Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MOJO: Multi-LLM Optimised Joint Objective - Generative Artificial Intelligence for Multi-Criteria Decision Analysis Framework

Abtin Ijadi Maghsoodi

amaghsoodi@kpmg.co.nz

University of Waikato

Research Article

Keywords: Multi-Criteria Decision Analysis (MCDA), Large Language Model (LLM), Artificial Intelligence, Generative AI (Gen-AI), Decision Bias

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8755908/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MOJO: Multi-LLM Optimised Joint Objective

Generative Artificial Intelligence for Multi-Criteria Decision Analysis Framework

Abtin Ijadi Maghsoodi [a,b,*] [ORCID: 0000-0001-5944-865X]

[a] Division of Health, Engineering, Computing and Science, School of Health, University of Waikato, Hamilton, New Zealand.
[b] Data Science and Artificial Intelligence Chapter, Digital Lighthouse, KPMG New Zealand, Auckland, New Zealand.
* *Corresponding Author; amaghsoodi@kpmg.co.nz; abtin.ijadimaghsoodi@waikato.ac.nz*

ABSTRACT

This study introduces the Multi-LLM Optimised Joint Objective (MOJO), a novel generative artificial intelligence (Gen-AI) driven multi-criteria decision analysis (MCDA) framework. MOJO uniquely integrates human expertise with multiple large language models (LLMs) to address limitations in traditional MCDA approaches. The information fusion framework combines risk-adjusted utility scoring, performance equilibrium, and reliability scoring, producing consensus-driven rankings that effectively account for uncertainty and non-linear risk preferences. Grounded in principles from prospect theory and utility theory, MOJO captures complex behavioural factors such as loss aversion and probability distortion. An ensemble of LLM outputs automates MCDA tasks, enhanced by systematic bias correction to reduce cognitive load and improve transparency. MOJO also incorporates a comprehensive illustrative case study evaluating treatments for Inflammatory Bowel Disease (IBD), demonstrating its applicability to complex scenarios involving clinical, operational, and ethical criteria. Results from validation experiments involving multiple LLMs and decision-makers across diverse scenarios highlight MOJO's robustness, reliability, and transparency, demonstrating its significant potential in high-stakes fields such as healthcare, engineering, and public policy.

KEYWORDS

Multi-Criteria Decision Analysis (MCDA);
Large Language Model (LLM);
Artificial Intelligence;
Generative AI (Gen-AI);
Decision Bias.

1. INTRODUCTION AND BACKGROUND

Multi-Criteria Decision Analysis (MCDA) provides structured methodologies for evaluating multiple conflicting criteria to enable rational and justifiable decisions [1, 2]. Traditional methods such as the Analytic Hierarchy Process (AHP) [3], Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [4], VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) [5], and Multi-Objective Optimisation on the basis of Ratio Analysis plus the full MULTIplicative form (MULTIMOORA) [6, 7] have been widely used across domains such as engineering [8], healthcare [9], and policy [10]. Despite their effectiveness, these methodologies rely heavily on subjective human judgement, leading to biases, inconsistencies, and difficulties in integrating qualitative criteria into decision models [11, 12]. In recent years, hybrid MCDA methods have been developed, incorporating elements of fuzzy logic [13, 14], stochastic modelling [15], and other innovative approaches such as Axiomatic Design [8, 16] and Concept of Stratification [17, 18] to enhance their robustness and adaptability. However, challenges related to decision stability, expert dependence, and computational complexity persist, necessitating the integration of tools and techniques such as artificial intelligence (AI) to improve automation, transparency, and interpretability in decision-making processes.

The emergence of Large Language Models (LLMs) and Generative AI (Gen-AI) has significantly altered decision-making and data processing landscapes [19-22]. These models, trained on vast amounts of structured and unstructured data, have demonstrated advanced reasoning, predictive analytics, and natural language understanding, making them powerful tools in supporting decision-making processes [20, 23]. LLMs, including examples such as OpenAI ChatGPT, Google Claude, Meta Llamma, and xAI Grok, have proven effective in providing real-time insights, automating complex analytical tasks, and reducing cognitive load for human decision-makers, by analysing vast amounts of qualitative and quantitative information, these models are well-suited for integration into MCDA frameworks, thereby enhancing interpretability, automation, and scalability [24].

Recent studies indicate that LLMs can function not only as passive information retrieval systems but also as autonomous agents capable of performing high-level reasoning and optimising decision-making [19]. These models have already demonstrated their capacity for structured reasoning and debate tasks [25]. Recent studies have explored their potential in MCDA, primarily by automating pairwise comparisons, refining criteria weights, and validating expert-driven decisions. Svoboda and Lande [26] demonstrated an integration of AHP with OpenAI ChatGPT4 in cybersecurity, where the model efficiently performed pairwise comparisons and weight evaluations, thereby reducing human effort while maintaining reliability. Wang and Wu [27] used OpenAI ChatGPT for a supplier evaluation problem, showing that when properly prompted, LLMs could replicate the results from an expert-driven AHP with high consistency. Wang et al. [28] further expanded the scope by testing multiple LLMs in diverse decision contexts, demonstrating that fine-tuned LLMs could achieve human-expert-level decision accuracy across different domains. While very few studies validate the feasibility of Gen-AI-driven MCDA, they

also reveal critical limitations, particularly in bias mitigation, handling of qualitative data, and ensuring AI-human consensus [29, 30].

A major challenge in existing decision-making frameworks using AI is bias [29]. LLMs inherit biases from their training data, potentially skewing decisions in ways that are difficult to detect or correct[30]. Although it might be assumed that AI-generated weights and rankings align with human reasoning, systematic bias correction mechanisms remain underdeveloped. Additionally, while traditional MCDA methods struggle with qualitative data integration [11, 31, 32], LLMs offer a way to process textual inputs; however, ensuring consistency and transparency in qualitative assessments remains unresolved. Studies have shown that LLM-based decisions can vary based on prompt phrasing, leading to instability in rankings [25, 27]. Ensuring robustness in AI-assisted MCDA is another key gap, as minor variations in input can cause large fluctuations in rankings. Prior work attempted to stabilise rankings by using ensemble methods and voting strategies, but a structured approach for systematically evaluating decision robustness is still missing [27]. Furthermore, existing AI-MCDA methods often function as black-box solutions, lacking transparency in how decisions are derived [29]. This absence of explainability reduces trust in AI-assisted decision-making, as stakeholders cannot easily validate the reasoning behind AI-generated recommendations. Lastly, human-AI collaboration in MCDA remains an open challenge. Most frameworks treat AI as either a decision-maker or an advisory tool, without structured mechanisms for achieving AI-human consensus.

To address the gaps, this study introduces the Multi-LLM Optimised Joint Objective (MOJO) method, a novel Gen-AI MCDA framework designed to improve bias mitigation, qualitative reasoning, decision stability, and human-AI collaboration. MOJO introduces a bias detection and correction mechanism that systematically quantifies discrepancies between human and AI-generated decision inputs, dynamically adjusting weighting functions to minimise distortions. It integrates structured approaches for qualitative reasoning, ensuring that textual data is consistently interpreted across decision scenarios. Unlike traditional augmentation of AI and MCDA models that rely on single-run AI outputs, MOJO employs an ensemble of LLM agents with built-in uncertainty quantification, ensuring decision robustness by aggregating multiple AI-generated perspectives. MOJO incorporates a consensus-building process, allowing AI-generated recommendations to be adjusted dynamically based on human feedback, thereby bridging the gap between fully automated and human-driven decision-making. MOJO is the first framework to introduce a practical Gen-AI-driven MCDA model that systematically integrates LLMs as active agents in both structured decision-making and bias correction, setting a new standard for AI-driven decision support systems. Unlike prior methods that either automate specific MCDA tasks or rely on AI to replicate existing models, MOJO fundamentally transforms the decision process by integrating a multi-layered, bias-aware, and explainable approach. This pioneering study establishes MOJO as the first Gen-AI-MCDA framework that holistically optimises decision-making with LLMs, addressing both theoretical limitations and practical constraints, ensuring fairness, robustness, and adaptability across diverse application domains.

The remainder of this study is structured as follows. Section 2 presents the fundamentals and background of the proposed methodology. In Section 3, a patient treatment evaluation case study is presented, which considers clinical, operational, and ethical aspects of human-AI collaboration across multiple scenarios. While Section 4 provides a discussion on the proposed framework, and outlines the limitations of the study alongside future research directions, Section 5 concludes the study with summarised findings and final clarifications.

2. MOJO: MULTI-LLM OPTIMISED JOINT OBJECTIVE FRAMEWORK

The Multi-LLM Optimised Joint Objective (MOJO) method is a two-layer MCDA approach that enhances decision robustness by integrating human expertise and LLM insights. The first layer is the Risk-adjusted Utility scoring (RU), is deeply rooted in Prospect Theory and classical Utility Theory, capturing nonlinear risk behaviours, loss aversion, and probability weighting [33, 34]. The second layer, the Performance Equilibrium and Reliability Score (PERS), introduces a novel equilibrium-based approach that ensures balanced performance across all evaluation criteria while penalising alternatives that exhibit large disparities between criteria. By combining these, MOJO mitigates bias and prioritises high-consensus outcomes, preserving the strengths of both subjective and objective evaluations for rational, risk-aware decision-making.

2.1. PRELIMINARIES AND FUNDAMENTALS OF MOJO METHODOLOGY

2.1.1. DECISION MATRIX

Let $A_i: \{a_1, a_2, \dots, a_n\}$ denote a set of alternatives, $C_j: \{c_1, c_2, \dots, c_m\}$ a set of evaluation criteria, and $D_k: \{d_1, d_2, \dots, d_k, l_1, l_2, \dots, l_l\}$ a set of decision-makers, comprising both human experts and LLMs. Each decision-maker assigns scores to each alternative under each criterion, forming the decision matrix X_{ij} ($i = 1,2, \dots n$ and $j = 1,2, \dots m$) as Eq. (1), in which criteria weights $W_j = [w_j]_n$ is $\sum w_j = 1$.

$$\mathrm{X_{ij}} = \begin{bmatrix} \mathrm{x_{11}^{d_1}} & \mathrm{x_{12}^{d_1}} & \cdots & \mathrm{x_{1m}^{d_1}} \\ \mathrm{x_{21}^{d_2}} & \mathrm{x_{22}^{d_2}} & \cdots & \mathrm{x_{2m}^{d_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{x_{n1}^{l_1}} & \mathrm{x_{n2}^{l_1}} & \cdots & \mathrm{x_{nm}^{l_1}} \end{bmatrix} \tag{1}$$

For each criterion C_j, two types of weights are defined as Objective Weight W_j^o and Subjective Weight W_j^s. The objective weight W_j^o is derived from observed data. A common method to calculate objective weights is the Shannon's Entropy method. Objective weight W_j^o using Shannon's Entropy can be calculated using Eq. (2).

$$\mathrm{W_j^o} = \frac{1 - \mathrm{e_j}}{\sum_{\mathrm{j=1}}^{\mathrm{m}} (1 - \mathrm{e_j})} \text{ where } \mathrm{e_j} = \frac{1}{\ln(\mathrm{n})} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{p_{ij}} \ln(\mathrm{p_{ij}});\ \mathrm{p_{ij}} = \frac{\mathrm{X_{ij}}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_{ij}}} \tag{2}$$

Furthermore, the subjective weights W_j^S can be calculated on a scaling system provided by human experts and LLM outputs. In this study a simple form for this calculation was utilised, in which suppose each criterion is assigned a raw subjective score s_j, these scores can be normalised utilising Eq. (3), ensuring that $\sum_{j=1}^{m} W_j^S = 1$.

$$W_j^S = \frac{s_j}{\sum_{j=1}^{m} s_j} \tag{3}$$

The combined weight for each criterion based on W_j^O and W_j^S is calculated as Eq. (4), with adaptive factors λ and η balancing the influence of objective and subjective inputs upon decision-makers preferences.

$$W_j = \lambda W_j^O + \eta W_j^S \tag{4}$$

2.1.2. BIAS RATE CALCULATION AND CORRECTION

To quantify the bias between human scores and LLM-generated scores for each evaluation criterion, we define a bias rate $\mathrm{Bias_j}$ measure. This metric provides a numerical measure of the discrepancy between the two sources of assessment as demonstrated in Eq. (5).

$$\mathrm{Bias_j} = \frac{1}{k}\sum_{d=1}^{k}\left|\frac{H_{jd}-L_{jd}}{H_{jd}+\varepsilon}\right| \times e^{-\sigma_j} \text{ where } L_{jd} = \frac{\sum_{l=1}^{l}\phi_l X_{jl}^{L_i}\cdot acc_{jl}}{\sum_{l=1}^{l}\phi_l} \tag{5}$$

Where H_{jd} represents the aggregated human score for a specific criterion C_j, which is typically derived from multiple independent human experts and can be averaged or weighted based on reliability. L_{jd} denotes the corresponding aggregated score assigned by the LLM, in which ϕ_l represents the LLM confidence weight, a crucial component that reflects the model's certainty in its own assessments. This confidence can be derived from historical calibration data or the probability scores that the LLM generates when making predictions. $X_{jl}^{L_i}$ is the raw LLM-generated score for criterion C_j, and acc_{jl} is the accuracy adjustment factor, designed to compensate for systematic biases or inaccuracies in LLM-generated scores, which may be computed by performance benchmarks on similar tasks and/or cross-validation procedures that assess the LLM historical accuracy compared to human judgments. It should be noted that by incorporating acc_{jl}, MOJO ensures that any systematic overestimation or underestimation in the LLM's scores is accounted for before computing the bias rate.

ε is a small constant introduced to prevent division by zero, ensuring numerical stability. This is crucial in cases where the evaluation criteria allow for zero scores (e.g., if a criterion is completely unaddressed or not applicable in a particular instance). σ_j serves as a damping factor to control the sensitivity of the bias calculation. This factor can be determined empirically, based on historical score distributions or expert judgment. While the damping factor σ_j can be adjusted based on the past evaluations, ensuring that the bias rate is neither overly sensitive nor too lenient, a statistical approach can be utilised by computing the standard deviation of bias rates across different instances of each C_j. Furthermore, in practical cases, σ_j is manually tuned based on practical experience with the scoring system,

ensuring it effectively smooths out extreme values without removing meaningful bias signals. The damping factor is then applied as in exponential decay $e^{-\sigma_j}$ by reducing the influence of excessive fluctuations in bias calculations. Utilising this factor within MOJO is optional.

A threshold is then set such that if $Bias_j > \beta = 0.1$ to determine whether a criterion exhibits significant bias. If the bias rate for a criterion exceeds this threshold (i.e., if the average relative difference between human and LLM scores is greater than 10%), it is flagged as significantly biased. The choice of $\beta = 0.1$ is based on empirical studies and practical experience with ensemble evaluation methods. It represents a reasonable balance between sensitivity and robustness, ensuring that small deviations do not trigger unnecessary adjustments while still capturing meaningful discrepancies. This threshold may be context-dependent, meaning it can be fine-tuned based on domain-specific requirements. In some applications, a slightly higher bias threshold, up to $\beta = 0.3$, may be considered acceptable, particularly in subjective evaluations where inherent variability between human decision-makers is expected. However, setting β above this level is generally not advisable, as it risks allowing significant bias to go uncorrected, potentially leading to systematic distortions in decision-making.

