Group Decision and Negotiation (2026) 35:45
https://doi.org/10.1007/s10726-026-10004-1

Agentic AI Architecture for Multi-Criteria Decision-Making: A Collaborative Human-AI Framework

Rui Ferreira[1,2] · **Marco Araújo**[3,4] · **Anabela Tereso**[1] · **Paulo Novais**[1]

Received: 14 July 2025 / Accepted: 19 May 2026

Abstract

Group decision-making and negotiation increasingly take place in settings where stakeholders hold divergent objectives, values, and interpretations of evidence. However, Large Language Models (LLMs) integration in collective decision processes remains constrained by limited traceability, weak procedural control, and ambiguity regarding the role of human judgment. This conceptual paper proposes a reference architecture for agentic Artificial Intelligence (AI) in Group Decision and Negotiation (GDN) that integrates language-based reasoning with formal Multi-Criteria Decision-Making (MCDM) procedures. The architecture assigns two complementary classes of specialized agents to discrete stages of the process: generative agents, responsible for interpretative tasks such as problem structuring, criteria definition, and preference elicitation, and logical agents, responsible for deterministic operations including weighting, aggregation, and ranking. A human-in-the-loop (HITL) governance layer supervises tasks requiring subjective judgment or domain expertise, ensuring consistency, transparency, and auditability throughout the decision workflow. The primary contribution is a modular reference architecture, grounded in design science principles, that decouples generative interpretation from formal evaluation within a unified and auditable decision pipeline. The framework is illustrated through a representative multi-stakeholder scenario demonstrating the coordination of agents and human oversight across all stages of the MCDM process.

Keywords Agentic AI · Multi-criteria decision-making (MCDM) · Group decision and negotiation (GDN) · Human-in-the-loop (HITL) · Large language models (LLMs)

Extended author information available on the last page of the article

Published online: 04 June 2026

1 Introduction

In contemporary decision-making environments, the participation of multiple stakeholders has become a defining characteristic of complex organizational, societal, and strategic problems. Stakeholders typically contribute heterogeneous perspectives, values, expertise, and objectives that must be reconciled under conditions of uncertainty, limited information, and conflicting priorities. Multi-Criteria Decision-Making (MCDM) provides a methodological framework for structuring such problems by enabling the systematic evaluation and ranking of alternatives across multiple, often conflicting, criteria (Roy 1990). In this context, group decision-making and negotiation are central, as decision quality relies not only on analytical rigor but also on integrating divergent viewpoints into outcomes that are both methodologically valid and socially acceptable.

Although classical MCDM approaches are methodologically rigorous, they face persistent challenges in large-scale or multi-stakeholder decision scenarios. As the number of participants and criteria increases, preference divergence intensifies, complicating consensus formation and elevating the risk of conflict (Kirschner 2002). Simultaneously, inconsistencies in individual judgments, arising from bounded rationality, limited expertise, or cognitive effort, can undermine the internal coherence of pairwise comparisons and aggregation procedures (Kuller et al. 2023). These challenges are further amplified in negotiation contexts, where strategic behavior, information asymmetries, and communication barriers often distort preference elicitation and compromise trust in the decision process. Therefore, there is a need for decision-support mechanisms that not only compute outcomes but also structure interaction, validation, and justification throughout MCDM-supported group decision processes.

To address these challenges, recent research in decision science has increasingly focused on integrating Artificial Intelligence (AI) techniques into traditional decision-support systems (Liao et al. 2023). Data-driven models have shown potential to enhance consistency, scalability, and operational efficiency by automating repetitive tasks, supporting preference elicitation, and extracting insights from large and heterogeneous datasets. The advent of Large Language Models (LLMs) has further expanded these capabilities (Harsha et al. 2024). Due to their advanced natural language understanding, reasoning, and generative abilities, LLMs facilitate more intuitive human–machine interaction, enable the interpretation of unstructured inputs, and support interactive decision assistance throughout complex analytical workflows.

Nevertheless, integrating LLMs into decision-support systems introduces significant limitations. LLMs function as probabilistic generative models with largely opaque internal reasoning processes, which raises concerns about consistency, reproducibility, and trustworthiness, especially in contexts requiring methodological rigor and auditability. Minor variations in prompts can result in divergent recommendations, and without explicit constraints, LLM-generated judgments may violate the theoretical assumptions of established MCDM methods. Consequently, fully automated or monolithic LLM-based decision systems do not sufficiently meet the requirements of high-stakes or collaborative decision environments, where human oversight, traceability, and formal correctness are critical.

In this evolving context, Agentic AI has emerged as a promising paradigm for developing more reliable and controllable AI-assisted decision-support systems (Acharya et al. 2025). Agentic AI encompasses architectures in which multiple autonomous yet coordinated agents are assigned explicit roles, objectives, and operational constraints, allowing them to collaboratively execute complex workflows (Sapkota et al. 2025). Unlike purely reactive generative systems, agentic architectures prioritize task decomposition, role specialization, and controlled interaction, properties that align with the sequential and modular structure of MCDM methodologies.

Accordingly, an Agentic AI MCDM framework is proposed in which the sequential stages of the decision process, including problem formulation, domain and criterion elicitation, weighting, evaluation, aggregation, and interpretation, are supported by specialized agents with clearly defined methodological roles. Generative agents, instantiated using LLMs, are responsible for interpretative and semantic tasks that require contextual reasoning and explanation, while logical agents execute formally specified MCDM procedures to ensure determinism, consistency, and theoretical validity.

To ensure transparency, robustness, and human oversight, the framework incorporates structured human-in-the-loop (HITL) mechanisms at predefined decision stages that require subjective judgment or domain expertise. Human feedback is documented, versioned, and explicitly validated, preventing uncontrolled refinement loops and preserving accountability and traceability. This design maintains the theoretical rigor of classical MCDM methods while leveraging the flexibility and expressive capabilities of LLMs in a disciplined and auditable manner.

Methodologically, this research follows a design science research orientation (Hevner et al. 2004), in which the primary artifact is a reference architecture, that is, a normative and prescriptive model that specifies structural components, their relationships, and coordination principles, rather than an implemented system subject to empirical performance evaluation. This positioning is deliberate, as the framework's contribution lies in establishing architectural constraints and design rationale that can guide subsequent implementation, empirical validation, and domain-specific adaptation.

The main contributions of this paper are threefold. First, it introduces a conceptual architecture for Agentic AI in Group Decision and Negotiation (GDN), providing a modular and normative blueprint that structurally integrates specialized agent roles, formal MCDM procedures, and HITL governance within a unified and auditable decision workflow. Second, it establishes a principled decoupling of generative interpretation and formal evaluation as a core architectural principle, demonstrating how LLM-based agents can augment rather than replace human judgment while preserving methodological traceability and reproducibility throughout the MCDM pipeline. Third, it formalizes a structured HITL governance model that embeds validation, consistency-checking, and versioned human input at predefined decision stages, which ensures accountability, transparency, and alignment with the procedural requirements of multi-stakeholder decision and negotiation contexts.

The remainder of this paper is organized as follows. Section 2 reviews related work on AI-enhanced decision-support systems, with emphasis on LLM integration into MCDM and group decision-making. Section 3 introduces the proposed Agen-

tic AI framework, covering both the internal agent architecture and the end-to-end MCDM workflow orchestration. Section 4 illustrates the framework through a representative scenario and discusses its limitations. Finally, Section 5 summarizes the contributions and outlines future research directions.

