean. Table 7 presents the collective evaluation matrix for the alternatives.

The final step is calculating the weighted total scores of the alternatives. We applied a weighting method, where the criteria weights were multiplied by the corresponding alternative's score. After weighting the scores, we summed them to obtain the final score for each alternative (see Table 8).

According to the results of the second phase, GPT-4o (A1) is the leading LLM with the highest weighted trustworthiness score of 6.77. GPT-3.5 (A2) maintains a strong secondary position with a score of 5.82, though a notable gap exists between it and GPT-4o. These results confirm that GPT-4o surpasses its predecessor, validating the assumption that newer iterations enhance trustworthiness. Claude 3.5 Sonnet (A4) and Gemini 1.5 (A5) demonstrate comparable performance with scores of 5.30 and 5.63, respectively, yet both remain behind the top closed-source models.

In the open-source category, Llama 3.1 (A3) emerges as the most trustworthy with a score of 5.38. By contrast, Mistral Large 2 (A6) and DeepSeek V3 (A7) received the lowest scores (4.94 and 4.99, respectively), revealing significant weaknesses, particularly regarding safety measures. These findings highlight a persistent trustworthiness gap between proprietary and open-source LLMs. Despite progress in the open-source domain, substantial improvements in trustworthiness criteria are still required to match standards established by leading proprietary models.

Across all evaluated models, safety scores were particularly concerning. Current LLMs continue to struggle with mitigating risks, preventing harmful outputs, and ensuring responsible deployment. DeepSeek V3's notably low safety score (3.00) raises significant concerns about its suitability for high-stakes applications. The universal challenge of explainability is evident, with scores ranging from 4.29 to 5.29 across all models. These scores suggest that LLMs still struggle to provide transparent reasoning and interpretable outputs, which are essential for fostering user trust. This highlights an industry-wide issue that requires further innovation. Additionally, the large variation in robustness scores (5.57–8.14) suggests that performance consistency under diverse conditions remains a key differentiator in real-world applications.

The responses of the alternatives to changes in criterion weights were analyzed under 11 different scenarios. The weights obtained from expert evaluations, along with the resulting values, are presented in Table 9 as Hesitant Fuzzy-AHP weights. From this point onward, the reported values reflect the scenario results, showing the scores of the alternatives based on the dominant criterion.

When the weight of the dominant criterion was set to 0.99 or 0.90, the weights of the remaining criteria were calculated by subtracting this value from 1 and distributing the remainder equally among them. This allowed for a complete dominance assessment. A weight of 0.90 represented a case where dominance was reduced by approximately 10%. To completely remove the influence of dominance, an additional scenario was applied in which all criteria were assigned equal weights.

Fig. 3 presents the alternative rankings across different scenarios. GPT-4o (A1) consistently emerged as the best-performing alternative in all cases, representing the most robust solution to the decision problem. GPT-3.5 (A2) and Gemini 1.5 (A5) closely followed GPT-4o (A1) under various criterion scenarios, indicating that they can be considered strong alternatives. In contrast, Mistral Large 2 (A6) was generally ranked among the lowest across all scenarios.

$$\begin{aligned} P^{-}_{C12} &= \Delta\left(\frac{1}{7}\left(\Delta^{-1}(mi,3)+\Delta^{-1}(hi,4)+\Delta^{-1}(ni,0)+\Delta^{-1}(li,2)+\Delta^{-1}(li,2)+\Delta^{-1}(hi,4)+\Delta^{-1}(vhi,5)\right)\right) \\ &= \Delta\left(\frac{1}{7}(3+4+2+2+4+5)\right) = \Delta(2.86) = (mi,-.14) \end{aligned} \tag{4}$$

$$\begin{aligned} P^{+}_{C12} &= \Delta\left(\frac{1}{7}\left(\Delta^{-1}(ai,6)+\Delta^{-1}(ai,6)+\Delta^{-1}(mi,3)+\Delta^{-1}(hi,4)+\Delta^{-1}(hi,4)+\Delta^{-1}(ai,6)+\Delta^{-1}(ai,6)\right)\right) \\ &= \Delta\left(\frac{1}{7}(6+6+3+4+4+6+6)\right) = \Delta(5.00) = (vhi,+.00) \end{aligned} \tag{5}$$

TABLE 3. Expert evaluations of criteria and their enveloped matrix forms.

E-1	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Greater than li	Between li and mi	Greater than mi	Greater than mi
Robustness		-	At least vhi	Greater than mi	Between mi and hi
Integrity			-	Greater than mi	At most hi
Explainability				-	Lower than hi
Safety					-

E-2	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Greater than mi	Lower than hi	Greater than mi	Between li and mi
Robustness		-	At most mi	At least mi	At most mi
Integrity			-	Lower than mi	At least mi
Explainability				-	At most mi
Safety					-

E-3	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Lower than hi	Between mi and hi	At least vhi	Greater than hi
Robustness		-	At least vhi	Greater than hi	Between hi and ai
Integrity			-	Greater than mi	Between mi and hi
Explainability				-	Lower than mi
Safety					-

E-4	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Between li and hi	Between vli and mi	mi	Greater than li
Robustness		-	At least hi	At least hi	At least hi
Integrity			-	At least hi	Greater than li
Explainability				-	mi
Safety					-

E-5	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Between li and hi	At least vhi	At most hi	Between hi and ai
Robustness		-	At most hi	Between mi and hi	Greater than mi
Integrity			-	Lower than mi	Between hi and ai
Explainability				-	At least mi
Safety					-

E-6	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	Greater than mi	At least hi	Between mi and hi	Greater than mi
Robustness		-	Between li and mi	At most mi	Lower than hi
Integrity			-	greater than li	At least vhi
Explainability				-	Lower than hi
Safety					-

E-7	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	At least vhi	At least vhi	Lower than mi	At most li
Robustness		-	Greater than mi	Lower than mi	At most li
Integrity			-	Lower than mi	At most li
Explainability				-	Lower than mi
Safety					-

TABLE 4. Pessimistic collective preferences.

