

Enhancing Group Decision-Making Through Large Language Models for Semantic Filtering of Expert Opinions

José Ramón Trillo University of Granada jrtrillo@ugr.es	Julia García-Cabello University of Granada cabello@ugr.es	Juan Miguel Tapia University of Granada jmtaga@ugr.es	Sergio Alonso University of Granada zerjioi@ugr.es	Juan Antonio Morente-Molinera University of Granada jamoren@ugr.es
---	--	---	---	---

Abstract

In modern decision-making contexts involving multiple experts, the use of natural language to express opinions introduces considerable difficulties for traditional analytical methods. With the growth of online collaboration, expert evaluations have become increasingly diverse in both format and scale, especially when communicated through unstructured textual comments. While conventional Group Decision-Making techniques can handle varied evaluation metrics, they cannot often fully interpret the semantic richness of human language. This work presents a novel Group Decision-Making framework that leverages a Large Language Model to semantically analyse and filter expert commentary, specifically DeepSeek LLM, identifying and removing remarks that may negatively influence others. Unlike standard sentiment analysis approaches based on fixed lexical rules, the Large Language Model captures subtle linguistic features such as tone, context, and implied meaning, allowing for a more accurate extraction of genuine expert preferences. The filtered and interpreted inputs are then aggregated to produce a consensus that better reflects the true collective judgment. This integration of advanced natural language processing enhances the interpretability, fairness, and reliability of group decisions in complex and dynamic environments. This constitutes a departure from traditional sentiment analysis approaches, as the LLM is used not only for interpretation but also for active semantic filtering of expert discourse.

Keywords: Group Decision-Making, Large Language Models, Semantic Analysis, DeepSeek LLM, Consensus Methods

1. Introduction

Group Decision-Making (GDM) involves the collective evaluation of a finite set of alternatives by a panel of experts, each contributing unique knowledge and perspectives (Zadeh, 1978). Traditionally, such processes have relied on structured and quantifiable data, such as numerical ratings or predefined scales, which facilitate the aggregation and analysis of individual preferences (Pérez et al., 2013; Urena et al., 2016). Nevertheless, in many real-world contexts, particularly those marked by high complexity and diversity, expert opinions are frequently expressed through natural language comments, rich in nuance, justification, and implicit reasoning that are difficult to effectively codify or summarise using conventional methods.

This shift toward environments dominated by discursive contributions presents significant challenges: unstructured information complicates the coherent integration of diverse viewpoints and introduces risks related to disproportionate influence from certain participants, as well as the presence of irrelevant or biased remarks that may distort the deliberation process. Consequently, the central problem we address is how to reliably extract authentic expert preferences from complex and heterogeneous textual inputs, while ensuring that the process remains fair, interpretable, and resilient to undue influence (Bueno et al., 2022).

To tackle this challenge, we propose an innovative approach leveraging advanced Large Language Models (LLMs) capable of semantically analysing expert commentary. Our methodology incorporates an intelligent filtering mechanism that identifies and removes comments potentially exerting a negative or distorting influence on other participants. This prevents the propagation of biases or irrelevant information that

could compromise the quality of group consensus.

Following this filtering stage, the LLM is employed to accurately extract the implicit preferences embedded within the remaining texts, converting complex linguistic expressions into structured, quantifiable representations. This capability to interpret natural language with rich contextual understanding far exceeds the limitations of simple sentiment analyses or constrained lexicons, thereby capturing the true intent and subtlety of each expert's contribution.

Subsequently, these extracted preferences are aggregated using operators designed to uphold fairness and faithfully reflect the genuine consensus of the group, yielding a robust and well-founded collective decision. This framework not only enhances the quality and transparency of decisions in complex, dynamic settings but also facilitates the inclusion of qualitative information traditionally marginalised or difficult to exploit through classical techniques.

Unlike existing sentiment analysis or rule-based NLP methods, which typically classify the polarity of comments or cluster opinions, our approach introduces an active semantic filtering mechanism powered by LLMs. This mechanism not only interprets expert discourse but also removes undue rhetorical influence, thereby ensuring that extracted preferences more faithfully represent authentic expertise.

This paper is organised into six main sections. Section 2 introduces the conceptual background, focusing on the principles underlying GDM and the novel use of LLMs to interpret expert-generated content. Section 3 outlines the operational design of the proposed model, with emphasis on its procedural flow and the integration of semantic capabilities for expert clustering. A proof-of-concept case study is described in Section 4, designed to illustrate the functioning of the proposed model in a controlled decision-making scenario. Section 5 reflects on the system's outcomes, assessing both its advantages and its constraints in comparison with existing methods. The manuscript concludes in Section 6 with a summary of contributions and potential directions for future development.

2. Preliminaries

The following section lays the groundwork for the methodological approach adopted in this study. It begins with Section 2.1, which explores the fundamentals of LLMs, highlighting their design, capabilities, and relevant use cases that reinforce the proposed system. This is followed by Section 2.2, where the core concepts and theoretical foundations of GDM are examined, providing the conceptual scaffolding for the model's

integration.

