Evolving Systems (2026) 17:47
https://doi.org/10.1007/s12530-026-09813-1

ORIGINAL PAPER

Mitigating Linguistic Aggression in Group Decision-Making: A Comparative Analysis of AI-Driven Hostility Detection

José Ramón Trillo[1] · Juan Carlos González-Quesada[1] · Francisco Javier Cabrerizo[1] · Ignacio Javier Pérez[1]

Received: 2 March 2025 / Revised: 20 August 2025 / Accepted: 20 February 2026

Abstract

The process of group decision-making is an integral component not only for quotidian interactions but also for strategic deliberations. However, it is profoundly shaped by the inherent semantic indeterminacy of natural language. This linguistic ambiguity starkly contrasts the syntactic and semantic precision characteristic of machine-generated language. Furthermore, the conveyance of affective states–such as aggressiveness or elation–via natural language introduces a layer of complexity that can significantly perturb the equilibrium of the group decision-making process. In response to these challenges, we propose an advanced consensus-reaching methodology based on sentiment analysis to quantify and mitigate aggressiveness in discourse. This study conducts a comparative evaluation of three state-of-the-art large language models: Gemini, Copilot, and ChatGPT for their efficacy in detecting and assessing hostility. By calibrating the influence of individual participants based on their degree of linguistic aggression, the proposed framework attenuates the disproportionate impact of dominant voices, thus fostering a more balanced and equitable deliberative environment. This methodological innovation not only incentivizes the adoption of a more dispassionate and constructive linguistic register but also safeguards the integrity of collective decision-making processes against the distortive effects of undue emotional influence. Across five repeated evaluations per comment, ChatGPT and Gemini exhibited $< 5\%$ variance, while Copilot showed $\approx 8 - 12\%$; in all cases, hostility-aware weighting reduced the most aggressive expert's influence by $\approx 27 - 29\%$, yielding robust group rankings. These mechanisms improve consensus quality by reducing bias from aggressive discourse, and they are expected to foster higher group satisfaction through perceived fairness in deliberation. Potential improvements include benchmarking against gold standards, extending to multilingual and multimodal contexts, and enhancing transparency for end-users.

Keywords Sentiment Analysis · Group Decision-Making · Large Language Model · Consensus Methods

1 Introduction

Decision-making, a common activity that spans both professional and personal domains, is deeply influenced by emotions (Sayegh et al. 2004; Gaudine and Thorne 2001). This emotional connection can significantly affect the quality of the decisions made, sometimes leading to suboptimal outcomes. This can occur due to factors such as the individual's mood or the mood of others involved in the process (Simon 1987). For example, emotions such as anxiety, enthusiasm, or even anger can cloud judgement, leading to impulsive, biased decisions or choices that result in regret and discomfort later (Paulus and Angela 2012). This issue is especially relevant in group decision-making (GDM) contexts, where emotional dynamics can be amplified by

These authors contributed equally to this work.

✉ Ignacio Javier Pérez
ijperez@decsai.ugr.es

José Ramón Trillo
jrtrillo@ugr.es

Juan Carlos González-Quesada
juancarlosgq@ugr.es

Francisco Javier Cabrerizo
cabrerizo@decsai.ugr.es

[1] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence, DaSCI, University of Granada, Granada 18071, Spain

Published online: 14 March 2026

interactions between participants (Morente-Molinera et al. 2019; Ruz et al. 2020; Trillo et al. 2024).

Fuzzy logic-based approaches have been widely used to address the inherent uncertainty and subjectivity present in GDM. In particular, soft consensus models aim to balance flexibility and rigor in decision-making by allowing varying degrees of agreement rather than enforcing strict unanimity. This approach acknowledges that absolute consensus is often unrealistic and instead focuses on reaching a level of agreement that is acceptable to all participants (Cabrerizo et al. 2010). By leveraging fuzzy sets, these models provide a more nuanced and interpretable framework for measuring consensus, ensuring that diverse perspectives are integrated while mitigating conflicts that may arise during discussions.

In GDM processes, a group of individuals, usually experts in the problem being analyzed, participates in a collaborative debate. During this debate, each participant expresses his/her opinions and proposes different alternatives or courses of action, bolstered by strong arguments based on data or relevant experiences (Yager 1981). The goal is not only to find the best solution to the problem but also to integrate and fully leverage the ideas, perspectives, and knowledge of all group members (Herrera-Viedma et al. 2002).

One of the main objectives of GDM is to reach a consensual solution. This consensus not only ensures that the final decision reflects a balance of different opinions but also promotes greater acceptance and commitment from the participants to implement the agreed-upon decisions. The underlying premise is that individuals are more likely to feel committed to decisions that have been consensually agreed upon, as these reflect a more democratic and participatory process (Herrera-Viedma et al. 2014). Additionally, consensus has the potential to reduce conflicts within the group, strengthening cohesion and increasing the effectiveness of future collaborations.

However, achieving consensus in a group setting is not without its challenges. Emotions can play both a constructive and destructive role in this process. On the one hand, positive emotions such as empathy and cooperation can facilitate the search for common ground, while negative emotions such as frustration or wounded pride can hinder negotiation and open dialogue. These complex emotional dynamics underscore the importance of effectively managing emotions during group decision-making processes.

In group decision-making (GDM) processes, experts use natural language, which is inherently ambiguous compared to machine language, making mutual understanding difficult (Pilehvar and Camacho-Collados 2021). Moreover, when experts express their opinions, they not only communicate objective ideas but also emotions, such as aggression, which can influence how others perceive and react to their arguments (Wankhade et al. 2022). This emotional component can introduce biases into the decision-making process, favoring the opinions of those who present their comments more aggressively rather than relying solely on the validity of the perspectives offered (Pérez et al. 2013; Cabrerizo et al. 2014).

This phenomenon highlights how emotions and communication style can negatively affect the deliberation process and the quality of group decisions (Svenson et al. 2024). Emotions such as aggression can shift attention away from rational arguments toward more subjective aspects, which can lead to suboptimal decisions. To mitigate these biases, the use of technological tools such as sentiment analysis and proper moderation of interactions can be useful, allowing for a more equitable decision-making process focused on the content of the ideas rather than the emotions involved. Linguistic aggression, in particular, exerts a profound influence on group decision-making dynamics. When participants express their views in a hostile or confrontational manner, they may dominate the discourse and intimidate others, thereby reducing the willingness of less assertive members to contribute. This imbalance not only diminishes the diversity of perspectives considered but also fosters conflict escalation and undermines cooperative negotiation. Consequently, outcomes can be biased toward the preferences of aggressive individuals rather than reflecting the collective rationality of the group. Addressing this phenomenon is central to our proposal, which introduces a hostility-aware weighting mechanism to mitigate the undue influence of aggressive discourse and to ensure that decisions are based on the substantive validity of arguments rather than on communicative dominance.

To avoid undue influence from one expert over others, we propose a new linguistic extension of a consensus model for group decision-making, incorporating sentiment analysis to detect the degree of hostility of each expert. This approach addresses the psychology of negotiation and uses the power of a fuzzy ontology as a tool for influence, enhancing the precision and realism of group decision-making scenarios. We employ generative artificial intelligence (GAI), specifically an advanced large language model (LLM), which receives the expert's comment as input and returns an evaluation of the degree of hostility. This comment is processed using a regular expression to extract the level of aggression.

In such a way, we will conduct tests using three different artificial intelligence platforms. Copilot, ChatGPT, and Gemini. Each "comment" will be sent five times through these AIs to obtain, according to each model, the level of aggressiveness. This repeated evaluation will allow us to obtain a more accurate and reliable assessment of the hostility level in the comments, minimizing any bias that might arise from individual interpretations. In addition, it will help adjust the weight of each expert in the decision-making

process, so that those who present more aggressive comments will see their influence reduced in the group. By performing this repeated analysis, we can establish a more robust and equitable decision-making system in which emotions do not disproportionately interfere with the final outcome.

Despite significant advances in sentiment analysis and consensus-reaching models, prior studies have primarily relied on supervised classifiers, dictionary-based methods, or manual moderation, all of which face scalability and adaptability limitations. In addition, existing approaches seldom incorporate the outputs of advanced Large Language Models (LLMs) into the mathematical core of decision-making frameworks. Motivated by these gaps, the objective of this study is to design and validate a novel methodology that integrates hostility detection into group decision-making processes, thereby ensuring more equitable participation and minimizing the disproportionate influence of aggressive communicative styles.

Guided by this objective, the study addresses the following research questions: (i) To what extent can state-of-the-art LLMs–specifically ChatGPT, Copilot, and Gemini–reliably detect and quantify hostility in deliberative exchanges? (ii) How can hostility scores be systematically incorporated into consensus-reaching models through a weighting mechanism that moderates the influence of aggressive participants? and (iii) What are the comparative strengths and limitations of these LLMs in shaping the outcomes of group decision-making? These questions frame the analytical trajectory of the paper and ground its contributions within both theoretical and applied perspectives.

The novelty of this study lies in the systematic integration of advanced LLMs–namely ChatGPT, Copilot, and Gemini–into group decision-making frameworks. Unlike previous approaches, which relied on supervised classifiers, sentiment dictionaries, or manual moderation, our methodology embeds hostility detection directly into consensus models by means of a weighting mechanism that penalizes aggressive contributions. Furthermore, by comparing multiple state-of-the-art LLMs and employing repeated evaluations to mitigate stochastic variability, this work provides the first rigorous analysis of LLM-based hostility detection in deliberative contexts.

While the framework draws inspiration from established models of fuzzy consensus and behavioral weighting, the originality of this work lies in extending such models by embedding hostility detection derived from advanced LLMs. This integration creates a novel methodological bridge between natural language processing and consensus-reaching theory.