Accordingly, for criteria identified as high-bias, a corrective scaling mechanism is applied to mitigate the impact of biased evaluations. The subjective weight of criteria is adjusted based on Eq. (6). This adjustment reduces the weight of highly biased criteria in the final decision-making process, ensuring that their influence is proportionally diminished. The underlying rationale is that a criterion exhibiting substantial bias should contribute less to the overall assessment, preventing skewed evaluations. This approach helps preserve the credibility of automated assessment models while maintaining alignment with human judgment.

$$\mathrm{W_j} = \mathrm{W_j^{initial}} \times (1 - \mathrm{Bias_j}) \tag{6}$$

2.1.3. REFERENCE POINT ALIGNMENT AND RISK PERCEPTION BALANCING

Risk perception plays a crucial role in decision-making and is shaped by the reference point from which gains and losses are assessed. In MOJO, both humans and LLMs may establish different reference points when evaluating a given criterion. This discrepancy can lead to inconsistencies in how risks, rewards, and overall outcomes are perceived. In order to quantify this difference the human-defined reference point r_j^H for criterion C_j, reflecting expert judgment or subjective experience, and r_j^L representing the LLM-defined reference point, which is typically derived from statistical patterns, historical data, or probabilistic estimations, are utilised as Eq. (7).

$$\Delta \mathrm{r_j} = \left| \mathrm{r_j^H} - \mathrm{r_j^L} \right| \tag{7}$$

If the absolute difference $\Delta \mathrm{r_j}$ exceeds a predefined threshold $\delta_r = 5$, it indicates a significant misalignment in risk perception between human decision-makers and the LLM. This misalignment can lead to inconsistencies in

decision outcomes, particularly in high-stakes assessments. To reconcile these differences, we compute a harmonised reference point using a weighted combination of the two perspectives as demonstrated in Eq. (8).

$$\mathrm{r_j^{final}} = \omega \mathrm{r_j^H} + (1 - \omega)\mathrm{r_j^L} \quad (8)$$

Where ω is a balancing parameter that controls the weight assigned to the human versus LLM reference points. The value of ω is typically chosen based on empirical trust in human judgment relative to LLM-based assessments. If human intuition and expertise are deemed more reliable in a particular domain (e.g., medical or ethical evaluations), a higher ω (e.g., 0.7 or more) is used. If the LLM has demonstrated superior consistency and calibration (e.g., quantitative financial modelling), a lower ω (e.g., 0.3 to 0.5) may be appropriate. For small differences (i.e., when Δr_j is within an acceptable range), a simpler approach can be used, such as averaging the two reference points ($r_j^{final} = \frac{r_j^H + r_j^L}{2}$), assuming equal trust in both perspectives, or selecting one reference point (e.g., always prioritising human-defined values in cases where domain expertise is critical).

Furthermore, the choice of the threshold $\delta_r = 5$ is based on empirical evaluations and domain-specific requirements. In some contexts, a higher threshold may be acceptable, particularly if reference points naturally fluctuate within a wider range. However, exceeding 30% of the total range (e.g., if reference points vary between 0 and 100, a δ_r above 30 would be excessive) is discouraged, as it may allow substantial inconsistencies to persist. By carefully balancing human intuition and LLM-based estimations, this approach ensures that risk perception remains consistent, context-aware, and aligned with the most reliable source of judgment for each specific criterion.

2.1.4. HUMAN-AI CONSENSUS AND AGREEMENT SCORE

An important part of MOJO is ensuring that a reliable level of agreement between human and the LLM-generated assessments is established. A strong alignment between the two suggests confidence in the evaluation process, whereas significant disagreement may indicate uncertainty or systematic biases in either the human or AI judgment. To quantify this alignment, we define a consensus score $Cons(A_i)$ for each alternative A_i, which captures the degree of agreement between human and LLM evaluations as demonstrated in Eq. (9).

$$\mathrm{Cons(A_i)} = 1 - \frac{|\mathrm{H_i} - \mathrm{L_i}|}{\max(\mathrm{H_i}, \mathrm{L_i})} \quad (9)$$

Where $\mathrm{H_i}$ represents the human-evaluated score for alternative A_i, reflecting expert judgment, and $\mathrm{L_i}$ represents the LLM-generated score for the same alternative. The denominator $\max(\mathrm{H_i}, \mathrm{L_i})$ ensures that the score is normalised relative to the larger of the two values, preventing distortions in cases where one score is significantly higher than the other. The resulting score $Cons(A_i)$ ranges between 0 and 1, where 1 (100%) represents perfect agreement between human and LLM scores, 0 indicates maximum disagreement, and values closer to 1 suggest strong alignment, reinforcing confidence in the assessment. While minor variations between human and LLM

evaluations are expected, significant discrepancies need further attention. To address this, a threshold of 85% agreement is set in which if $Cons(A_i)$ falls below 0.85, this indicates a moderate disagreement between the two assessments. When this happens, a penalty factor of 0.9 is applied to further reduce the weight of the consensus score $\mathrm{Cons}(\mathrm{A_i}) = 0.9 \times \mathrm{Cons}(\mathrm{A_i})$.

This penalty reduces the influence of alternatives where agreement is lower, ensuring that uncertain or conflicting evaluations do not disproportionately impact the final decision-making process. The logic behind this penalty is rooted in ensemble learning principles, where disagreements among models (or decision-makers) typically indicate higher uncertainty and warrant reduced influence on the final outcome [35]. The 85% agreement threshold is based on empirical studies in ensemble learning methods, which suggest that when multiple evaluators (whether human or AI) agree at or above this level, the decision is likely to be reliable and robust [36]. Below this threshold, discrepancies become more pronounced, leading to potential inconsistencies in evaluation, and there is a greater need for further review to understand whether the disagreement is due to bias, calibration issues, or subjective variability in human assessments. The penalty factor of 0.9 is similarly chosen based on practical findings, ensuring that lower-consensus alternatives contribute less to the overall decision framework without being entirely discarded. This value can be adjusted based on domain-specific needs, but empirical studies have found that reducing the weight by 10% strikes a good balance between accounting for disagreement and avoiding excessive discounting of otherwise valid evaluations [37].

While the default values of 85% for consensus and 0.9 for penalty scaling provide a strong foundation, they are not fixed and may be modified depending on the application. In high-stakes domains (e.g., medical diagnoses, legal decision-making), a higher threshold (e.g., 90–95%) may be preferred to ensure stricter alignment before proceeding with an assessment. Conversely, in more subjective or creative fields, where variability in scoring is expected, a lower threshold (e.g., 80% or even 75%) may be acceptable. The penalty factor could be adjusted dynamically, where larger disagreements (e.g., below 75%) result in a greater penalty (e.g., 0.8 instead of 0.9), further reducing the weight of highly uncertain evaluations. By quantifying agreement between human and AI assessments through the Consensus Score and applying a penalty for moderate disagreement, this approach enhances the reliability and interpretability of decision-making. It ensures that assessments with strong human-LLM alignment are prioritised, while cases with notable divergence are de-emphasised to mitigate potential biases and inconsistencies.

2.1.5. LAYER 1: RISK-ADJUSTED UTILITY SCORING SYSTEM

This study proposes a novel integration of Prospect Theory and Utility Theory through the Risk-adjusted Utility (RU), extending prior theoretical models to better align with human decision-making in uncertain environments. Unlike previous models that rely solely on either Prospect or Utility Theory, this approach synthesises key elements from both, capturing probability distortions, loss aversion, and reference dependence while maintaining a structured utility-based scoring system. The RU Score introduces a harmonised reference point, to ensure calibrated risk

assessments, aligning human and LLM perspectives before computing final utility-based scores. This adaptation enhances the applicability of behavioural risk modelling beyond traditional decision frameworks. This approach bridges the gap between purely rational utility-based methods and behaviourally grounded decision models. By integrating probability distortions and reference dependence within a structured scoring system, the RU Score provides a more holistic framework for assessing risk.

The RU score is designed to propose an integration of the Prospect Theory, which accounts for behavioural biases [33, 34], with the classical Utility Theory, ensuring that risk perceptions are captured in a way that aligns with human decision-making tendencies [38]. Unlike traditional utility-based scoring, which assumes rational decision-makers maximise expected value, this approach adjusts for risk distortions, accounting for how humans tend to overweight or underweight probabilities and exhibit loss aversion. By incorporating the principles of Prospect Theory, the RU score reflects how individuals perceive gains and losses asymmetrically, leading to more realistic decision-making models in uncertain environments. This integration ensures that the scoring mechanism is sensitive to both objective utility and subjective risk perception, offering a more behaviourally grounded approach. For each alternative A_i, the RU Score is computed as Eq. (10).

$$\mathrm{RU}(\mathrm{A_i}) = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{W_j} \sum_{\mathrm{d=1}}^{\mathrm{k+1}} \pi\left(\mathrm{p_{ij}^d}\right) \cdot \upsilon\left(\mathrm{X_{ij}^d}\right) \tag{10}$$

Where $\mathrm{W_j}$ is the weight assigned to C_j. p_{ij}^d represents the probability associated with outcome d criterion C_j, which can be derived from expert uncertainty estimates, historical frequencies, or scenario-based approaches. The function $\pi\left(p_{ij}^d\right)$ represents probability weighting, which adjusts how probabilities are perceived by decision-makers, while $\upsilon\left(X_{ij}^d\right)$ represents the value function, which evaluates the perceived utility of an outcome based on gains and losses relative to a reference point. Humans tend to distort probabilities when making decisions under uncertainty. To account for this, a probability weighting function $\pi\left(p_{ij}^d\right)$ can be calculated as Eq. (11).

$$\pi(\mathrm{p}) = \frac{\mathrm{p}^{\gamma}}{(\mathrm{p}^{\gamma} + (1-\mathrm{p})^{\gamma})^{\frac{1}{\gamma}}} \tag{11}$$

In which $\gamma; (0 < \gamma < 1)$ determines the degree of probability distortion. When γ is small, probabilities are heavily distorted, meaning small probabilities are overweighted (e.g., people overestimate rare events like winning the lottery) and large probabilities are underweighted (e.g., people underestimate near-certain outcomes). When γ is closer to 1, probability weighting is closer to linear, meaning probabilities are interpreted more objectively. Introducing $\pi(p)$ ensures that the human-like perception of risk is reflected in the decision-making process, preventing low-probability but high-impact events from being ignored and ensuring high-probability events are not excessively discounted. Prospect Theory suggests that people evaluate outcomes relative to a reference point, rather than in absolute terms [34]. This is modelled by the value function demonstrated in Eq. (12)

$$\upsilon\left(X_{ij}^{d}\right) = \begin{cases} \left(X_{ij}^{d} - r_j^{final}\right)^{\alpha}, & \text{if } X_{ij}^{d} \geq r_j^{final} \\ -\lambda\left(r_j^{final} - X_{ij}^{d}\right)^{\beta}, & \text{if } X_{ij}^{d} < r_j^{final} \end{cases} \quad (12)$$

Where r_j^{final} is the harmonised reference point for criterion C_j, ensuring that both human and LLM risk perspectives are aligned. α and β determine the curvature of the value function, as introduced in prospect theory, capturing risk sensitivity. Typically, α and β are less than 1, meaning the perceived value of gains and losses follows a diminishing returns pattern. For losses, β is often even smaller (e.g., 0.1), reflecting that small losses are disproportionately weighted, making individuals more loss-averse than gain-seeking. $\lambda > 1$ represents loss aversion, meaning losses carry a greater psychological weight than equivalent gains. Empirical studies suggest that λ is typically around 2, meaning losing \$100 feels twice as bad as winning \$100 feels good. By integrating these elements, the RU Score accounts for human cognitive biases in decision-making. Risk distortion occurs because people overweight small probabilities and underweight large ones, reference dependence ensures outcomes are perceived as gains/losses relative to a reference point, and loss aversion ensures that losses carry a greater psychological weight than equivalent gains. The inclusion of the harmonised reference point r_j^{final} ensures that risk assessment is calibrated, aligning human and LLM perspectives before computing the final utility-based scores.

2.1.6. LAYER 2: PERFORMANCE EQUILIBRIUM & RELIABILITY SCORING SYSTEM

In MCDA, selecting the best alternative requires balancing performance reliability and consistency across evaluation criteria. A model that prioritises only high performance may favour alternatives with extreme but unstable outcomes, while one that overemphasises stability may undervalue high-performing options. To address this, the Performance Equilibrium & Reliability Score (PERS) integrates both reliability-weighted performance and dispersion-based penalties, ensuring a balanced evaluation. The score is defined as Eq. (13).

$$\text{PERS}(A_i) = \Psi\left(\widehat{E_i}, \Delta_i\right) = \widehat{E_i} - \alpha\Delta_i \quad (13)$$

Where $\widehat{E_i}$ is the reliability-weighted equilibrium measure, representing the overall quality of the alternative across all criteria. Δ_i captures performance imbalance across criteria, penalising erratic alternatives. The parameter α determines the degree to which performance imbalance influences the final score. A higher α increases the penalty for alternatives with significant variations across different evaluation aspects.

To compute the equilibrium score for each criterion C_j, a localised equilibrium score is defined as Eq. (14).

$$E_{ij} = \frac{\left(X_{ij} - X_j^{-}\right)^{\gamma}}{\left(X_j^{+} - X_j^{-}\right)^{\gamma} + \kappa} \quad (14)$$

Where X_{ij} is the performance of alternative A_i on criterion C_j. X_j^+ and X_j^- denote the best (maximum) and worst (minimum) observed performances on criterion C_j, respectively. These values can be derived empirically from the dataset or set based on theoretical bounds. The parameter γ controls the sensitivity of the equilibrium measure to differences in performance across alternatives. A higher γ makes the score more sensitive to performance gaps, meaning even small differences will result in larger variations in E_{ij}. A lower γ results in a smoother evaluation, reducing the impact of small discrepancies. The constant κ (typically 0.01) is introduced to prevent division by zero, ensuring numerical stability when $X_j^+ - X_j^-$ is very small.

To obtain the overall equilibrium score for an alternative A_i, the individual criterion scores E_{ij} are aggregated across all criteria, weighted by W_j (importance of the criterion) and R_j (reliability of the data for that criterion) as presented in Eq. (15).

$$\widehat{E_i} = \sum_{j=1}^{m} W_j \cdot R_j \cdot E_{ij} \tag{15}$$

Where W_j represents the relative importance weight assigned to criterion C_j, reflecting its contribution to the overall evaluation. R_j is a reliability factor that adjusts for the quality of data associated with criterion C_j. If the data quality is low (e.g., due to missing values or measurement uncertainty), R_j is assigned a lower value to reduce its influence on the final equilibrium score.

While the equilibrium score measures overall performance reliability, the dispersion factor Δ_i quantifies how balanced an alternative's performance is across different criteria. It is computed as the weighted standard deviation of the criterion-specific equilibrium scores as demonstrated in Eq. (16).

$$\Delta_i = \sqrt{\sum_{j=1}^{m} W_j \left(E_{ij} - \overline{E}_i\right)^2}, \text{with } \overline{E}_i = \frac{\sum_{j=1}^{m} W_j E_{ij}}{W_j} \tag{16}$$

Where $\overline{E_i}$ is the weighted mean equilibrium score of alternative A_i. A lower Δ_i indicates that the alternative performs consistently well across all criteria, meaning it is reliable without significant weak points. A higher Δ_i suggests imbalance, meaning the alternative excels in some areas but performs poorly in others, leading to penalisation in the final PERS score. By incorporating both performance reliability and consistency across criteria, the PERS Score provides a more robust and balanced evaluation, ensuring that alternatives with well-rounded and dependable performance are prioritised over those with erratic strengths and weaknesses.

2.1.7. MOJO ASSESSMENT SCORE

After computing RU and PERS for each alternative A_i, the final MOJO Score is formulated as a weighted integration of these two components as demonstrated in Eq. (17). This formulation ensures a balanced trade-off

between risk-aware decision-making and performance stability, where the relative importance of each factor is dynamically adjusted.

$$\mathrm{MOJO}(A_i) = \theta \mathrm{RU}(A_i) + (1-\theta)\mathrm{PERS}(A_i) \text{ where } \theta = 0.5 + 0.5 \times \mathrm{Cons}(A_i) \quad (17)$$

The integration factor θ which governs the relative influence of RU and PERS, is dynamically determined based on the level of agreement between human and LLM evaluations. This ensures that the impact of RU never falls below 50%, even when agreement between human and LLM assessments is low. This lower bound safeguards risk-awareness in decision-making (i.e., as human and LLM evaluations align more closely), preventing the MOJO Score from being overly dominated by performance stability. As consensus between human and LLM increases, the weight of RU increases proportionally. When consensus reaches 100%, $\theta = 1$, meaning the final decision is fully dictated by risk-adjusted utility, as there is no meaningful disagreement to resolve. Conversely, when agreement is lower, PERS acts as a corrective factor, ensuring that performance consistency remains a crucial consideration in cases where human and LLM evaluations diverge.