2.1. Large languages models

LLMs constitute the current state of the art in natural language processing, and their operation goes far beyond statistical associations between words. At their core, LLMs are built upon the Transformer architecture (Vaswani et al., 2017), whose key innovation is the self-attention mechanism. This mechanism computes context-dependent weights across all tokens in a sequence, enabling the model to capture both short-range and long-range semantic dependencies. Unlike recurrent or convolutional models, Transformers process input in parallel, refining contextual representations across multiple layers. Through this pipeline—tokenisation, vector embeddings, attention-based contextualisation, and autoregressive or masked prediction—LLMs can represent meaning at a high semantic and pragmatic level. These features are essential in our context, as they allow the extraction of subtle evaluative signals and implied preferences from expert commentary.

Beyond their technical design, LLMs have demonstrated remarkable versatility. They are capable of producing coherent text, translating across languages, summarising long documents, answering questions, and performing complex reasoning tasks with minimal task-specific engineering (Devlin et al., 2019). Such adaptability arises from large-scale pretraining on diverse corpora combined with transfer learning and fine-tuning, which permit efficient adaptation to specialised domains. This flexibility makes LLMs uniquely suited for contexts where unstructured, heterogeneous textual input must be integrated into formal decision processes.

In the literature, several applications of LLMs to decision-support and knowledge integration have been explored. (Wang et al., 2023) propose interactive frameworks in which LLMs adapt dynamically to user input and shifting contextual variables, fostering more responsive decision environments. (Zhou et al., 2024) show how LLMs can distil diverse expert perspectives into unified recommendations, particularly in policy design. (Zeng et al., 2024) demonstrate that LLMs can autonomously build domain-aware ontologies, facilitating more precise clustering and semantic representation. While these contributions highlight the growing role of LLMs, they tend to remain descriptive: most focus on showcasing technical capabilities without addressing how undue rhetorical influence in expert discourse may distort consensus formation.

This omission motivates our contribution. We do not merely exploit LLMs to summarise or cluster expert input; rather, we employ them as active semantic filters, identifying persuasive or manipulative commentary that could bias group deliberation. In doing so, our approach extends the role of LLMs from powerful language processors to safeguards of fairness in collective decision-making, ensuring that consensus outcomes reflect genuine expertise rather than rhetorical dominance.

2.2. Group decision making

In our framework, GDM is defined as the evaluation of a finite set of alternatives by a panel of experts, whose preferences are expressed in free-form textual comments. Instead of relying solely on predefined scales, these comments are semantically processed by an LLM, which extracts structured preference relations for subsequent aggregation.

Let us define the set of decision agents as $\Theta = \{\theta_1, \theta_2, \dots, \theta_\nu\}$, where $\nu \in \mathbb{N}$ indicates the total number of experts. The available alternatives to be examined are denoted by $R = \{\rho_1, \rho_2, \dots, \rho_\kappa\}$, with $\kappa \in \mathbb{N}$ representing the cardinality of the alternatives (Kobashikawa et al., 2008).

A key challenge addressed in this work is the mismatch between natural language judgments provided by experts and the structured numerical inputs required for computation. Our framework resolves this by using an LLM to translate qualitative discourse into reciprocal preference matrices, thus preserving semantic richness while enabling formal analysis.

To transcend the limitations inherent in traditional scale-based approaches, we adopt a paradigm wherein experts are no longer bound to fixed evaluative schemes. Instead, they are invited to express their comparative judgments freely in natural language, without being constrained to predefined numerical labels. The task of interpreting these relational statements is subsequently delegated to an LLM, whose operational foundations have been delineated in prior sections (Lin and Hung, 2011). In collaborative decision-making environments, LLMs offer novel capabilities for processing qualitative expert contributions—free-form comments, reflective judgments, and argumentative discourse—by converting this input into structured formats that align with computational models (Wang et al., 2023; Zhou et al., 2024). This translation enhances the incorporation of nuanced human input into data-driven decision processes and reinforces the reliability of collective outcomes. In particular, recent studies have shown how LLMs can support decision frameworks by dynamically

adapting to user input (Wang et al., 2023) and by distilling heterogeneous perspectives into unified and interpretable outputs (Zhou et al., 2024).

Once the key concepts have been established, the main stages of a GDM method can be summarised as follows:

- **Opinion Collection:** Experts express judgments on alternatives via reciprocal preference relations.
- **Consensus Evaluation:** Agreement is checked against threshold α , with up to r feedback rounds if unmet, else a ranking with note on lack of consensus.
- **Information Aggregation:** Preferences are merged into a collective matrix \mathcal{G} using the weighted average (WA) operator.
- **Ranking Derivation:** The QGDD operator processes \mathcal{G} to obtain the final ranking of alternatives.