2 Background & Related Work

The complexity of current MCDM problems has increased substantially, primarily due to the challenges of reconciling diverse individual preferences and mitigating cognitive biases associated with processing large volumes of data for comprehensive analysis and insight generation. These challenges, in turn, increase the risk of inconsistent or suboptimal recommendations (Djartov et al. 2024). Consequently, considerable efforts have been made in the literature to enhance the accuracy and robustness of recommendations.

The integration of AI into MCDM has progressively enabled the development of more data-driven and objective decision-support frameworks (Nguyen et al. 2024). AI-driven systems can enhance both group and individual performance by facilitating joint value creation, improving the consistency of judgments, and supporting informed consensus-building. Moreover, AI has the potential to improve operational efficiency by automating repetitive tasks, optimizing communication flows, and dynamically adapting to the evolving needs of stakeholders.

These capabilities extend across critical stages of group decision-making and negotiation processes, including problem structuring, preference elicitation, alternative generation, concession modeling, and conflict management. Among these stages, determining reliable criteria weights plays a pivotal role, as it directly influences the outcomes of multi-criteria evaluations. In this context, AI-driven methodologies have demonstrated promising results in extracting objective weights from historical data (Zhao et al. 2024). A prominent line of research focuses on leveraging feature importance metrics from ML models to determine criteria weights (Arabameri et al. 2019). The studies conducted in this scope suggest a concrete improvement over traditional expert-based methods, reducing dependence on subjective human judgments (Abdulla et al. 2023).

At the same time, as decision-making scenarios increasingly apply a larger number of evaluation criteria, managing the complexity of decision problems has become a critical challenge. To address this, dimensionality reduction techniques, such as Principal Component Analysis (PCA), have been used to identify and retain the most important criteria while eliminating redundancy and noise within the set of evaluation criteria (Costa et al. 2024). These advancements have been crucial for simplifying the structure of the decision problem, enabling more efficient analysis and allowing decision-makers to focus on the most relevant information without compromising the quality of the analysis. For example, Yang et al. (2008) applied PCA to streamline the evaluation of sustainable supplier performance, successfully reducing the number of criteria without impacting the quality of the final selection.

Deep learning architectures have further extended these capabilities by enabling the analysis of unstructured and high-volume data, supporting more comprehensive

decision analyses (Yang et al. 2020; LeCun et al. 2015). Complementary work has integrated Bayesian networks with MCDM to address incomplete expert knowledge (Kaya et al. 2023), while sentiment analysis of user-generated content has been incorporated into multi-criteria recommender systems to capture implicit stakeholder preferences (Angamuthu and Trojovský 2023). Despite these advances, the quality and reliability of outcomes remain sensitive to judgment inconsistencies, communication inefficiencies, and cognitive biases, which represent limitations that motivate the integration of LLMs into decision-support workflows.

In this way, the integration of LLMs into decision science represents a promising step toward the next generation of decision-support systems (Hendriksen 2023). Particularly, their core capabilities for understanding and generating natural human language open new avenues for improving communication clarity, fostering consensus-building, and facilitating the structured gathering and aggregation of preferences among diverse stakeholder groups.

In this regard, these architectures can be further used to support advanced decision-support workflows through mechanisms such as in-context learning, instruction following, and tool use. In-context learning allows models to acquire new task skills from just a few examples provided in the prompt (Dong et al. 2024). By operationalizing the expected outputs through prompt-based demonstrations, this approach avoids the need for extensive fine-tuning or retraining. Consequently, it facilitates rapid adaptation to evolving decision-making scenarios and supports flexible generalization across varying task specifications. On the other hand, instruction-following assigns LLMs to perform a wide range of tasks based on natural language instructions, supporting flexible, user-friendly, and adaptable decision-making procedures (Lou et al. 2024). Ultimately, tool use enables LLMs to interface with external systems, APIs and knowledge bases, facilitating the efficient retrieval, processing, and integration of domain-specific information (Carolan et al. 2024).

In line with this potential, recent research has started to explore the use of LLMs in decision-support operations. As a result, there is an increasing focus on developing domain-specific LLMs, which are fine-tuned with specialized vocabularies and knowledge bases tailored to sectors or disciplines. These models provide a deeper understanding of context and can generate more accurate and relevant insights when applied to decision-support tasks within specific domains. For instance, Tariq et al. (2024) developed a domain-specific LLM for clinical applications related to prostate cancer. The model was evaluated on tasks that involved predicting clinical information and answering questions. Then, it was compared to a similarly sized general-purpose model (GPT-2) and a larger domain-specialized model (BioGPT). The domain-specific LLM outperformed GPT-2 in both tasks and even outperformed BioGPT in predicting clinical information, while also showing advantages in question answering. These results highlight that targeted training on specialized data and vocabulary can yield more accurate and contextually relevant outputs for decision support in specific domains (Chen et al. 2024).

The integration of LLMs as virtual experts into MCDM frameworks is also an emerging research area. In this case, LLMs function as scalable knowledge bases for identifying decision criteria and can be prompted to generate quantitative judgments, such as pairwise comparisons or performance scores. Additionally, LLMs can

simulate various expert personas, enabling the rapid formation of virtual panels. This capability facilitates a more detailed and efficient exploration of the decision-making landscape. A pioneering study by Wang and Wu (2024) explored the application of ChatGPT as an MCDM tool for supplier evaluation. According to the proposed methodology, supplier evaluations were initially conducted using a traditional AHP and Fuzzy Comprehensive Evaluation approach, based on data collected from expert surveys. ChatGPT, specifically the GPT-3.5-turbo model, was then used to generate comparable evaluations via targeted prompts designed to mimic expert assessments. The findings revealed a strong alignment between the evaluations generated by ChatGPT and those produced by human experts. In addition, the authors demonstrated that these results could be further improved through the implementation of advanced prompting strategies, including chain-of-thought reasoning, demonstrations, and voting ensembles.

Building on this recent research, Wang et al. (2025) introduced an LLM-based framework designed to automate and enhance the performance of MCDM systems across various domains. The authors evaluated several open-source and commercial LLMs based on three representative applications: supplier evaluation, customer satisfaction, and air quality assessment. The results showed that when LLMs were applied without specialized prompting techniques, they achieved a moderate accuracy of 60%. However, this improved significantly to 70% when methods such as few-shot learning and chain-of-thought reasoning were used. To further bridge the gap between machine and expert performance, the authors fine-tuned open-source models using Low-Rank Adaptation (LoRA) methods.

Lu et al. (2024) introduced a novel approach that combines LLMs with the Analytic Hierarchy Process (AHP) for the multi-criteria evaluation of open-ended responses. In the initial phase, these models automatically generate multiple evaluation criteria specific to the given question. In the subsequent stage, LLMs perform pairwise comparisons of candidate answers based on each criterion. The outcomes are then aggregated using the AHP to produce a final ranking or score for each answer. The results validate the proposed approach by demonstrating a high degree of alignment with human judgment when compared against four established baseline methods across four distinct datasets.

Another study conducted by Zuheros et al. (2024) analyzed the integration of ChatGPT into crowd decision-making processes, focusing on prompt design strategies to extract structured evaluations from unstructured social media reviews. For this purpose, the research evaluated five distinct models with different prompt strategies to gather assessments from user-generated content. These models encompassed polarity classification, numerical and linguistic scoring, multi-criteria evaluation using category ontologies, and comprehensive end-to-end decision systems capable of providing an overall opinion and score for various alternatives. The experimental results demonstrated consistent rankings across most models, which aligned well with traditional sentiment analysis-based baselines. However, the article also identified critical challenges related to the consistency, sensitivity, and explainability of ChatGPT outputs. To address these issues, the authors highlighted the importance of prompt engineering as a core element for the continued development of more transparent and explainable decision-support systems based on LLMs.