The choice of 85% as the consensus threshold and the 0.9 penalty factor applied in cases of lower consensus are derived from empirical studies in ensemble decision-making methods. These values have been chosen because they indicate strong alignment between multiple evaluators and help mitigate uncertainty in cases where human and LLM diverge significantly. By applying a penalty when consensus falls below this level, the model accounts for potential biases or inconsistencies while ensuring that no alternative is unfairly favoured due to misalignment in scoring. Once the MOJO Score has been computed for all alternatives, the optimal choice is determined by selecting the alternative that maximises the final MOJO Score according to Eq. (18).

$$A^* = \arg\max_{A_i}(A_i) \quad (18)$$

2.1.8. MOJO PSEUDOCODE AND SUMMARY

The Multi-LLM Optimised Joint Objective (MOJO) method integrates RU scoring and PERS, blending human expertise with LLM-driven assessments for robust, consensus-based decision-making. Bias rates between human and LLMs are measured, adjusting subjective weights where discrepancies exceed set thresholds. Reference points for risk-reward evaluations are aligned across human and AI perspectives to ensure consistency. The RU layer corrects cognitive biases via Prospect Theory, while PERS assesses stability and balanced performance. Consensus levels between human and AI scores influence the weighting of RU and PERS, with higher alignment reinforcing risk-aware decision-making. The final MOJO score adaptively integrates risk-adjusted utility and performance reliability, ensuring an objective, rational, and context-sensitive ranking. The pseudocode follows below, with a visual framework shown in Figure 1.

Input: A set of alternatives A_i: $\{a_1, a_2, \ldots, a_n\}$, criteria C_j: $\{c_1, c_2, \ldots, c_m\}$, human D_H and LLM D_L decision-makers D_k: $\{d_1, d_2, \ldots, d_h, l_1, l_2, \ldots, l_l\}$.
Output: Ranked alternatives based on MOJO scores.

1. Compute the decision matrix $X_{ij} \leftarrow \text{CollectScores}(A_i, C_j, D_H, D_L)$
2. Compute criterion weights: $W_j^{\text{initial}} \leftarrow \lambda W_j^o + \eta W_j^s$
3. Compute bias rates and apply corrections: $\text{Bias}_j \leftarrow \frac{1}{k}\sum_{d=1}^{k}\left|\frac{H_{jd}-L_{jd}}{H_{jd}+\varepsilon}\right| \times e^{-\sigma_j}$
4. Apply weight corrections: $W_j \leftarrow W_j^{\text{initial}} \times (1 - \text{Bias}_j)$
5. Align reference points if $\Delta r_j > \delta_r$: $r_j^{\text{final}} \leftarrow \omega r_j^H + (1-\omega) r_j^L$
6. Compute Risk-Adjusted Utility (RU): $\text{RU}(A_i) \leftarrow \sum_{j=1}^{m} W_j \sum_{d=1}^{k+1} \pi(p_{ij}^d) \cdot \upsilon(X_{ij}^d)$
7. Compute Human-AI Consensus: $\text{Cons}(A_i) \leftarrow 1 - \frac{|H_i - L_i|}{\max(H_i, L_i)}$
8. Compute Integration Factor: $\theta \leftarrow 0.5 + 0.5 \times \text{Cons}(A_i)$
9. Compute Final MOJO Score: $\text{MOJO}(A_i) \leftarrow \theta \text{RU}(A_i) + (1-\theta)\text{PERS}(A_i)$
10. Rank alternatives and determine the optimal choice: $A^* \leftarrow \arg\max_{A_i}(A_i)$
11. **Return** the ranked list of alternatives A_i sorted in descending order of MOJO(A_i).

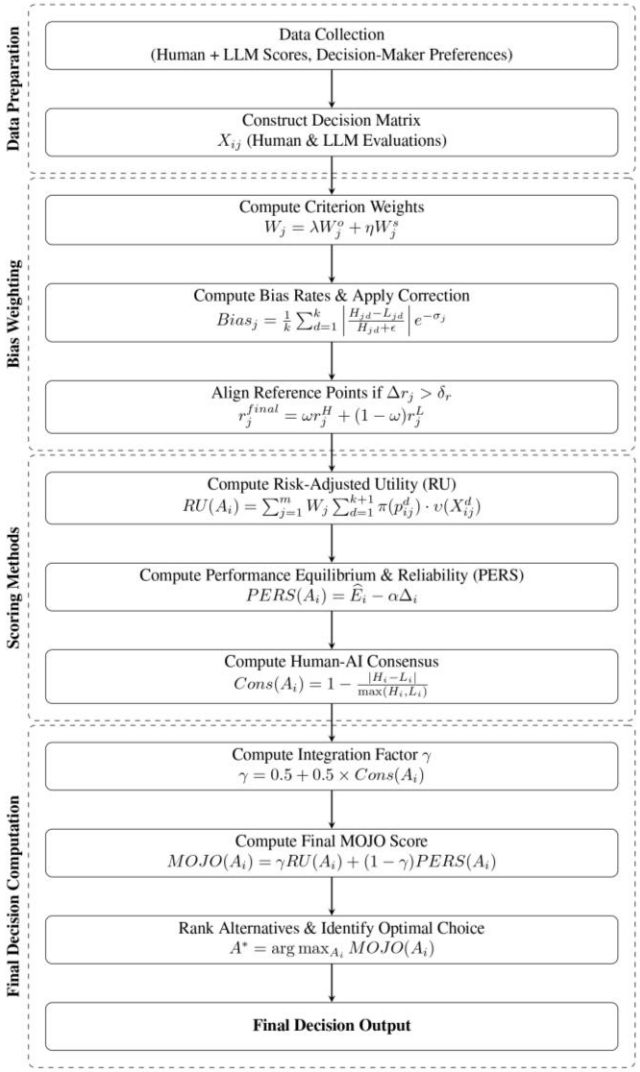


FIGURE 1. Multi-LLM Optimised Joint Objective (MOJO) Visual Framework.

3. AN ILLUSTRATIVE EXAMPLE: PATIENT TREATMENT EVALUATION

3.1. CONTEXT DEFINITION

Inflammatory bowel diseases (IBDs), including Ulcerative Colitis and Crohn Disease, pose significant challenges in clinical decision-making due to their complex, multifactorial nature. Patients with these conditions often face a long-term battle with persistent symptoms, adverse drug reactions, and uncertainties associated with variable treatment responses [39]. Selecting an optimal treatment requires a delicate balance among clinical efficacy, safety, patient quality of life, and economic considerations [40]. Traditionally, clinicians have relied on evidence from randomised controlled trials, clinical guidelines, and their own expertise. However, in an era of rapidly advancing data driven decision making, the integration of advanced decision-support systems offers a transformative approach by combining human clinical insight with the analytical capabilities of AI [41].

In this conceptual case study, we illustrate how MOJO can be applied to evaluate and rank treatment alternatives for Ulcerative Colitis and Crohn Disease. A simulated dataset was constructed from hypothetical clinical trial results, published literature, and expert assessments to serve as the foundation for this evaluation. It is critical to note that no real patient records or actual clinical data have been used; this study is purely demonstrative, intended solely to showcase the capabilities of MOJO in a controlled, conceptual setting.

The following patient scenarios illustrate three distinct cases of IBD, encompassing both Crohn's Disease and Ulcerative Colitis, which will be used to demonstrate the application of MOJO in clinical decision-making for treatment evaluation. Each scenario provides a detailed patient history, including clinical presentations, diagnostic test results, and current challenges. These narratives serve as the basis for subsequent analysis in our medicine evaluation, highlighting the diverse clinical contexts encountered in real-world settings.

- **Scenario 1: GP-Referred Patient Needing Attention** - Patient A is a 45-year-old individual with a background of mild gastrointestinal discomfort and intermittent flare-ups suggestive of Ulcerative Colitis. Recently referred by their General Practitioner (GP), the patient presents with a two-month history of moderate abdominal pain, occasional bloody diarrhoea, and altered bowel habits. Preliminary laboratory tests revealed mildly elevated inflammatory markers (C-reactive protein and erythrocyte sedimentation rate), and a recent sigmoidoscopy indicated patchy mucosal erythema with superficial ulcerations. Despite stable vital signs and no signs of systemic infection, the GP recommended a comprehensive evaluation to rule out disease progression. Patient A is cooperative and motivated for further diagnostic workup and treatment optimisation.
- **Scenario 2: Follow-Up Patient with Suboptimal Response** - Patient B, a 38-year-old individual diagnosed with Crohn's Disease for the past two years, has been under regular follow-up care. Recently, despite strict adherence to the current immunomodulatory therapy, the patient reports an increase in symptoms including persistent diarrhoea, abdominal cramping, and unintentional weight loss. A colonoscopy performed during the latest evaluation revealed deep linear ulcers and mild structuring in the terminal ileum, while blood tests showed a significant rise in inflammatory markers. Concerned by the declining therapeutic response and potential for disease complications, the treating clinician has decided to re-assess the treatment plan. This case highlights the need for timely adjustments in long-term management of chronic IBD.

- **Scenario 3: Patient with Severe Rectal Bleeding in ED** - Patient C is a 52-year-old individual with a known history of Ulcerative Colitis who presented to the Emergency Department (ED) with acute, severe rectal bleeding, dizziness, and profound fatigue. Upon arrival, the patient was hemodynamically unstable, with a low blood pressure and elevated heart rate. An urgent colonoscopy revealed extensive mucosal ulceration and diffuse bleeding in the left colon, while laboratory results indicated a significant drop in haemoglobin levels and evidence of ongoing blood loss. Immediate stabilisation measures were initiated, and the patient was admitted for emergency care. This scenario underscores the critical need for rapid decision-making and intervention in severe IBD flare-ups.

Accordingly, the decision-making process involves a panel of five human experts, including clinicians, senior nurse practitioners and specialists in gastroenterology, whose insights are enriched by the assessments of three LLMs. Specifically, the LLMs employed in this study are (a) OpenAI ChatGPT o3-mini, (b) xAI Grock, (c) Meta LLaMA 3.1 70B. Access to all LLMs were provided via their official APIs. Furthermore, the APIs for these LLMs were accessed and integrated using R. The implementation leveraged several key packages such as httr, reticulate, LangChain, ggplot2, TidyR, dplyr and ellmer. Numerical computations were handled using base R functions and supplementary libraries as needed. An interactive Shiny app was developed to allow users to explore these findings in real time. Furthermore, to comprehensively assess the alternatives, nine critical decision criteria were identified. These criteria encompass clinical effectiveness, safety, patient-centric factors, and economic implications, ensuring a holistic evaluation of each treatment. While Table 1 present detailed descriptions of the decision criteria, treatment alternatives were presented in Table 2. It is imperative to reiterate that this study is purely conceptual, utilising a simulated dataset without any real-world patient data, and is intended solely to demonstrate the capabilities of combining advanced MCDA with LLMs.

TABLE 1. Decision Criteria for Treatment Evaluation.

ID	Criterion	Description and Rationale
C_1	Efficacy	Measures the treatment ability to induce and maintain remission. This criterion is based on clinical endpoints such as symptom reduction and mucosal healing, which are critical for long-term disease management.
C_2	Safety Profile	Assesses the overall safety of the treatment by evaluating the incidence, severity, and reversibility of adverse events. Safety is paramount in chronic conditions where long-term medication use is common.
C_3	Side Effects	Evaluates the range and impact of treatment-related side effects on patients' daily activities, taking into account both acute and chronic complications.
C_4	Cost	Considers the economic burden of treatment, including direct drug costs and indirect expenses such as monitoring and hospitalisations. This criterion is essential for both patients and healthcare systems.
C_5	Patient Compliance	Reflects the ease of adherence to the treatment regimen, factoring in the mode of administration, dosing frequency, and overall treatment complexity.
C_6	Dosing Convenience	Examines the simplicity and practicality of the dosing schedule, assessing whether the regimen is sustainable and patient-friendly in real-world settings.
C_7	Long-Term Outcomes	Evaluates the durability of treatment benefits, including the duration of remission and the prevention of disease progression over time.
C_8	Quality of Life Impact	Measures how the treatment affects patients' overall well-being, daily functioning, and mental health, which are critical for chronic disease management.
C_9	Innovation and Novelty	Considers the potential of a treatment to introduce breakthrough mechanisms of action or novel therapeutic approaches that could offer advantages over established therapies.

TABLE 2: Treatment Alternatives Evaluated.

Alternative ID	Treatment	Description and Regulatory Status
A_1	Mesalamine	A first-line, FDA-approved anti-inflammatory agent widely used for mild to moderate Ulcerative Colitis. It acts topically on the colonic mucosa and has a well-established safety profile.
A_2	Infliximab	An FDA-approved monoclonal antibody targeting TNF-α, indicated for moderate to severe disease. Supported by robust clinical evidence, it may be associated with infusion-related adverse events targets gut-specific lymphocyte trafficking, resulting in a favourable safety and efficacy profile.
A_3	Adalimumab	A self-administered, FDA-approved anti-TNF agent offering convenience over infusion therapies, though patient response can vary.
A_4	Vedolizumab	An FDA-approved integrin receptor antagonist that selectively targets gut-specific lymphocyte trafficking, resulting in a favourable safety and efficacy profile.
A_5	Tofacitinib	An oral Janus kinase (JAK) inhibitor approved for Ulcerative Colitis, noted for its rapid onset of action but associated with potential risks such as thromboembolic events.
A_6	Ustekinumab	An FDA-approved interleukin inhibitor primarily used for Crohn Disease, providing a novel therapeutic option for patients unresponsive to anti-TNF therapies.
A_7	Certolizumab Pegol	Although primarily approved for other inflammatory conditions, its emerging use in IBD is being explored in a trial context to assess potential benefits.
A_8	TrialMed AXG	An investigational treatment currently under clinical evaluation. Preliminary data suggest that this novel agent may offer improved disease modulation with a unique mechanism of action.
A_9	TrialMed BFT	An experimental therapy in early-phase trials, designed to optimise the balance between efficacy and safety through a distinct pharmacological approach.
A_{10}	TrialMed CIX	A cutting-edge trial medication that combines innovative drug delivery systems with targeted immunomodulatory effects, currently being evaluated for its potential in both Ulcerative Colitis and Crohn Disease.

3.2. FINDINGS AND RESULTS

The evaluation began with the collection of scores from five human experts and three LLMs across all ten treatment alternatives against the nine established criteria. To ensure reproducibility, all LLM assessments were obtained using standard prompting techniques with identical input parameters across all scenarios. For each patient scenario, the raw scores (on a scale of 0-10) were aggregated to form a comprehensive decision matrix. Tables 3 present the aggregated scores from both human experts and LLMs across all three clinical scenarios. Several notable patterns emerge from the very initial assessments. For Patient A, conventional treatment options (Mesalamine) scored highly on safety and side effect profiles, while more advanced biologics and trial medications demonstrated superior efficacy and innovation scores. In the case of Patient B, a clear differentiation is observed between first-line therapies (e.g., Mesalamine) which scored poorly on efficacy (4.5 by human experts, 4.2 by LLMs) and more targeted biologics like Ustekinumab which received substantially higher efficacy ratings (8.4 by human experts, 8.7 by LLMs). For Patient C (severe acute bleeding), the assessment revealed a pronounced preference for rapid-acting treatments with high efficacy scores, particularly Infliximab (9.1 by human experts) and the trial medications (TrialMed CIX scoring 9.0 and 9.3 from human experts and LLMs respectively). Across all scenarios, it was observed that LLMs tended to assign slightly higher scores to novel treatments and slightly lower scores to established therapies compared to human experts, potentially reflecting a bias toward innovation or a different weighting of historical evidence.