The rest of the paper is organized as follows. Section 2 delves into the foundational principles that underpin the GDM system. Subsequently, Section 3 articulates the proposed framework in detail. In Section 4, a concrete case study is presented to elucidate its practical implementation. Section 5 offers a critical examination of the system's advantages and limitations, contextualizing its performance through a comparative analysis with existing methodologies in the literature. Section 6 encapsulates the principal contributions of this work, providing a synthesis of the proposed approach. Finally, the conclusions of this study are presented in Section 7.

2 Preliminaries

In this section, the fundamental concepts linked to the proposed method will be presented. Section 2.1 will address the concepts associated with group decision-making (GDM), while Section 2.2 will explore the aspects related to Generative Artificial Intelligence (GAI), with a special focus on the essential concepts related to Large Language Models (LLMs).

2.1 Group Decision-Making Methods

In this section, we delve into the fundamental concepts of GDM within a specified framework. We consider a finite set of individuals, denoted as $\Xi = \xi_1, \ldots, \xi_M$, tasked with selecting from a finite set of alternatives, denoted as $\Gamma = \gamma_1, \ldots, \gamma_N$ (Nurmi and Kacprzyk 1991; Kacprzyk et al. 2019; Trillo et al. 2023). Each individual, drawing upon their knowledge, evaluates pairs of alternatives and designates their preference for one over the other. This evaluation involves the provision of input, which may manifest as linguistic labels (Herrera-Viedma et al. 2014; Kabak and Ervural 2017) or numerical arrays (Taghavi et al. 2020). The input, assumed to be reciprocal preference relations (Pramanik and Mukhopadhyaya 2011), is expressed as P_s, where s ranges from 1 to M. In this context, each individual $e_s \in E$ compares pairs of alternatives γ_i and γ_j, with the outcome denoted as $p_{ij}^s \in [0, 1]$.

Having established the foundational concepts, we can now delineate the various components integral to a GDM method:

- **Debate and Opinion Formation**: In the initial phase, individuals engage in discussions encompassing the array of available alternatives. Within this discourse, they articulate their thoughts and preferences, elucidating the rationale behind their choices. After this deliberation, the individuals consolidate their insights into reciprocal preference relations.

- **Consensus Analysis**: In this phase, the reciprocal preference relations provided by individuals become the basis for assessing disparities in viewpoints. The consensus threshold, a critical benchmark for consensus, is set, and its attainment determines whether a consensus in the group's opinions is established (Qin et al. 2022; Taghavi et al. 2020). Exceeding this threshold leads to the aggregation of reciprocal preference relations. Otherwise, if the threshold is not reached, a feedback process is initiated, prompting individuals to engage again in discussions to achieve mutual agreement (Trillo et al. 2023; Morente-Molinera et al. 2022).
- **Aggregation of Individual Information**: Utilizing insights contributed by individuals, information is aggregated to formulate a unified collective reciprocal preference relation denoted as P. This aggregation employs operators, with our method utilizing the Weighted Average (WA) operator.
- **Ranking of Alternatives**: The final step involves leveraging the collective reciprocal preference relation to determine the preferred alternative(s) among individuals. This necessitates the application of an operator on the collective reciprocal preference relation (Yager and Kacprzyk 2012). In our method, we employ the Quantifier-Guided Degree of Dominance (QGDD) operator (Trillo et al. 2022; Yager 1996).

The outlined process establishes a structured methodology for effectively navigating group decision-making scenarios. These components collectively ensure that the preferences of the group are considered and harmonized, facilitating the identification of the optimal alternatives. Furthermore, they provide a mechanism for handling disagreements and fostering consensus, which is crucial in collaborative decision-making environments.

2.2 Generative Artificial Intelligence and Large Language Models

Generative Artificial Intelligence (GAI) refers to systems that can generate new and original content, such as text, images, music, etc. Goodfellow et al. (2014). One of the most popular approaches is the use of generative neural networks, such as Generative Adversarial Networks (GANs) (Brown et al. 2020) and Autoregressive Generative Models (Magalhães et al. 2023).

Large Language Models (LLMs) are advanced natural language processing systems that use deep learning techniques to understand and generate text in a contextually relevant way. To improve clarity, the main features of LLMs are grouped into three categories: *Architectural Foundations*, *Training and Adaptation*, and *Capabilities* (Devlin et al. 2018):

Architectural Foundations. From an architectural perspective, LLMs rely on several mechanisms that allow them to capture long-term dependencies and contextual relationships in text:

- **Transformer architecture:** The Transformer architecture employs attention mechanisms to process information in parallel and capture long-term dependencies in text. It comprises attention layers and feed-forward layers (Vaswani et al. 2017).
- **Multi-headed attention:** This mechanism enables the model to focus on different parts of the input simultaneously, improving its ability to process context (Buehler 2023).
- **Transformer layers:** Each Transformer layer consists of attention and feed-forward sublayers. These components capture contextual relationships and model complex, non-linear dependencies (Miao et al. 2023). In our case, we rely on the pre-trained embeddings natively used by Gemini, Copilot, and ChatGPT, which ensure semantic consistency across comments without requiring additional training.***Training and Adaptation.*** In terms of training, LLMs typically follow a two-stage process and employ different techniques:

- **Pre-training and fine-tuning:** LLMs are trained in two stages. During pre- training, the model learns language structures and representations using large-scale unlabeled data. Fine-tuning adapts the model to specific tasks using labeled datasets, such as text summarization or question answering (Ozdemir 2023). It is important to emphasize that no fine-tuning was applied to any of the evaluated LLMs. All models were used in their native pre-trained form, accessed through their official APIs. This design choice ensures that the methodology is reproducible under real-world conditions, where users typically interact with LLMs as off-the-shelf systems without direct access to model parameters.
- **Employing word embeddings:** In the pre-training stage, dense vector representations are created for each word, capturing semantic and syntactic relationships. Fine-tuning adjusts these embeddings to optimize performance for specific tasks (Levy and Goldberg 2014).
- **Autoregressive modeling:** Models like the GPT series generate text sequentially, predicting each word based on preceding context. This sequential approach ensures the coherence of generated text (Alberts et al. 2023).***Capabilities.*** Finally, regarding their functional abilities, LLMs demonstrate remarkable strengths:

- **Generation and comprehension capabilities:** LLMs can generate text relevant to a given task while understanding input context, enabling applications such as translation, summarization, and question answering (Nicula et al. 2023).
- **Contextual capacities:** LLMs excel at capturing context over extended sequences, making them capable of generating more coherent and contextually appropriate responses (Huang et al. 2023).

To reinforce the theoretical grounding and demonstrate alignment with current academic progress, we have expanded the literature review by integrating recent high-impact contributions. Zhang et al. (2023) conduct a comprehensive evaluation of large language models (LLMs) across 26 datasets for diverse sentiment-analysis tasks, revealing strong performance in simpler settings and few-shot learning, albeit limitations in more complex scenarios. In the realm of toxicity and hate-speech detection, Zhuo et al. (2025) review LLM-based mitigation strategies within software-engineering contexts and empirically show that rewriting techniques can effectively reduce toxic language, while (Kumarage et al. 2024) assess the strengths and challenges of LLMs in hate-speech classification, highlighting both their promise and current limitations. Additionally, Pangtey et al. (2025) contribute a systematic survey of stance detection methods powered by LLMs, offering a valuable taxonomy and identifying emerging evaluation challenges. Finally, Krugmann and Hartmann (2024) benchmark GPT−3.5, GPT-4, and Llama 2 for sentiment analysis accuracy, interpretability, and reproducibility, discussing critical considerations such as dataset characteristics, model biases, and computational cost. Complementing these advances, recent studies have also explored deep learning approaches beyond purely text-based analysis. For instance, the FSTL-SA model (Meena et al. 2024) introduces a few-shot transfer learning strategy for sentiment analysis from facial expressions, showing how multimodal deep architectures can capture affective signals with limited training data. Such contributions highlight the broader applicability of deep learning to sentiment and hostility detection, and reinforce the importance of investigating LLMs as part of a wider ecosystem of advanced AI methodologies.

By integrating ChatGPT, Copilot, and Gemini into GDM processes, the potential for automating and moderating group interactions becomes evident. These three LLMs have been selected due to their distinct architectures, capabilities, and widespread adoption, making them ideal candidates for comparative analysis in decision-making contexts (Brown 2020). ChatGPT is renowned for its conversational depth and contextual adaptability, Copilot is optimized for task-oriented assistance and code generation, and Gemini stands out for its multimodal processing and integrative reasoning. Their differences allow for a rigorous evaluation of how various LLM-driven approaches impact group discussions, particularly in assessing the influence of communication styles, including aggressive language, on decision dynamics. Such insights could significantly enhance the understanding and implementation of GDM methodologies across multiple domains.

In current literature, there has been a notable surge in the development and application of algorithms associated with Large Language Models (LLMs) and Artificial Intelligence (AI). These technologies are being increasingly utilized to tackle complex challenges and drive innovation across diverse fields. For instance, in Mohammed and Venkataraman (2023), an advanced AI mechanism is proposed for monitoring Parkinson's disease. This system analyzes patient data using LLM-driven algorithms, identifying nuanced patterns that traditional methods might overlook. By providing precise insights into disease progression, this approach enhances diagnostic accuracy and supports more personalized healthcare solutions.

Similarly, Pearce et al. (2023) presents an algorithm that leverages LLMs to assist in software development by detecting and correcting coding errors. This capability improves real-time debugging efficiency, reducing development time and enabling the creation of more robust software. Furthermore, Strader et al. (2020) demonstrates the use of generative AI in market analysis, where models are employed to generate insights, predict consumer behaviour, and guide strategic business decisions. These examples highlight the versatility of LLMs and generative AI, showcasing their ability to address domain-specific challenges and deliver transformative results across healthcare, software engineering, and business intelligence.