Extending this line of research, Aljohani et al. (2025) proposed an innovative approach that combines LLMs with the Fuzzy Best-Worst Method (FBWM) to improve agile requirements change management in global software development. This research addresses the limitations of traditional expert-driven MCDM techniques, particularly the subjectivity and uncertainty that can arise in dynamic environments. In this study, GPT-4 was used as a global project management expert to conduct Best-Worst method comparisons. It identified the most and least important factors and rated the other factors in relation to them through structured prompt engineering. The findings suggest a high level of alignment and consistency between the prioritization outcomes of human experts and those of LLMs.

Although promising advancements have been noted in the literature regarding AI-based decision-support frameworks, several limitations remain, and further research is needed to address them. Among these challenges, a primary concern consists of ensuring the reliability and validity of outputs generated by LLMs Bender et al. (2021). Due to their black-box nature and opaque internal knowledge representation and reasoning processes, the acceptability of LLM-generated outputs in decision analysis can be compromised, particularly in scenarios that require interpretability, transparency, and user trust. To address these challenges, prior research has emphasized the integration of validation mechanisms, such as expert review, structured verification procedures, and HITL validation, in order to ensure the reliability and credibility of LLM-assisted decision outcomes before their full adoption (Amirizaniani et al. 2024). Despite recent advancements, several challenges persist regarding the consistency, bias, and sensitivity of LLM-based assessments. Notably, small variations in prompt formulation can generate divergent recommendations (Loya et al. 2023). In addition, latent biases embedded in the training data of these models may inadvertently influence the outputs, potentially compromising the fairness and objectivity of decision support. To mitigate these concerns, current research has explored the incorporation of feedback loops aimed at evaluating the logical coherence and validity of LLM outputs, either through secondary LLMs or rule-based validation mechanisms (Bilal et al. 2025). In this context, ChatGPT has been integrated into the MCDM Python library pyDecision to interpret and elucidate the results of various MCDM techniques, offering interactive, natural language explanations that enhance the interpretability and accessibility of complex analytical outputs (Pereira et al. 2024).

Furthermore, researchers are also analyzing techniques such as fuzzy logic and uncertainty quantification to evaluate the confidence levels of LLM-generated judgments and, consequently, improve the robustness of decision-making processes (Ling et al. 2024). For example, Dong et al. (2024) proposed integrating verbal uncertainty estimation into LLM-based evaluations to identify outputs associated with low confidence. The study concludes that filtering these uncertain responses can enhance the consistency of LLM-generated decisions and improve their alignment with human ground truth, especially in personalization tasks.

As LLMs continue to evolve within decision-support domains, this research aims to address the persistent challenges associated with integrating them into structured decision-making frameworks. Building upon recent advancements and the current state-of-the-art, the proposed framework introduces a modular decision-support sys-

tem based on the Agentic AI concept. In this architecture, specialized AI agents are assigned to discrete stages of the decision-making pipeline, enhancing transparency and mitigating a critical limitation of conventional LLM-based systems, namely the propensity to generate inaccurate or logically inconsistent outputs (Huang et al. 2025).

Notable progress in this direction has already been demonstrated. For instance, Svoboda and Lande (2024) developed a custom ChatGPT agent supported by a set of virtual expert personas, each tailored with domain-specific expertise and personality features. These agents collaboratively executed all stages of the AHP, including criteria elicitation, pairwise comparisons, and consistency verification, to identify optimal strategies for mitigating social engineering risks in corporate data centers. Additionally, the study explored prompt engineering, the use of custom GPT instances, and aggregation strategies, setting a precedent for how generative AI can be embedded into structured MCDM workflows. In comparison with the approach presented in this study, although both methodologies employ the AHP as the core mechanism for deriving criterion weights, several methodological distinctions emerge. Svoboda and Lande (2024) assume a prompt-driven orchestration strategy in which domain-specific virtual experts are instantiated via manually configured GPT-based agents. These agents function independently and are activated sequentially, lacking inter-agent coordination or access to shared memory. In contrast, this research introduces an Agentic AI-based architecture characterized by modular agent roles, task-specific memory, and structured inter-agent collaboration. This design enables dynamic orchestration and preserves contextual coherence throughout the MCDM pipeline. Moreover, while Svoboda and Lande (2024) predominantly rely on subjective scoring and aggregation procedures, the proposed framework integrates iterative HITL feedback and consistency validation mechanisms to enhance both transparency and decision robustness. By decomposing the decision process into agentic modules, the architecture facilitates traceability, adaptability, and improved scalability in complex decision-making environments.

In summary, despite the advancements made in LLM-Based MCDM, several gaps remain in the literature, specifically in the areas of model interpretability, dynamic coordination among agents, and robust validation of generated judgments. First, many current systems still depend on static prompt engineering and sequential agent interactions, which limit adaptability and reduce the contextual coherence of decisions across stages. This undermines the potential of fully Agentic AI architectures, where modular agents should dynamically share memory, coordinate tasks, and adapt their reasoning based on the evolving state of the problem. Furthermore, while initial attempts to simulate expert panels through virtual personas have demonstrated promising results, these agents often lack the capacity for genuine deliberation, reflexive adjustment, or negotiation when conflicting perspectives emerge, which represent key aspects for group decision-making scenarios.

Moreover, the literature often overlooks rigorous mechanisms for ensuring traceability, logical consistency, and reproducibility in multi-agent decision outputs. Most LLM-generated recommendations still operate within black-box abstraction pipelines, raising concerns about their reliability, particularly in sensitive domains where auditability is critical.

Likewise, despite some integration of HITL feedback and uncertainty estimation, there is limited research on how to systematically calibrate confidence levels across distributed agent contributions, especially when agents exhibit divergent levels of task-specific competence. Moreover, although few-shot prompting and instruction-following techniques have enhanced LLM performance in structured tasks, they remain sensitive to slight changes in input phrasing, usually producing volatile or biased outcomes. This sensitivity poses challenges for maintaining stability and fairness in longitudinal decision-support applications. Another underexplored area involves the ethical and procedural implications of delegating authority to autonomous agents in collective decision-making processes, including accountability attribution and preserving user agency. Finally, while the literature highlights the benefits of domain-adapted LLMs, few studies rigorously assess how domain-specific fine-tuning impacts the collaborative behavior of multi-agent systems or scales across diverse decision environments.

3 Conceptual Framework: An Agentic AI Architecture for Collaborative MCDM

This section presents the conceptual foundations of the proposed Agentic AI architecture for collaborative MCDM. Section 3.1 introduces the agent model architecture, detailing the internal design principles, reasoning mechanisms, memory structures, and role constraints that support the behavior of individual agents. Section 3.2 then describes the overall system architecture, explaining how specialized agents are orchestrated and coordinated to implement an end-to-end MCDM workflow.

3.1 Conceptual Agent Architecture and Functional Components

The proposed framework adopts a role-based multi-agent paradigm as a structured mechanism for organizing, executing, and validating the sequential stages of the MCDM process. This approach leverages the decomposability of established MCDM methods, such as the AHP and related hierarchical frameworks, which segment the process into analytically distinct yet interdependent stages: problem formulation, criteria elicitation, weighting, evaluation, and aggregation. Each stage requires heterogeneous reasoning, combining formally specified analytical operations with subjective, interpretative, and collaborative judgments. As a result, a single uniform reasoning model cannot adequately support the entire decision workflow. The agent-based abstraction enables these diverse reasoning requirements within dedicated computational entities, while preserving transparency, traceability, and methodological control throughout the decision process (Guo et al. 2024).