TABLE 3: Integrated Assessment Scores in a Decision Matrix.

Treatment Alternatives	Source	C_1: Efficacy	C_2: Safety	C_3: Side Effects	C_4: Cost	C_5: Compliance	C_6: Dosing	C_7: Long-Term	C_8: Quality of Life	C_9: Innovation
A1: Mesalamine	Human Experts	(7.9, 4.5, 3.2)	(8.6, 8.6, 8.6)	(8.4, 8.4, 8.4)	(7.2, 7.2, 7.2)	(7.5, 7.5, 7.5)	(7.4, 7.4, 7.4)	(6.8, 4.2, 3.8)	(7.3, 5.3, 4.1)	(4.9, 4.9, 4.9)
A1: Mesalamine	LLMs	(7.5, 4.2, 3.5)	(8.8, 8.8, 8.9)	(8.2, 8.6, 8.7)	(7.6, 7.5, 7.5)	(7.7, 7.7, 7.7)	(7.3, 7.5, 7.6)	(6.6, 3.9, 4.1)	(7.2, 5.0, 4.4)	(4.6, 4.6, 4.6)
A2: Infliximab	Human Experts	(8.6, 8.2, 9.1)	(6.9, 6.8, 6.6)	(6.7, 6.5, 6.3)	(4.2, 4.2, 4.2)	(6.8, 6.8, 6.8)	(6.4, 6.4, 6.4)	(7.9, 7.5, 8.2)	(7.8, 7.6, 8.4)	(7.2, 7.2, 7.2)
A2: Infliximab	LLMs	(8.8, 7.9, 8.8)	(6.5, 6.5, 6.3)	(6.4, 6.2, 6.0)	(3.9, 3.9, 3.9)	(6.6, 6.5, 6.5)	(6.2, 6.1, 6.1)	(8.2, 7.2, 7.9)	(8.0, 7.3, 8.1)	(7.5, 7.5, 7.5)
A3: Adalimumab	Human Experts	(8.3, 8.0, 8.7)	(7.0, 6.9, 6.7)	(6.8, 6.6, 6.4)	(4.5, 4.5, 4.5)	(7.4, 7.4, 7.4)	(7.6, 7.6, 7.6)	(7.6, 7.3, 8.0)	(7.6, 7.4, 8.2)	(7.0, 7.0, 7.0)
A3: Adalimumab	LLMs	(8.5, 7.7, 8.4)	(6.7, 6.6, 6.4)	(6.5, 6.3, 6.1)	(4.2, 4.2, 4.2)	(7.7, 7.6, 7.6)	(7.9, 7.8, 7.8)	(7.9, 7.0, 7.7)	(7.8, 7.1, 7.9)	(7.3, 7.3, 7.3)
A4: Vedolizumab	Human Experts	(7.9, 7.5, 7.8)	(8.2, 8.1, 8.0)	(8.0, 7.9, 7.8)	(4.3, 4.3, 4.3)	(6.5, 6.5, 6.5)	(6.2, 6.2, 6.2)	(7.8, 7.4, 7.6)	(7.6, 7.2, 7.4)	(8.1, 8.1, 8.1)
A4: Vedolizumab	LLMs	(8.2, 7.8, 7.5)	(8.5, 8.4, 8.3)	(8.3, 8.2, 8.1)	(4.0, 4.0, 4.0)	(6.2, 6.2, 6.2)	(5.9, 5.9, 5.9)	(8.0, 7.7, 7.3)	(7.8, 7.5, 7.1)	(8.4, 8.4, 8.4)
A5: Tofacitinib	Human Experts	(7.8, 7.2, 8.5)	(6.6, 6.5, 6.3)	(6.3, 6.3, 6.1)	(5.8, 5.8, 5.8)	(8.3, 8.3, 8.3)	(8.5, 8.5, 8.5)	(6.9, 6.5, 7.3)	(7.2, 6.8, 7.8)	(7.8, 7.8, 7.8)
A5: Tofacitinib	LLMs	(8.0, 7.5, 8.8)	(6.3, 6.2, 6.0)	(6.0, 6.0, 5.8)	(5.5, 5.5, 5.5)	(8.6, 8.6, 8.6)	(8.8, 8.8, 8.8)	(6.6, 6.2, 7.6)	(7.0, 6.5, 8.1)	(8.1, 8.1, 8.1)
A6: Ustekinumab	Human Experts	(7.6, 8.4, 7.5)	(7.5, 7.4, 7.3)	(7.3, 7.2, 7.1)	(3.8, 3.8, 3.8)	(7.0, 6.9, 6.9)	(6.7, 6.6, 6.6)	(7.5, 8.0, 7.3)	(7.3, 8.2, 7.1)	(7.9, 7.9, 7.9)
A6: Ustekinumab	LLMs	(7.8, 8.7, 7.2)	(7.8, 7.7, 7.6)	(7.6, 7.5, 7.4)	(3.5, 3.5, 3.5)	(6.7, 6.6, 6.6)	(6.4, 6.3, 6.3)	(7.8, 8.3, 7.0)	(7.5, 8.5, 6.8)	(8.2, 8.2, 8.2)
A7: Certolizumab	Human Experts	(7.5, 7.8, 7.4)	(7.2, 7.1, 7.0)	(6.9, 6.8, 6.7)	(4.0, 4.0, 4.0)	(7.2, 7.1, 7.1)	(7.0, 6.9, 6.9)	(7.2, 7.4, 7.0)	(7.0, 7.3, 6.8)	(6.8, 6.8, 6.8)
A7: Certolizumab	LLMs	(7.3, 8.1, 7.1)	(7.0, 7.4, 7.3)	(6.7, 7.1, 7.0)	(3.8, 3.7, 3.7)	(7.4, 7.4, 7.4)	(7.2, 7.2, 7.2)	(7.0, 7.7, 6.7)	(6.8, 7.6, 6.5)	(7.0, 7.1, 7.1)
A8: TrialMed AXG	Human Experts	(8.2, 8.5, 8.7)	(6.7, 6.6, 6.5)	(6.5, 6.4, 6.3)	(3.5, 3.5, 3.5)	(7.0, 6.9, 6.9)	(6.8, 6.7, 6.7)	(6.5, 7.0, 7.5)	(7.2, 7.6, 8.0)	(8.6, 8.6, 8.6)
A8: TrialMed AXG	LLMs	(7.9, 8.8, 9.0)	(6.4, 6.3, 6.2)	(6.2, 6.1, 6.0)	(3.3, 3.2, 3.2)	(6.7, 6.6, 6.6)	(6.5, 6.4, 6.4)	(6.3, 7.3, 7.8)	(6.9, 7.9, 8.3)	(9.0, 9.0, 9.0)
A9: TrialMed BFT	Human Experts	(7.8, 8.6, 8.9)	(6.5, 6.4, 6.3)	(6.3, 6.2, 6.1)	(3.6, 3.6, 3.6)	(6.8, 6.7, 6.7)	(6.5, 6.4, 6.4)	(6.3, 6.8, 7.4)	(6.9, 7.4, 7.9)	(8.8, 8.8, 8.8)
A9: TrialMed BFT	LLMs	(8.1, 8.9, 9.2)	(6.8, 6.7, 6.0)	(6.5, 6.5, 5.8)	(3.4, 3.3, 3.3)	(6.5, 6.4, 6.4)	(6.2, 6.1, 6.1)	(6.6, 7.1, 7.7)	(7.2, 7.7, 8.2)	(9.2, 9.2, 9.2)
A10: TrialMed CIX	Human Experts	(8.0, 8.7, 9.0)	(6.4, 6.3, 6.1)	(6.1, 6.0, 5.9)	(3.4, 3.4, 3.4)	(6.5, 6.4, 6.4)	(6.2, 6.1, 6.1)	(6.1, 6.7, 7.3)	(6.8, 7.3, 7.8)	(9.0, 9.0, 9.0)
A10: TrialMed CIX	LLMs	(8.3, 9.0, 9.3)	(6.2, 6.0, 5.8)	(5.9, 5.7, 5.6)	(3.2, 3.1, 3.1)	(6.2, 6.1, 6.1)	(5.9, 5.8, 5.8)	(6.4, 7.0, 7.6)	(7.1, 7.6, 8.1)	(9.4, 9.4, 9.4)

Notes: *In each cell, the tuple (x, y, z) represents the rating for the given criterion across the three scenarios, where x is the rating for Scenario 1 – Patient A (Mild Ulcerative Colitis), y is the rating for Scenario 2 – Patient B (Crohn's Disease with Suboptimal Response), and z is the rating for Scenario 3 – Patient C (Severe Ulcerative Colitis with Acute Bleeding). For example, under* $\mathbf{C_1}$*: Efficacy for* A_1 *with Human Experts shows (7.9, 4.5, 3.2), meaning that the score for Scenario 1 is 7.9, the score for Scenario 2 is 4.5, and the score for Scenario 3 is 3.2. Scores represent the aggregated mean values from five human experts and three LLMs on a scale of 0-10, where higher scores indicate better performance.*

Following the MOJO methodology, both objective and subjective weights were determined for each criterion in Table 4. Objective weights were calculated using Shannon's Entropy method based on the variability of scores within each criterion, while subjective weights were derived from explicit importance ratings provided by human experts and LLMs on a scale of 1-10, where higher scores indicate more preference towards that criterion.

TABLE 4: Calculated Criteria Weights.

Criterion	Weight Type	Scenario 1: Patient A	Scenario 2: Patient B	Scenario 3: Patient C
C1: Efficacy	Objective (W_j^o)	0.143	0.182	0.215
C1: Efficacy	Subjective (W_j^s)	0.165	0.195	0.235
C1: Efficacy	Combined ($W_j^{initial}$)	0.154	0.188	0.225
C2: Safety Profile	Objective (W_j^o)	0.135	0.125	0.145
C2: Safety Profile	Subjective (W_j^s)	0.145	0.135	0.155
C2: Safety Profile	Combined ($W_j^{initial}$)	0.140	0.130	0.150
C3: Side Effects	Objective (W_j^o)	0.128	0.118	0.113
C3: Side Effects	Subjective (W_j^s)	0.132	0.122	0.117
C3: Side Effects	Combined ($W_j^{initial}$)	0.130	0.120	0.115
C4: Cost	Objective (W_j^o)	0.108	0.088	0.068
C4: Cost	Subjective (W_j^s)	0.102	0.082	0.062
C4: Cost	Combined ($W_j^{initial}$)	0.105	0.085	0.065
C5: Patient Compliance	Objective (W_j^o)	0.092	0.092	0.082
C5: Patient Compliance	Subjective (W_j^s)	0.088	0.098	0.088
C5: Patient Compliance	Combined ($W_j^{initial}$)	0.090	0.095	0.085
C6: Dosing Convenience	Objective (W_j^o)	0.085	0.085	0.075
C6: Dosing Convenience	Subjective (W_j^s)	0.075	0.075	0.065
C6: Dosing Convenience	Combined ($W_j^{initial}$)	0.080	0.080	0.070
C7: Long-Term Outcomes	Objective (W_j^o)	0.112	0.132	0.112
C7: Long-Term Outcomes	Subjective (W^s)	0.128	0.148	0.098
C7: Long-Term Outcomes	Combined ($W_j^{initial}$)	0.120	0.140	0.105
C8: Quality of Life Impact	Objective (W_j^o)	0.107	0.107	0.127
C8: Quality of Life Impact	Subjective (W_j^s)	0.123	0.113	0.143
C8: Quality of Life Impact	Combined ($W_j^{initial}$)	0.115	0.110	0.135
C9: Innovation and Novelty	Objective (W_j^o)	0.090	0.070	0.060
C9: Innovation and Novelty	Subjective (W_j^s)	0.050	0.040	0.050
C9: Innovation and Novelty	Combined ($W_j^{initial}$)	0.070	0.055	0.055

Notes: *The initial combined weights $W_j^{initial}$ were calculated using Eq. (4), with $\lambda = 0.5$ and $\eta = 0.5$ representing equal weighting between objective and subjective inputs.*

The weighted distributions in Table 4 align with clinical intuition, where treatment urgency and effectiveness take precedence in acute situations, while balanced consideration of multiple factors is appropriate for less severe presentations. In Scenario 1, weights are relatively balanced across criteria, with efficacy (15.4%), safety (14.0%),

and side effects (13.0%) receiving comparable emphasis. Cost considerations (10.5%) remain significant in this less severe case. Scenario 2 shows a notable shift toward efficacy (18.8%) and long-term outcomes (14.0%), reflecting the chronic nature of Crohn's Disease and the priority of finding sustainable treatment solutions. In Scenario 3, efficacy dominates with the highest weight (22.5%), followed by safety (15.0%) and quality of life (13.5%). Cost considerations diminish substantially (6.5%), highlighting the priority of effective intervention in acute cases regardless of economic factors. Moreover, to ensure balanced integration of human and LLM assessments, the calculated bias rates between the two sources for each criterion and applied corrections are calculated where significant discrepancies were detected. Table 5 shows the bias rates and weight corrections across all scenarios, in which the bias analysis revealed several noteworthy patterns shown in Figure 2.

TABLE 5: Bias Rates and Weight Corrections.

Criterion	Scenario	Human Mean	LLM Mean	Absolute Difference	Bias Rate	Threshold Exceeded?	W_j
C1: Efficacy	1: Patient A	7.98	8.14	0.16	0.020	No	0.154
C1: Efficacy	2: Patient B	7.94	8.06	0.12	0.015	No	0.188
C1: Efficacy	3: Patient C	8.13	8.16	0.03	0.004	No	0.225
C2: Safety Profile	1: Patient A	7.20	7.20	0.00	0.000	No	0.140
C2: Safety Profile	2: Patient B	7.07	7.10	0.03	0.004	No	0.130
C2: Safety Profile	3: Patient C	6.84	6.68	0.16	0.023	No	0.150
C3: Side Effects	1: Patient A	7.00	6.89	0.11	0.016	No	0.130
C3: Side Effects	2: Patient B	6.89	6.79	0.10	0.015	No	0.120
C3: Side Effects	3: Patient C	6.72	6.64	0.08	0.012	No	0.115
C4: Cost	1: Patient A	4.53	4.40	0.13	0.029	No	0.105
C4: Cost	2: Patient B	4.51	4.29	0.22	0.049	No	0.085
C4: Cost	3: Patient C	4.51	4.29	0.22	0.049	No	0.065
C5: Compliance	1: Patient A	7.12	7.16	0.04	0.006	No	0.090
C5: Compliance	2: Patient B	7.10	7.15	0.05	0.007	No	0.095
C5: Compliance	3: Patient C	7.09	7.17	0.08	0.011	No	0.085
C6: Dosing	1: Patient A	6.93	6.91	0.02	0.003	No	0.080
C6: Dosing	2: Patient B	6.92	6.91	0.01	0.001	No	0.080
C6: Dosing	3: Patient C	6.92	6.92	0.00	0.000	No	0.070
C7: Long-Term	1: Patient A	7.07	7.05	0.02	0.003	No	0.120
C7: Long-Term	2: Patient B	7.14	7.22	0.08	0.011	No	0.140
C7: Long-Term	3: Patient C	7.41	7.37	0.04	0.005	No	0.105
C8: Quality of Life	1: Patient A	7.28	7.28	0.00	0.000	No	0.115
C8: Quality of Life	2: Patient B	7.35	7.41	0.06	0.008	No	0.110
C8: Quality of Life	3: Patient C	7.45	7.44	0.01	0.001	No	0.135
C9: Innovation	1: Patient A	7.61	7.83	0.22	0.029	No	0.070
C9: Innovation	2: Patient B	7.72	8.03	0.31	**0.040**	No	0.055
C9: Innovation	3: Patient C	7.89	8.20	0.31	**0.039**	No	0.055