Recent years (2023–2025) have witnessed an acceleration in research on sentiment-aware recommendation systems, introducing novel architectures and application domains. For instance, in Meena et al. (2024) proposed a BiLSTM-based sentiment classifier combined with an LSTM-based POI recommendation engine applied to Foursquare data, reporting excellent classification results but more modest recall in POI ranking. Similarly, in Gajula (2025) presented a comprehensive survey of sentiment-aware recommender systems, covering transformer-based methods, graph neural networks, and conversational recommenders, outlining key methodological trends. In Darraz et al. (2025) integrated BERT-based sentiment analysis with recommendation strategies in the hospitality sector, showing improvements in personalized services for Yelp datasets (see Table 1).

In summary, while these works primarily focus on enhancing recommendation quality in consumer-oriented domains, our study departs from this trajectory by embedding

Table 1 Comparison with recent sentiment-aware recommender systems (2023–2025)

Study (Year)	Approach / Techniques	Dataset	Metrics / Highlights	Comparison with Our Work
Meena et al. (2024)	BiLSTM for sentiment + LSTM for POI recommendation	Foursquare (POI data)	Classification accuracy, precision, recall, $F1 \approx 99.5\%$; POI recall 48.5%, precision 85%	Focuses on consumer POI recommendation; our work targets fairness in group decision-making via hostility detection
Gajula (2025)	Survey of sentiment-aware recommenders: transformers, GNNs, conversational recommenders	–	Synthesis of methodological trends (2023–25)	Provides broad taxonomy; our work extends this landscape by applying LLM-based hostility detection to decision-making
Darraz et al. (2025)	BERT-based sentiment integration for recommendations	Yelp (restaurants, hotels)	Improved personalization in hospitality recommendation	Domain-specific (hospitality); our work diverges by embedding hostility detection into fuzzy consensus models
Our Work (2025)	LLM-based hostility detection + fuzzy consensus model	Group decision-making dialogues	Robustness to aggression-induced bias; equitable weighting	First to integrate hostility-aware weighting into consensus-reaching mechanisms

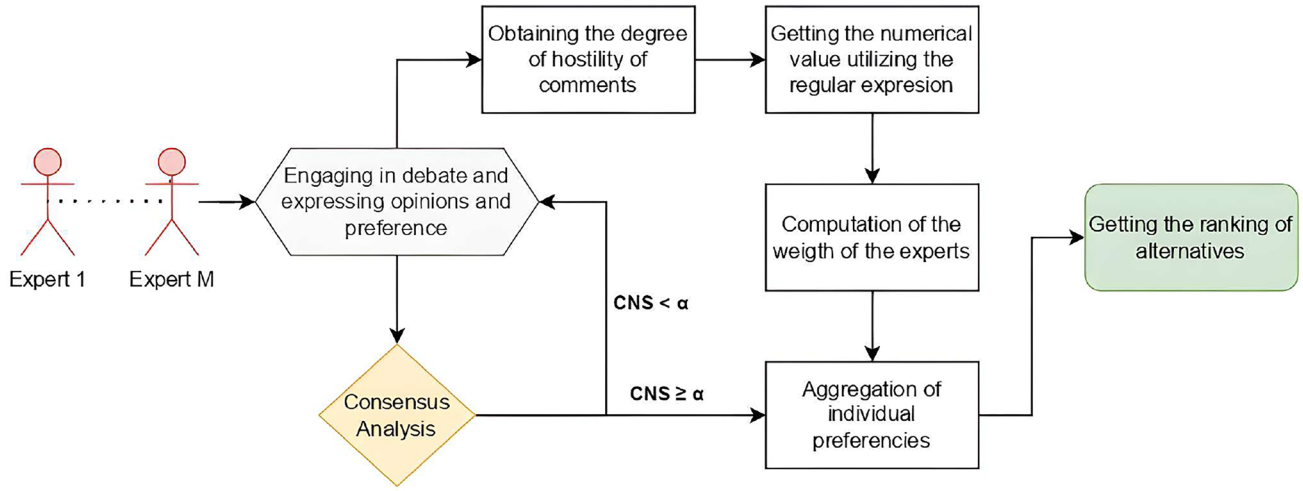

Fig. 1 Diagram of the proposed method. The notation "$CNS > \alpha$" indicates the condition that the computed consensus level (*CNS*) must exceed the predefined threshold (α) for the decision-making process to proceed

sentiment (specifically hostility detection) directly into group decision-making frameworks. This methodological distinction situates our contribution within the broader landscape of sentiment-aware systems, but emphasizes fairness and equity in collective decisions rather than individual consumer satisfaction. Thus, our framework represents the first integration of hostility-aware weighting into fuzzy consensus models, complementing prior advances while addressing a distinct class of decision-making challenges.

3 Evaluating Hostility Detection in GDM

This section introduces a novel group decision-making (GDM) approach that incorporates generative artificial intelligence (GAI) to detect hostility in comments and optimize decision-making processes. The proposed method consists of seven systematic steps, as illustrated in Fig. 1. The rationale for structuring the methodology into seven steps lies in the need to ensure both granularity and transparency in each phase of the decision-making process. By isolating hostility detection, weighting, and consensus analysis, the framework allows each component to be rigorously validated. The choice of Gemini, ChatGPT, and Copilot responds to their architectural diversity and complementary capabilities: while Gemini emphasizes multimodality, ChatGPT is recognized for conversational coherence, and Copilot for task-specific guidance. This variety strengthens the robustness of the comparative analysis and enhances the generalizability of our findings across different LLM families. It is important to note that our comparative analysis does not involve direct access to or examination of the internal parameters of

Table 2 Experts and Their Corresponding Comments

Expert	Comment
ξ_1	I'll start by saying that coal is unacceptable It has no place in any serious discussion about modern heating systems
ξ_3	Here we go again... Sure, coal is bad, but do you have any realistic idea for replacing it without bankrupting people?
ξ_2	Yes, I have an idea: solar and aerothermal But I suppose you're going to say it's "too expensive" as always

Gemini, ChatGPT, or Copilot, as these remain proprietary. Instead, the comparison relies exclusively on their observable outputs, obtained under identical experimental settings through official APIs. By standardizing the prompts, repetition protocol, and preprocessing steps, we ensure that the evaluation focuses on the consistency, reliability, and accuracy of the hostility scores generated by each model, rather than on inaccessible architectural or training parameters. Below, we elaborate on each step in detail:

- **Engaging in Debate and Expressing Opinions and Preferences**: In the initial phase, participating experts engage in a structured debate, allowing them to articulate and exchange their ideas regarding the available alternatives. Following the discussion, each expert quantitatively expresses their preferences through a reciprocal preference relation, represented as a numerical set within the interval [0, 1]. This enables a structured evaluation of pairwise comparisons among the alternatives, serving as the foundation for subsequent analysis. Hostility values, initially expressed as percentages by the LLMs, are normalized to the range [0, 1] in order to standardize subsequent computations.
- **Obtaining the Degree of Hostility in Comments**: In this step, the comments made by each expert during the debate are collected and sent to the three LLMs considered in this study—Gemini 1.5 Flash, ChatGPT 4.0, and Copilot— application via its API. They use advanced generative AI to analyze the text and return a detailed assessment of the aggressiveness level in each comment. This automated analysis ensures unbiased and consistent evaluation of hostility levels across all participants, which is critical for maintaining a constructive decision-making environment.
- **Extracting Numerical Values with Regular Expressions**: Once the three LLMs considered in this study have returned the textual descriptions of the aggressiveness levels, these outputs are parsed to extract numerical values using regular expressions. The extracted values represent the hostility level of each comment on a numerical scale. This step converts qualitative assessments into quantitative data that can be incorporated into the weighting calculations for the experts.
- **Computation of the Weights of the Experts**: With the numerical hostility values obtained, the next step involves calculating the weight assigned to each expert. The weights are inversely proportional to the hostility levels detected in their comments, reflecting their constructive contributions to the group discussion. Experts with lower hostility levels are given higher weights, as their inputs are deemed more conducive to fostering consensus and rational decision-making. This step ensures that the decision-making process prioritizes contributions that enhance collaboration and reduce conflict.
- **Consensus Analysis**: After assigning weights, the system evaluates the degree of agreement among the experts' preferences to determine whether a consensus has been reached. A predefined consensus threshold is established, which must be surpassed for the group to proceed to the next stage. If the consensus level falls short of the threshold, a feedback loop is triggered, prompting the experts to revisit their preferences and engage in further discussions aimed at aligning their viewpoints.
- **Aggregation of Individual Preferences**: Once the consensus threshold is met, the individual preferences are aggregated to form a collective preference relation. This aggregation is conducted using the Weighted Aggregation (WA) operator, which incorporates the experts' weights and their individual preference relations. The resulting collective preference matrix reflects the group's overall evaluation of the alternatives, integrating the weighted influence of each expert's contributions.
- **Getting the Ranking of Alternatives**: In the final step, the alternatives are ranked based on the collective preference matrix. The Quantifier-Guided Degree of Dominance (QGDD) operator is applied to assign a score to each alternative, reflecting its relative dominance within the group. These scores are then sorted in descending order to produce a prioritized ranking of alternatives. This ranking provides a clear representation of the group's preferences, facilitating informed decision-making.

The proposed methodology leverages the capabilities of generative artificial intelligence (GAI) to ensure that the decision-making process is not only efficient but also constructive and collaborative. By incorporating hostility detection, this method addresses a key challenge in group decision-making–managing conflicts and fostering productive discussions. Additionally, this approach compares the effectiveness of three advanced GAI tools–Gemini, ChatGPT, and Copilot–to determine which system provides the

most accurate and consistent measurement of aggression in comments. This innovative comparative analysis showcases the potential of GAI to enhance traditional decision-making frameworks, paving the way for more inclusive and effective collaborative processes.