	C1	C2	C3	C4	C5
C1	-	*mi*, -.14	*mi*, -.14	*mi*, -.29	*mi*, +.14
C2	*vli*, +.00	-	*mi*, -.14	*mi*, -.29	*li*, +.14
C3	*li*, -.43	*vli*, +.14	-	*li*, +.14	*mi*, -.43
C4	*li*, -.43	*vli*, +.29	*li*, -.29	-	*vli*, -.14
C5	*vli*, +.00	*li*, -.29	*vli*, +.14	*mi*, -.14	--

TABLE 5. Optimistic collective preferences.

	C1	C2	C3	C4	C5
C1	-	*vhi*, +.00	*hi*, +.43	*hi*, +.43	*vhi*, +.00
C2	*mi*, +.14	-	*vhi*, -.14	*vhi*, -.29	*hi*, +.29
C3	*mi*, +.14	*mi*, +.14	-	*hi*, +.29	*vhi*, -.14
C4	*mi*, +.29	*mi*, +.29	*hi*, -.14	-	*mi*, +.14
C5	*mi*, -.14	*hi*, -.14	*mi*, +.43	*vhi*, +.14	-

TABLE 6. Calculated linguistic intervals, interval utilities, mid-points, and weights.

	Linguistic intervals	Interval utilities		Mid points	Weights
C1	[(*mi*, - .11), (*vhi*, - .29)]	2.89	4.71	3.80	0.25
C2	[(*li*, + .18), (*hi*, + .25)]	2.18	4.25	3.21	0.21
C3	[(*li*, - .14), (*hi*, - .14)]	1.86	3.86	2.86	0.19
C4	[*vli*, + .36), (*mi*, + .39)]	1.36	3.39	2.38	0.16
C5	[(*li*, - .32), (*hi*, - .18)]	1.68	3.82	2.75	0.18

V. DISCUSSION

This study presents a novel trust-based evaluation of LLMs by leveraging an MCDM methodology, specifically a hesitant fuzzy extension of the AHP. The proposed approach addresses the key limitations found in existing trust evaluation frameworks, such as the inability to systematically incorporate expert uncertainty and the lack of nuanced weighting for trust dimensions. A key strength of the proposed approach is its ability to consider multiple trust-related criteria simultaneously and consolidate them into a single overall score, allowing for a clear and interpretable ranking of alternatives. This method not only reflects the relative importance of each criterion but also incorporates the varying expertise levels of the evaluators, ensuring that more informed judgments carry greater weight in the decision-making process. By integrating hesitant fuzzy linguistic term sets, the methodology accommodates expert hesitation and ambiguity in pairwise comparisons, offering a more human-aligned and flexible evaluation framework.

In this study, the use of the same expert group in both stages was preferred due to the specific field of expertise the study focused on and the subject-specific knowledge required. These experts consist of individuals with the technical knowledge and contextual expertise necessary for the evaluation process. However, various measures were taken to minimize the potential risks of subjectivity and bias. First, the evaluation metrics were defined in a clear and measurable manner, and the evaluation process was standardized. Additionally, experts were asked to make independent and individual assessments, and joint decision-making processes were deliberately avoided. The experts' independent evaluation results were then aggregated using arithmetic averages to reduce the effect of individual differences. The highly specific nature of the subject under study makes it challenging to find alternative experts in this field, which further justifies the use of the same expert group for the evaluations.

The results from the criteria weighting phase reveal that the experts prioritized fairness the most, while they placed the least emphasis on explainability. This ranking underscores an increasing awareness of the societal implications of LLM behavior, especially in high-stakes applications. Fairness is emerging as a central concern, aligning with broader discourses on AI ethics, bias mitigation, and equitable access. This prioritization of fairness aligns with recent literature emphasizing the growing concern over biased outcomes in AI systems, especially LLMs. Several studies have documented that unfairness in LLM outputs can propagate discrimination and harm marginalized groups, particularly in domains like healthcare, law, and education [11], [26]. Moreover, large-scale benchmarks such as TRUSTLLM [26] and DecodingTrust [28], XTRUST [30] highlight fairness as one of the most critical and frequently evaluated dimensions in trustworthiness assessments of LLMs. Additionally, user-centered studies like [39] empirically identified perceived credibility, neutrality, and bias minimization as key determinants of trust in LLMs. During our expert evaluation process, multiple experts explicitly indicated that fairness was their top priority due to its direct societal impact and its role in ensuring ethical AI deployment. Therefore, the higher weight assigned to fairness reflects both empirical evidence from the literature and the subjective importance perceived by domain experts in the context of LLM deployment in high-stakes scenarios.

On the other hand, the relatively lower weight given to explainability may reflect either a perceived maturity of this dimension or, conversely, an ongoing challenge in making LLMs truly interpretable, which might discourage experts from expecting high standards in this area. This finding is also consistent with prior research that highlights the ongoing challenges in operationalizing explainability for LLMs. Studies [26], [27] report that while explainability is recognized as an important trust dimension, it often receives less emphasis in practical evaluations due to the lack of standardized metrics and the difficulty in producing human-interpretable explanations from large-scale neural models.

TABLE 7. Collective evaluation matrix for alternatives.

	A1	A2	A3	A4	A5	A6	A7
Fairness	7.14	5.71	5.43	5.71	6.29	5.43	5.00
Robustness	8.14	7.57	6.29	5.86	6.00	5.57	7.00
Integrity	6.71	5.57	5.57	5.57	5.57	5.29	5.00
Explainability	5.29	4.86	4.29	4.43	4.57	4.29	4.57
Safety	6.00	5.00	5.00	4.57	5.29	3.71	3.00

TABLE 8. Alternative scores (collective and weighted collective).