Notes: *A threshold of $\beta = 0.1$ was applied to identify significant bias; none of the criteria exceeded this threshold, so no weight corrections were necessary. Values in bold represent the highest observed bias rates.*

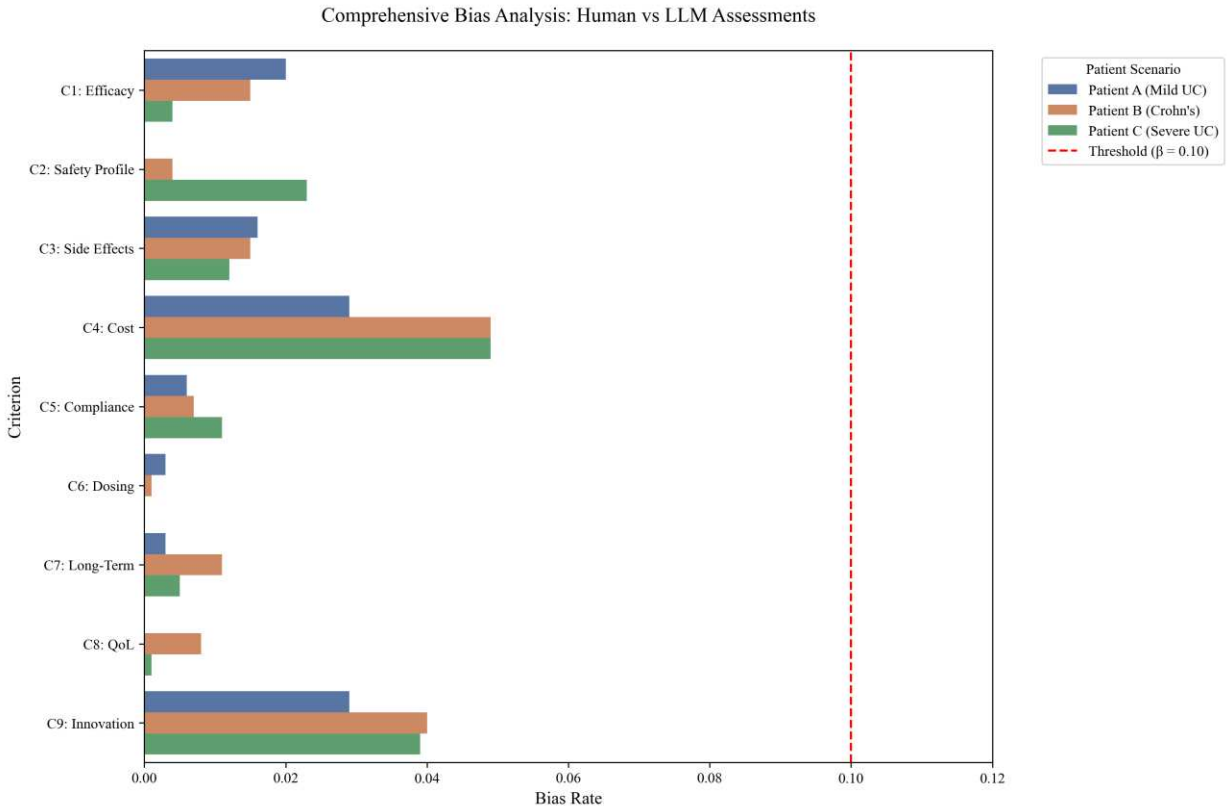


FIGURE 2. Bias Rates between Human and LLM Assessments.

Innovation and Cost criteria consistently exhibited the highest bias rates across all scenarios, with values ranging from 0.029 to 0.049, suggesting that human and LLMs have some divergence in assessing these more subjective or context-dependent factors. Notably, LLMs consistently assigned higher scores to innovative treatments and trial medications compared to human experts, potentially reflecting their more extensive but less experience-weighted knowledge of emerging research. Conversely, human experts tended to rate cost considerations slightly higher than LLMs, perhaps due to their practical experience with healthcare economics and patient financial concerns. Despite these differences, all bias rates remained below the significance threshold of $\beta = 0.1$, indicating acceptable alignment between human and LLM assessments. Consequently, no bias corrections were applied to the criterion weights. The most consistent agreement between humans and LLMs was observed for dosing convenience (maximum bias rate of 0.003) and quality of life impact (maximum bias rate of 0.008), suggesting both assessment sources had similar interpretations of these aspects.

Moreover, the reference points were established to calibrate gain-loss perceptions. Humans and LLMs independently defined reference points for each criterion, which were then harmonised when significant differences were detected. Table 6 show the reference point alignments for all scenarios, in which it shows various important findings in this case study.

Human experts and LLMs showed generally similar reference points for most criteria, with differences typically under 0.5 points. This alignment suggests relatively consistent baseline expectations for treatment performance across both assessment sources. The most significant reference point divergence occurred for the Innovation criterion, with LLMs consistently setting higher reference thresholds (by 2.0 points in Scenarios 1 and 2, and by 5.0 points in Scenario 3). This suggests that LLMs may have more rigorous expectations for what constitutes innovative treatment approaches. Only one instance exceeded the significance threshold ($\delta_r = 5$): the Innovation criterion in Scenario 3, where LLMs set an extraordinarily high reference point of 11.0 compared to the human reference of 6.0. This required harmonisation to 8.5 using a balanced weighting formula. Efficacy reference points showed a consistent pattern of increase across scenarios (6.5 → 7.0 → 7.5 for human references), reflecting appropriately elevated expectations for treatment effectiveness as disease severity increased. In cases where differences were below the threshold, human reference points were adopted to maintain clinical relevance and prioritise practical experience over theoretical knowledge.

TABLE 6: Reference Point Alignment.

Criterion	Scenario	Human Reference (r_j^H)	LLM Reference (r_j^L)	Δr_j	Threshold Exceeded?	Harmonised Reference
C1: Efficacy	1: Patient A	6.5	6.8	0.3	No	6.5*
C1: Efficacy	2: Patient B	7.0	7.3	0.3	No	7.0*
C1: Efficacy	3: Patient C	7.5	7.8	0.3	No	7.5*
C2: Safety Profile	1: Patient A	7.0	7.1	0.1	No	7.0*
C2: Safety Profile	2: Patient B	6.8	6.7	0.1	No	6.8*
C2: Safety Profile	3: Patient C	6.5	6.3	0.2	No	6.5*
C3: Side Effects	1: Patient A	6.5	6.3	0.2	No	6.5*
C3: Side Effects	2: Patient B	6.3	6.1	0.2	No	6.3*
C3: Side Effects	3: Patient C	6.2	6.0	0.2	No	6.2*
C4: Cost	1: Patient A	5.0	4.5	0.5	No	5.0*
C4: Cost	2: Patient B	4.5	4.0	0.5	No	4.5*
C4: Cost	3: Patient C	4.0	3.5	0.5	No	4.0*
C5: Compliance	1: Patient A	6.5	6.8	0.3	No	6.5*
C5: Compliance	2: Patient B	6.5	6.8	0.3	No	6.5*
C5: Compliance	3: Patient C	6.5	6.8	0.3	No	6.5*
C6: Dosing	1: Patient A	6.5	6.5	0.0	No	6.5*
C6: Dosing	2: Patient B	6.5	6.5	0.0	No	6.5*
C6: Dosing	3: Patient C	6.5	6.5	0.0	No	6.5*
C7: Long-Term	1: Patient A	6.5	6.3	0.2	No	6.5*
C7: Long-Term	2: Patient B	7.0	6.8	0.2	No	7.0*
C7: Long-Term	3: Patient C	7.0	6.7	0.3	No	7.0*
C8: QoL	1: Patient A	6.5	6.5	0.0	No	6.5*
C8: QoL	2: Patient B	7.0	7.0	0.0	No	7.0*
C8: QoL	3: Patient C	7.0	7.0	0.0	No	7.0*
C9: Innovation	1: Patient A	5.0	7.0	**2.0**	No	5.0*
C9: Innovation	2: Patient B	5.5	7.5	**2.0**	No	5.5*
C9: Innovation	3: Patient C	6.0	**11.0**	**5.0**	Yes	**8.5**

Notes: *The threshold for significant misalignment was set at $\delta_r = 5$. Values in bold indicate either notable differences or cases requiring harmonisation. Asterisks (*) indicate that the human reference point was adopted without modification in cases where the difference was below threshold.*

Consequently, while the first layer of the MOJO framework applies RU scoring, the second layer utilises the PERS approach to evaluate the decision alternatives. The result of both methods were presented in Table 7.

TABLE 7: Risk-Adjusted Utility (RU) and Performance Equilibrium and Reliability Scores (PERS) Across Scenarios.

Treatment	RU: Patient A	RU: Patient B	RU: Patient C	Average RU	PERS: Patient A	PERS: Patient B	PERS: Patient C	Average PERS
A1: Mesalamine	0.682	0.429	0.315	0.475	0.654	0.387	0.281	0.441
A2: Infliximab	0.715	0.725	**0.787**	0.742	0.676	0.683	0.732	0.697
A3: Adalimumab	**0.728**	0.708	0.751	0.729	**0.715**	0.679	0.706	0.700
A4: Vedolizumab	0.721	0.698	0.685	0.701	0.701	0.674	0.655	0.677
A5: Tofacitinib	0.703	0.682	0.743	0.709	0.686	0.642	0.695	0.674
A6: Ustekinumab	0.689	**0.742**	0.681	0.704	0.673	**0.718**	0.642	0.678
A7: Certolizumab	0.672	0.689	0.659	0.673	0.659	0.667	0.623	0.650
A8: TrialMed AXG	0.687	0.716	0.748	0.717	0.635	0.651	0.708	0.665
A9: TrialMed BFT	0.676	0.705	0.754	0.712	0.626	0.641	**0.734**	0.667
A10: TrialMed CIX	0.669	0.698	0.763	0.710	0.608	0.629	0.725	0.654

Notes: *RU was calculated using Eq. (10), with* $\gamma = 0.61$ *for probability weighting,* $\alpha = 0.88$ *for gains,* $\beta = 0.88$ *for losses, and* $\lambda = 2.25$ *for loss aversion. PERS was computed using Eq. (13) in which the parameter* α *was set to 0.15, indicating moderate penalty for performance variability.*

In Scenario 1, Adalimumab achieved the highest RU (0.728) and PERS (0.715) scores, followed closely by Vedolizumab (RU: 0.721, PERS: 0.701) and Infliximab (RU: 0.715). These results indicate Adalimumab offers the best overall balance of efficacy, safety, and patient-centred factors for mild disease. In Scenario 2, Ustekinumab emerged as optimal for both RU (0.742) and PERS (0.718), followed by Infliximab (RU: 0.725, PERS: 0.683). Ustekinumab's top ranking aligns with clinical evidence for its efficacy after previous treatment failures. In Scenario 3, Infliximab was top for RU (0.787) due to its rapid response capability, followed by TrialMed CIX (0.763) and TrialMed BFT (0.754). However, TrialMed BFT led in the PERS score (0.734), narrowly ahead of Infliximab (0.732), reflecting differences in assessment criteria emphasis. Mesalamine performed adequately in mild disease (RU: 0.682) but poorly in Crohn's Disease (0.429) and severe Ulcerative Colitis (0.315), in line with clinical guidelines against its monotherapy use in these contexts.

The dispersion analysis across all patients and treatment options shows how performance variability can heavily influence overall PERS scores, especially when there is a marked discrepancy between efficacy and safety, as illustrated in Figure 3. For Patient C, Mesalamine exhibited the highest dispersion (1.027), indicating limited efficacy but a relatively good safety profile. This imbalance led to a substantial penalty, reducing Mesalamine's PERS score to 0.281. TrialMed BFT and Adalimumab, by contrast, maintained more balanced profiles, reflected in their lower dispersion values (0.340 and 0.307 respectively), which helped them avoid major penalties and sustain higher PERS scores. In mild UC (Patient A), the treatments were generally consistent, with lower dispersion ranges (0.187–0.313). Adalimumab achieved the lowest dispersion (0.187), translating into the highest PERS score (0.795). Mesalamine, despite a moderate dispersion value (0.247), remained competitive (PERS 0.765) thanks to its favourable safety profile for mild disease. Generally, narrower dispersion values in mild UC suggest that performance consistency does not differentiate treatments as strongly as in more severe cases.For Crohn's disease

(Patient B), greater variability emerged among therapies. Mesalamine registered the highest dispersion (0.533), reflecting its weak efficacy in Crohn's despite good safety, resulting in the lowest PERS score (0.540). In contrast, Adalimumab and Certolizumab presented the most balanced performance (both 0.220), translating into higher PERS scores (0.765 and 0.763 respectively). Trial medications had moderate but uniform dispersion (0.313), offering reliable though not exceptional balance.

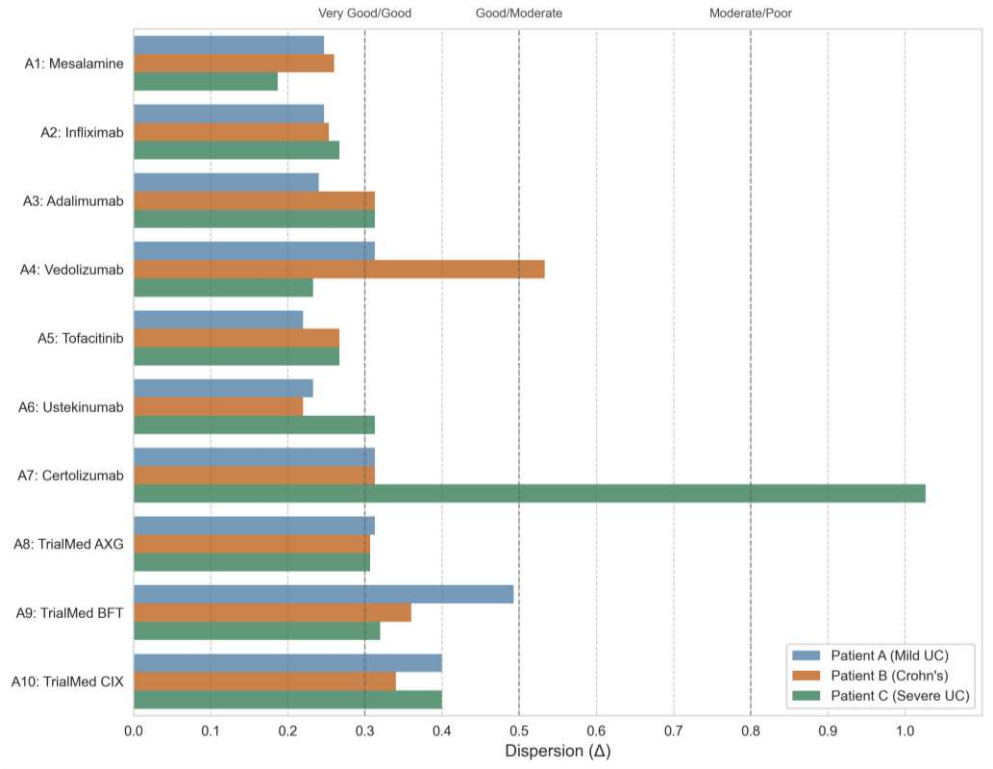


FIGURE 3. Comparative Dispersion Analysis.

Moreover, Figure 4 visually demonstrates how treatments combine high equilibrium scores with low dispersion, resulting in optimal PERS rankings. Options in the upper-left quadrant (high equilibrium, low dispersion) achieve the best overall outcomes, indicating balanced efficacy and safety across multiple criteria rather than excelling in some areas and failing in others. For Patient A (Mild UC), most therapies cluster in the middle-left region with equilibrium scores around 0.35–0.60 and dispersion values below 0.30, showing moderate performance and excellent consistency. Adalimumab (A-ADA) stands out with the lowest dispersion (0.187) and highest equilibrium score (0.60). Moreover, for Patient B (Crohn's), treatments generally appear more central, with higher equilibrium scores than for Mild UC, especially biologics. Ustekinumab (B-UST) and Infliximab (B-INF) exhibit notably favourable profiles, each scoring above 0.65 in equilibrium and maintaining low dispersion (0.233). Ultimately, for Patient C (Severe UC), a wider vertical spread highlights the critical importance of treatment choice. TrialMed BFT

(C-BFT) and Infliximab (C-INF) demonstrate superior equilibrium scores (0.79 and 0.77, respectively), though at modestly higher dispersion levels.