3.1 Engaging in Debate and Expressing Opinions and Preferences

During the debate phase, participants, represented as $\xi_s \in \Xi$ where $\Xi = \xi_1, \ldots, \xi_M$, express their opinions and discuss various alternatives, denoted as $\gamma_i \in \Gamma$, with $\Gamma = \gamma_1, \ldots, \gamma_N$. Each participant contributes comments to comprehensively evaluate and explore the available options.

At the conclusion of the discussion, participants articulate their preferences through reciprocal preference relations, denoted as P_s for each participant $\xi_s \in \Xi$, as described in Section 2.1. Participants are not required to evaluate all pairwise comparisons; instead, they may selectively assess pairs of alternatives that align with their priorities.

These reciprocal preference relations are organized as $N \times N$ matrices with an empty main diagonal. Each matrix numerically represents the participant's pairwise preferences among the alternatives. Participants retain the flexibility to use individualized scales for expressing preferences, allowing for tailored evaluations while maintaining the coherence of the collective decision-making process. In line with the methodological outline presented in Section 3, the hostility values associated with each comment are normalized to the interval [0, 1]. This normalization step ensures consistency across different outputs and facilitates their integration in the subsequent weighting and consensus analysis.

3.2 Obtaining the Degree of Hostility of Comments

The second step involves recording all comments made by participants in chronological order to preserve the sequence of interventions. This record forms the basis for analysing the level of hostility in the comments. To achieve this, the hostility detection capabilities of Gemini, ChatGPT, and Copilot are tested and compared to determine which tool provides the most reliable and accurate results.

The process begins by establishing a connection to the respective APIs of Gemini, ChatGPT, and Copilot. These tools were selected due to their advanced LLM-based architectures and accessibility. Once connected, structured prompts are provided to ensure consistent outputs across all systems, thereby facilitating the extraction of numerical values for hostility. The prompt format is standardized as follows:

The sentence - The percentage of aggressiveness of this comment is *X*.

Here, X represents the percentage of aggressiveness detected in the comment. Each tool processes the input text and outputs a value for X, which is subsequently recorded for analysis.

By comparing the results from Gemini, ChatGPT, and Copilot, this methodology aims to identify the most accurate system for quantifying hostility. The comparison ensures that the decision-making framework integrates the most effective hostility detection tool, enhancing the accuracy and reliability of the analysis. This structured and comparative approach not only simplifies the integration of GAI into the decision-making process but also ensures replicability and transparency, key factors for fostering confidence in the system's outcomes. Each comment was submitted five times to each model in order to reduce the stochastic variability inherent in LLM outputs. This repeated sampling strategy ensures that the hostility score attributed to each comment reflects a stable central tendency rather than a single model instance. By averaging across repetitions, the influence of outlier responses or occasional inconsistencies is minimized, thereby improving the reliability of the analysis.

Once the prompt has been set, the comments are submitted for analysis. However, prior to sending the comments to the API, a preprocessing step is performed to address potential formatting issues that might interfere with the extraction of aggressiveness values. Specifically, any comment containing the % symbol undergoes a transformation where the symbol is replaced with the phrase "*per cent*". This ensures the reliability of the extraction process, as the % symbol could otherwise lead to parsing errors or inaccuracies in the returned response. For example, the sentence "*90 % of people think the same as me*" is transformed into "*90 per cent of people think the same as me*". It should be noted that the reliance on a single prompt formulation may constrain the reproducibility of the experiments. Slight variations in wording can influence the responses generated by LLMs, potentially introducing variability. In this study, the prompt was intentionally kept fixed across Gemini 1.5 Flash, ChatGPT 4.0, and Copilot in order to ensure comparability, but it should not be regarded as a definitive or optimized formulation.

Once this transformation is complete, the comments are sent to the API, which connects to the server and processes the input text according to the predefined prompt structure. The API subsequently returns the output in the desired format, facilitating the next stages of analysis.

It is important to note that hostility detection in this framework relies on the internal embedding mechanisms of the selected LLMs. Each comment is first transformed into dense vector representations by the models' embedding layers, which capture semantic and syntactic properties of the text. These embeddings serve as the foundation for

subsequent model operations, enabling the detection of hostile or aggressive expressions. While we did not explicitly extract or fine-tune embedding vectors, the hostility scores returned by Gemini, ChatGPT, and Copilot are computed on the basis of these representations. Thus, embeddings are implicitly leveraged within our methodology as the latent structure underpinning the models' judgments. It is important to note that the hostility scores generated by the LLMs are not interpreted as precise or absolute numerical truths, but as relative indicators that can be compared and aggregated. This mitigates the inherent limitations of LLMs in handling numbers as tokens without intrinsic mathematical grounding.

3.3 Getting the Numerical Value Utilizing the Regular Expression

The comments returned by the LLMs (e.g., Gemini, ChatGPT, and Copilot) follow the format specified in the prompt. However, it is essential to account for the possibility of *hallucinations*–responses that deviate from the expected structure due to the model generating inaccurate or irrelevant content. These hallucinations can lead to errors in the analysis pipeline if not managed effectively.

To mitigate this risk and ensure the accurate extraction of the aggressiveness value X_s^t (where s represents the expert and t the comment number, with T_s total comments per expert), a robust regular expression is implemented. The proposed regular expression is as follows:

$$r'(\backslash d+)\%' \tag{1}$$

This regular expression is designed to search for text patterns containing one or more digits, followed by the % symbol. The numeric value captured by the pattern "$\backslash d$" represents the measure of the aggressiveness of the comment, denoted as X_s^t. This value lies in the [0, 100] range, reflecting the aggressiveness scale associated with each expert's comment. To reduce the stochastic variability of LLM outputs, each query was repeated five times, and the final weight was computed as the mean of these five attempts. It is worth noting that the use of regular expressions for extracting hostility percentages from LLM outputs, although functional, is inherently fragile and may be prone to parsing errors. This strategy was adopted as a pragmatic choice to demonstrate feasibility. In future developments, more reliable extraction techniques—such as structured prompting, function calling, or requiring JSON-formatted outputs—will be incorporated to ensure robustness and eliminate potential inconsistencies.

Implementing this regular expression allows for the accurate and consistent extraction of the relevant information, ensuring that any aggressive sentiment in the comments is quantified precisely. By leveraging this approach, we can handle unexpected variations in the API's output (such as hallucinations or formatting inconsistencies) and maintain the robustness of the analytical process. Since no model fine-tuning was involved, the outputs correspond to raw inferences from the respective APIs, ensuring that the comparative analysis reflects the baseline performance of each system

3.4 Computation of the Weight of the Experts

Once the value of each comment has been extracted, the weight of each expert is calculated based on the aggressiveness level of their comments. The procedure for determining the weight of experts is divided into two distinct parts. The first part involves computing the raw weight, which is individually assigned to each participant, denoted as $W_s \in \mathbb{R}$. The weighting scheme is designed to inversely correlate hostility levels with expert influence. This principle is grounded in consensus theory, where aggressive interventions may distort deliberative balance. By penalizing higher hostility scores, the method privileges constructive participation, aligning the process with democratic decision-making norms. The calculation for this weight is outlined as follows:

$$W_s = \frac{\sum_{t=1}^{T_s} 1 - \frac{X_s^t}{100}}{T_s};\ s = 1, \ldots, M \tag{2}$$

In this equation, W_s represents the raw weight of the expert s. The value X_s^t corresponds to the arithmetic mean, after repeating the experiment five times, of the level of aggressiveness of the t-th comment made by expert s. The weight is computed by summing the inverse of the aggressiveness for each comment, which helps to ensure that experts who make more hostile or aggressive comments are given lower weight, reflecting their less constructive contributions to the group discussion. This computation accounts for the overall impact of each expert's level of hostility on the decision-making process, ensuring that the final collective preferences are influenced more by those who contribute in a more neutral or cooperative manner.

Once the raw weights are calculated for each expert, the next steps involve refining the weights further to enhance the fairness and precision of the decision-making process.

After determining the gross weight of each individual, we proceed to the calculation of their relative weight, denoted as $w_\iota \in [0, 1]$. This relative weight represents the weighting of an individual relative to all other participants, thus ensuring that the contribution of each participant is

proportionate to their calculated weight. Consequently, the sum of all relative weights w_ι must equal 1, maintaining the total weight as a normalized value. This normalization ensures that the collective decision-making process remains balanced and reflects the varying degrees of influence that different experts have on the final outcome.

The formula for calculating the relative weight w_ι of expert ι is as follows:

$$w_\iota = \frac{W_\iota}{\sum_{s=1}^{M} W_s}; \ \iota = 1, \ldots, M \tag{3}$$

Where w_ι denotes the non-hostility weight associated with expert ι, computed as $(1 - h_\iota)$, and $\sum_{s=1}^{M} W_s$ represents the total sum of such weights across all *M* experts. Accordingly, w_ι expresses the normalized weight of expert ι. This formula ensures that the relative weight is calculated as the proportion of each expert's gross weight relative to the total weight, which will later be used in the aggregation process.

3.5 Consensus Analysis

Once the reciprocal preference relationships have been established by the experts, the next step in the methodology is the consensus analysis. The purpose of this analysis is to measure the degree of agreement or disagreement among the experts' preferences. To do this, a consensus threshold $\alpha \in [0, 1]$ is set. The consensus value, denoted as $CNS \in [0, 1]$, must exceed this threshold for the group to be considered in agreement. If the consensus value falls below the threshold, it indicates that there is insufficient agreement among the experts, triggering a feedback process where experts are asked to revisit and possibly revise their preferences in order to reach a higher level of consensus. To avoid an unproductive cycle of repeated revisions, a predefined number of rounds, denoted as ρ, is specified to limit the number of times the experts can return to the discussion. For the purposes of this study, ρ is set to 10, which ensures that the feedback process does not become unnecessarily infinite.