It is important to note that the literature on GDM has already proposed numerous sophisticated frameworks to cope with uncertainty, heterogeneity, and incomplete information. For example, Li et al., 2020 incorporates experts' confidence levels into the aggregation process, thereby refining the influence of individual judgments and improving consensus quality. Likewise, Zhang et al., 2021 introduce interval-valued intuitionistic fuzzy sets to capture hesitation and ambiguity, enhancing robustness in contexts where preferences are not crisply defined. Advanced consensus-reaching models such as the one by Singh et al., 2019 further allow dynamic adjustment of expert influence based on iterative feedback, promoting convergence while preserving diversity. Existing approaches address uncertainty and incomplete information but do not explicitly handle the semantic and rhetorical influence present in expert discourse. Our framework introduces LLM-based semantic filtering to address this gap, ensuring that consensus reflects genuine expertise rather than persuasive dominance.

3. Enhancing GDM through LLM for semantic filtering of expert opinions

This section details the methodological procedure that supports the GDM model. The distinctive contribution of this framework lies in combining semantic filtering with a novel weighting mechanism, where expert influence is adjusted according to the filtering ratio of their comments, and in embedding this adjustment into the consensus-verification stage.

The proposal is based on the integration of LLM, taking advantage of their capacity to interpret, debug and condense the contributions made by experts. In this study, we specifically employ the DeepSeek LLM, whose advanced contextual reasoning capabilities support the semantic filtering and preference extraction stages. This approach allows for the establishment of a sophisticated deliberative environment. The methodology is organised in a series of sequential steps (see Figure 1):

- Debate and filtering of Influential Expert Comments via LLM: LLMs detect and remove potentially influential remarks to ensure fair deliberation.
- Extraction of Expert Preferences from Filtered Comments: Genuine expert preferences are derived by analysing the semantic content of refined comments.
- Calculation of Expert Weights Based on Comment Filtering Ratio: Expert influence is weighted inversely to the proportion of filtered comments.
- Consensus Verification Among Experts: Agreement and alignment are assessed through consensus metrics on weighted preferences.
- Aggregation of Weighted Preferences: Weighted individual preferences are combined to form a collective preference structure.
- Derivation of the Final Ranking of Alternatives: The aggregated preferences yield a final prioritised ranking of alternatives.

3.1. Debate and filtering of influential expert comments via LLM

At the outset of the decision-making process, each expert $\theta_i \in \Theta = \{\theta_1, \theta_2, \dots, \theta_\nu\}$ engages in a deliberative discussion concerning the set of available alternatives $R = \{\rho_1, \rho_2, \dots, \rho_\kappa\}$. Experts articulate their evaluations through natural language comments, denoted by $c_{i,j}$, where each comment corresponds to the j -th contribution of expert θ_i . The complete set of comments produced across all experts is expressed as:

$$C = \bigcup_{i=1}^{\nu} \{c_{i,1}, c_{i,2}, \dots, c_{i,m_i}\}, \quad (1)$$

where m_i is the number of comments made by expert θ_i . These comments serve as the primary

medium through which preference information and argumentative reasoning are communicated.

To ensure the deliberation remains constructive, objective, and free of undue influence, we deploy an LLM to evaluate the semantic content of each comment. The LLM acts as a dynamic filter, identifying and excluding those comments that exhibit indicators of bias, manipulation, redundancy, lack of relevance, or non-cooperativity. Let the binary filtering function be denoted as:

$$\varphi(c_{i,j}) = \begin{cases} 1 & \text{if } c_{i,j} \text{ is accepted by the LLM,} \\ 0 & \text{if } c_{i,j} \text{ is filtered out.} \end{cases} \quad (2)$$

The subset of accepted (i.e., non-filtered) comments for expert θ_i is thus defined as:

$$\widehat{C}_i = \{c_{i,j} \in C : \varphi(c_{i,j}) = 1\}. \quad (3)$$

This filtered corpus forms the basis for reliable preference extraction, mitigating epistemic noise and ensuring that only semantically coherent and argumentatively valid contributions influence the group decision.

To further guarantee the correctness of the filtering process, the system applies a redundancy mechanism whereby each comment is evaluated through multiple parallel runs of the LLM. Only classifications that remain stable across these iterations are retained, reducing the likelihood of spurious outputs. In addition, the filtering step assigns a reliability indicator that can be monitored to identify potential anomalies.

3.2. Extraction of expert preferences from filtered comments

Following the filtering stage, the subset \widehat{C}_i is semantically analysed to extract implicit or explicit preferences expressed by expert θ_i toward each alternative $\rho_k \in R$. The LLM is employed again—this time for preference mining—leveraging its contextual reasoning capabilities to identify evaluative intent and comparative judgments embedded in natural language.

The extraction process is also accompanied by a reliability score, reflecting the consistency of the inferred preferences across repeated evaluations. This score allows stakeholders to detect unstable outputs and to interpret the robustness of the generated preference matrices.