Criteria and their weights	Collective scores of A1	The weighted scores of A1	Collective scores of A2	The weighted scores of A2	Collective scores of A3	The weighted scores of A3	Collective scores of A4	The weighted scores of A4	Collective scores of A5	The weighted scores of A5	Collective scores of A6	The weighted scores of A6	Collective scores of A7	The weighted scores of A7
Fairness (0.25)	7.14	1.81	5.71	1.45	5.43	1.38	5.71	1.45	6.29	1.59	5.43	1.38	5.00	1.27
Robustness (0.21)	8.14	1.74	7.57	1.62	6.29	1.35	5.86	1.26	6.00	1.29	5.57	1.19	7.00	1.50
Integrity (0.19)	6.71	1.28	5.57	1.06	5.57	1.06	5.57	1.06	5.57	1.06	5.29	1.01	5.00	0.95
Explainability (0.16)	5.29	0.84	4.86	0.77	4.29	0.68	4.43	0.70	4.57	0.72	4.29	0.68	4.57	0.72
Safety (0.18)	6.00	1.10	5.00	0.92	5.00	0.92	4.57	0.84	5.29	0.97	3.71	0.68	3.00	0.55
The sum of weighted scores of alternatives:		**6.77**		**5.82**		**5.38**		**5.30**		**5.63**		**4.94**		**4.99**

Note: GPT-4o (A1), GPT-3.5 (A2), Llama 3.1 (A3), Claude 3.5 Sonnet (A4), Gemini 1.5 (A5), Mistral Large 2 (A6), and DeepSeek V3 (A7).

TABLE 9. Sensitivity analysis.

			The weights of the dominant criterion (DC): 0.99					The weights of the dominant criterion (DC): 0.90				
Alternatives	**Hesitant Fuzzy-AHP Weights**	**Equal**	**DC: C1**	**DC: C2**	**DC: C3**	**DC: C4**	**DC: C5**	**DC: C1**	**DC: C2**	**DC: C3**	**DC: C4**	**DC: C5**
A1	6.77	6.66	7.14	8.12	6.71	5.30	6.01	7.08	7.96	6.71	5.46	6.08
A2	5.82	5.74	5.71	7.55	5.57	4.87	5.01	5.72	7.34	5.59	4.97	5.09
A3	5.38	5.31	5.43	6.27	5.57	4.30	5.00	5.41	6.16	5.54	4.41	5.04
A4	5.30	5.23	5.71	5.85	5.57	4.44	4.58	5.65	5.78	5.53	4.53	4.65
A5	5.63	5.54	6.28	5.99	5.57	4.58	5.29	6.19	5.94	5.57	4.69	5.32
A6	4.94	4.86	5.42	5.56	5.28	4.29	3.73	5.36	5.48	5.23	4.36	3.86
A7	4.99	4.91	5.00	6.97	5.00	4.58	3.02	4.99	6.74	4.99	4.61	3.24

Note: GPT-4o (A1), GPT-3.5 (A2), Llama 3.1 (A3), Claude 3.5 Sonnet (A4), Gemini 1.5 (A5), Mistral Large 2 (A6), and DeepSeek V3 (A7). Fairness(C1), Robustness(C2), Integrity (C3), Explainability (C4), Safety (C5).

In the second phase of the analysis, the application of the weighted scoring method enabled a comparative evaluation of prominent LLMs, including both proprietary and open-source models. GPT-4o ranked highest in terms of overall trustworthiness, followed by GPT-3.5. This finding supports the notion that newer iterations of models tend to show incremental improvements in alignment, robustness, and ethical behavior. The performance gap between GPT-4o and earlier versions is also consistent with technical reports highlighting advancements in architecture, fine-tuning, and training data coverage.

Open-source alternatives such as Llama 3.1 and Mistral Large 2 displayed relatively weaker performance, particularly in the safety dimension. This gap may stem from resource constraints in the open-source development pipeline, including limitations in data filtering, alignment tuning, and post-training evaluation. Notably, DeepSeek V3, despite its sophisticated architecture and high parameter count, received the lowest trustworthiness score among the evaluated models. This outcome suggests that technical complexity alone is insufficient for fostering trust; rather, comprehensive safeguards and maturity in deployment processes play a crucial role.

The observed performance gap between proprietary and open-source models can be better understood in light of their underlying training methodologies and deployment strategies. Proprietary models such as GPT-4o, Claude 3.5, and Gemini 1.5 are typically developed with access to significantly larger and more diverse training corpora, extensive compute resources, and rigorous post-training alignment pipelines (e.g., multi-stage RLHF, large-scale human preference data, and continuous deployment feedback loops). These investments enable stronger safeguards in terms of safety, robustness, and fairness. In contrast, open-source models like Llama 3.1 and Mistral Large 2, while technically competitive in architecture, are often constrained by more limited compute budgets and smaller-scale or less diverse alignment data, which can impact their safety and explainability scores. Similarly, DeepSeek V3, despite its large parameter count and innovative mixture-of-experts design, highlights

TABLE 10. Enveloped matrices.

E-1	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[mi, ai]	[li, mi]	[hi, ai]	[hi, ai]
Robustness	[ni, mi]	-	[vhi, ai]	[hi, ai]	[mi, hi]
Integrity	[mi, hi]	[ni, vli]	-	[hi, ai]	[ni, hi]
Explainability	[ni, li]	[ni, li]	[ni, li]	-	[ni, mi]
Safety	[ni, li]	[li, mi]	[li, ai]	[mi, ai]	-

E-2	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[hi, ai]	[ni, mi]	[hi, ai]	[li, mi]
Robustness	[ni, li]	-	[ni, mi]	[mi, ai]	[ni, mi]
Integrity	[mi, ai]	[mi, ai]	-	[ni, li]	[mi, ai]
Explainability	[ni, li]	[ni, mi]	[hi, ai]	-	[ni, mi]
Safety	[mi, hi]	[mi, ai]	[ni, mi]	[mi, ai]	-

E-3	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[ni, mi]	[mi, hi]	[vhi, ai]	[vhi, ai]
Robustness	[mi, ai]	-	[vhi, ai]	[vhi, ai]	[hi, ai]
Integrity	[li, mi]	[ni, vli]	-	[hi, ai]	[mi, hi]
Explainability	[ni, vli]	[ni, vli]	[ni, li]	-	[ni, li]
Safety	[ni, vli]	[ni, li]	[li, mi]	[hi, ai]	-