The consensus threshold was set at $\alpha = 0.9$ in order to guarantee a high degree of agreement among participants, reflecting the standard adopted in similar GDM frameworks. Meanwhile, the maximum number of revision rounds ($\rho = 10$) was introduced to avoid infinite feedback loops while still allowing sufficient opportunities for alignment. This balance ensures both methodological rigor and practical feasibility.

The consensus value is calculated using the Euclidean distance formula, which measures the collective differences in preferences among the experts. Specifically, the formula for calculating the consensus value *CNS* is as follows:

$$CNS = 1 - \frac{2 \cdot \sum_{s=1}^{M-1} \sum_{k=1;s>k}^{M} \frac{\sqrt{\sum_{i=1}^{N} \sum_{j=1;i\neq j}^{N} (p_{ij}^{s} - p_{ij}^{k})^2}}{N \cdot N - N}}{(M-1) \cdot M} \tag{4}$$

In this formula, p_{ij}^{s} and p_{ij}^{k} represent the pairwise preferences expressed by experts *s* and *k*, respectively, for alternatives *i* and *j*. The Euclidean distance between each pair of experts' preference matrices is calculated, and the consensus value *CNS* is derived from the normalized sum of these distances. A high *CNS* value indicates a strong agreement among the experts, while a low value suggests a need for further discussion and alignment.

3.6 Aggregation of Individual Preferences

After the determination of the relative weights for each expert and the completion of the consensus analysis, the next step involves aggregating the individual preferences to form a collective preference relation. This is done by combining the preferences of all experts in a weighted manner, where the influence of each expert is proportional to their calculated relative weight. The collective reciprocal preference relation, denoted as $P = (p_{ij}; \ i \neq j = 1, \ldots, N)$, is represented as an $N \times N$ matrix, where each element p_{ij} indicates the degree of preference of alternative *i* over alternative *j*. The main diagonal is excluded, as it would represent self-preference.

To perform this aggregation, we utilize the Weighted Average (WA) operator. This operator takes into account both the reciprocal preference relations expressed by each expert and their respective weights, ensuring that more influential experts have a greater impact on the final aggregated preferences. The computation for each element of the collective preference relation p_{ij} is as follows:

$$p_{ij} = \sum_{s=1}^{M} p_{ij}^{s} \cdot w_s \tag{5}$$

Here, p_{ij}^{s} represents the pairwise preference of expert *s* for alternatives *i* and *j*, and w_s is the relative weight of expert *s*. By applying this aggregation formula, we obtain a collective preference matrix that reflects the weighted contributions of all experts, with each individual's preferences being adjusted according to their relative importance in the decision-making process.

This aggregated preference matrix serves as the foundation for the subsequent steps, where it will be used to determine the final ranking of the alternatives, ensuring that the group decision-making process is both fair and representative of the collective preferences of the experts.

3.7 Getting the Ranking of Alternatives

In this final phase of the decision-making process, after establishing the collective reciprocal preference relationship, the next crucial step involves determining the ranking of alternatives. To accomplish this task, it is essential to utilize an appropriate operator. Among the available choices, we specifically select the Quantifier Guided Degree of Dominance (QGDD) operator. The QGDD operator is particularly advantageous because it allows for a precise and quantitative assessment of the degree to which one alternative is preferred over another, taking into account the pairwise preferences established by the experts.

Additionally, to compute the value of each alternative in comparison to others, we adopt the average operator. This ensures that the preferences across all individuals are evenly integrated into the calculation. As a result, to determine the degree of dominance of a given alternative γ_i over the remaining alternatives, we calculate the value $QGDD_{\gamma_i}$, which is computed using the following formula:

$$QGDD_{\gamma_i} = \frac{\sum_{j=1;i\neq j}^{N} p_{ij}}{N-1} \tag{6}$$

In this equation, p_{ij} represents the preference value of alternative γ_i over alternative γ_j, as derived from the collective preference matrix, and N denotes the total number of alternatives under consideration. The summation is performed for all alternatives j that are different from i. This calculation produces the degree of dominance of alternative γ_i over all others.

Once the values for each alternative have been obtained, the next step is to validate the results for consistency and accuracy. This validation is performed using Trillo's theorem (Trillo et al. 2022), which provides a method for verifying the logical coherence of the results, ensuring that no contradictions or logical errors exist within the decision-making process. If the results pass the validation process, they are deemed reliable.

The final action is to determine which alternative holds the highest degree of dominance, based on the previously computed $QGDD_{\gamma_i}$ values. This can be formalized as follows:

$$\gamma_{QGDD} = \{\gamma_i \in X \mid QGDD_{\gamma_i} = \max_{\gamma_j \in X} QGDD_{\gamma_j}\} \tag{7}$$

In this equation, γ_{QGDD} represents the alternative with the highest dominance value, which corresponds to the collective preference of all the participants in the decision-making process. The alternative with the maximum *QGDD* value is considered the most preferred choice, reflecting the consensus of the group based on their preferences and the analysis conducted.

This final ranking step ensures that the decision-making process culminates in a clear and justified selection of the best alternative, taking into account all expert preferences, conflict resolutions, and aggregated judgments.

4 Illustrative Example

In this section, we delineate a case study designed to exemplify the capabilities of our proposed innovative methodology. Specifically, we consider a panel of three experts, $\Xi = \{\xi_1, \xi_2, \xi_3\}$, tasked with formulating strategic investment decisions to enhance heating systems. The objective of their deliberation is to optimize thermal efficiency, ensure economic feasibility, and uphold environmental sustainability. The experts are required to choose from a finite set of alternatives, $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$, comprising a coal-based heating system (γ_1), a solar-powered solution (γ_2), a gas-fueled system (γ_3), and an aerothermal system (γ_4). The preferences articulated by each expert within this decision-making framework serve to underscore the robustness of our method in addressing the multidimensional optimization of thermal performance, ecological responsibility, and economic prudence in heating system improvements.

With both sets rigorously defined, the experts engage in a structured deliberative process. Each utterance made by the participants is systematically recorded and stored within the system. Upon the conclusion of the discourse, the system evaluates the degree of hostility exhibited by each expert using the following approach. Initially, the recorded dialogue is organized into a tabular format as illustrated below (It is possible to watch the full conversation at the following link[1]):

A connection is then established with the APIs for Gemini 1.5 Flash, ChatGPT 4.o, and Copilot. The following prompt is transmitted to each system: *"I want you to determine the percentage of aggressiveness for the following list of comments using the structure: 'sentence' - The percentage of aggressiveness is $X\%$."* Next, the second row from the previously mentioned table is incorporated in its entirety into the prompt. This query is applied to all comments, resulting

[1] https://github.com/jrtrillo/Evaluating-Hostility-Detection-in-Group-Decision-Making.git

in each system returning responses formatted according to the specified structure. Consequently, the output consists of all comments paired with their respective aggressiveness percentages. Below, examples are provided of comments made by experts ξ_1, ξ_3, and ξ_2, along with the corresponding responses generated by the systems:

Table 3 The raw weight of each expert. Each reported weight corresponds to the mean of five independent attempts (i.e., repeated queries) to the respective LLM for the same input.

LLM	Expert	Attempt	Raw weight	Weight Average
GEMINI 1.5	ξ_1	**1**	0.372040628	0.376020354
		2	0.381191223	
		3	0.372607251	
		4	0.379771488	
		5	0.37449118	
	ξ_2	**1**	0.35636751	0.356550133
		2	0.361128527	
		3	0.351906848	
		4	0.35730979	
		5	0.356037992	
	ξ_3	**1**	0.271591862	0.267429513
		2	0.257680251	
		3	0.275485901	
		4	0.262918722	
		5	0.269470828	
ChatGPT 4.o	ξ_1	**1**	0.367232518	0.374825668
		2	0.382184379	
		3	0.367310283	
		4	0.378299120	
		5	0.379102041	
	ξ_2	**1**	0.353838966	0.350473843
		2	0.35107254	
		3	0.337674001	
		4	0.351906158	
		5	0.357877551	
	ξ_3	**1**	0.278928516	0.274700489
		2	0.266743082	
		3	0.295015716	
		4	0.269794721	
		5	0.263020408	
Copilot	ξ_1	**1**	0.359626943	0.358465809
		2	0.363829787	
		3	0.356807512	
		4	0.347722904	
		5	0.364341898	
	ξ_2	**1**	0.35946114	0.344626621
		2	0.34893617	
		3	0.330516432	
		4	0.340223556	
		5	0.343995805	
	ξ_3	**1**	0.280911917	0.29690757
		2	0.287234043	
		3	0.312676056	
		4	0.31205354	
		5	0.291662297	

- "I'll start by saying that coal is unacceptable. It has no place in any serious discussion about modern heating systems." - The percentage of aggressiveness of this comment is 15%.
- "Here we go again... Sure, coal is bad, but do you have any realistic idea for replacing it without bankrupting people?" - The percentage of aggressiveness of this comment is 40%.
- "I have an idea: solar and aerothermal. But I suppose you're going to say it's "too expensive" as always" - The percentage of aggressiveness of this comment is 90%.

By employing the devised regular expression, the individual values associated with each expert are meticulously extracted in a systematic manner. Once these values have been retrieved, the preliminary weight assigned to each expert is calculated with precision. The resulting data is succinctly consolidated and presented in the table below (refer to Table 3):

The values reported in Table 3 were computed following the mathematical framework described in Sect. 3.3 and Sect. 3.4. Specifically, the numerical outputs were obtained using regular expressions in conjunction with Equation (1), whereas the behavioral weights were derived from Equation (2) and Equation (3), which incorporate the hostility scores generated by the LLMs. The results presented in the table, therefore, represent the aggregated outcomes of applying these formulas to the full set of expert comments.