Let the function $\psi : \widehat{C}_i \rightarrow \mathbb{R}^{\kappa \times \kappa}$ map the filtered comments of expert θ_i to a pairwise preference matrix \mathcal{G}_i , where each entry quantifies the expressed preference of one alternative over another. Formally,

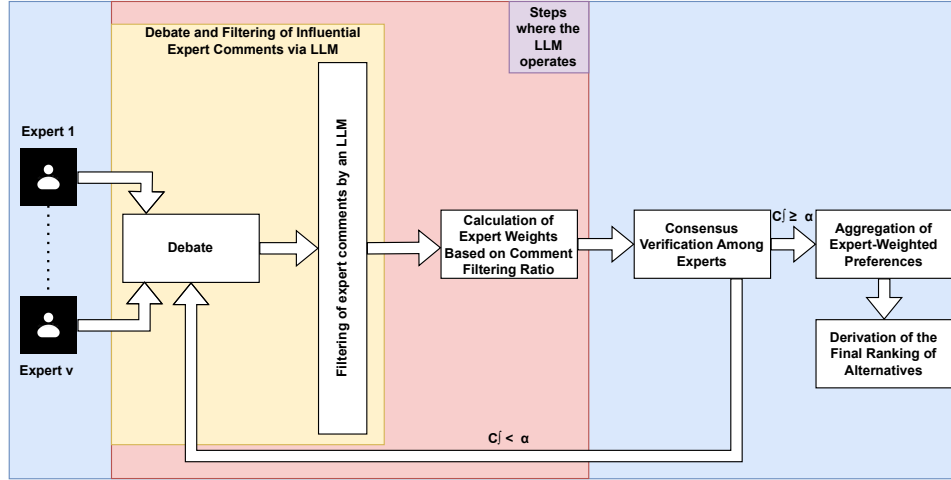


Figure 1. Workflow of the proposed framework. The red-highlighted blocks indicate the stages where the LLM operates (semantic filtering of expert comments and extraction of preferences), whose outputs feed into weight calculation, consensus verification, and aggregation.

$$\mathcal{G}_i = \psi(\widehat{C}_i) = (p_{s,k}^i) \in [0, 1]^{\kappa \times \kappa}, \quad \text{for } i = 1, \dots, \nu, \quad (4)$$

where each element $p_{s,k}^i \in [0, 1]$ reflects the strength of preference that expert θ_i expresses in favour of alternative ρ_s over ρ_k . In this study, we adopt reciprocal preference relations, defined as

$$p_{s,k}^i + p_{k,s}^i = 1, \quad \forall s, k \in \{1, \dots, \kappa\}, s \neq k. \quad (5)$$

While we acknowledge that many GDM approaches do not enforce reciprocity and allow more general preference structures (e.g., interval-valued intuitionistic fuzzy relations (Zhang et al., 2021), hesitant fuzzy ontologies (Morente-Moliner et al., 2020)), reciprocity is assumed here for two main reasons. First, it ensures internal consistency in pairwise comparisons, which simplifies the aggregation process and reduces computational ambiguity. Second, because our framework already introduces a novel semantic filtering stage via LLMs, imposing reciprocity provides a stable and interpretable foundation for preference extraction, making it easier to trace the impact of the filtering mechanism on the final group consensus.

The full matrix of preferences $\mathcal{G}_i \in \mathbb{R}^{\kappa \times \kappa}$ is constructed as:

$$\mathcal{G}_i = \begin{bmatrix} p_{1,1}^i & p_{1,2}^i & \cdots & p_{1,\kappa}^i \\ p_{2,1}^i & p_{2,2}^i & \cdots & p_{2,\kappa}^i \\ \vdots & \vdots & \ddots & \vdots \\ p_{\kappa,1}^i & p_{\kappa,2}^i & \cdots & p_{\kappa,\kappa}^i \end{bmatrix}; \quad i = 1, \dots, \nu \quad (6)$$

This matrix encodes the filtered and inferred preferences of all experts, serving as the semantic foundation for the final aggregation and consensus estimation.

3.3. Calculation of expert weights based on comment filtering ratio

To modulate the influence of each expert in the final decision, we introduce a weighting mechanism based on the quality of their contributions. Specifically, an expert's weight is defined as inversely proportional to the fraction of their comments that were filtered out by the LLM. This strategy is consistent with established practices in GDM where weights are adjusted according to the confidence level or consistency of experts' assessments (Zhang et al., 2021). In our framework, the semantic filtering ratio serves as a proxy for contribution quality, ensuring that the aggregation process privileges informative and unbiased input, thereby enhancing the robustness and fairness of the final group consensus.