E-4	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[li, hi]	[vli, mi]	[mi, mi]	[mi, ai]
Robustness	[li, hi]	-	[hi, ai]	[hi, ai]	[hi, ai]
Integrity	[mi, vhi]	[ni, li]	-	[hi, ai]	[mi, ai]
Explainability	[mi, mi]	[ni, li]	[ni, li]	-	[mi, mi]
Safety	[ni, mi]	[ni, li]	[ni, mi]	[mi, mi]	-

E-5	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[li, hi]	[vhi, ai]	[ni, hi]	[hi, ai]
Robustness	[li, hi]	-	[ni, hi]	[mi, hi]	[hi, ai]
Integrity	[ni, vli]	[li, ai]	-	[ni, li]	[hi, ai]
Explainability	[li, ai]	[li, mi]	[hi, ai]	-	[mi, ai]
Safety	[ni, li]	[ni, li]	[ni, li]	[ni, mi]	-

E-6	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[hi, ai]	[hi, ai]	[mi, hi]	[hi, ai]
Robustness	[ni, li]	-	[li, mi]	[ni, mi]	[ni, mi]
Integrity	[ni, li]	[mi, hi]	-	[mi, ai]	[vhi, ai]
Explainability	[li, mi]	[mi, ai]	[ni, mi]	-	[ni, mi]
Safety	[ni, li]	[mi, ai]	[ni, vli]	[mi, ai]	-

E-7	Fairness	Robustness	Integrity	Explainability	Safety
Fairness	-	[vhi, ai]	[vhi, ai]	[ni, li]	[ni, li]

TABLE 10. *(Continued.)* **Enveloped matrices.**

Robustness	[ni, vli]	-	[hi, ai]	[ni, li]	[ni, li]
Integrity	[ni, vli]	[ni, li]	-	[ni, li]	[ni, li]
Explainability	[hi, ai]	[hi, ai]	[hi, ai]	-	[ni, li]
Safety	[hi, ai]	[hi, ai]	[hi, ai]	[hi, ai]	-

that architectural sophistication alone is insufficient without equally mature training and alignment practices. These differences suggest that resource allocation, alignment strategies, and deployment maturity are key drivers behind the trustworthiness disparities observed between proprietary and open-source LLMs.

A critical insight from the study is the persistently low scores across all models in the safety and explainability dimensions. This pattern signals an industry-wide challenge that has yet to be fully addressed. Ensuring that LLMs produce non-harmful, transparent, and controllable outputs remains one of the most pressing obstacles in trustworthy AI development. Moreover, the broad range of robustness scores among the models indicates that performance consistency under diverse or adverse conditions continues to be a key differentiator in practical settings. While leading models already employ advanced alignment techniques such as RLHF, guardrails, and large-scale filtering, our findings indicate that these measures remain insufficient to fully address the challenges of safety and explainability. To strengthen safety, future LLM development should integrate domain-specific guardrails, dynamic risk monitoring, and continuous post-deployment auditing mechanisms to capture emerging threats that static training cannot anticipate. For explainability, promising directions include interpretable reasoning traces, attribution-based explanations, and user-facing transparency tools such as model cards and visualization dashboards. These complementary strategies, combined with existing alignment methods, can help close the critical gaps identified in this study and foster greater trust in deploying LLMs within high-stakes domains.

Overall, the findings of this study demonstrate that trust in LLMs is not determined by a single metric but rather emerges from the interplay of multiple interdependent criteria. The MCDM-based framework adopted here offers a structured and transparent means of navigating these complexities. It enables informed selection among alternatives and supports the development of models that align with both technical excellence and societal values. As LLMs become increasingly integrated into critical decision-making contexts, such frameworks will be essential for guiding responsible and trustworthy deployment.

VI. LIMITATIONS AND FUTURE STUDIES

While this study provides a structured expert-based evaluation of LLM trustworthiness, some limitations should be acknowledged, which also open promising avenues for future research.

The method used in this study is an MCDM technique that has been validated in the literature and successfully applied in various fields. Therefore, our statement that the method is generally applicable to different decision problems is based on its structural flexibility and the diversity of its applications in previous studies. However, in this study, the method was applied to a single, specific decision problem to enable an in-depth analysis of the problem's unique complexity and its application context. The multi-criteria nature of the problem, the use of expert opinion-based data, and its grounding in a real-world scenario demonstrate that the method is an effective tool for such decision-making contexts. Although the method has been widely used in the literature to solve diverse problems and is therefore considered applicable across different domains, in this study it was applied to a single scenario due to scope limitations.

The evaluation was restricted to seven LLMs. While these models were selected to represent both leading proprietary and competitive open-source alternatives, the limited scope may reduce the generalizability of the findings. Future studies could address this limitation by expanding the set of evaluated LLMs to include a broader and more diverse selection, thereby improving representativeness and strengthening the robustness of the results.

The study adopted only the five FRIES-based dimensions (fairness, robustness, integrity, explainability, and safety). While these criteria represent a widely recognized and frequently applied set in the literature, other relevant dimensions, such as privacy, transparency, and accountability, were not explicitly included. Future studies could extend the framework by incorporating additional criteria such as privacy, transparency, and accountability, enabling more comprehensive and nuanced evaluations.

The framework was applied to a general trustworthiness evaluation of LLMs, without tailoring criteria or weights to domain-specific contexts, which may reduce its applicability in highly specialized areas such as healthcare or law. Future studies may adapt the framework to domain-specific applications, where the relative importance of trust dimensions varies, thereby increasing contextual relevance and flexibility.

The analysis relied on a single expert group. Although this was justified by the area-specific expertise required, it may not fully capture the diversity of perspectives across different application contexts. Future studies could involve

TABLE 11. Experts' evaluations of alternatives based on criteria.