The primary methodological contribution of the framework is the explicit separation of distinct reasoning regimes through two complementary agent classes: logical agents and generative agents (Chen et al. 2025). Logical agents are assigned to tasks in which correctness is defined by formal specification rather than contextual interpretation, encompassing those stages of the MCDM process that require deterministic, rule-based, or mathematically formalized reasoning. These agents operate exclu-

sively through deterministic computational procedures, implemented as discrete and verifiable units designed to execute established MCDM algorithms with controlled and reproducible behavior. Generative agents, in contrast, employ an LLM as their core reasoning engine and are designed to support tasks requiring semantic interpretation, abstraction, and controlled content generation, including problem structuring, criteria definition, and the articulation of subjective judgments into operationalizable inputs for subsequent analytical processing. This separation is grounded in the recognized limitation that LLM outputs are non-deterministic by nature, rendering generative inference unsuitable as a sole mechanism for tasks where consistency and reproducibility are required conditions rather than desirable properties (Mohammadi et al. 2025). The architecture operationalizes this separation through three explicit mechanisms. First, all formally specified analytical operations are confined to logical agents, which execute them through deterministic procedures whose inputs and outputs are independently verifiable, ensuring that identical validated inputs produce identical results regardless of the language model employed by generative agents (Sureshkumar 2026). Second, generative agents are restricted to interpretative and communicative tasks, and their outputs are subject to structured HITL validation before entering the analytical pipeline, preventing unverified LLM-generated content from directly influencing computational results (Lazaros et al. 2026). Third, all human inputs, agent outputs, and inter-stage transitions are explicitly recorded and versioned, that is, each validated state is stored as an immutable snapshot indexed by stage, iteration, and timestamp, enabling complete post-hoc traceability of the decision process (Ojewale et al. 2026). Together, these mechanisms provide the architectural basis for auditability and reproducibility within the proposed framework.

To ensure structured and verifiable operation, generative agents incorporate an explicit internal architecture comprising a modular memory subsystem, constrained tool access, and predefined interaction protocols. The memory subsystem is designed to differentiate among short-term working memory, long-term validated knowledge, entity memory, contextual memory, and user-specific memory, reflecting the heterogeneous informational requirements that arise across the distinct stages of the decision process. Critically, memory updates occur only upon task completion and validation, ensuring that persistent knowledge reflects confirmed decisions rather than provisional or speculative content. Building on this internal structure, the external information access and augmentation capabilities of generative agents are strictly regulated. Tool usage, including retrieval-augmented generation over scientific literature, knowledge bases, and shared repositories, is governed by explicit constraints based on agent roles and assigned tasks at the architectural level. These constraints prevent uncontrolled information access and speculative reasoning, ensuring that generative outcomes remain consistent with the objectives of structured decision support.

Furthermore, HITL interaction is architecturally integrated as a stage-specific governance mechanism, exclusively associated with generative agents. Consistent with the validation and traceability mechanisms described above, human intervention is invoked only at predefined decision stages where subjective judgment or domain expertise constitutes an indispensable input. Upon acceptance, human inputs acquire the status of authoritative constraints that can be neither modified nor overridden by

any agent, whether generative or logical. This governance principle prevents circular reasoning and retroactive rationalization of prior decisions, while reinforcing accountability and internal consistency across the decision workflow. The computationally intensive stages of the pipeline, including weight derivation, normalization, aggregation, and ranking, are executed autonomously by logical agents without dependence on human response time. Human involvement is therefore confined to the architecturally bounded set of validation checkpoints defined above, representing a structurally delimited fraction of the overall decision process rather than a throughput constraint.

Although precise latency measurements require empirical evaluation, a structural analysis of the proposed workflow permits a preliminary estimation of the HITL impact on operational throughput. The end-to-end pipeline comprises twelve discrete functional stages, of which six involve HITL validation checkpoints: problem formalization, domain confirmation, domain weight validation, criteria validation, criteria weight confirmation, and decision matrix review. The remaining stages, including pairwise comparison computation, consistency ratio evaluation, normalization, weighted aggregation, global ranking, and result explanation, are executed autonomously by logical agents at computational speed. Consequently, HITL interaction is structurally confined to approximately half of the pipeline stages. Furthermore, within each checkpoint, interaction is bounded by predefined iteration limits (e.g., a maximum of two to three revision cycles), which constrains the maximum time overhead per validation gate. In practical terms, the autonomous computational stages would execute in the order of seconds to minutes, while HITL stages would depend on stakeholder availability, potentially ranging from minutes in synchronous settings to hours or days in asynchronous multi-stakeholder configurations. Nevertheless, given that this work advances a conceptual reference architecture grounded in design science principles, a precise quantitative characterization of interaction latency is deferred to future empirical investigation, as discussed in Section 5.2.

From an implementation perspective, the proposed agent model is compatible with contemporary role-based agent orchestration paradigms. As a concrete illustrative example, the CrewAI framework, developed by Moura (2024) as an open-source platform for orchestrating role-playing autonomous AI agents, provides a realization of the abstract agent model described in this section, in which agents are defined through explicit role, goal, and task specifications, as shown in Figure 1. This clear separation between conceptual design and implementation ensures that the proposed framework remains extensible and adaptable across various orchestration platforms and execution environments, of which CrewAI represents one possible instantiation.

3.2 Overall System Architecture: Multi-Agent Orchestration and End-to-End Workflow

The proposed framework operationalizes the MCDM methodology by orchestrating the agents introduced in Section 3.1 into a structured, interpretable, and methodologically controlled analytical pipeline. Rather than acting as a passive execution environment in which MCDM steps unfold independently, the framework serves as an active orchestrator that regulates information flow, enforces methodological con-

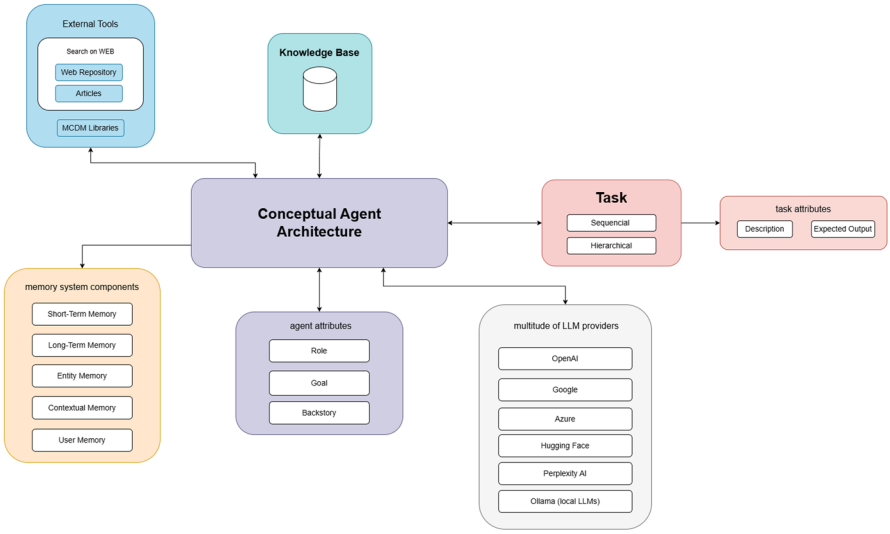

Fig. 1 Conceptual agent architecture and functional components. The diagram depicts the internal structure of a generative agent, including its attributes, memory subsystem, external tool access, and task specification

straints, and activates each stage of the decision process through agents instantiated with their designated reasoning regimes. Formally specified analytical stages are executed by logical agents, while interpretative, semantic, and collaborative stages are delegated to generative agents, ensuring a strict alignment between agent capabilities and methodological requirements.