Let η_i denote the total number of comments made by expert θ_i , and $\hat{\eta}_i = |\widehat{C}_i|$ the number of comments retained after filtering. The filtering ratio δ_i is given by:

$$\delta_i = \frac{\hat{\eta}_i}{\eta_i}, \quad (7)$$

and the corresponding weight $\mathcal{W}_i \in [0, 1]$ for expert θ_i is computed as:

$$\mathcal{W}_i = \frac{1 - \delta_i}{\sum_{j=1}^{\nu} (1 - \delta_j)}. \quad (8)$$

This normalisation ensures that:

$$\sum_{i=1}^{\nu} \mathcal{W}_i = 1. \quad (9)$$

In essence, this weighting scheme reduces the epistemic impact of experts whose discourse failed to meet the semantic thresholds of the LLM, thus reinforcing the integrity and reliability of the collective decision.

3.4. Consensus verification among experts

Once the individual weights \mathcal{W}_i associated with each expert θ_i have been established, the subsequent phase focuses on evaluating the degree of consensus within the panel. This stage is pivotal for identifying the level of coherence in the experts' preferences and for diagnosing potential discrepancies that may compromise the reliability of the collective outcome. Rather than assuming uniformity, the framework rigorously quantifies divergences in opinion through precise computational mechanisms. A threshold α guides consensus validation, triggering feedback if unmet, while a maximum of $r = 10$ iterations ensures efficiency and prevents endless loops.

The consensus index is derived through a double-layered distance-based formulation. Let $\mathcal{G}_i = (p_{s,k}^i)$ and $\mathcal{G}_j = (p_{s,k}^j)$ denote the preference matrices from experts θ_i and θ_j , respectively. The expression for computing the consensus degree is defined as:

$$\mathcal{C}f = 1 - \frac{2 \cdot \sum_{i=1}^{\nu-1} \sum_{j=1; i>j}^{\nu} \sqrt{\frac{\sum_{s=1}^{\kappa} \sum_{k=1; s \neq k}^{\kappa} (p_{s,k}^i - p_{s,k}^j)^2}{\kappa \cdot \kappa - \kappa}}}{(\nu - 1) \cdot \nu} \quad (10)$$

Here, $p_{s,k}^i$ represents the strength of preference of expert θ_i for alternative ρ_s over ρ_k , and the inner term reflects the average pairwise discrepancy between two experts' evaluations over all alternative combinations. The outer sum aggregates such deviations across all expert pairs.

The resulting index $\mathcal{C}f \in [0, 1]$ acts as a bounded consensus metric: values approaching 1 indicate high agreement, while values close to 0 signal substantial divergence. The iterative revision process is repeated until either a satisfactory consensus level is attained ($\mathcal{C}f \geq \alpha$) or the feedback limit r is reached, ensuring a rational and cohesive aggregation of expert knowledge.

3.5. Aggregation of expert-weighted preferences

Once the individual weights associated with each expert have been determined and all preference information has been collected, the subsequent step involves synthesizing this information into a single collective structure. This is accomplished through a weighted aggregation process that yields the collective reciprocal preference relation, denoted as $\mathcal{G} = (p_{sk}; s \neq k = 1, \dots, \kappa)$. The resulting matrix has dimensions $\kappa \times \kappa$, omitting the main diagonal as self-comparisons are not considered meaningful within this framework.

To effectively consolidate the individual assessments, the use of an appropriate aggregation operator is essential. In this context, the WA operator is employed due to its capacity to account for varying degrees of influence among experts. This operator integrates each expert's reciprocal preference values with their corresponding weights, thereby reflecting both the content and the credibility of the input data.

The aggregation process is formally defined as follows: for each pair of alternatives (i, j) , the corresponding entry in the collective preference matrix is computed using:

$$p_{sk} = \sum_{i=1}^{\nu} \mathcal{W}_i \cdot p_{sk}^i \quad (11)$$

This formulation ensures that the contribution of each expert to the collective matrix is proportional to their assigned relevance. It also preserves the structure and interpretability of the initial data, as the aggregated preference values remain within the original bounded domain $[0, 1]$.

The output matrix \mathcal{G} serves as a unified representation of the group's judgments and forms the basis for subsequent steps such as consensus measurement, consistency checking, and final decision-making. By relying on a rigorously weighted aggregation strategy, the model enhances the robustness and transparency of the overall decision process, especially in scenarios involving heterogeneous expertise and multidimensional evaluation criteria.

3.6. Deriving the final ranking of alternatives

In the concluding phase of the decision-making process, and once the collective reciprocal preference matrix has been fully constructed, the next task is to generate a final ranking of the available alternatives. To achieve this, we employ a quantitative approach that evaluates each option's relative strength compared to

the rest. Among the possible techniques, we adopt the *Quantifier Guided Degree of Dominance* (QGDD) method, which is particularly well-suited for measuring how dominant a given alternative is with respect to the others based on the aggregated preference values.