E-1	ALTERNATIVES						
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	High	Medium	Medium	Slightly High	Slightly High	Medium	Medium
Robustness	Very High	Very High	Slightly High	High	High	Slightly High	Very High
Integrity	Slightly High	Medium	Medium	Slightly High	Slightly High	Medium	Slightly High
Explainability	Slightly Low	Slightly Low	Very Low	Very Low	Very Low	Slightly Low	Slightly Low
Safety	Slightly High	Medium	Medium	Low	Low	Slightly Low	Very Low

E-2	ALTERNATIVES						
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	High	Slightly High	Slightly High	Medium	High	Medium	Slightly High
Robustness	High	Slightly High	Slightly High	Medium	Slightly High	Medium	Slightly High
Integrity	Very High	High	High	Slightly High	High	Slightly High	High
Explainability	Extremely Low	Extremely Low	Slightly Low	Slightly Low	Extremely Low	Slightly Low	Very Low
Safety	Very High	Slightly High	Medium	Medium	High	Low	Slightly Low

E-3	ALTERNATIVES						
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	High	Slightly High	Medium	Slightly High	High	Slightly High	Medium
Robustness	Extremely High	Extremely High	High	High	Medium	Medium	Very High
Integrity	High	Slightly High	Slightly High	Slightly High	Medium	Slightly High	Medium
Explainability	High	Slightly High	Slightly High	Slightly High	Medium	Slightly High	Medium

TABLE 11. *(Continued.)* **Experts' evaluations of alternatives based on criteria.**

Safety	Medium	Medium	Low	Low	Medium	Slightly Low	Very Low
E-4				**ALTERNATIVES**			
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	Slightly High	Medium	Low	Slightly Low	Very Low	Slightly Low	Very Low
Robustness	Very High	High	Slightly High	Low	Medium	Slightly High	Slightly High
Integrity	Low	Slightly Low	Medium	Low	Slightly Low	Medium	Low
Explainability	Medium	Medium	Medium	Slightly Low	Low	Medium	Medium
Safety	Low	Slightly Low	Very Low	Very Low	Very Low	Very Low	Very Low
E-5				**ALTERNATIVES**			
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	Extremely High	High	Slightly High	Very High	Extremely High	High	Very High
Robustness	Very High	High	Slightly High	High	Very High	Slightly High	Very High
Integrity	Extremely High	Slightly High	Medium	High	High	Slightly High	Slightly High
Explainability	Very High	High	Slightly High	Very High	Extremely High	Medium	High
Safety	Very High	Slightly High	High	Very High	Slightly High	High	Medium
E-6				**ALTERNATIVES**			
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	High	Slightly High	High	Slightly High	High	High	Medium
Robustness	Extremely High	Very High	High	Medium	Slightly High	Medium	Very High
Integrity	High	Slightly High	Slightly High	Medium	Slightly High	Medium	Low
Explainability	Low	Low	Slightly Low	Low	Low	Slightly Low	Low

TABLE 11. *(Continued.)* **Experts' evaluations of alternatives based on criteria.**

Safety	High	Slightly High	Slightly High	Medium	High	Slightly Low	Low

E-7	**ALTERNATIVES**						
CRITERIA	A1	A2	A3	A4	A5	A6	A7
Fairness	High	Medium	Medium	Slightly High	Slightly High	Medium	Low
Robustness	Very High	Very High	Slightly High	Slightly High	Medium	Slightly High	Medium
Integrity	Slightly High	Slightly High	Medium	Medium	Medium	Low	Slightly Low
Explainability	Extremely High	Very High	Medium	Medium	High	Medium	Slightly High
Safety	Low	Low	Slightly High	Low	Slightly High	Low	Slightly Low

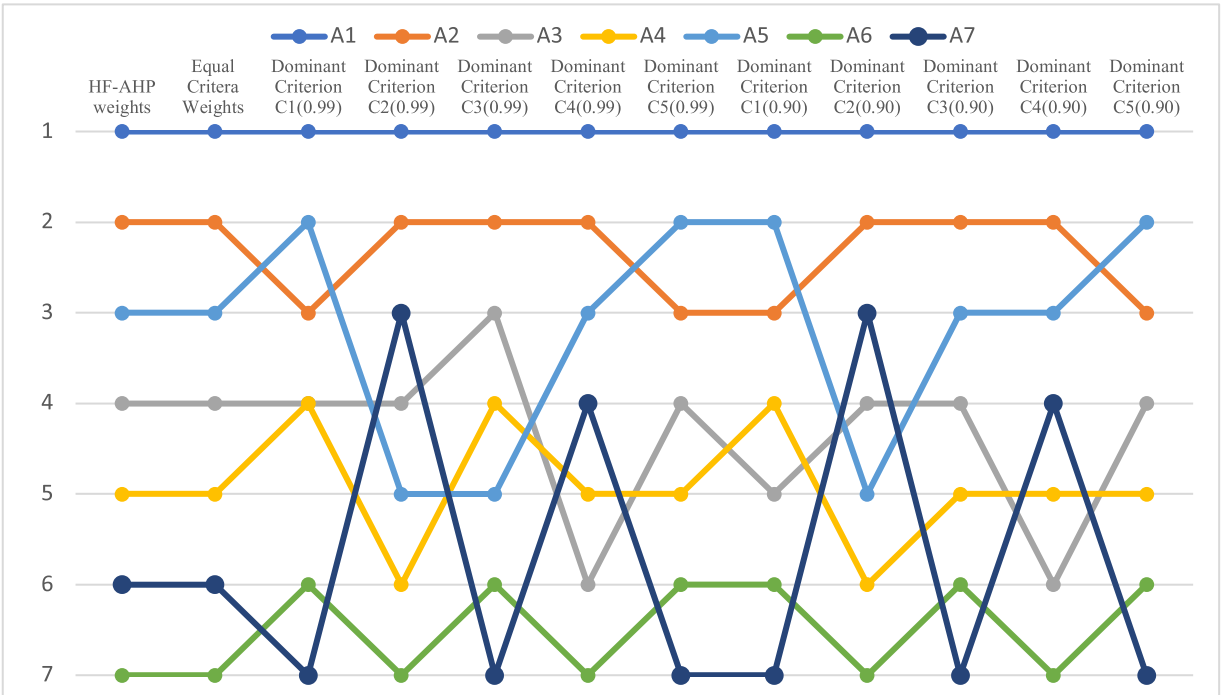


FIGURE 3. Illustration of alternative rankings across various scenarios.

additional experts from diverse professional and regional backgrounds to provide more heterogeneous perspectives and reduce potential biases. Thus, the evaluation was conducted by experts with the necessary technical and contextual knowledge. To minimize subjectivity and bias, the evaluation metrics were clearly defined, the process was standardized, and experts provided independent assessments without group influence. Their results were aggregated using arithmetic

averages to reduce individual differences. For future work, this approach could be further strengthened by reporting inter-rater reliability measures.