The workflow begins with problem formalization, a stage that involves semantic interpretation and contextual understanding. A dedicated generative interpretation agent transforms raw stakeholder inputs into a structured representation encoding the decision objective, candidate alternatives, relevant constraints, and contextual assumptions. This representation constitutes the semantic backbone of the entire methodology and is stored in contextual memory, enabling subsequent agents to retrieve, refine, and operationalize it as the decision process evolves. At this stage, a controlled HITL validation mechanism is activated, allowing stakeholders to confirm or correct the interpretation. Consistent with the framework's design principles, this interaction is bounded to a limited number of iterations, and all validated corrections are persistently recorded to prevent repeated misinterpretations or uncontrolled refinement loops.

Building on the validated problem representation, the methodology advances to the domain elicitation and weighting stage. Generative agents propose an initial set of decision domains inferred from the structured problem representation and provide contextual justifications for their relevance. These proposals are subsequently validated by users through a bounded interaction process. The elicitation phase terminates once the user either explicitly confirms the proposed domain set or no further modifications are introduced within the predefined iteration limit, at which point the validated domains are treated as fixed inputs for subsequent analytical stages.

Once validated, domain weighting is conducted using an agentic AHP-based procedure. Generative agents produce candidate pairwise comparisons accompanied by concise semantic justifications that reflect subjective judgment and contextual interpretation. These comparisons are then passed to a logical weighting agent, which deterministically computes the priority vector and evaluates the consistency ratio using formally specified algorithms. HITL interaction is invoked only after a full comparison and consistency evaluation cycle has been completed, rather than after individual comparisons. To prevent infinite refinement cycles, the framework enforces explicit iteration limits and escalation mechanisms, such as targeted revision of specific comparisons or direct human correction of inconsistent entries. All validated adjustments are stored in memory, ensuring that generative agents do not reintroduce previously rejected judgments.

A similar agent-driven procedure governs the elicitation and weighting of criteria at subordinate hierarchical levels. Generative agents propose domain-specific evaluation criteria based on contextual memory, while logical agents verify structural validity. After stakeholder validation, criteria weights are computed using the same deterministic AHP workflow applied at the domain level. Versioned memory states enable rollback, targeted re-evaluation, and the prevention of circular or self-contradictory updates, maintaining internal mathematical consistency across the hierarchy.

Once the hierarchical structure is finalized, the methodology proceeds to the evaluation stage. At this point, the decision matrix is instantiated, and performance values for each alternative–criterion pair are populated. These values may be provided directly by stakeholders or, when appropriate, inferred or extracted by specialized intelligence agents using available datasets, analytical models, or contextual knowledge sources. This hybrid strategy supports efficient decision matrix construction while preserving transparency into data provenance. All agent-generated or inferred values are explicitly recorded and, when required, subject to human validation.

Normalization of performance values is performed exclusively by a logical agent using predefined, rule-based procedures associated with the selected MCDM method. Human oversight is invoked only in exceptional cases where attribute semantics (e.g., cost-versus-benefit interpretation) require explicit confirmation.

Following normalization, evaluation and aggregation are conducted through logical agents that compute weighted scores for each alternative by combining normalized performance values with validated criteria and domain weights. These computations are deterministic and reproducible, requiring no iterative human refinement. If stakeholders introduce contextual constraints or preferences that affect interpretation, such input is mediated through bounded generative-agent workflows without altering the underlying analytical computations.

The ranking and synthesis phase is executed by a logical synthesis agent that aggregates weighted evaluations into global priority scores, fully respecting the validated hierarchical structure. In addition to producing a final global ranking, the framework may generate domain-level rankings to support comparative analysis and negotiation. Generative agents may subsequently be invoked to explain, contextualize, or visualize ranking outcomes for stakeholders; however, these explanations do not modify the analytical results produced by logical agents.

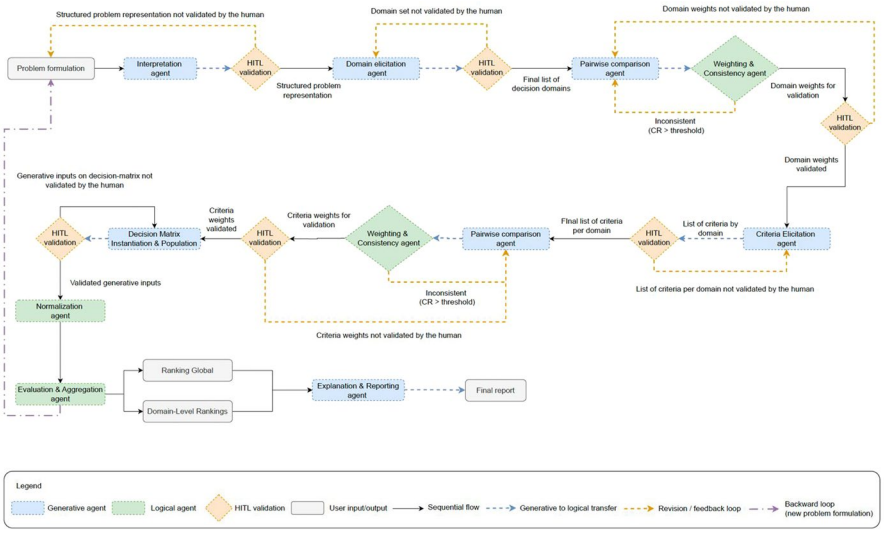

Fig. 2 Conceptual end-to-end orchestration of the proposed agentic collaborative MCDM framework. The diagram distinguishes three functional layers. The generative agent layer (dashed borders) is responsible for semantic interpretation, criteria elicitation, and stakeholder communication. The logical agent layer (solid borders) executes deterministic AHP-based weighting, normalization, aggregation, and ranking. The HITL governance layer (diamond nodes) is activated at predefined validation checkpoints where human judgment constitutes an authoritative and versioned input

Throughout all stages, the proposed framework depends on explicit agent coordination and control mechanisms to maintain analytical rigor, interpretability, and operational robustness within the MCDM workflow. As illustrated in Figure 2, the workflow employs four distinct line conventions to represent the nature of inter-stage connections: solid arrows denote sequential flow within the same functional layer; dashed arrows indicate generative-to-logical transfers across layers; dotted arrows represent revision and feedback loops triggered by consistency violations or stakeholder corrections; and the backward loop arrow captures the possibility of returning to problem formulation when structural revisions are required. Intermediate outputs, such as the structured problem representation, the final list of decision domains, or the list of criteria per domain, mediate the transitions between agent stages and HITL validation gates, reflecting the sequential and validated nature of the workflow.

Logical agents execute formally specified decision-analytic procedures to ensure determinism and methodological correctness, while generative agents are selectively activated to facilitate semantic interpretation, explanation, and deliberate human interaction.

Given the conceptual scope of this work, Figure 2 is intended as an architectural schematic rather than a process-flow specification. Precise sequencing, exception handling, and runtime behavior are subject to implementation-level design decisions that fall outside the scope of the current framework.