Once the percentage corresponding to each expert has been determined, the next step involves calculating their respective significance within the process. This procedure entails establishing the weight attributed to each expert in relation to each Large Language Model. Subsequently, after completing the calculation of individual weights for each expert, their preferences are scrutinized, uncovering the reciprocal preference relationships. These relationships are meticulously detailed in the following section.

$$P_1 = \begin{pmatrix} - & 0.30 & 0.50 & 0.80 \\ 0.70 & - & 0.75 & 0.70 \\ 0.50 & 0.25 & - & 0.60 \\ 0.20 & 0.30 & 0.40 & - \end{pmatrix} P_2 = \begin{pmatrix} - & 0.55 & 0.75 & 0.70 \\ 0.45 & - & 0.60 & 0.70 \\ 0.25 & 0.40 & - & 0.60 \\ 0.30 & 0.30 & 0.40 & - \end{pmatrix}$$

$$P_3 = \begin{pmatrix} - & 0.59 & 0.70 & 0.75 \\ 0.41 & - & 0.60 & 0.60 \\ 0.30 & 0.40 & - & 0.50 \\ 0.25 & 0.40 & 0.50 & - \end{pmatrix}$$

The presence of a consensus among the experts is assessed by establishing a consensus threshold of $\alpha = 0.9$. Given that the calculated consensus value is $CNS = 0.9618$, it can be affirmed with confidence that a sufficient degree of agreement exists among the experts. Therefore, it is appropriate

to proceed with the computation of the collective reciprocal preference relation for each LLM:

$$P_{ChatGPT} = \begin{pmatrix} - & 0.4673 & 0.6426 & 0.7512 \\ 0.5327 & - & 0.6562 & 0.6725 \\ 0.3574 & 0.3438 & - & 0.5725 \\ 0.2488 & 0.3275 & 0.4275 & - \end{pmatrix}$$

$$P_{Ge\min i} = \begin{pmatrix} - & 0.4667 & 0.6426 & 0.7510 \\ 0.5333 & - & 0.6564 & 0.6733 \\ 0.3574 & 0.3436 & - & 0.5733 \\ 0.2490 & 0.3267 & 0.4267 & - \end{pmatrix}$$

$$P_{Copilot} = \begin{pmatrix} - & 0.4723 & 0.6455 & 0.7507 \\ 0.5277 & - & 0.6538 & 0.6703 \\ 0.3545 & 0.3462 & - & 0.5703 \\ 0.2493 & 0.3297 & 0,4297 & - \end{pmatrix}$$

Subsequently, the QGDD is applied to each Large Language Model using the average operator within the matrix. This process yields the ranking of the alternatives:

To conclude, we employ the Trillo theorem (Trillo et al. 2022) as a robust methodological tool to verify the accuracy, consistency, and reliability of the entire execution process. This theorem provides a theoretical foundation for assessing the integrity of the procedural steps and the validity of the outcomes derived from the applied methodology. By applying this theorem, it is conclusively affirmed that, in all three cases analyzed, the process has been executed correctly, ensuring that the results are both methodologically sound and free from procedural inaccuracies.

Moreover, the analysis reveals a significant and noteworthy observation regarding the influence of the selected LLM on the decision-making process. Specifically, when the ChatGPT and Copilot models are employed, the alternative γ_1 emerges as the preferred choice. Conversely, when the Gemini model is utilized, the selection shifts, and γ_2 is identified as the most suitable alternative. This divergence in outcomes highlights the inherent variability introduced by different LLMs, emphasizing the role that model-specific characteristics, underlying algorithms, and interpretative frameworks play in shaping the final decisions.

Such findings underscore the critical importance of selecting an appropriate LLM tailored to the specific context and objectives of the decision-making scenario. It also reinforces the necessity of thorough validation methods, such as the application of the Trillo theorem, to ensure that the processes underpinning these decisions remain both accurate and reliable, regardless of the variability introduced by different computational models. This exploration not only illustrates the practical implications of LLM diversity but also enriches our understanding of how advanced AI systems can influence and shape critical decision-making processes across varying contexts.

This illustrative case underscores that the proposed methodology can be readily applied to real-world deliberative settings, such as corporate strategic planning, public policy debates, or collaborative research projects. By quantifying hostility and incorporating it into consensus models, the framework provides not only theoretical insights but also actionable tools to enhance fairness and efficiency in decision-making environments.

5 Analysis of Results

This study implemented a methodology aimed at evaluating the impact of hostility in group decision-making processes. The evaluation considered three state-of-the-art language models: *ChatGPT 4.o*, *Gemini 1.5 Flash*, and *Copilot*, with a detailed analysis of participant comments and divergences in the results generated by each model.

As shown in Table 5, ChatGPT and Gemini exhibit closer alignment in moderate-hostility contexts, while Copilot tends to overestimate hostility levels. Despite these divergences, the proposed weighting mechanism consistently reduces the influence of highly aggressive experts, ensuring stability across models. Notably, the divergence in final rankings highlights the importance of incorporating hostility-aware weighting, as it absorbs model-specific variability and reinforces the robustness of the group decision-making process.

5.1 Accuracy in Hostility Evaluation

Representative comments made by three experts during the deliberative process were analysed. Each comment was evaluated by the three models, generating associated hostility scores. Table 6 summarizes some key examples:

The other evaluations can be viewed at the following link.[2] The results indicate that *ChatGPT 4.o* and *Gemini 1.5 Flash* exhibit greater concordance in detecting hostility in comments with low to moderate levels, whereas *Copilot* tends to assign higher percentages, especially in comments perceived as critical or confrontational. For instance, in the case of the comment $\ddot{H}ere\ we\ go\ again\ldots\ddot{}$, *Copilot* detected a hostility level of 70%, significantly higher than the 30% and 35% estimated by *ChatGPT 4.o* and *Gemini*, respectively.

Across five repeated evaluations per comment, ChatGPT and Gemini yielded low within-model variance ($< 5\%$), while Copilot showed moderate variance ($\approx 8 - 12\%$). Copilot also tended to overestimate hostility for moderately critical statements. Despite this, the

[2] https://github.com/jrtrillo/Evaluating-Hostility-Detection-in-Group-Decision-Making.git

hostility-aware weighting remained stable: the most aggressive expert's weight decreased by $\approx 27-29\%$ under all three models, indicating that minor score shifts do not overturn the intended moderating effect.

5.2 Weighting of Experts and Influence in the Process

The hostility evaluation enabled the calculation of relative weights for each expert based on the average aggressiveness of their interventions. Experts with less aggressive comments received higher weights, as shown in Table 7:

These results reflect how hostility modulates the influence of experts in the group process. Expert 1, who demonstrated the lowest average hostility (15% in *ChatGPT 4.o*), received the highest relative weight, while Expert 3, with higher hostility levels, saw their weighting reduced.

This moderation effect directly enhances consensus quality: by curbing the disproportionate weight of aggressive participants, the aggregated preferences better reflect the substantive validity of arguments. As shown in Tables 5 and 7, the relative weight adjustments are consistent across models, providing robustness against LLM-specific variability.

5.3 Impact on the Ranking of Alternatives

The weighted aggregation process led to consistent results in the selection of alternatives, though with subtle divergences among the models. *ChatGPT 4.o* and *Copilot* agreed in prioritizing alternative γ_1 (coal-based heating system), while *Gemini 1.5 Flash* favored γ_2 (solar solution). Table 4 summarizes the dominance scores. These differences highlight the importance of selecting appropriate models based on the specific decision-making context, as algorithmic particularities can influence the final results. However, the overall consistency in dominance values reinforces the validity of the approach. Although the top alternative differs between Gemini (γ_2) and the other two models (γ_1), the QGDD dominance values are close, indicating that the relative ordering of the leading options is robust to plausible inter-model differences in hostility scoring.

6 Discussion

In this study, the analysis serves as a proof-of-concept demonstration rather than an exhaustive evaluation. The implementation of these models enables a rigorous and systematic evaluation of the tone of comments expressed in deliberative contexts, allowing for an objective quantification of the level of aggressiveness exhibited by each participant. This approach significantly contributes to mitigating biases in interaction and promoting equity in decision-making.

One of the primary contributions of this methodology lies in its ability to minimize the disproportionate influence of individuals with an aggressive communicative style. By weighting their impact within the deliberative process, a more balanced discussion environment is fostered, aligning with the principles of deliberative justice. Additionally, the implementation of a feedback mechanism based on sentiment analysis encourages more respectful communication, prioritizing argument robustness over the mere manner in which ideas are expressed.

The comparative evaluation of hostility detection capabilities across the assessed models reveals notable differences in terms of sensitivity and consistency in aggressiveness assessments. While all models demonstrated robust performance in identifying and quantifying hostile expressions, it was observed that redundancy in evaluation–submitting each comment five times to each model–helped reduce result variability and enhance the reliability of the analytical process. These comparative findings confirm that although the three LLMs exhibit distinct biases, the proposed hostility-aware weighting framework effectively stabilizes the decision-making process, mitigating model-specific distortions and reinforcing robustness.

Nevertheless, applying generative models in this context also presents methodological challenges. One such challenge is the possibility of hallucinations in model-generated outputs, which could compromise the precision of hostility assessments. To mitigate this issue, a strategy based on regular expressions has been incorporated to systematically extract numerical values, thereby ensuring the coherence and standardization of processed information.