The evaluation primarily relied on expert-based judgments, without integrating objective task-based metrics or user-centered feedback, potentially limiting the comprehensiveness of the assessment. Future studies could develop hybrid approaches that combine expert judgments with task-based performance metrics and user feedback, leading to more comprehensive and scalable trustworthiness assessments.

VII. CONCLUSION

The rapid advancement of artificial intelligence is driving the widespread adoption of LLMs. This ongoing trend has enabled LLMs to secure a strong foothold not only in professional business environments but also in everyday life. A variety of alternatives with diverse technical foundations have already been developed and continue to emerge. At this point, it is important to consider how the relative superiority of these models can be effectively evaluated. While such evaluations can certainly be based on various parameters, such as performance or cost, this study focuses primarily on assessing these models through the lens of user trust. Although there is a substantial and growing body of literature addressing trust in LLMs, no prior research has employed MCDM techniques to evaluate these models specifically from the perspective of trustworthiness.

Furthermore, to the best of our knowledge, the integration of a technologically grounded method-utilizing hesitant fuzzy linguistic term sets based on word calculation-for determining the importance of trust-related evaluation criteria, particularly as judged by multiple decision-makers, has yet to be explored in the existing literature. In this study, we designed a two-stage evaluation process to assess various LLM alternatives with respect to perceived trust. As additional expert opinions were incorporated into the process, GPT-4o emerged as the consistently most prominent and trusted alternative.

Notably, despite its rapid emergence and significant recent impact, DeepSeek ranked second lowest among the evaluated alternatives in terms of trust. This outcome may suggest that cultivating user trust in an LLM is influenced not only by technical performance but also by the model's maturity and the amount of time it has spent within the product life cycle. This study highlights the growing importance of trust in the evaluation of LLMs, especially as their presence expands across various domains. The insights of this paper provide a foundation for future studies aiming to design more reliable and user-centric AI systems.

APPENDIX
SUPPLEMENTARY EVALUATION TABLES

See Tables 10 and 11.

REFERENCES

[1] D. Yuan, E. Rastogi, G. Naik, S. P. Rajagopal, S. Goyal, F. Zhao, B. Chintagunta, and J. Ward, "A continued pretrained LLM approach for automatic medical note generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Jun. 2024, pp. 565–571, doi: 10.18653/v1/2024.naacl-short.47.

[2] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individual Differences*, vol. 103, Apr. 2023, Art. no. 102274, doi: 10.1016/j.lindif.2023.102274.

[3] I. Cheong, K. Xia, K. J. K. Feng, Q. Z. Chen, and A. X. Zhang, "(A)I am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2024, pp. 2454–2469, doi: 10.1145/3630106.3659048.

[4] M. A. Haque, "LLMs: A game-changer for software engineers?" *BenchCouncil Trans. Benchmarks, Standards Evaluations*, vol. 5, no. 1, Mar. 2025, Art. no. 100204, doi: 10.1016/j.tbench.2025.100204.

[5] A. Roush, E. Zakirov, A. Shirokov, P. Lunina, J. Gane, A. Duffy, C. Basil, A. Whitcomb, J. Benedetto, and C. DeWolfe, "LLM as an art director (LaDi): Using LLMs to improve text-to-media generators," 2023, *arXiv:2311.03716*.

[6] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, Supryadi, L. Yu, Y. Liu, J. Li, B. Xiong, and D. Xiong, "Evaluating large language models: A comprehensive survey," 2023, *arXiv:2310.19736*.

[7] R. Doshi and A. Kaleel, "Bibliometric analysis of generative AI and large language models in the scopus database: Trends, insights, and research landscape," *Appl. Data Sci. Anal.*, vol. 2025, pp. 7–18, Mar. 2025, doi: 10.58496/adsa/2025/003.

[8] M. Aksoy and A. Bush, "A bibliometric analysis of trust in conversational agents over the past fifteen years," in *Proc. 11th Int. Conf. Comput. Artif. Intell. (ICCAI)*, Kyoto, Japan, Mar. 2025, pp. 451–464, doi: 10.1109/ICCAI66501.2025.00076.

[9] L. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022, *arXiv:2203.02155*.

[11] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu, "Bias in large language models: Origin, evaluation, and mitigation," 2024, *arXiv:2411.10915*.

[12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surveys*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[13] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: Opportunities and challenges," *AI Mag.*, vol. 45, no. 3, pp. 354–368, Sep. 2024, doi: 10.1002/aaai.12188.

[14] N. Pantha, M. Ramasubramanian, I. Gurung, M. Maskey, and R. Ramachandran, "Challenges in guardrailing large language models for science," 2024, *arXiv:2411.08181*.

[15] J. Rutinowski, S. Klüttermann, J. Endendyk, C. Reining, and E. Müller, "Benchmarking trust: A metric for trustworthy machine learning," *Commun. Comput. Inf. Sci.*, vol. 2153, pp. 287–307, May 2024, doi: 10.1007/978-3-031-63787-2_15.

[16] E. K. Zavadskas, Z. Turskis, and S. Kildienė, "State of art surveys of overviews on MCDM/MADM methods," *Technological Econ. Develop. Economy*, vol. 20, no. 1, pp. 165–179, Mar. 2014, doi: 10.3846/20294913.2014.892037.