From a GDN perspective, the mechanisms outlined in this section together constitute a structured negotiation support environment. Problem formalization, combined with bounded HITL validation, supports preference revelation under facilitated

and controlled conditions. At the same time, the stages devoted to domain and criteria elicitation reflect the agenda-setting and issue-framing functions that typically precede substantive negotiation. The AHP-based weighting process introduces a disciplined approach to preference refinement, reducing the influence of position-driven negotiation dynamics, whereas domain-level rankings enable the comparative analysis of trade-offs across competing stakeholder priorities. By integrating these GDN-consistent properties into a formally auditable multi-agent architecture, the framework provides a principled basis for AI-assisted negotiation support, ensuring that normative and preference-based assessments remain under stakeholder control while analytically well-defined computations are delegated to deterministic agents.

4 Discussion

This section illustrates the proposed framework through a representative scenario and examines its current limitations. Section 4.1 presents a conceptual walkthrough of a multi-stakeholder decision and negotiation case. Section 4.2 discusses the structural and methodological boundaries of the current design.

4.1 Illustrative Application: A Group Negotiation Scenario

This subsection presents a conceptual walkthrough rather than an empirical evaluation, tracing how each agent, coordination mechanism, and HITL interaction would function within a representative multi-stakeholder scenario. No system has been implemented, no data has been collected, and no performance metrics are reported. This deliberate abstraction reflects the design science orientation of the paper, wherein the primary artifact is the reference architecture itself rather than a deployed system subject to performance benchmarking. Formal empirical validation, including prototype deployment, user studies, and comparative benchmarking against existing decision-support systems, is explicitly identified as a primary direction for future research, as discussed in Section 5.2. The aim of this section is solely to demonstrate how the Agentic AI MCDM framework structures interactions, coordinates heterogeneous reasoning tasks, and integrates human judgment throughout a complex multi-stakeholder decision pipeline, providing a concrete operational illustration of the architectural principles introduced in Section 3.

The scenario considers a regional transportation authority tasked with selecting the most suitable location for a new multimodal mobility hub. The decision involves multiple stakeholder groups, including municipal planners, environmental agencies, local businesses, and citizen representatives, each contributing distinct priorities and evaluative perspectives. The problem requires balancing economic feasibility, environmental impact, accessibility, and long-term development potential.

The process begins with problem interpretation and structuring. Stakeholders provide qualitative descriptions of objectives, constraints, and contextual considerations. A generative interpretation agent transforms these inputs into a structured representation encoding the decision objective (e.g., “select the optimal location for the mobility hub”), the set of alternatives (e.g., Sites A, B, C, and D), and an initial set of

decision domains (e.g., economic, environmental, social, and infrastructural dimensions). The HITL validation stage then allows stakeholders to confirm or refine this structure before it is fixed in contextual memory, ensuring alignment and preventing downstream misinterpretation.

After problem structuring, the framework advances to domain elicitation and validation following the bounded interaction process described in Section 3.2. In this scenario, generative agents propose decision domains grounded in the validated problem representation, including the introduction of "future scalability" as an infrastructural consideration based on long-term mobility forecasts. Once stakeholders confirm the domain set, the validated domains are treated as authoritative inputs for subsequent analytical stages.

Domain weighting is conducted following the agentic AHP-based workflow described in Section 3.2. Generative agents produce candidate pairwise comparisons with contextual justifications. For instance, environmental impact may be assessed as moderately more important than economic cost, reflecting the sustainability-oriented priorities of the planning context. These comparisons are then passed to a logical weighting agent, which deterministically computes priority vectors and evaluates the consistency ratio.

Once the logical agent confirms that the consistency ratio falls within the conventionally accepted threshold of 0.10, stakeholders review the resulting priority vectors and may selectively revise specific judgments. In this scenario, for instance, a municipal planner might challenge the relative weighting assigned to economic cost, prompting a targeted revision that preserves the remaining validated comparisons.

Criteria elicitation and weighting at subordinate hierarchical levels follow the same procedure. Within the environmental domain, for example, generative agents propose criteria such as CO_2 emissions, biodiversity disruption, and land-use efficiency, which stakeholders validate before logical agents compute their relative weights.

Once the hierarchical structure is finalized, the framework proceeds to alternative evaluation. Performance values for each alternative–criterion pair may be provided directly by stakeholders or inferred by specialized intelligence agents using available datasets, reports, or contextual knowledge (e.g., estimated construction costs, projected emission levels, accessibility indices, or expected travel-time reductions). At this stage, the framework prioritizes transparency over automation. Inferred values are explicitly identified and may be subject to human validation.

Following normalization and aggregation, performed deterministically by logical agents as specified in Section 3.2, the framework may produce weighted scores and global rankings. In addition to a final global ranking (e.g., Site C emerging as the preferred option), the framework supports domain-level rankings (e.g., Site B ranking highest in environmental performance while underperforming economically). These disaggregated results enable stakeholders to identify trade-offs and explore compromise solutions during negotiation. Generative agents may subsequently produce explanatory summaries or justification reports; however, these explanations do not alter the analytical results.

In summary, it is important to note that this illustrative scenario is intentionally framed at a conceptual and operational level. While the framework specifies agent

roles, decision stages, validation logic, and termination conditions, it deliberately abstracts from concrete software architectures, interface design, deployment strategies, and runtime optimization. Consequently, the scenario highlights the coordination of structured reasoning and human supervision, rather than technical implementation details. This abstraction reflects current methodological boundaries, including reliance on validated human judgments, deterministic aggregation procedures, and the absence of explicit uncertainty propagation within the analytical core. These limitations are discussed in the following subsection.

4.2 Structural and Methodological Limitations of the Current Framework

Despite its contributions, the proposed framework demonstrates several limitations that define its current boundaries. Structurally, the use of multiple specialized agents introduces coordination complexity. While agent modularity improves transparency and methodological control, it also increases orchestration overhead, including memory management and inter-agent communication. At the conceptual level, these challenges are acknowledged but not yet resolved through concrete optimization strategies.

Methodologically, the framework remains partially dependent on the quality and consistency of human input, particularly during pairwise comparisons and validation stages. Subjective biases and inconsistent judgments may still influence outcomes, even when bounded interaction protocols and escalation mechanisms are in place. In addition, while generative agents enhance interpretative capacity, their outputs may vary depending on the underlying language model, introducing residual variability that is not fully controlled within the current design. Additionally, while Section 2 identified the ethical and procedural implications of delegating authority to autonomous agents as an underexplored area in the literature, the current framework does not yet incorporate explicit mechanisms for modeling accountability attribution beyond the HITL validation checkpoints already defined. Although human authority is preserved at predefined decision stages, the framework does not formally specify how responsibility is distributed between human and agent contributions in the final decision outcome. Addressing this dimension would require extending the governance layer to account for normative constraints on agent autonomy, an aspect that is deferred to future research.

Furthermore, uncertainty associated with inferred or estimated performance values is not yet explicitly propagated through the aggregation and ranking stages. The framework currently assumes deterministic inputs once validated, which limits its applicability in highly uncertain or data-sparse environments. Lastly, the framework is primarily designed for discrete MCDM problems and deterministic aggregation methods. Dynamic decision contexts, continuous optimization problems, or scenarios with rapidly evolving criteria may necessitate methodological extensions beyond the current scope. Extending the framework to accommodate alternative multi-criteria procedures would require verifying the compatibility of their specific procedural structures with the proposed agent roles and inter-stage coordination mechanisms. Similarly, while the current weighting process is grounded in pairwise comparison logic, incorporating AI-driven weight derivation techniques, such as those based

on machine learning feature importance or data-driven preference learning, could reduce dependence on subjective human judgments and enhance scalability in data-rich environments. These extensions remain subject to future empirical validation.