Furthermore, the interpretation of hostility levels remains dependent on contextual complexity and semantic nuances that the models the models may not fully capture. Although automating aggressiveness analysis enhances objectivity in the decision-making process, the inclusion of human moderators could serve as a valuable complement for validating obtained results and fine-tuning the weighting assigned to each participant with greater precision. An important consideration concerns the reliability of sentiment analysis in detecting hostile or aggressive language. While traditional approaches based on lexicons or supervised classifiers often failed to capture contextual nuances, recent advances in

Table 4 QGDD results

QGDD	γ_1	γ_2	γ_3	γ_4
QGDD (CHAT GPT4.o)	**0.620352584**	0.620490733	0.424582514	0.334574169
QGDD (Gemini 1.5 flash)	0.620096346	**0.620,989,337**	0.42474352	0.334170797
QGDD (Copilot)	**0.622,829,993**	0.617273088	0.423667067	0.336229852

transformer-based LLMs have demonstrated strong capabilities in distinguishing between neutral assertiveness and genuine hostility. In our framework, this reliability is enhanced by three methodological safeguards: (i) the use of multiple state-of-the-art LLMs (ChatGPT, Copilot, and Gemini) to cross-validate outputs, (ii) repeated evaluations of each comment to reduce stochastic variability, and (iii) the transformation of textual outputs into structured numerical hostility scores through regular expressions. Together, these measures mitigate model-specific biases and ensure that hostility detection is consistent, reproducible, and suitable for integration into consensus-reaching mechanisms.

Compared to previous studies, the proposed methodology offers significant advancements. It should be noted that the consensus model employed here is adapted from established approaches in fuzzy decision-making. However, unlike previous adaptations, the present study introduces hostility-aware behavioral weighting based on LLM predictions, which constitutes a novel methodological synthesis not found in the existing literature. For instance, in works such as Hashmi et al. (2025), a supervised learning-based approach for hostility detection has been implemented, requiring large volumes of labelled data and exhibiting limitations in adaptability to different deliberative contexts. Similarly, Trillo et al. (2023) describes a classification system based on dictionaries of aggressive words, which, although efficient, lacks the flexibility inherent in the generative models employed in this study. Finally, Sadiq et al. (2021) explores manual discourse moderation as a strategy to reduce hostility in debates, an effective method but one with high operational costs and limited scalability. In contrast, the methodology proposed herein achieves a balance between automation and adaptability, enabling real-time hostility detection without necessitating constant human intervention.

In contrast to previous approaches, which relied either on pre-defined dictionaries, supervised models, or manual moderation, the present study constitutes the first attempt to systematically integrate multiple state-of-the-art LLMs into a consensus-reaching process. Our framework not only detects hostility but also operationalizes it within the weighting mechanism of experts, thereby introducing an innovative linkage between natural language processing and decision-theoretic modelling.

Concisely, this study represents a significant step forward in the application of artificial intelligence for regulating communication in group decision-making contexts. The presented methodology could be extended to large-scale decision-making environments, where managing hostility in interactions is a critical factor in achieving equitable and sustainable consensus. Future research could explore the integration of language models with advanced discourse analysis techniques and natural language processing to refine the detection of intentions and emotions in group interactions.

The scope of the present analysis is restricted to a single illustrative dataset, without manual or expert annotation and without comparison to alternative baseline methods. This design was sufficient to demonstrate the methodological integration of LLM outputs into fuzzy consensus models, but it does not provide empirical generalizability. Future research should therefore expand the dataset substantially, incorporate human labelling to validate hostility scores, and benchmark the proposed framework against existing sentiment or hostility detection approaches. These steps would provide stronger evidence of the robustness and external validity of the method.

A known limitation of using LLMs lies in their treatment of numbers as symbolic tokens rather than grounded quantities, which can affect the reliability of numeric outputs. In our framework, this issue is partially mitigated by repeated querying and aggregation across multiple models, reducing stochastic inconsistencies. Nevertheless, future studies could adopt comparative prompting strategies (e.g., asking which comment is more aggressive between two options) to evaluate whether LLMs capture relative distinctions more reliably than absolute scoring. Such enhancements would further strengthen the methodological robustness of hostility detection.

From a practical standpoint, the framework could be integrated into online platforms that support group deliberation, serving as an automated moderator to reduce the dominance of aggressive participants and to encourage constructive dialogue. Such integration would be particularly valuable in large-scale participatory governance or corporate decision-making processes where hostility can compromise efficiency and fairness.

Despite the encouraging results, several improvements could further strengthen the proposed framework. First, a human-in-the-loop component could be integrated to validate borderline cases where models struggle to differentiate assertiveness from genuine aggression. Second, benchmarking against gold-standard annotated corpora would provide a more objective measure of detection accuracy. Third, the methodology could be extended to multilingual and multimodal deliberations, incorporating not only textual inputs but also speech and paralinguistic cues such as tone or facial expressions. Fourth, more sophisticated discourse analysis techniques could refine the detection of pragmatic intent, reducing false positives in critical but non-hostile statements. Finally, enhancing the interpretability and transparency of hostility scores would help participants better understand how their contributions are weighted, fostering trust in the system.

Table 5 Comparative analysis of LLM performance in hostility detection and decision impact

Criterion	ChatGPT 4.0	Gemini 1.5 Flash	Copilot	Observations
Sensitivity to low hostility	Moderate (10–20%)	Moderate (10–15%)	Overestimates (up to 50%)	Copilot systematically exaggerates hostility
Stability across repetitions	High (variance <5%)	High (variance <5%)	Moderate (variance 8–12%)	Gemini and ChatGPT more consistent
Weighting impact on Expert 3	Reduced by ∼27%	Reduced by ∼27%	Reduced by ∼ 29%	Weight reduction stable across models
Final ranking outcome	γ_1 prioritized	γ_2 prioritized	γ_1 prioritized	Demonstrates variability across LLMs

Table 6 Levels of hostility detected by the linguistic model in an attempt

Expert	Commentary	ChatGPT 4.o	Gemini 1.5 Flash	Copilot
ξ_1	*I'll start by saying that coal is unacceptable...*	15%	10%	50%
ξ_3	*Here we go again...*	30%	35%	70%
ξ_2	*Yes, I have an idea: solar and aerothermal...*	40%	40%	65%

Table 7 Relative weights of experts based on hostility levels

Expert	ChatGPT 4.o	Gemini 1.5 Flash	Copilot
Expert 1	37.6%	37.4%	35.8%
Expert 2	35.0%	35.6%	34.5%
Expert 3	27.4%	27.0%	29.7%

Beyond technical stability, the integration of hostility-aware weighting mechanisms is expected to positively influence group satisfaction. Prior studies in consensus theory indicate that participants are more likely to endorse and adhere to decisions when the process is perceived as equitable and inclusive. By systematically reducing the influence of disproportionately aggressive interventions, our framework fosters a perception of procedural fairness. While satisfaction was not directly measured in this study, we argue that the documented improvements in consensus quality (Tables 5 and 7,) provide a strong proxy for enhanced acceptance of the final outcomes. Future work should complement these findings with user-centered evaluations involving real deliberative groups.

Looking ahead, several promising avenues emerge from our findings. First, empirical user studies should be conducted to assess not only consensus quality but also perceived fairness and satisfaction among participants. Second, creating annotated corpora of deliberative hostility would provide a benchmark for systematically evaluating LLM performance. Third, extending the framework to multilingual and multimodal contexts, including speech and non-verbal cues, would enhance its ecological validity. Fourth, incorporating hybrid approaches that combine LLM-based hostility detection with symbolic or ontology-based reasoning may increase precision. Finally, greater emphasis on explainability mechanisms will be necessary to foster transparency and trust when applying the framework in real-world decision-making platforms.

Finally, our evaluation emphasizes comparative and operational effectiveness rather than absolute accuracy. We did not employ human-annotated gold standards, and therefore we refrain from claims about true detection accuracy in an absolute sense. Future work will benchmark LLM outputs against expert annotations and examine domain- and culture-specific norms to better separate firm but civil assertiveness from genuine hostility.

7 Conclusions

This study has presented a novel methodology for evaluating hostility detection in group decision-making processes, utilizing the comparative analysis of three advanced Large Language Models (LLMs): Gemini 1.5 Flash, ChatGPT 4.o, and Copilot. By integrating generative artificial intelligence into the decision-making framework, we aimed to address the challenges posed by linguistic aggression and its potential to bias collective decisions. Moreover, the proposed approach effectively quantifies the degree of aggressiveness in expert comments, ensuring that emotional influences are mitigated during decision-making. This enables a more balanced and constructive dialogue among participants.

The method not only demonstrates the potential of generative artificial intelligence to enhance group decision-making but also sheds light on the critical interplay between advanced AI systems and human dynamics in collaborative environments. The findings advocate for further exploration of LLM-based frameworks, emphasizing their adaptability to diverse decision-making contexts and their capacity to foster equitable and efficient deliberations. Our findings further suggest that hostility-aware weighting not only improves the quality of consensus but also lays the groundwork for enhancing participant satisfaction, insofar as decisions emerge from a more balanced and procedurally fair process.

Future research should build on these findings by (i) validating the framework in real deliberative environments with human participants, (ii) benchmarking against annotated datasets for hostility detection, (iii) expanding into multilingual and multimodal settings, (iv) testing domain-specific adaptations, and (v) improving transparency for

end-users. Together, these lines of work will refine both the scientific foundations and the practical impact of hostility-aware consensus models.Furthermore, future enhancements such as human-in-the-loop validation, benchmarking against annotated datasets, and multimodal hostility detection represent natural extensions of this work, aimed at improving accuracy, fairness, and user trust. Therefore, the proposed approach should not be regarded merely as a theoretical contribution, but as a methodological advancement with concrete applicability across diverse decision-making domains. A limitation of the present study is the absence of empirical validation against human-labelled datasets. While the proposed framework demonstrates methodological feasibility, future research will focus on benchmarking the models against annotated corpora in order to assess accuracy, inter-rater reliability, and alignment between LLM predictions and human judgment. Such an extension will provide a rigorous empirical grounding and further substantiate the applicability of hostility-aware consensus models. Additionally, the current reliance on regular expressions for parsing LLM outputs constitutes a methodological limitation. Future implementations will address this issue by enforcing structured outputs (e.g., JSON schemas) to improve robustness and reproducibility. Another limitation concerns the reliance on a specific prompt. While this decision guaranteed consistency across the three LLMs, it also constrains the reproducibility of the results. Future research will therefore include systematic prompt engineering, the evaluation of multiple prompt variants, and sensitivity analyses to quantify how results may vary under alternative formulations. In addition, standardized structured prompting protocols (e.g., JSON-based instructions) will be adopted to further enhance reproducibility and robustness. The methodological novelty of this study lies in coupling LLM-based hostility detection with fuzzy consensus models, an integration that has not been previously explored in the literature.