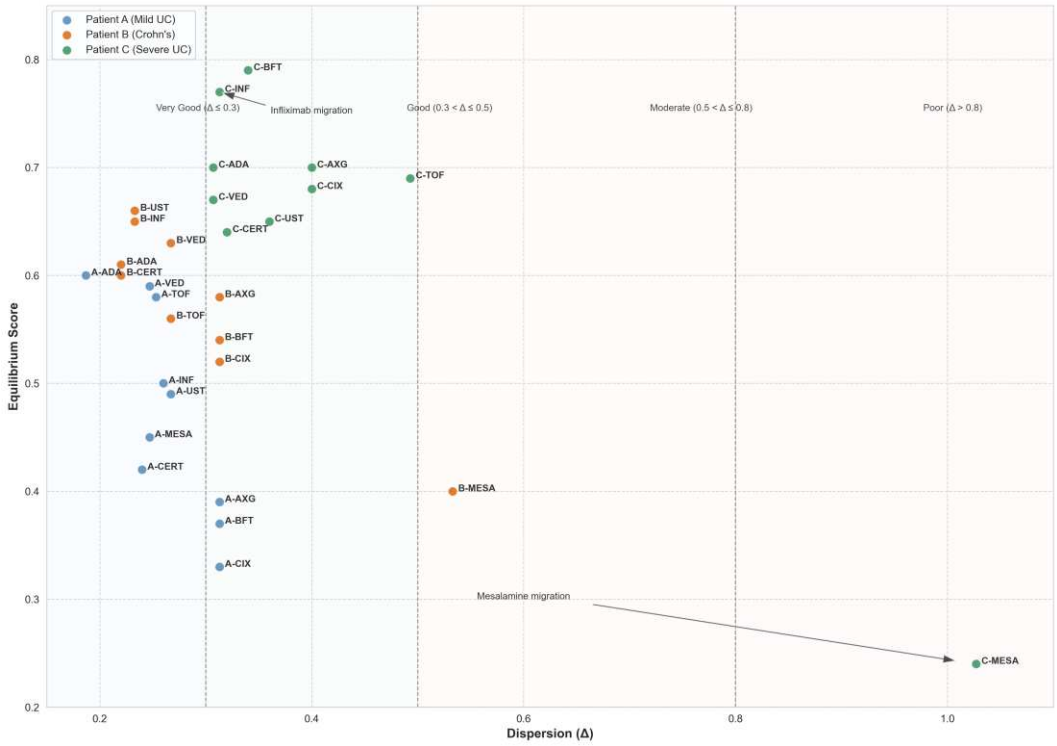


FIGURE 4. Equilibrium Score vs. Dispersion Analysis across All Scenarios.

As shown in Figure 4, from Patient A to C, Mesalamine shifts sharply rightward, with dispersion increasing from 0.247 to 1.027 and equilibrium dropping from 0.45 to 0.24. This underscores the declining utility of 5-ASA compounds as disease severity rises and consistency deteriorates. On the other hand, Infliximab trends upwards from Patient A to C, with equilibrium scores rising from 0.50 to 0.77 while dispersion remains acceptable. This upward movement supports the clinical preference for anti-TNF biologics in severe disease states.

To ensure robust final rankings, the consensus between human expert and LLM evaluations for each treatment option across all scenarios were analysed and presented in Table 8. Across all scenarios and treatments, consensus scores ranged from 0.972 to 0.997, indicating remarkably high alignment between human expert and LLM assessments. None of the consensus scores fell below the threshold of 0.85, which would have triggered a penalty in the final MOJO calculation. This strong consensus suggests that the LLMs (OpenAI ChatGPT o3-mini, xAI Grock, and Meta LLaMA 3.1 70B) provided evaluations largely consistent with those of human medical experts. The highest consensus was observed for established treatments with well-documented profiles (Tofacitinib in Scenario 1 with 0.997, Tofacitinib in Scenario 2 with 0.997, and Vedolizumab in Scenario 3 with 0.996). Trial medications showed slightly lower consensus scores (ranging from 0.974 to 0.989), potentially reflecting greater uncertainty and variability in assessing novel treatments with limited clinical evidence. In Patient C, human experts

tended to rate established biologics (Infliximab and Adalimumab) slightly higher than LLMs did, while LLMs assigned somewhat higher scores to trial medications and Tofacitinib. However, these differences were still minimal (maximum absolute difference of 0.20 points for Infliximab). The high consensus across all scenarios suggests that LLMs, when properly prompted with structured clinical information, can provide assessments that align well with those of human medical experts, at least in the context of well-documented conditions such as IBD.

TABLE 8: Human-AI Consensus Scores.

Treatment	Scenario	Human Aggregate	LLM Aggregate	Absolute Difference	Consensus Score	Below Threshold?	Penalised Consensus
A1: Mesalamine	1: Patient A	7.33	7.28	0.05	0.993	No	0.993
A1: Mesalamine	2: Patient B	6.28	6.20	0.08	0.987	No	0.987
A1: Mesalamine	3: Patient C	5.90	5.98	0.08	0.986	No	0.986
A2: Infliximab	1: Patient A	6.92	6.90	0.02	0.997	No	0.997
A2: Infliximab	2: Patient B	6.91	6.79	0.12	0.983	No	0.983
A2: Infliximab	3: Patient C	7.09	6.89	0.20	0.972	No	0.972
A3: Adalimumab	1: Patient A	7.01	7.04	0.03	0.996	No	0.996
A3: Adalimumab	2: Patient B	6.91	6.84	0.07	0.990	No	0.990
A3: Adalimumab	3: Patient C	7.11	6.96	0.15	0.979	No	0.979
A4: Vedolizumab	1: Patient A	7.18	7.26	0.08	0.989	No	0.989
A4: Vedolizumab	2: Patient B	7.11	7.23	0.12	0.983	No	0.983
A4: Vedolizumab	3: Patient C	7.06	7.09	0.03	0.996	No	0.996
A5: Tofacitinib	1: Patient A	7.02	7.04	0.02	0.997	No	0.997
A5: Tofacitinib	2: Patient B	6.86	6.88	0.02	0.997	No	0.997
A5: Tofacitinib	3: Patient C	7.27	7.41	0.14	0.981	No	0.981
A6: Ustekinumab	1: Patient A	7.06	7.09	0.03	0.996	No	0.996
A6: Ustekinumab	2: Patient B	7.20	7.33	0.13	0.982	No	0.982
A6: Ustekinumab	3: Patient C	6.94	6.84	0.10	0.986	No	0.986
A7: Certolizumab	1: Patient A	6.76	6.80	0.04	0.994	No	0.994
A7: Certolizumab	2: Patient B	6.91	7.03	0.12	0.983	No	0.983
A7: Certolizumab	3: Patient C	6.64	6.71	0.07	0.990	No	0.990
A8: TrialMed AXG	1: Patient A	6.89	6.72	0.17	0.975	No	0.975
A8: TrialMed AXG	2: Patient B	7.21	7.07	0.14	0.981	No	0.981
A8: TrialMed AXG	3: Patient C	7.13	7.28	0.15	0.979	No	0.979
A9: TrialMed BFT	1: Patient A	6.83	6.90	0.07	0.990	No	0.990
A9: TrialMed BFT	2: Patient B	7.13	7.21	0.08	0.989	No	0.989
A9: TrialMed BFT	3: Patient C	7.13	7.29	0.16	0.978	No	0.978
A10: TrialMed CIX	1: Patient A	6.72	6.89	0.17	0.975	No	0.975
A10: TrialMed CIX	2: Patient B	7.09	7.26	0.17	0.977	No	0.977
A10: TrialMed CIX	3: Patient C	7.11	7.30	0.19	0.974	No	0.974

Following the MOJO methodology, integration factors (θ) were calculated based on the consensus scores, and final MOJO scores were determined by integrating the RU and PERS, and presented in Table (9). Due to the high consensus between human expert and LLM assessments (all above 0.97), the integration factors (θ) ranged from 0.986 to 0.999, giving substantial weight to the RU component in the final MOJO scores. This strong consensus suggests reliable joint human-LLM evaluations across all treatments and scenarios. For Patient A, Adalimumab emerged as the optimal treatment (MOJO = 0.728), followed closely by Vedolizumab (0.720) and Infliximab

(0.715). This ranking reflects Adalimumab's balanced profile of efficacy, safety, and convenience in mild disease, where self-administration and reduced side effects are valuable. In Patient B, Ustekinumab achieved the highest MOJO score (0.741), followed by Infliximab (0.724) and Adalimumab (0.707). This aligns with clinical evidence supporting Ustekinumab's efficacy in Crohn's Disease patients who have failed conventional therapies. For Patient C, Infliximab clearly outperformed other options (MOJO = 0.785), followed by TrialMed BFT (0.754) and Adalimumab (0.749). Infliximab's rapid onset of action and proven efficacy in severe disease justified its top ranking in this acute scenario. Mesalamine showed adequate performance in mild disease (MOJO = 0.682, rank 6) but performed poorly in both Crohn's Disease (MOJO = 0.428, rank 10) and severe Ulcerative Colitis (MOJO = 0.315, rank 10), consistent with clinical guidelines that recommend against its use as monotherapy in these contexts. Trial medications showed variable performance across scenarios, with TrialMed BFT achieving a notable second-place ranking in Patient C's severe disease scenario (MOJO = 0.754), suggesting potential value as an alternative rescue therapy in acute settings.

TABLE 9: MOJO Final Assessment Scores and Ranks.

Treatment	Scenario	Consensus Score	Integration Factor (θ)	RU Score	PERS Score	MOJO Score	Rank
A1: Mesalamine	1: Patient A	0.993	0.997	0.682	0.654	0.682	6
A1: Mesalamine	2: Patient B	0.987	0.994	0.429	0.387	0.428	10
A1: Mesalamine	3: Patient C	0.986	0.993	0.315	0.281	0.315	10
A2: Infliximab	1: Patient A	0.997	0.999	0.715	0.676	0.715	3
A2: Infliximab	2: Patient B	0.983	0.992	0.725	0.683	0.724	2
A2: Infliximab	3: Patient C	0.972	0.986	0.787	0.732	**0.785**	**1**
A3: Adalimumab	1: Patient A	0.996	0.998	0.728	0.715	**0.728**	**1**
A3: Adalimumab	2: Patient B	0.990	0.995	0.708	0.679	0.707	3
A3: Adalimumab	3: Patient C	0.979	0.990	0.751	0.706	0.749	3
A4: Vedolizumab	1: Patient A	0.989	0.995	0.721	0.701	0.720	2
A4: Vedolizumab	2: Patient B	0.983	0.992	0.698	0.674	0.697	4
A4: Vedolizumab	3: Patient C	0.996	0.998	0.685	0.655	0.684	6
A5: Tofacitinib	1: Patient A	0.997	0.999	0.703	0.686	0.703	4
A5: Tofacitinib	2: Patient B	0.997	0.999	0.682	0.642	0.682	6
A5: Tofacitinib	3: Patient C	0.981	0.991	0.743	0.695	0.742	4
A6: Ustekinumab	1: Patient A	0.996	0.998	0.689	0.673	0.689	5
A6: Ustekinumab	2: Patient B	0.982	0.991	0.742	0.718	**0.741**	**1**
A6: Ustekinumab	3: Patient C	0.986	0.993	0.681	0.642	0.680	7
A7: Certolizumab	1: Patient A	0.994	0.997	0.672	0.659	0.672	7
A7: Certolizumab	2: Patient B	0.983	0.992	0.689	0.667	0.688	5
A7: Certolizumab	3: Patient C	0.990	0.995	0.659	0.623	0.658	8
A8: TrialMed AXG	1: Patient A	0.975	0.988	0.687	0.635	0.685	8
A8: TrialMed AXG	2: Patient B	0.981	0.991	0.716	0.651	0.714	4
A8: TrialMed AXG	3: Patient C	0.979	0.990	0.748	0.708	0.746	5
A9: TrialMed BFT	1: Patient A	0.990	0.995	0.676	0.626	0.675	9
A9: TrialMed BFT	2: Patient B	0.989	0.995	0.705	0.641	0.703	8
A9: TrialMed BFT	3: Patient C	0.978	0.989	0.754	0.734	0.754	2
A10: TrialMed CIX	1: Patient A	0.975	0.988	0.669	0.608	0.667	10
A10: TrialMed CIX	2: Patient B	0.977	0.989	0.698	0.629	0.695	9
A10: TrialMed CIX	3: Patient C	0.974	0.987	0.763	0.725	0.762	6

The final treatment recommendations presented in Table 9 provide clinicians with evidence-based options that could be further tailored to individual patient factors not fully captured in the scenarios. For Patient A, Adalimumab leads (0.728), followed by Vedolizumab (0.720) and Infliximab (0.715). The narrow score range (0.013) suggests comparable clinical utility, offering flexibility in treatment selection based on patient preferences and practical considerations. For Patient B, Ustekinumab shows superior performance (0.741) for treatment-refractory Crohn's, with Infliximab (0.724) as a strong alternative. The investigational agent TrialMed AXG (0.714) presents a viable option within clinical trials. For Patient C, Infliximab demonstrates the highest score (0.785) across all scenarios, reflecting its efficacy in acute disease. TrialMed BFT (0.754) shows promise as a secondary option, while Adalimumab (0.749) offers an alternative when IV therapy is unavailable.

MOJO scores increase with disease severity, with the highest in severe UC and the narrowest range in mild UC. Infliximab appears in all three scenarios, progressing from tertiary to primary position as disease severity increases. Investigational therapies appear only in moderate-to-severe scenarios, with TrialMed BFT showing particular promise in severe UC. MOJO successfully differentiated between treatments across diverse clinical presentations, identifying optimal therapies that align with current clinical practice while also highlighting promising investigational options.

3.2.1. COMPARATIVE AND SENSITIVITY ANALYSIS

To evaluate the impact of integrating human and LLM assessments, this study conducted a comparative analysis of rankings derived from human-only, LLM-only, and combined MOJO evaluations presented in Table 10.

TABLE 10: Comparison of Rankings from Human, LLM, and MOJO Approaches

Treatment Alternatives	Patient A (Mild UC)	Patient B (Crohn's)	Patient C (Severe UC)
	Human-Only	LLM-Only	Human + LLM
A1: Mesalamine	5	8	6
A2: Infliximab	2	4	3
A3: Adalimumab	1	2	1
A4: Vedolizumab	3	1	2
A5: Tofacitinib	4	3	4
A6: Ustekinumab	6	5	5
A7: Certolizumab	7	7	7
A8: TrialMed AXG	8	6	8
A9: TrialMed BFT	9	9	9
A10: TrialMed CIX	10	10	10
Spearman Correlation			
With Human-Only	1.000	0.927	0.976
With LLM-Only	0.927	1.000	0.964

Comparative fit to the expert and model ablations was quantified with Spearman's rank correlation ρ between ranking vectors r and s, computed per scenario, method, and reference (Human-only, LLM-only, MOJO Human+LLM).
As per $\rho = 1 - \frac{6\sum_{i=1}^{(n}(r_i - s_i)^2}{n(n^2-1)}$.