This approach involves computing a dominance score for each alternative, capturing the average intensity with which it is preferred over all other alternatives. Specifically, for each alternative ρ_s , we determine the score $QGDD_{\rho_s}$ using the following expression:

$$QGDD_{\rho_s} = \frac{\sum_{\substack{k=1 \\ k \neq s}}^{\kappa} p_{sk}}{\kappa - 1} \quad (12)$$

Here, κ denotes the total number of alternatives, and $p_{ij} \in [0, 1]$ corresponds to the collective preference of alternative ρ_s over ρ_k as derived from the previously aggregated preference matrix.

Following validation, the final step is to identify the most dominant alternative. This is accomplished by selecting the option whose dominance score is the highest among all candidates, as indicated in the formal definition below:

$$\rho_{QGDD} = \left\{ \rho_i \in R \mid QGDD_{\rho_s} = \max_{\rho_k \in R} QGDD_{\rho_k} \right\} \quad (13)$$

The selected alternative ρ_{QGDD} thus reflects the optimal choice according to the collective preferences provided by the expert panel. This ensures that the final decision is aligned with the aggregated judgments and satisfies the logical principles established throughout the evaluation process.

4. Case study: illustrative proof-of-concept

To illustrate the practical impact of the proposed decision-making framework, we consider a scenario involving a public health advisory board tasked with prioritising among five policy interventions in response to an emerging infectious disease. The panel is composed of four experts, denoted as $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. It is important to note that this case does not represent a real-world deployment, but rather a controlled illustrative setting aimed at demonstrating the operation of the proposed framework.

The set of strategic alternatives is represented by $R = \{\rho_1, \rho_2, \rho_3, \rho_4, \rho_5\}$, where each element ρ_k corresponds to a distinct project category. Specifically, these alternatives include a mass

vaccination programme, expansion of intensive care capacity, investment in rapid diagnostic testing, public information and awareness campaigns, and targeted restrictions in high-risk regions.

During deliberation, the LLM (DeepSeek in the case study) averages expert ratings, analyses comments for coherence, extracts structured preferences, and groups experts by similarity to track preference shifts and influence.

$$\mathcal{G}_1 = \begin{pmatrix} - & 0.28 & 0.23 & 0.34 & 0.15 \\ 0.72 & - & 0.56 & 0.55 & 0.34 \\ 0.77 & 0.44 & - & 0.34 & 0.36 \\ 0.66 & 0.45 & 0.66 & - & 0.29 \\ 0.85 & 0.66 & 0.64 & 0.71 & - \end{pmatrix}$$

The study shows that after LLM-based filtering of expert comments, both the consensus level and the relative weights of experts are recalculated. The mechanism reduces the impact of filtered remarks, thereby adjusting expert weights. This demonstrates the key contribution of the approach: the LLM not only removes biased inputs but also reshapes the aggregation process through reliability-adjusted weighting.

During the validation, the LLM was prompted in a zero-shot setting to identify comments containing absolutist framing, emotional language, or redundant and irrelevant statements. No fine-tuning was applied, and each comment was evaluated through multiple runs to ensure consistency in the classification. For instance, remarks such as “*It is obvious that option ρ_3 is the only rational choice, and everyone should agree*” or “*I strongly insist that ρ_2 must be selected, regardless of other considerations*” were removed because they displayed persuasive or coercive framing rather than neutral evaluation. By filtering these elements, the system reduced rhetorical influence and enhanced the fairness of the aggregation process.

The level of agreement within the expert panel is assessed by applying a previously established consensus threshold, set at ($\alpha = 0.90$). In this case, a consensus index ($\mathcal{C}f = 0.9225$) was obtained, which is well above this threshold, allowing us to conclude with confidence that there is a satisfactory degree of convergence in the assessments made by the experts. This evidence supports the validity of the assessment process and ensures that preferences reflect shared and consistent criteria. From the comments made by the experts, weighted weights were calculated for each expert, resulting in values of $\mathcal{W}_1 = 0.267$, $\mathcal{W}_2 = 0.236$, $\mathcal{W}_3 = 0.369$, and $\mathcal{W}_4 = 0.128$, respectively, allowing the relative influence of each expert to be incorporated into the final analysis.

$$\mathcal{G} = \begin{pmatrix} - & 0.55 & 0.45 & 0.44 & 0.23 \\ 0.45 & - & 0.35 & 0.36 & 0.33 \\ 0.55 & 0.65 & - & 0.42 & 0.28 \\ 0.56 & 0.64 & 0.58 & - & 0.25 \\ 0.77 & 0.67 & 0.72 & 0.75 & - \end{pmatrix}$$

The ultimate ranking of the alternatives is obtained by employing the Quantified Generalized Dominance Degree (QGDD) approach, where the arithmetic mean operator is applied to the overall preference matrix. The calculated dominance scores are presented in Table 1, enabling a systematic comparison among the candidate options:

Table 1. Dominance indices computed via the QGDD framework, comparing the baseline case (without LLM filtering) and the proposed model (with semantic filtering).