[17] A. Adem, A. Çolak, and M. Dağdeviren, "An integrated model using SWOT analysis and hesitant fuzzy linguistic term set for evaluation occupational safety risks in life cycle of wind turbine," *Saf. Sci.*, vol. 106, pp. 184–190, Jul. 2018, doi: 10.1016/j.ssci.2018.02.033.

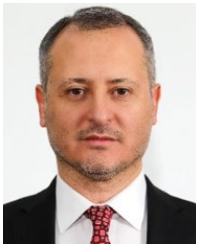
[18] S. K. Sahoo and S. S. Goswami, "A comprehensive review of multiple criteria decision-making (MCDM) methods: Advancements, applications, and future directions," *Decis. Making Adv.*, vol. 1, no. 1, pp. 25–48, Dec. 2023, doi: 10.31181/dma1120237.

[19] W. Ho, "Integrated analytic hierarchy process and its applications—A literature review," *Eur. J. Oper. Res.*, vol. 186, no. 1, pp. 211–228, Apr. 2008, doi: 10.1016/j.ejor.2007.01.004.

[20] M. R. Asadabadi, E. Chang, and M. Saberi, "Are MCDM methods useful? A critical review of analytic hierarchy process (AHP) and analytic network process (ANP)," *Cogent Eng.*, vol. 6, no. 1, Jan. 2019, Art. no. 1623153, doi: 10.1080/23311916.2019.1623153.
[21] N. Subramanian and R. Ramanathan, "A review of applications of analytic hierarchy process in operations management," *Int. J. Prod. Econ.*, vol. 138, no. 2, pp. 215–241, Aug. 2012, doi: 10.1016/j.ijpe.2012.03.036.
[22] M. Yavuz, B. Oztaysi, S. C. Onar, and C. Kahraman, "Multi-criteria evaluation of alternative-fuel vehicles via a hierarchical hesitant fuzzy linguistic model," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2835–2848, Apr. 2015, doi: 10.1016/j.eswa.2014.11.010.
[23] C. Acar, A. Beskese, and G. T. Temur, "Sustainability analysis of different hydrogen production options using hesitant fuzzy AHP," *Int. J. Hydrogen Energy*, vol. 43, no. 39, pp. 18059–18076, Sep. 2018, doi: 10.1016/j.ijhydene.2018.08.024.
[24] A. Beskese, A. Camci, G. T. Temur, and E. Erturk, "Wind turbine evaluation using the hesitant fuzzy AHP-TOPSIS method with a case in Turkey," *J. Intell. Fuzzy Syst.*, vol. 38, no. 1, pp. 997–1011, Jan. 2020, doi: 10.3233/jifs-179464.
[25] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.
[26] L. Sun et al., "TrustLLM: Trustworthiness in large language models," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024, p. 105. [Online]. Available: https://dl.acm.org/doi/10.5555/3692070.3692883
[27] P. Liang et al., "Holistic evaluation of language models," 2022, *arXiv:2211.09110*.
[28] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li, "DecodingTrust: A comprehensive assessment of trustworthiness in GPT models," 2023, *arXiv:2306.11698*.
[29] L. Mo, B. Wang, M. Chen, and H. Sun, "How trustworthy are open-source LLMs? An assessment under malicious demonstrations shows their vulnerabilities," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Jun. 2024, pp. 2775–2792, doi: 10.18653/v1/2024.naacl-long.152.
[30] Y. Li, Y. Wang, Y. Chang, and Y. Wu, "XTRUST: On the multilingual trustworthiness of large language models," 2024, *arXiv:2409.15762*.
[31] M. Song, S. H. Sim, R. Bhardwaj, H. L. Chieu, N. Majumder, and S. Poria, "Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025, pp. 1–11.
[32] A. H. Alamoodi, O. Zughoul, D. David, S. Garfan, D. Pamucar, O. S. Albahri, A. S. Albahri, S. Yussof, and I. M. Sharaf, "A novel evaluation framework for medical LLMs: Combining fuzzy logic and MCDM for medical relation and clinical concept extraction," *J. Med. Syst.*, vol. 48, no. 1, p. 81, Aug. 2024, doi: 10.1007/s10916-024-02090-y.
[33] W. Zheng, L. Turner, J. Kropczynski, M. Ozer, T. Nguyen, and S. Halse, "LLM-as-a-fuzzy-judge: Fine-tuning large language models as a clinical evaluation judge with fuzzy logic," 2025, *arXiv:2506.11221*.
[34] I. Svoboda and D. Lande, "Enhancing multi-criteria decision analysis with AI: Integrating analytic hierarchy process and GPT-4 for automated decision support," 2024, *arXiv:2402.07404*.
[35] R. K. Chakrabortty, M. Abdel-Basset, and A. M. Ali, "A multi-criteria decision analysis model for selecting an optimum customer service chatbot under uncertainty," *Decis. Analytics J.*, vol. 6, Mar. 2023, Art. no. 100168, doi: 10.1016/j.dajour.2023.100168.
[36] H. M. Alabool, "Large language model evaluation criteria framework in healthcare: Fuzzy MCDM approach," *Social Netw. Comput. Sci.*, vol. 6, no. 1, pp. 1–28, Jan. 2025, doi: 10.1007/s42979-024-03533-6.
[37] V. S. Barletta, D. Caivano, L. Colizzi, G. Dimauro, and M. Piattini, "Clinical-chatbot AHP evaluation based on 'quality in use' of ISO/IEC 25010," *Int. J. Med. Informat.*, vol. 170, Feb. 2023, Art. no. 104951, doi: 10.1016/j.ijmedinf.2022.104951.
[38] M.-C. Hsu, "The construction of critical factors for successfully introducing chatbots into mental health services in the army: Using a hybrid MCDM approach," *Sustainability*, vol. 15, no. 10, p. 7905, May 2023, doi: 10.3390/su15107905.
[39] M. Pawlowska-Nowak, "Identification of trust determinants in LLM technology using the DEMATEL method," *Eur. Res. Stud. J.*, vol. 27, no. 2, pp. 694–711, May 2024.
[40] M. Dağdeviren, S. Yavuz, and N. Kılınç, "Weapon selection using the AHP and TOPSIS methods under fuzzy environment," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8143–8151, May 2009.
[41] A. T. Eshlaghy and M. Homayonfar, "MCDM methodologies and applications: A literature review from 1999 to 2009," *Res. J. Int. Stud.*, vol. 21, no. 21, pp. 86–137, 2011.
[42] C. Kahraman, S. C. Onar, and B. Oztaysi, "Fuzzy multicriteria decision-making: A literature review," *Int. J. Comput. Intell. Syst.*, vol. 8, no. 4, pp. 637–666, 2015.
[43] Y. Liu, C. M. Eckert, and C. Earl, "A review of fuzzy AHP methods for decision-making with subjective judgements," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113738.
[44] B. Singh, "Analytical hierarchical process (AHP) and fuzzy AHP applications—A review paper," *Int. J. Pharm. Technol.*, vol. 8, no. 4, pp. 4925–4946, 2016.
[45] S. Sipahi and M. Timor, "The analytic hierarchy process and analytic network process: An overview of applications," *Manage. Decis.*, vol. 48, no. 5, pp. 775–808, Jun. 2010.
[46] T. L. Saaty, "How to make a decision: The analytic hierarchy process," *Eur. J. Oper. Res.*, vol. 48, pp. 9–26, Jan. 1970.
[47] T. L. Saaty, *The Analytic Hierarchy Process*. New York, NY, USA: McGraw-Hill, 1980.
[48] S. C. Onar, B. Oztaysi, and C. Kahraman, "Strategic decision selection using hesitant fuzzy TOPSIS and interval type-2 fuzzy AHP: A case study," *Int. J. Comput. Intell. Syst.*, vol. 7, no. 5, pp. 1002–1021, 2014.
[49] F. Herrera, L. Martínez, and R. M. Rodríguez, "Hesitant fuzzy linguistic term sets," *Adv. Intell. Soft Comput.*, vol. 122, pp. 287–295, Aug. 2011.
[50] D. Yildiz, G. T. Temur, A. Beskese, and F. T. Bozbura, "Evaluation of positive employee experience using hesitant fuzzy analytic hierarchy process," *J. Intell. Fuzzy Syst.*, vol. 38, no. 1, pp. 1043–1058, Jan. 2020.
[51] R. M. Rodriguez, L. Martinez, and F. Herrera, "Hesitant fuzzy linguistic term sets for decision making," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 109–119, Feb. 2012.
[52] A. Adem, E. Çakıt, and M. Dağdeviren, "A fuzzy decision-making approach to analyze the design principles for green ergonomics," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1373–1384, Jan. 2022.
[53] F. Herrera and L. Martinez, "A 2-tuple fuzzy linguistic representation model for computing with words," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 6, pp. 746–752, Dec. 2000.
[54] OpenAI. (May 2024). *Hello GPT-4o*. Accessed: Jun. 2025. [Online]. Available: https://openai.com/index/hello-gpt-4o/
[55] OpenAI. (Nov. 2022). *Introducing ChatGPT*. Accessed: Jun. 2025. [Online]. Available: https://openai.com/blog/chatgpt
[56] MetaAI. (Jul. 2024). *Introducing Llama 3.1: Our Most Capable Models to Date*. Accessed: Jun. 2025. [Online]. Available: https://ai.meta.com/blog/meta-llama-3-1/
[57] Anthropic. (Jun. 2024). *Claude 3.5 Sonnet*. Accessed: Jun. 2025. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet
[58] S. Pichai and D. Hassabis. (Feb. 2024). *Our Next-Generation Model: Gemini 1.5*. Accessed: Jun. 2025. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/
[59] Mistral AI Team. (Jul. 2024). *Large Enough*. Accessed: Jun. 2025. [Online]. Available: https://mistral.ai/en/news/mistral-large-2407
[60] A. Liu et al., "DeepSeek-V3 technical report," 2024, *arXiv:2412.19437*.
[61] LMSYS. *ChatBot Arena Leaderboard*. Accessed: Jun. 2025. [Online]. Available: https://chat.lmsys.org/
[62] Hugging Face. *Open LLM Leaderboard*. Accessed: Jun. 2025. [Online]. Available: https://huggingface.co/open-llm-leaderboard
[63] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment," 2023, *arXiv:2308.05374*.