Taken together, the limitations outlined above do not undermine the validity of the proposed framework. Rather, they delineate its current boundaries and reflect deliberate design choices aimed at establishing a rigorous conceptual–operational foundation for Agentic AI–enhanced MCDM, while deferring implementation-level optimization, uncertainty modeling, and large-scale deployment concerns for future work.

5 Conclusion and Future Directions

This section summarizes the contributions of the proposed framework and outlines directions for future research. Section 5.1 presents the conceptual and architectural contributions, and Section 5.2 identifies open research questions for further refinement.

5.1 Summary of Contributions

As reviewed in Section 2, the integration of LLMs into decision-support systems has progressed substantially, yet persistent limitations in traceability, consistency, and auditability continue to constrain their adoption in structured multi-criteria decision environments. These limitations motivated the hybrid approach adopted in this work, wherein LLMs are positioned as reasoning components embedded within a formally specified decision methodology rather than as autonomous decision-makers.

To this end, the paper introduced an Agentic AI MCDM collaborative framework based on a multi-agent architecture that operationalizes this principle through the three mechanisms established in Section 3.1: the confinement of formal computations to deterministic logical agents, the restriction of generative agents to interpretative tasks subject to structured HITL validation, and the versioned recording of all human inputs and agent outputs throughout the decision pipeline. The coordinated operation of these agents, governed by stage-specific human oversight, ensures that methodological validity, transparency, and alignment with stakeholder expectations are preserved across all stages of the MCDM workflow.

Overall, the proposed framework advances the field of AI-augmented decision analytics by demonstrating how LLM-based agents can augment rather than replace human judgment within a formally specified and human-governed decision pipeline. This hybrid paradigm, which confines analytical computation to deterministic agents and reserves evaluative authority for human stakeholders, establishes a foundation for scalable and trustworthy decision-support systems and opens new avenues for research at the intersection of decision science, AI, and human–AI collaboration. In doing so, the framework addresses key literature gaps identified in Section 2, namely the lack of dynamic inter-agent coordination, the absence of rigorous traceability and reproducibility mechanisms, and the limited integration of structured human oversight within LLM-based MCDM pipelines.

5.2 Future Research

While the proposed Agentic AI MCDM framework provides a structured and human-aligned approach to decision support, several research avenues remain open, particularly regarding its application in real-world group decision-making contexts.

Future work should focus on extending the framework beyond theoretical validation toward operational deployments, where consensus building, uncertainty management, and stakeholder heterogeneity introduce additional complexity.

Although the current methodology enables multi-stakeholder input, validation, and structured aggregation, it does not yet explicitly model consensus-building processes or negotiation dynamics. Future research should explore how specialized agents can identify conflicting preferences, highlight trade-offs between stakeholder priorities, and propose compromise solutions that guide stakeholders toward mutually acceptable outcomes. Extending the framework in this direction would address the methodological limitation of relying on subjective human judgments by providing structured agent-supported mechanisms for convergence while retaining human deliberative control over outcomes.

In parallel, an important research direction concerns the variability introduced by instantiating agents with different LLMs. Preliminary observations suggest that LLMs may differ substantially in their reasoning styles, levels of determinism, and susceptibility to inconsistent or hallucinated outputs. More expressive models often exhibit stronger contextual reasoning and richer justifications but may introduce greater variability across runs, whereas more constrained or instruction-tuned models tend to produce more stable and repeatable outputs at the cost of reduced semantic flexibility.

Another important research direction arises from the partial dependence on inferred or agent-generated data when populating the decision matrix. While agent-assisted filling improves efficiency, Section 4.2 highlights the absence of explicit uncertainty modeling within the current aggregation and ranking stages. Future work should investigate how uncertainty and confidence levels associated with agent-generated or estimated performance values can be formally represented, propagated through the MCDM workflow, and reflected in the final recommendations. Incorporating uncertainty-aware aggregation or confidence-sensitive ranking mechanisms would strengthen the framework's robustness, particularly in data-sparse or highly uncertain decision environments.

Further research is also required to address structural scalability and adaptability. As noted in the limitations analysis, orchestrating multiple specialized agents introduces coordination overhead that may affect performance as the number of criteria, alternatives, or stakeholders increases. Empirical validation across diverse application domains, such as public policy planning, infrastructure investment, or sustainability assessment, would enable systematic evaluation of scalability, computational efficiency, and user interaction complexity.

In summary, these future research directions aim to advance the proposed framework beyond its current theoretical and methodological boundaries toward empirical validation, strengthening its contribution to the broader fields of group decision-making, human–AI collaboration, and intelligent decision analytics.

Author Contributions A.B. and C.D. wrote the main manuscript text, and A. prepared figures 1, 2. All authors reviewed the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on).

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abdulla A, Baryannis G, Badi I (2023) An integrated machine learning and MARCOS method for supplier evaluation and selection. Decis Anal J 9:100342. https://doi.org/10.1016/j.dajour.2023.100342

Acharya DB, Kuppan K, Divya B (2025) Agentic AI: autonomous intelligence for complex goals–a comprehensive survey. IEEE Access 13:18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

Aljohani B, Aljuhani A, Alsanoosy T (2025) Enhancing agile requirements change management: Integrating LLMs with fuzzy best-worst method for decision support. Int J Adv Comput Sci Appl 16(3)

Amirizaniani M, Yao J, Lavergne A, Okada ES, Chadha A, Roosta T, Shah C (2024) LLMAuditor: a framework for auditing large language models using human-in-the-loop. arXiv:2402.09346

Angamuthu SK, Trojovský P (2023) Integrating multi-criteria decision-making with hybrid deep learning for sentiment analysis in recommender systems. PeerJ Comput Sci 9:1497. https://doi.org/10.7717/peerj-cs.1497

Arabameri A, Yamani M, Pradhan B, Melesse A, Shirani K, Bui DT (2019) Novel ensembles of COPRAS multi-criteria decision-making with logistic regression, boosted regression tree, and random forest for spatial prediction of gully erosion susceptibility. Sci Total Environ 688:903–916. https://doi.org/10.1016/j.scitotenv.2019.06.205

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bilal A, Ebert D, Lin B (2025) LLMs for explainable AI: a comprehensive survey. arXiv:2504.00125

Carolan K, Fennelly L, Smeaton A (2024) A review of multi-modal large language and vision models. arXiv:2404.01322

Chen ZZ, Ma J, Zhang X, Hao N, Yan A, Nourbakhsh A, Yang X, McAuley J, Petzold L, Wang WY (2024) A survey on large language models for critical societal domains: finance, healthcare, and law. arXiv:2405.01769

Chen S, Liu Y, Han W, Zhang W, Liu T (2025) A survey on llm-based multi-agent system: recent advances and new frontiers in application. arXiv preprint arXiv:2412.17481

Costa APDA, Choren R, Pereira DADM, Terra AV, Costa IPDA, Junior CDSR, Santos MD, Gomes CFSO, Moreira MAL (2024) Integrating multicriteria decision making and principal component analysis: a systematic literature review. Cogent Eng 11(1):2374944. https://doi.org/10.1080/23311916.2024.2374944