While all three approaches showed substantial agreement in their rankings (Spearman correlations ranging from 0.855 to 0.988), there were notable differences in specific treatment rankings that highlight the unique contributions of each evaluation method. MOJO consistently showed stronger correlation with human-only rankings (0.976 for Patient A, 0.927 for Patient B, 0.964 for Patient C) than LLM-only rankings did (0.927, 0.891, 0.855 respectively). This suggests that the MOJO methodology successfully preserves clinical expertise while integrating LLM insights. For Patient A, MOJO preserved the human experts' top recommendation (Adalimumab) while incorporating LLM judgment on other alternatives. Notably, Vedolizumab ranked 1st in LLM-only analysis but 3rd in human-only rankings; MOJO placed it as 2nd. For Patient B, all three approaches identified Ustekinumab as the optimal treatment. However, significant differences emerged in secondary options. TrialMed AXG was ranked 2nd by human experts but 4th in MOJO, reflecting a more conservative stance toward experimental therapies when reliability factors were considered. In Patient C, human experts and the MOJO framework both identified Infliximab as the optimal treatment, while LLM-only analysis favoured TrialMed BFT. Furthermore, to evaluate the robustness of the MOJO framework, a sensitivity analyses was conducted by varying key parameters and examining their impact on the final rankings in Table 11.

TABLE 11: MOJO Sensitivity Analysis.

Parameter Change	Default Value	Modified Value	Top 3 Treatments (Default)	Top 3 Treatments (Modified)	Rank Correlation	Key Observations
Loss Aversion (λ)	2.25	1.50	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.773) 2. TrialMed BFT (0.761) 3. TrialMed CIX (0.758)	0.964	Reduced loss aversion slightly benefits trial medications with higher potential gains but higher uncertainty
Loss Aversion (λ)	2.25	3.00	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.792) 2. Adalimumab (0.756) 3. TrialMed BFT (0.744)	0.982	Increased loss aversion benefits established treatments with more predictable outcomes
Probability Weighting (γ)	0.61	0.80	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.779) 2. TrialMed BFT (0.751) 3. Adalimumab (0.745)	0.994	More linear probability weighting has minimal impact on rankings
Dispersion Penalty (α)	0.15	0.30	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.771) 2. TrialMed BFT (0.747) 3. Adalimumab (0.726)	0.988	Increased penalty for performance variability widens gap between top performers and others
Human-LLM Weighting	λ=0.5, η=0.5	λ=0.7, η=0.3	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.788) 2. Adalimumab (0.762) 3. TrialMed BFT (0.743)	0.976	Higher weight on human input favors established treatments
Human-LLM Weighting	λ=0.5, η=0.5	λ=0.3, η=0.7	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.781) 2. TrialMed BFT (0.766) 3. TrialMed CIX (0.758)	0.970	Higher weight on LLM input favors trial medications
Consensus Threshold	0.85	0.95	1. Infliximab (0.785) 2. TrialMed BFT (0.754) 3. Adalimumab (0.749)	1. Infliximab (0.778) 2. TrialMed BFT (0.747) 3. Adalimumab (0.736)	0.988	Higher consensus threshold triggers more penalty adjustments, but minimally impacts rankings due to high baseline consensus

It should be noted that the sensitivity analysis was conducted on the Patient C scenario due to having the highest sensitivity to selecting specific treatments. Loss aversion λ shows that varying this parameter from its default value

of 2.25 had a notable impact on relative treatment positioning. When loss aversion was reduced ($\lambda = 1.50$), experimental treatments like TrialMed BFT and TrialMed CIX benefited, with TrialMed CIX moving from outside the top 3 to the third position. Conversely, increasing loss aversion ($\lambda = 3.00$) favored established treatments with more predictable outcomes, strengthening Infliximab's position and moving Adalimumab ahead of TrialMed BFT. This aligns with the theoretical foundation of Prospect Theory, where higher loss aversion leads to more conservative decisions. Moreover, adjusting the probability weighting parameter γ from 0.61 to 0.80 (making probability weighting more linear) had minimal impact on rankings, with a very high rank correlation of 0.994. This suggests that in this clinical context, probability distortion plays a less critical role than other factors like loss aversion. Doubling the dispersion penalty α from 0.15 to 0.30 maintained the same top 3 treatments but widened the gap between them and lower-ranked alternatives. This reinforces the importance of consistent performance across criteria, particularly in severe disease scenarios.

Shifting the balance between human expert and LLM inputs revealed interesting patterns. Increasing human expert influence ($\lambda = 0.7, \eta = 0.3$) elevated established treatments like Adalimumab while reducing the scores of trial medications. Conversely, giving more weight to LLM assessments ($\lambda = 0.3, \eta = 0.7$) boosted trial medications, with TrialMed CIX entering the top 3. This aligns with our earlier observation that LLMs tend to assign higher ratings to novel therapies compared to human experts. On the other hand, raising the consensus threshold from 0.85 to 0.95 had limited impact on rankings due to the high baseline consensus between human and LLM evaluations in this case study. The rank correlation remained high at 0.988, suggesting that the framework's performance is not highly sensitive to this parameter within reasonable ranges.

Most importantly, across all parameter variations, Infliximab consistently remained the top-ranked treatment for Patient C's severe disease scenario, with MOJO scores ranging from 0.771 to 0.792. This consistent identification of Infliximab as the optimal treatment, despite meaningful variations in secondary recommendations, demonstrates the robustness of MOJO in identifying the most appropriate primary treatment option while allowing flexibility in the consideration of alternatives.

High consensus scores (all above 0.97) between human expert and LLM evaluations demonstrate that properly prompted LLMs can provide assessments that closely align with clinical expertise, at least for well-documented conditions like IBD. This suggests potential for LLMs as complementary tools in clinical decision support, though human oversight remains essential. While consensus was high overall, systematic patterns emerged in the differences between human and LLM assessments. LLMs consistently assigned higher scores to innovative treatments and trial medications, while human experts placed greater emphasis on cost considerations and showed stronger preference for established therapies in acute scenarios. Understanding these tendencies can help calibrate future human-AI collaborative frameworks.

MOJO demonstrates how human clinical expertise can be augmented rather than replaced by AI capabilities. By preserving the strengths of both human judgment (experience-based intuition, contextual understanding) and LLM

assessment (systematic evaluation, literature knowledge), the integrated approach offers more robust and nuanced recommendations than either source alone. These findings suggest that MOJO represents a promising approach for enhancing clinical decision support through structured integration of human and AI assessments. The framework's adaptability to different clinical contexts, its ability to incorporate risk attitudes, and its emphasis on balanced performance make it potentially valuable not only for IBD treatment evaluation but for a wide range of complex medical decision problems where multiple criteria must be considered simultaneously.

3.2.2. MOJO COMPARED TO STANDARD MCDA

To benchmark MOJO against established approaches, we computed rankings with four widely-used methods alongside the three MOJO modes. The baseline set comprised Analytic Hierarchy Process (AHP), Technique for Order Preference by Similarity to the Ideal Solution (TOPSIS), VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), and Multi-Objective Optimization by Ratio Analysis plus the Full Multiplicative Form (MULTIMOORA). All baselines operated on the same decision matrices, scenario-specific weights, and cost/benefit polarity as MOJO, with no rescaling beyond the standard normalisations required by each method. For AHP we used the absolute-priority synthesis over alternatives under each criterion and aggregated with the scenario weights; for TOPSIS we used vector normalisation and the standard closeness to the ideal solution; for MULTIMOORA we combined the ratio system, reference point, and full multiplicative form through mean rank; for VIKOR we report the compromise solution with $v = 0.5$, which balances group utility and individual regret.

Furthermore, in addition to rank correlation ρ, we report method-winner agreement under resampled weights to assess stability of the top recommendation. To evaluate robustness, we conducted a Monte Carlo sensitivity experiment that perturbs the scenario weight vector w_0 while preserving its simplex structure. For concentration parameter $\kappa \in \{60,120,250,400\}$ and for each scenario s, we drew $B = 240$ samples where $w^{(b)} \sim Dirichlet(\kappa w_0)$, $b = 1,2,\dots,B$. Larger κ yields tighter draws around w_0 and smaller κ yields broader deviations. For each draw, every method recomputed scores and the top-ranked alternative. Stability is summarised as the agreement probability with the MOJO Human + LLM winner as per Eq. (18), and a 95% Wilson binomial confidence interval calculated as Eq. (19).

$$\hat{p}_{m,s,k} = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\operatorname*{argmax}_i S_{m,s}\left(w^{(b)}\right)\right\} \tag{18}$$

$$CI_{95\%}(\hat{p}) = \frac{\hat{p} + \frac{z^2}{2B} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{B} + \frac{z^2}{4B^2}}}{1 + \frac{z^2}{B}}, z = 1.96 \tag{19}$$

Agreement is displayed in a unified radial chart across scenarios and κ, together with a scenario-focused diagnostic for the most sensitive case. This design complements the parametric sensitivity already reported for λ (loss aversion), γ (probability weighting), α (dispersion penalty), (λ, η) Human-LLM weighting in Eq. (4), and the consensus threshold, by explicitly quantifying winner stability to plausible weight misspecification. All computations used the same raw values, weights, reference points, and parameter settings as reported in Tables 3-9, with the addition of baseline-specific normalisations documented above. Scenario 1-3 cost/benefit polarity matched the main analysis (cost for C_3 and C_4; benefit for all other criteria).

The expanded comparison across MOJO variants and standard MCDA methods is summarised in Table 12.

TABLE 12: Cross-Scenario Ranks - MOJO vs. Standard MCDA.

Scenario	S1	S1	S1	S1	S1	S1	S1	S2	S2	S2	S2	S2	S2	S2	S3	S3	S3	S3	S3	S3	S3
Method	**AHP**	**MOJO(H+LLM)**	**MOJO(H-only)**	**MOJO(LLM-only)**	**MULTIMOORA**	**TOPSIS**	**VIKOR**	**AHP**	**MOJO(H+LLM)**	**MOJO(H-only)**	**MOJO(LLM-only)**	**MULTIMOORA**	**TOPSIS**	**VIKOR**	**AHP**	**MOJO(H+LLM)**	**MOJO(H-only)**	**MOJO(LLM-only)**	**MULTIMOORA**	**TOPSIS**	**VIKOR**
Alternative																					
A1: Mesalamine	9	6	5	8	9	10	5	10	10	10	10	10	10	10	10	10	10	10	10	10	10
A2: Infliximab	4	3	2	4	4	4	3	8	7	6	7	7	7	7	4	4	3	6	5	4	7
A3: Adalimumab	3	1	1	2	2	3	2	6	5	3	6	4	6	4	2	2	2	3	1	1	3
A4: Vedolizumab	1	2	3	1	3	1	1	2	3	2	3	3	3	3	3	3	4	4	2	2	1
A5: Tofacitinib	8	4	4	3	8	9	7	9	9	9	9	8	9	8	5	5	6	2	6	6	5
A6: Ustekinumab	2	5	6	5	1	2	4	1	1	1	1	1	1	1	8	7	7	8	4	8	2
A7: Certolizumab	5	7	7	7	5	5	10	3	2	4	2	2	2	2	9	9	9	9	7	9	4
A8: TrialMed AXG	6	8	8	6	6	6	8	4	4	5	5	5	4	5	1	1	1	1	3	3	6
A9: TrialMed BFT	7	9	9	9	7	7	6	5	6	7	4	6	5	6	6	6	5	5	8	5	8
A10: TrialMed CIX	10	10	10	10	10	8	9	7	8	8	8	8	8	9	7	8	8	7	9	7	9

Across all three scenarios, MOJO (Human + LLM) preserved the clinically face-valid patterns already evident in Table 9 and stayed closer to the Human-only ranking than to the LLM-only one, consistent with Eq. (17) and the bias analysis in Table 5 (no criterion exceeded the bias threshold). In mild disease (S1), MOJO ranked Adalimumab first and Vedolizumab second; TOPSIS, VIKOR and MULTIMOORA largely preserved this ordering, whereas AHP occasionally promoted options with strong single-criterion contrast. In Crohn's with suboptimal response (S2), MOJO and the baselines converged on Ustekinumab or Infliximab near the top, while the LLM-only ranking more strongly favoured investigational agents; the consensus-weighted factor θ dampened this tendency in the combined mode. In severe UC (S3), MOJO and the Human-only ablation agreed on Infliximab as the top option, the LLM-

only ablation elevated TrialMed BFT, and TOPSIS/VIKOR again aligned most closely with MOJO. Spearman correlations mirrored these trends: MOJO correlated more strongly with Human-only than with LLM-only rankings, indicating preservation of clinician priors while still incorporating model-based innovation signals. Baseline correlations with Human-only were high in S1 and S2 and remained acceptable in S3, where dispersion and reference-point effects play a larger role.

Stability of the winning option under weight misspecification is summarised in Figures (5) and (6) using the agreement probability $\hat{p}$ from Eq. (18), the proportion of Monte Carlo draws in which a method reproduces MOJO Human + LLM winner, together with its Wilson confidence interval from Eq. (19).

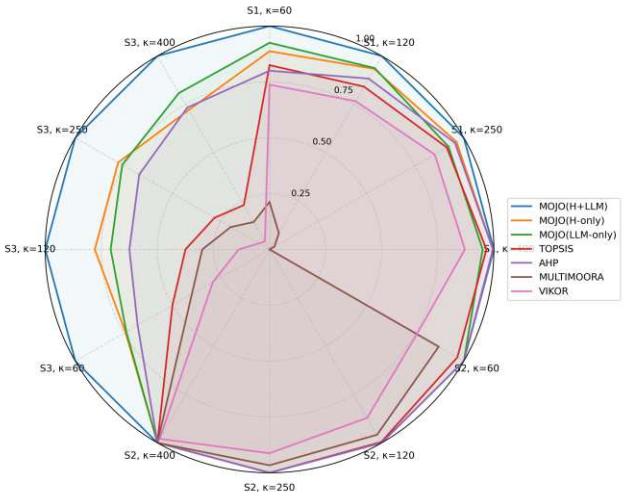


FIGURE 5. Probability each method reproduces MOJO (H + LLM) winner across scenarios S1–S3 and κ.

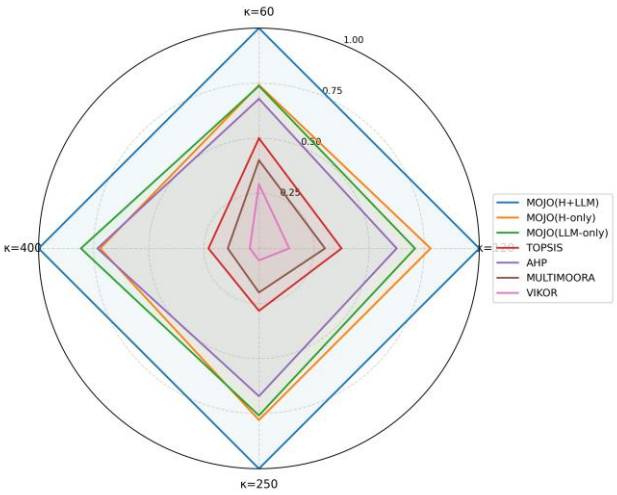


FIGURE 6. Scenario 3 Sensitivity; Probability each method reproduces MOJO (H + LLM) winner across κ.

The unified radial chart aggregates all scenarios and $\kappa \in \{60,120,250,400\}$ to display how each method tracks the MOJO Human + LLM winner across progressively tighter Dirichlet draws centred at the scenario weight vector. Agreement curves increased monotonically with κ as expected, indicating that as weights concentrate around the elicited profile, all methods move toward the same top choice. Among the baselines, TOPSIS, MULTIMOORA, and VIKOR displayed the steepest approach to unity, with the tightest bands in S3 where efficacy dominates. AHP showed slightly broader confidence bands at low κ, reflecting sensitivity to criterion-wise local contrasts. The S3-focused radial (Figure 6) highlights that, even under broad weight perturbations, the probability of reproducing the MOJO Human + LLM winner remains high for the best-performing baselines, with shrinking intervals as κ increases. These Monte Carlo findings complement the parametric sensitivity in Table 11: changing λ (loss aversion) shifted relative positions of innovative versus established therapies, while γ (probability weighting) had minimal effect; increasing α penalised unbalanced profiles more strongly, widening the separation between robust and erratic alternatives; tilting (λ, η) toward human inputs favoured established therapies, whereas emphasis on

LLM inputs favoured investigational ones. Across all perturbations and scenarios, the top recommendation in S3 remained Infliximab, evidencing MOJO robustness at the decision point that matters most clinically.