Alternative	Baseline QGDD	Proposed QGDD
ρ_1	0.2500	0.4175
ρ_2	0.5425	0.3725
ρ_3	0.4775	0.4750
ρ_4	0.5150	0.5075
ρ_5	0.7150	0.7275

To conclude the analysis, we employ a verification mechanism to confirm the soundness and consistency of the applied method. This step not only guarantees the methodological validity but also highlights a clear expert preference bias toward alternative ρ_5 , thereby reinforcing its position as the most dominant choice among the evaluated options.

5. Discussions

This research presents a novel approach to enhance GDM by employing LLMs to filter expert comments that could potentially bias the opinions of others. Instead of clustering expert inputs based on semantic similarity, the framework focuses on removing comments with undue influence—such as overly persuasive or assertive remarks—before extracting individual expert preferences. This strategy promotes a more impartial and balanced aggregation of judgments, preserving the diversity of expert perspectives without allowing dominant voices to overshadow minority opinions. This constitutes the core novelty of our proposal: moving beyond surface-level text analytics toward a proactive filtering and extraction process that directly safeguards fairness in consensus formation, distinguishing our method from conventional NLP or sentiment analysis techniques.

The analysis shows that semantic filtering with LLMs adds clear value to group decision-making. Compared to unfiltered deliberation and lexicon-based filtering, the LLM approach achieved a more balanced distribution of contributions and greater stability of the collective outcome. Qualitative examples reveal that the model captures not only polarity but also subtle persuasive framings that could bias evaluations. Overall, this reinforces the role of LLM-based filtering as a mechanism for more reliable and equitable group decisions.

The strength of this approach lies in the LLM’s capacity for deep contextual understanding of natural language, which allows it to distinguish subtle nuances in expert discourse beyond simple sentiment scores or polarity measures. Rather than relying on surface-level indicators of bias, such as word frequency or sentiment polarity, the model is capable of identifying complex rhetorical strategies, such as appeals to authority, absolutist framing, or emotionally charged language. This enables a more granular filtering process that targets the root causes of influence rather than its superficial expressions. This focus on integrating LLM-based semantic filtering with weighting and consensus mechanisms differentiates our work from prior NLP- or sentiment-based approaches, where language models are typically used only for auxiliary classification or clustering.

By selectively filtering biased comments, the system helps preserve the authenticity of each expert’s viewpoint, reducing social influence effects common in group deliberations, conformity pressure, or anchoring bias. In doing so, it encourages a decision-making environment where experts feel freer to express dissenting or unconventional opinions, thereby improving the epistemic quality of the group’s output. Key benefits of this method include: The system filters assertive, emotional, or dominant remarks to preserve authentic expert preferences. Furthermore, LLMs use semantic and pragmatic cues to detect subtle influence strategies beyond lexical or sentiment features. Finally, separating biasing from neutral comments ensures traceable, interpretable, and trustworthy decision-making.

These benefits can already be partially observed in the case study presented in Section 4. For example, the semantic filtering process reduced the influence of Expert 4 by lowering their weight due to a higher proportion of filtered comments, which demonstrates the framework’s capacity for bias reduction. Similarly, the removal of redundant or irrelevant remarks illustrates the context-aware nature of the filtering, while the explicit reporting of expert

weights and consensus indices reflects the transparency of the process. Although the case study is limited in scale, these outcomes provide initial evidence of the practical impact of the proposed mechanisms. While LLMs provide powerful semantic capabilities, they are not error-free. To address this limitation, our framework incorporates three safeguards: (i) redundancy through multiple LLM runs to ensure stable results, (ii) reliability scores associated with both the filtering ratio and the preference matrices, and (iii) the possibility of introducing a human-in-the-loop audit in critical contexts. These mechanisms strengthen the robustness and transparency of the model, ensuring that the outputs can be trusted even in high-stakes decision-making scenarios. Nonetheless, we acknowledge that the proof-of-concept case study cannot by itself fully demonstrate all the claimed benefits, especially in terms of scalability and generalisability. A larger and more diverse empirical validation will therefore be required to conclusively prove the improvements in fairness, robustness, and transparency introduced by the proposed model.

While the proposed approach highlights the potential of LLM-based semantic filtering in group decision-making, several limitations must be noted. One risk is the excessive removal of dissenting opinions, which may suppress valuable minority perspectives. The method is also sensitive to prompt design and model configuration, where small changes can alter outcomes. Additionally, in highly technical domains, LLMs may overlook domain-specific nuances, and in contexts with limited textual input, semantic filtering adds little value. Acknowledging these constraints is crucial for refining the method's robustness and adaptability.