Acknowledgements This work has been supported by the grant PID2022–139297OB-I00 funded by MICIU/AEI/10.13039/501,100,011,033 and by ERDF/EU. Moreover, it is part of the project C-ING-165-UGR23, co-funded by the Regional Ministry of University, Research and Innovation and by the European Union under the Andalusia ERDF Program 2021–2027.

Funding Funding for open access publishing: Universidad de Granada/CBUA.

Data Availability The experimental data and the simulation results that support the findings of this study are available at https://github.com/jrtrillo/Evaluating- Hostility-Detection-in-Group-Decision-Making.git. This paragraph has also been included in the manuscript.

Declarations

Conflict of interest Statement The authors declare that they have no Conflict of interest related to this research.

References

Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, Afshar-Oromieh A (2023) Large language models (llm) and chatgpt: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging 50(6):1549–1552

Brown TB (2020) otros: language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Buehler MJ (2023) Melm, a generative pretrained language modeling framework that solves forward and inverse mechanics problems. J Mech Phys Solids 181:105454

Cabrerizo FJ, Moreno JM, Pérez IJ, Herrera-Viedma E (2010) Analyzing consensus approaches in fuzzy group decision making: advantages and drawbacks. Soft Comput 14:451–463

Cabrerizo FJ, Ureña R, Pedrycz W, Herrera-Viedma E (2014) Building consensus in group decision making with an allocation of information granularity 255:115–127

Darraz M, Laour M, Guemghar S (2025) Integrated sentiment analysis with bert for enhanced recommendation systems. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2024.124657

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Gajula MVNS (2025) Sentiment-aware recommendation systems in e-commerce: A review from a natural language processing perspective. arXiv preprint arXiv:2505.03828 [cs.IR]

Gaudine A, Thorne L (2001) Emotion and ethical decision-making in organizations. J Bus Ethics 31:175–187

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Advances in neural information processing systems **27**

Hashmi E, Yayilgan SY, Yamin MM, Abomhara M, Ullah M (2025) Self-supervised hate speech detection in norwegian texts with lexical and semantic augmentations. Expert Syst Appl 264:125843

Herrera-Viedma E, Herrera F, Chiclana F (2002) A consensus model for multiperson decision making with different preference structures. IEEE Transactions on Systems Man and Cybernetics-Part A Systems and Humans 32(3):394–402

Herrera-Viedma E, Cabrerizo FJ, Kacprzyk J, Pedrycz W (2014) A review of soft consensus models in a fuzzy environment. Information Fusion 17:4–13

Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, Yin H, Xu C, Yang R, Zheng Q et al (2023) Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 15(1):29

Kabak Ö, Ervural B (2017) Multiple attribute group decision making: A generic conceptual framework and a classification scheme. Knowl-Based Syst 123:13–30

Kacprzyk J, Yager RR, Merigo JM (2019) Towards human-centric aggregation via ordered weighted aggregation operators and linguistic data summaries: A new perspective on zadeh's inspirations. IEEE Comput Intell Mag 14(1):16–30

Krugmann JO, Hartmann J (2024) Sentiment analysis in the age of generative ai. Cust Needs Solut 11(1):3

Kumarage T, Bhattacharjee A, Garland J (2024) Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. arXiv preprint arXiv:2403.08035

Levy O, Goldberg Y (2014) Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 302–308

Magalhães B, Neto A, Cunha A (2023) Generative adversarial networks for augmenting endoscopy image datasets of stomach precancerous lesions: A review. IEEE Access 11:136292–136307

Meena G, Mohbey KK, Lokesh K (2024) Point of interest recommendation system using sentiment analysis. Journal of Information Science Theory and Practice 12(2):65–78. https://doi.org/10.1633/JISTaP.2024.12.2.5

Meena G, Mohbey KK, Lokesh K (2024) Fstl-sa: Few-shot transfer learning for sentiment analysis from facial expressions. Multimedia Tools and Applications, 1–29

Miao H, Li C, Wang J (2023) A future of smarter digital health empowered by generative pretrained transformer. J Med Internet Res 25:49963

Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784

Mohammed IA, Venkataraman S (2023) An innovative study for the development of a wearable ai device to monitor parkinson's disease using generative ai and llm techniques. International Journal of Creative Research Thoughts (IJCRT) www. ijcrt. org, ISSN, 2320–2882

Morente-Molinera JA, Kou G, Samuylov K, Ureña R, Herrera-Viedma E (2019) Carrying out consensual group decision making processes under social networks using sentiment analysis over comparative expressions. Knowl-Based Syst 165:335–345

Morente-Molinera JA, Cabrerizo FJ, Trillo J, Pérez I, Herrera-Viedma E (2022) Managing group decision making criteria values using fuzzy ontologies. Procedia Computer Science 199:166–173

Nicula B, Dascalu M, Arner T, Balyan R, McNamara DS (2023) Automated assessment of comprehension strategies from self-explanations using llms. Information 14(10):567

Nurmi H, Kacprzyk J (1991) On fuzzy tournaments and their solution concepts in group decision making. Eur J Oper Res 51(2):223–232

Ozdemir S (2023) Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs

Pangtey L, Bhatnagar A, Bansal S, Dar SS, Kumar N (2025) Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. arXiv preprint arXiv:2505.08464

Paulus MP, Angela JY (2012) Emotion and decision-making: affect-driven belief systems in anxiety and depression. Trends Cogn Sci 16(9):476–483

Pearce H, Tan B, Ahmad B, Karri R, Dolan-Gavitt B (2023) Examining zero-shot vulnerability repair with large language models. In: 2023 IEEE Symposium on Security and Privacy (SP), 2339–2356 . IEEE

Pérez IJ, Wikström R, Mezei J, Carlsson C, Herrera-Viedma E (2013) A new consensus model for group decision making using fuzzy ontology. Soft Comput 17:1617–1627

Pilehvar MT, Camacho-Collados J (2021) Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning

Pramanik S, Mukhopadhyaya D (2011) Grey relational analysis based intuitionistic fuzzy multi-criteria group decision-making approach for teacher selection in higher education. International Journal of Computer Applications 34(10):21–29

Qin J, Li M, Liang Y (2022) Minimum cost consensus model for crp-driven preference optimization analysis in large-scale group decision making using louvain algorithm. Information Fusion 80:121–136

Ruz GA, Henríquez PA, Mascareño A (2020) Sentiment analysis of twitter data during critical events through bayesian networks classifiers. Futur Gener Comput Syst 106:92–104

Sadiq S, Mehmood A, Ullah S, Ahmad M, Choi GS, On B-W (2021) Aggression detection through deep neural model on twitter. Futur Gener Comput Syst 114:120–129

Sayegh L, Anthony WP, Perrewé PL (2004) Managerial decision-making under crisis: The role of emotion in an intuitive decision process. Hum Resour Manag Rev 14(2):179–199

Simon HA (1987) Making management decisions: The role of intuition and emotion. Acad Manag Exec 1(1):57–64

Strader TJ, Rozycki JJ, Root TH, Huang Y-HJ (2020) Machine learning stock market prediction studies: review and research directions. Journal of International Technology and Information Management 28(4):63–83

Svenson F, Peuser M, Çetin F, Aidoo DC, Launer MA (2024) Decision-making styles and trust across farmers and bankers: Global survey results. Decision Analytics Journal 10:100427

Taghavi A, Eslami E, Herrera-Viedma E, Ureña R (2020) Trust based group decision making in environments with extreme uncertainty 191:105168

Trillo JR, Cabrerizo FJ, Pérez IJ, Morente-Molinera JA, Herrera-Viedma E (2024) A new consensus reaching method for group decision-making based on the large language model gemini for detecting hostility during the discussion process. In: 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), 1–8 . IEEE

Trillo JR, Cabrerizo FJ, Chiclana F, Martínez MÁ, Mata F, Herrera-Viedma E (2022) Theorem verification of the quantifier-guided dominance degree with the mean operator for additive preference relations. Mathematics 10(12):2035

Trillo JR, Herrera-Viedma E, Morente-Molinera JA, Cabrerizo FJ (2023) A large scale group decision making system based on sentiment analysis cluster. Information Fusion 91:633–643

Trillo JR, Herrera-Viedma E, Morente-Molinera JA, Cabrerizo FJ (2023) A group decision-making method based on the experts' behavior during the debate. IEEE Transactions on Systems Man and Cybernetics Systems 53(9):5796–5808

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems **30**

Wankhade M, Rao ACS, Kulkarni C (2022) A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55(7):5731–5780

Yager RR, Kacprzyk J (2012) The Ordered Weighted Averaging Operators: Theory and Applications

Yager RR (1981) A procedure for ordering fuzzy subsets of the unit interval. Inf Sci 24(2):143–161

Yager RR (1996) Quantifier guided aggregation using OWA operators. Int J Intell Syst 11(1):49–73

Zhang W, Deng Y, Liu B, Pan SJ, Bing L (2023) Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005

Zhuo H, Yang Y, Peng K (2025) Combating toxic language: A review of llm-based strategies for software engineering. arXiv preprint arXiv:2504.15439