Djartov B, Oswald F, Jakobi J (2024) Through the psychological lens: unveiling biases in multi-criteria decision-making. In: Ahram T, Casarotto L, Costa P (eds) Human interaction and emerging technologies (IHIET 2024). AHFE International, New York

Dong YR, Hu T, Collier N (2024) Can LLM be a personalized judge?. arXiv:2406.11657

Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, Xia H, Xu J, Wu Z, Liu T, Chang B, Sun X, Li L, Sui Z (2024) A survey on in-context learning. arXiv:2301.00234

Guo T, Chen X, Wang Y, Chang R, Pei S, Chawla NV, Wiest O, Zhang X (2024) Large language model-based multi-agents: a survey of progress and challenges. In: Larson K (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp 8048–8057. International Joint Conferences on Artificial Intelligence Organization, Jeju, South Korea. https://doi.org/10.24963/ijcai.2024/890 . Survey Track

Harsha K, Tarun Kumar K, Sumathi D, Ajith Jubilson E (2024) In: Raza K, Ahmad N, Singh D (eds.) A survey on LLMs: evolution, applications, and future frontiers. Springer, Singapore, pp 289–327. https://doi.org/10.1007/978-981-97-8460-8_14

Hendriksen C (2023) Artificial intelligence for supply chain management: disruptive innovation or innovative disruption? J Supply Chain Manag 59(3):65–76. https://doi.org/10.1111/jscm.12304

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105. https://doi.org/10.2307/25148625

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T (2025) A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inform Syst 43(2):1–55. https://doi.org/10.1145/3703155

Kaya R, Salhi S, Spiegler V (2023) A novel integration of MCDM methods and Bayesian networks: the case of incomplete expert knowledge. Ann Oper Res 320:205–234. https://doi.org/10.1007/s10479-022-04996-7

Kirschner P (2002) Decision-support and complexity in decision making. Proceedings of the 3rd European Conference on Organisational Knowledge, Learning and Capabilities (OKLC). Athens

Kuller M, Beutler P, Lienert J (2023) Preference change in stakeholder group-decision processes in the public sector: Extent, causes and implications. Eur J Oper Res 308(3):1268–1285. https://doi.org/10.1016/j.ejor.2022.12.001

Lazaros K, Vrahatis AG, Kotsiantis S (2026) Human-in-the-loop artificial intelligence: a systematic review of concepts, methods, and applications. Entropy 28(4):377. https://doi.org/10.3390/e28040377

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Liao H, He Y, Wu X, Wu Z, Bausys R (2023) Reimagining multi-criterion decision making by data-driven methods based on machine learning: a literature review. Inf Fusion 100:101970. https://doi.org/10.1016/j.inffus.2023.101970

Ling C, Zhao X, Zhang X, Cheng W, Liu Y, Sun Y, Oishi M, Osaki T, Matsuda K, Ji J, Bai G, Zhao L, Chen H (2024) Uncertainty quantification for in-context learning of large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 3357–3370. https://doi.org/10.18653/v1/2024.naacl-long.184

Lou R, Zhang K, Yin W (2024) Large language model instruction following: a survey of progresses and challenges. Comput Linguist 50(3):1053–1095. https://doi.org/10.1162/coli_a_00523

Loya M, Sinha D, Futrell R (2023) Exploring the sensitivity of LLMs' decision-making capabilities: insights from prompt variations and hyperparameters. Findings of the Association for Computational Linguistics: EMNLP 2023. pp 3711–3716. https://doi.org/10.18653/v1/2023.findings-emnlp.241

Lu X, Li J, Takeuchi K, Kashima H (2024) AHP-powered LLM reasoning for multi-criteria evaluation of open-ended responses. arXiv:2410.01246

Mohammadi M, Li Y, Lo J, Yip W (2025) Evaluation and benchmarking of llm agents: a survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. KDD '25. Association for Computing Machinery, New York, pp 6129–6139. https://doi.org/10.1145/3711896.3736570

Moura J (2024) CrewAI: framework for orchestrating role-playing autonomous AI agents. https://github.com/crewAIInc/crewAI. Accessed: January 2025

Nguyen VTT, Vo NTM, Truong VC, Nguyen V-T (2024) Multi-criteria decision-making and optimum design with machine learning: a practical guide, 1st edn. CRC Press, Boca Raton

Ojewale V, Suresh H, Venkatasubramanian S (2026) Audit trails for accountability in large language models. arXiv preprint arXiv:2601.20727

Pereira V, Basilio MP, Santos CHT (2024) Enhancing decision analysis with a large language model: pyDecision a comprehensive library of MCDA methods in python. arXiv:2404.06370
Roy B (1990) Decision-aid and decision-making. Eur J Oper Res 45(2–3):324–331. https://doi.org/10.1016/0377-2217(90)90196-I
Sapkota R, Roumeliotis KI, Karkee M (2025) AI agents vs. agentic AI: a conceptual taxonomy, applications and challenges. Super Intell. https://doi.org/10.70777/si.v2i3.15161
Sureshkumar S (2026) R-lam: reproducibility-constrained large action models for scientific workflow automation. arXiv preprint arXiv:2601.09749
Svoboda I, Lande D (2024) Enhancing multi-criteria decision analysis with AI: integrating analytic hierarchy process and GPT-4 for automated decision support. arXiv:2402.07404
Tariq A, Urooj A, Das A, Jeong J, Trivedi S, Patel B, Banerjee I (2024) Domain-specific LLM development and evaluation – a case-study for prostate cancer. medRxiv. https://doi.org/10.1101/2024.03.15.24304362
Wang X, Wu X (2024) Can ChatGPT serve as a multi-criteria decision maker? A novel approach to supplier evaluation. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 10281–10285. https://doi.org/10.1109/ICASSP48485.2024.10447204
Wang H, Zhang F, Mu C (2025) One for all: a general framework of llms-based multi-criteria decision making on human expert level. https://doi.org/10.48550/arXiv.2502.15778
Yang Z, Xu Q, Qiu X, Wang H (2008) An applied study on the method for supplier selection with PCA and ELECTRE. In: 2008 IEEE International Conference on Service Operations and Logistics, and Informatics, vol 2, pp 2151–2156. https://doi.org/10.1109/SOLI.2008.4682890
Yang M, Nazir S, Xu Q, Ali S (2020) Deep learning algorithms and multicriteria decision-making used in big data: a systematic literature review. Complexity 2020:2836064. https://doi.org/10.1155/2020/2836064
Zhao LQ, Duynhoven A, Dragićević S (2024) Machine learning for criteria weighting in GIS-based multicriteria evaluation: a case study of urban suitability analysis. Land 13(8):1288. https://doi.org/10.3390/land13081288
Zuheros C, Herrera-Poyatos D, Montes R, Herrera F (2024) Large language models for crowd decision making based on prompt design strategies using ChatGPT: models, analysis and challenges. arXiv:2403.15587

Authors and Affiliations

Rui Ferreira[1,2] · Marco Araújo[3,4] · Anabela Tereso[1] · Paulo Novais[1]

✉ Rui Ferreira
r.pedropassos21@gmail.com; rui-pedro-passos.ferreira@capgemini.com

Marco Araújo
marco.araujo@upt.pt

Anabela Tereso
anabelat@dps.uminho.pt

Paulo Novais
pjon@di.uminho.pt

1 ALGORITMI Research Centre/LASI, University of Minho, Guimaraes, Portugal

2 Capgemini Engineering, Vila Nova de Gaia, Portugal

3 Universidade Portucalense, Porto, Portugal

4 Instituto de Telecomunicações, Aveiro, Portugal