MELTEM AKSOY received the Ph.D. degree in industrial engineering from Istanbul Technical University, Istanbul, Türkiye, in 2023. She is currently a Researcher with the Research Center Trustworthy Data Science and Security, Technical University Dortmund, Germany. Her research interests include natural language processing, trustworthy AI, data science, and optimization.

AYLIN ADEM received the Ph.D. degree in industrial engineering from Gazi University, Ankara, Türkiye, in 2020. She is currently an Associate Professor with the Department of Industrial Engineering, Gazi University. She was a Postdoctoral Researcher with the Department of Industrial Engineering and Management Systems, University of Central Florida, from January 2024 to January 2025. She has authored more than 40 publications, including peer-reviewed journal articles, book chapters, and conference proceedings. Her research interests include applications of human factors/ergonomics, multi-criteria decision making, decision making under fuzzy environments, process management, neuroergonomics, and human–computer interaction.



METIN DAĞDEVIREN received the Ph.D. degree in industrial engineering from Gazi University, Ankara, Türkiye, in 2005. He is currently a Professor with the Department of Industrial Engineering, Gazi University. His doctorate was about the performance evaluation of employees by utilizing multi-criteria decision-making techniques and their fuzzy extensions. He was an Executive and a Researcher in decision-making and human factors-based projects at Gazi University. He supervised more than ten Ph.D. students as an academician. He published more than 100 papers as peer-reviewed journal articles, book chapters, and conference proceedings. His research interests include ergonomics, human factors, process, human resources management, multi-criteria decision making, decision making under a fuzzy environment, occupational health, and safety.