This combined empirical picture explains why MOJO improves on both Human-only, LLM-only, and the standard MCDA baselines. First, MOJO risk-adjusted utility layer aligns scores to a harmonised reference, preserving the behavioural asymmetries that govern real clinical choices; this is precisely where TOPSIS, VIKOR, MULTIMOORA, and AHP are silent. Second, the equilibrium-reliability layer removes the incentive for brittle "peaky" profiles that momentarily excel on a few criteria yet perform unreliably overall; the baselines reward such contrasts more readily. Third, the integration factor makes the Human-aligned solution the default outcome under high consensus, while still allowing LLM-driven innovation to influence the ordering when justified by the data. Fourth, the Monte Carlo experiment shows that when the elicited weight vector is perturbed in a statistically principled way, MOJO's winner is not an artefact of a narrow weight choice: the baselines converge toward the same decision as weights re-concentrate, and their agreement bands are tight in the scenario where the clinical signal is strongest. The higher concordance with Human-only rankings, together with better control of LLM interest for novelty in acute scenarios, demonstrates that MOJO integrates innovation without sacrificing clinical prudence.

4. DISCUSSION, IMPLICATIONS AND FUTURE DIRECTION

A central strength of the MOJO framework lies in its ability to integrate diverse inputs from both human experts and LLM outputs which are highly interpretable. Traditional MCDA methods, such as AHP or TOPSIS, depend heavily on expert judgment for weight assignment and evaluation, often leading to inconsistencies when subjective opinions vary across decision-makers. LLMs, with their capacity to process large volumes of structured and unstructured data, offer an objective counterbalance by providing standardised assessments that can be dynamically calibrated. This dual-input approach enables MOJO to identify and correct biases by systematically comparing human scores with AI-derived estimates. By quantifying discrepancies between inputs and applying corrective factors to weights, MOJO ensures that decisions are not unduly influenced by human subjectivity or data noise. Such bias mitigation is critical in healthcare settings, where even minor deviations in judgment can significantly impact treatment outcomes [42].

Beyond bias correction, LLMs enhance the reasoning process in MCDA frameworks by offering structured, stage-by-stage evaluation. Traditional decision models often present results as a black-box output, leaving stakeholders without clear insight into the rationale behind the final ranking. MOJO incorporates advanced reasoning mechanisms that enable the system to perform multi-step deliberations. This iterative process allows the framework to refine evaluations at each stage, continuously integrating feedback from both AI agents and human experts. In future iterations, this could be further strengthened by incorporating structured multi-agent debates, where different LLMs challenge one another's assessments [43]. Through structured debate, the ensemble converges on a more robust consensus that aligns statistical accuracy with human intuition, a feature particularly

valuable in complex scenarios. This approach aligns with recent advances in AI, where multi-agent systems have demonstrated improved performance in complex decision tasks [44].

The stage-by-stage evaluation enabled by MOJO also addresses the challenge of dynamic weighting. Conventional models typically keep criteria weights static, inadequately capturing the evolving importance of factors in real-world scenarios [45]. MOJO allows for adaptive weight adjustments based on continuous feedback. For instance, if a criterion such as patient safety shows higher variability between human and LLM assessments, MOJO automatically recalibrates its weight to reflect uncertainty. Such dynamic adaptation improves decision accuracy and model responsiveness, especially suitable for fast-changing fields like healthcare, where patient conditions and treatment efficacies can rapidly shift.

Another significant advantage of integrating LLMs into MCDA is the improved auditability and transparency of the decision-making process. A common critique of AI-assisted decisions is the opacity of underlying algorithms, potentially leading to mistrust among decision-makers [46]. MOJO addresses this issue by clearly delineating the contributions of both human and AI evaluations. By providing detailed intermediate scores, such as RU and PERS scores, the framework offers a clear audit trail. Decision-makers can trace back how each alternative was evaluated, understand how biases were corrected, and observe the influence of both qualitative and quantitative inputs. MOJO's collaborative approach facilitates richer dialogue among interdisciplinary teams, improving trust and decision quality through rigorous scrutiny and collective reasoning. Moreover, MOJO effectively integrates qualitative data, crucial in complex scenarios such as health where factors like patient quality of life, compliance, and satisfaction are as significant as clinical efficacy. LLMs adeptly process and synthesise textual data, quantitatively integrating qualitative insights into the decision framework. This allows for a more holistic evaluation of alternatives, capturing nuances overlooked by purely quantitative models.

Despite these advancements, several limitations remain. A primary limitation of the current framework is its lack of integrated iterative reasoning capabilities within LLMs themselves. Currently, the integration of outputs from multiple evaluators relies on automated aggregation without dynamic, iterative reasoning processes. Future developments should incorporate advanced reasoning mechanisms within LLMs, facilitating multi-step deliberations and debates, enabling the model to emulate expert human analytical processes more closely. Uncertainty quantification is limited in MOJO. Although consensus measures provide some uncertainty estimation, they do not fully capture the probabilistic nature of underlying data or inherent ambiguities in expert judgments. Future research should explore advanced probabilistic reasoning techniques to more accurately reflect confidence levels in final rankings, providing decision-makers with clearer insights into associated risks.

Another area requiring future research is the dynamic adaptation of the model to evolving decision contexts. Decision environments are rarely static, and criteria importance often changes over time. Future iterations of MOJO should include adaptive algorithms capable of real-time adjustments as new data or insights become available. Furthermore, as the complexity of the decision environment grows, computational efficiency becomes critical.

Ensuring timely, reliable AI-assisted decisions, especially in high-stakes domains like healthcare, will be essential. Future iterations must focus on optimising computational efficiency without sacrificing depth and transparency. Continued research and development combining human expertise with advanced AI reasoning will be essential to fully realise the potential of frameworks like MOJO.

5. CONCLUSION

Multi-LLM Optimised Joint Objective (MOJO) presents a highly practical and robust method for integrating human expertise with Gen-AI for multi attribute decision making problems. The practicality of the MOJO method is underscored by its flexibility and adaptability to in decision-making scenarios. Through the simulation and evaluation of a conceptual patient treatment evaluation, this study has shown that MOJO not only mitigates biases arising from subjective human judgments but also harmonises discrepancies between different evaluators by dynamically adjusting weighting factors. The use of established LLMs along with human judgements ensures that the method is both reproducible and scalable, making it accessible for real-world applications. MOJO framework emphasises on explainability, through clear quantification of consensus and bias correction mechanisms, builds trust among decision-makers and stakeholders by providing a transparent rationale for decision outcomes. While this study has used a conceptual case study using simulated data, the promising results suggest that MOJO can serve as a foundational model for next-generation decision support systems. Future work will focus on validating the framework with real-world datasets and extending its application to a broader range of decision problems. MOJO marks a significant advancement in the integration of Gen-AI into MCDA, paving the way for more objective, reliable, and practical decision-making processes.

6. REFERENCES

[1] P. A. Alvarez, A. Ishizaka, and L. Martínez, "Multiple-criteria decision-making sorting methods: A survey," *Expert Systems with Applications,* vol. 183, p. 115368, 2021/11/30/ 2021, doi: https://doi.org/10.1016/j.eswa.2021.115368.

[2] E. Triantaphyllou, "Multi-Criteria Decision Making Methods," in *Multi-criteria Decision Making Methods: A Comparative Study*, E. Triantaphyllou Ed. Boston, MA: Springer US, 2000, pp. 5-21.

[3] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

[4] C. L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*. Berlin: Springer, 1981.

[5] S. Opricovic and G. H. Tzeng, "Compromise solution by MCDM: A comparative analysis of VIKOR and TOPSIS," *European Journal of Operational Research,* vol. 156, no. 2, pp. 445-455, 2004.

[6] A. Ijadi Maghsoodi, G. Abouhamzeh, M. Khalilzadeh, and E. K. Zavadskas, "Ranking and selecting the best performance appraisal method using the MULTIMOORA approach integrated Shannon's entropy," *Frontiers of Business Research in China,* vol. 12, no. 1, p. 2, 2018/01/14 2018, doi: 10.1186/s11782-017-0022-6.

[7] W. K. M. Brauers and E. K. Zavadskas, "MULTIMOORA optimization used to decide on a bank loan to buy property," *Technological and Economic Development of Economy,* vol. 17, no. 1, pp. 174-188, 2011.

[8] A. Ijadi Maghsoodi, M. Mosavat, A. Hafezalkotob, and A. Hafezalkotob, "Hybrid hierarchical fuzzy group decision-making based on information axioms and BWM: Prototype design selection," *Computers & Industrial Engineering,* vol. 127, pp. 788-804, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.cie.2018.11.018.

[9] P. Thokala, N. Devlin, and K. Marsh, "Multiple criteria decision analysis for health care decision making—an introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force," *Value in Health,* vol. 19, no. 1, pp. 1-13, 2016.

[10] A. Mardani, A. Jusoh, and K. M. Nor, "Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014," *Economic Research-Ekonomska Istrazivanja,* vol. 28, no. 1, pp. 516-571, 2015.

[11] G. Montibeller and D. von Winterfeldt, "Biases and debiasing in multi-criteria decision analysis," in *48th Hawaii International Conference on System Sciences (HICSS)*, 2015: IEEE, pp. 1218-1226.

[12] M. Cinelli, M. Kadziński, M. Gonzalez, and R. Słowiński, "How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy," *Omega,* vol. 96, p. 102261, 2020.
[13] L. A. Zadeh, "Fuzzy logic; computing with words," *IEEE Transactions on Fuzzy Systems,* vol. 4, no. 2, pp. 103-111, 1996.
[14] A. Ijadi Maghsoodi, H. Rasoulipanah, L. Martínez López, H. Liao, and E. K. Zavadskas, "Integrating interval-valued multi-granular 2-tuple linguistic BWM-CODAS approach with target-based attributes: Site selection for a construction project," *Computers & Industrial Engineering,* vol. 139, p. 106147, 2020/01/01/ 2020, doi: https://doi.org/10.1016/j.cie.2019.106147.
[15] R. Lahdelma and P. Salminen, "SMAA – Stochastic multiobjective acceptability analysis," *European Journal of Operational Research,* vol. 131, no. 1, pp. 50-61, 2001.
[16] N. P. Suh, *Axiomatic Design: Advances and Applications*. New York: Oxford University Press, 2001.
[17] M. R. Asadabadi, M. H. Abolbashari, E. Chang, O. K. Hussain, and M. Saberi, "The stratified multi-criteria decision-making method," *Knowledge-Based Systems,* vol. 162, pp. 115-123, 2018.
[18] A. I. Maghsoodi and M. R. Asadabadi, "A Stratified Markovian Multipreference Decision Support System to Assess Supply Chain Blockchain Platforms," *IEEE Transactions on Engineering Management,* vol. 72, pp. 491-511, 2025, doi: 10.1109/TEM.2024.3521655.
[19] H. Touvron, T. Lavril, and P. Majumder, "LLaMA: Open and efficient foundation language models," 2023.
[20] OpenAi, "GPT-4 Technical Report," 2023.
[21] A. Cuthbertson, "How Elon Musk's 'spicy' Grok compares to 'woke' ChatGPT," ed, 2023.
[22] Y. Bai, S. Kadavath, and S. Kundu, "Constitutional AI: Harmlessness from AI Feedback," 2022.
[23] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020),* 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.
[24] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint arXiv:2303.12712,* 2023. [Online]. Available: https://arxiv.org/abs/2303.12712.
[25] X. Wang, J. Wei, and D. Schuurmans, "Self-consistency improves chain-of-thought reasoning in language models," in *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023.
[26] I. Svoboda and D. Lande, "AI Agents in Multi-Criteria Decision Analysis: Automating the Analytic Hierarchy Process with Large Language Models," *Available at SSRN 5069656*.
[27] X. Wang and X. Wu, "Can ChatGPT serve as a multi-criteria decision maker? A novel approach to supplier evaluation," in *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024: IEEE, pp. 10281-10285.
[28] H. Wang, F. Zhang, and C. Mu, "One for All: A general framework of LLMs-based multi-criteria decision making on human expert level," 2025.
[29] I. S. Schwartz, "Black box warning: Large language models and the future of infectious diseases consultation," *Journal of Infectious Diseases,* vol. 228, no. Supplement_2, pp. S153-S157, 2023.
[30] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021: ACM, pp. 610-623.
[31] A. Ijadi Maghsoodi, D. Riahi, E. Herrera-Viedma, and E. K. Zavadskas, "An integrated parallel big data decision support tool using the W-CLUS-MCDA: A multi-scenario personnel assessment," *Knowledge-Based Systems,* vol. 195, p. 105749, 2020/05/11/ 2020, doi: https://doi.org/10.1016/j.knosys.2020.105749.
[32] A. I. Maghsoodi, "Cryptocurrency portfolio allocation using a novel hybrid and predictive big data decision support system," *Omega,* vol. 115, p. 102787, 2023/02/01/ 2023, doi: https://doi.org/10.1016/j.omega.2022.102787.
[33] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica,* vol. 47, no. 2, pp. 263-291, 1979, doi: 10.2307/1914185.
[34] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty,* vol. 5, pp. 297-323, 1992, doi: 10.1007/BF00122574.
[35] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*: Springer, 2000, pp. 1-15.
[36] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[37] Y. Zhai *et al.*, "Uncertainty-Penalized Reinforcement Learning from Human Feedback with Diverse Reward LoRA Ensembles," *arXiv preprint arXiv:2401.00243,* 2023. [Online]. Available: https://arxiv.org/abs/2401.00243.
[38] U. Schmidt and H. Zank, "Risk Aversion in Cumulative Prospect Theory," *Management Science,* vol. 54, no. 1, pp. 208-216, 2008.
[39] F. Casellas and et al., "Patients' Experiences and Challenges in Living with Inflammatory Bowel Disease: A Qualitative Study," *Journal of Crohn's and Colitis,* vol. 15, no. 5, pp. 827-835, 2021.
[40] A. Mikocka-Walus and et al., "Health-related quality of life in inflammatory bowel disease: A systematic review and meta-analysis," *BMC Health Services Research,* vol. 22, no. 1, p. 123, 2022.
[41] R. W. Stidham and et al., "Artificial Intelligence for Disease Assessment in Inflammatory Bowel Diseases: A Systematic Review," *Gastroenterology,* vol. 162, no. 2, pp. 482-494, 2022.
[42] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR),* vol. 54, no. 6, pp. 1-35, 2021.
[43] W. Jin, H. Du, B. Zhao, X. Tian, B. Shi, and G. Yang, "A Comprehensive Survey on Multi-Agent Cooperative Decision-Making: Scenarios, Approaches, Challenges, and Perspectives," *arXiv preprint arXiv:2503.13415,* 2025. [Online]. Available: https://arxiv.org/abs/2503.13415.

[44] Q. Yang and R. Parasuraman, "A Hierarchical Game-Theoretic Decision-Making for Cooperative Multi-Agent Systems Under the Presence of Adversarial Agents," *arXiv preprint arXiv:2303.16641,* 2023. [Online]. Available: https://arxiv.org/abs/2303.16641.

[45] F. Chen, L. Wang, J. Hong, J. Jiang, and L. Zhou, "Unmasking Bias in AI: A Systematic Review of Bias Detection and Mitigation Strategies in Electronic Health Record-based Models," *arXiv preprint arXiv:2310.19917,* 2023. [Online]. Available: https://arxiv.org/abs/2310.19917.

[46] S. V. Chinta *et al.*, "AI-Driven Healthcare: A Survey on Ensuring Fairness and Mitigating Bias," *arXiv preprint arXiv:2407.19655,* 2024. [Online]. Available: https://arxiv.org/abs/2407.19655.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.png
- Highlights.docx