Compared to existing literature, this framework offers significant advancements. For instance, Kauffmann et al., 2020 primarily focuses on detecting deceptive or fallacious comments but does not address broader social influence dynamics within group deliberations. Meanwhile, the approach presented by Morente-Molinera et al., 2020 applies sentiment analysis to select optimal alternatives across multiple rounds in decision processes involving numerous options; nonetheless, their method depends on a fixed bag-of-words model, which limits its ability to capture the full complexity of expert language. In contrast, our approach leverages the superior contextual understanding and processing capabilities of LLMs, enabling more flexible and accurate filtering without reliance on predefined lexical sets. Furthermore, approaches like Morente-Molinera et al., 2019 use sentiment lexicons and polarity scores that tend to

oversimplify expert discourse, failing to capture subtle evaluative nuances. Our model, by contrast, exploits the inferential and contextual reasoning power of LLMs to filter biased commentary and extract genuine expert preferences, providing a more nuanced, robust, and equitable framework for group decision-making.

6. Conclusions and future work

This work introduces a novel GDM framework that leverages LLMs to enhance the integrity of expert evaluations by filtering out comments that may bias other participants. Unlike other methods based on sentiment polarity, this approach prioritises the removal of statements with potential undue influence—such as highly assertive or persuasive remarks—before extracting individual expert preferences from the remaining unbiased input.

The model's strength lies in the LLMs ability to semantically analyse natural language with contextual awareness, adapting to domain-specific expert discourse. This alignment enables precise extraction of true expert preferences, beyond superficial sentiment analysis, ensuring a more faithful representation of each participant's viewpoint.

Moreover, the proposed framework facilitates a more equitable decision-making environment by actively mitigating social influence biases that often skew group judgments. By ensuring that dominant voices do not disproportionately affect the consensus, the system promotes diversity of thought and preserves the independence of expert opinions. This aspect is particularly relevant in contexts where balanced deliberation and fairness are paramount, such as healthcare panels or policy committees. We acknowledge as a limitation that the current evaluation is restricted to a proof-of-concept case study. A more extensive real-world validation, involving larger panels of experts and domain-specific datasets, will be pursued in future research to further substantiate the practical impact of the proposed framework.

Future research directions include incorporating additional data sources, such as temporal interaction patterns or paralinguistic cues, to refine the filtering mechanism further. Implementing continual learning paradigms could also allow the system to adapt dynamically as expert interactions evolve. Additionally, integrating explainability features within the LLM would promote transparency by clarifying the rationale behind filtering decisions and preference derivation, a critical factor in applications demanding accountability and interpretability.

7. Acknowledgements

This work has been supported by the grant PID2022-139297OB-I00 funded by MICIU/AEI/10.13039/501100011033 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.

References

- Bueno, I., Carrasco, R. A., Ureña, R., & Herrera-Viedma, E. (2022). A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations. *Information Sciences*, 589, 300–320.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.
- Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, 90, 523–537.
- Kobashikawa, C., Hatakeyama, Y., Dong, F., & Hirota, K. (2008). Fuzzy algorithm for group decision making with participants having finite discriminating abilities. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 39(1), 86–95.
- Li, H., Wang, X., & Liu, B. (2020). Incorporating experts' confidence levels into group decision making for weight determination. *Information Sciences*, 523, 42–56.
- Lin, K.-P., & Hung, K.-C. (2011). An efficient fuzzy weighted average algorithm for the military uav selecting under group decision-making. *Knowledge-Based Systems*, 24(6), 877–889.
- Morente-Molinera, J. A., Cabrerizo, F. J., Mezei, J., Carlsson, C., & Herrera-Viedma, E. (2020). A dynamic group decision making process for high number of alternatives using hesitant fuzzy ontologies and sentiment analysis. *Knowledge-Based Systems*, 195, 105657.
- Morente-Molinera, J. A., 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. *Knowledge-Based Systems*, 165, 335–345.
- Pérez, I. J., Cabrerizo, F. J., Alonso, S., & Herrera-Viedma, E. (2013). A new consensus model for group decision making problems with non-homogeneous experts. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4), 494–498.
- Singh, S., Gupta, D., & Singh, B. (2019). A dynamic consensus reaching model for group decision making with incomplete preference information. *IEEE Transactions on Cybernetics*, 49(2), 596–606.
- Urena, R., Cabrerizo, F. J., Morente-Molinera, J. A., & Herrera-Viedma, E. (2016). Gdm-r: A new framework in r to support fuzzy group decision making processes. *Information Sciences*, 357, 161–181.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, F.-Y., Yang, J., Wang, X., Li, J., & Han, Q.-L. (2023). Chat with chatgpt on industry 5.0: Learning and decision-making for intelligent industries. *IEEE/CAA Journal of Automatica Sinica*, 10(4), 831–834.
- Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1), 3–28.
- Zeng, W., Dou, Y., Pan, L., Xu, L., & Peng, S. (2024). Improving prediction performance of general protein language model by domain-adaptive pretraining on dna-binding protein. *Nature Communications*, 15(1), 7838.
- Zhang, Y., Xu, Z., & Wu, Q. (2021). Interval-valued intuitionistic fuzzy set based group decision making method for supplier selection. *Expert Systems with Applications*, 173, 114645.
- Zhou, G., Hong, Y., & Wu, Q. (2024). Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7), 7641–7649.