

Article

From Divergence to Alignment: Evaluating the Role of Large Language Models in Facilitating Agreement Through Adaptive Strategies

Loukas Triantafyllopoulos *  and Dimitris Kalles 

School of Science and Technology, Hellenic Open University, 263 31 Patras, Greece; kalles@eap.gr

* Correspondence: triantafyllopoulos.loukas@ac.eap.gr

Abstract

Achieving consensus in group decision-making often involves overcoming significant challenges, particularly reconciling diverse perspectives and mitigating biases hindering agreement. Traditional methods relying on human facilitators are usually constrained by scalability and efficiency, especially in large-scale, fast-paced discussions. To address these challenges, this study proposes a novel real-time facilitation framework, employing large language models (LLMs) as automated facilitators within a custom-built multi-user chat system. This framework is distinguished by its real-time adaptive system architecture, which enables dynamic adjustments to facilitation strategies based on ongoing discussion dynamics. Leveraging cosine similarity as a core metric, this approach evaluates the ability of three state-of-the-art LLMs—ChatGPT 4.0, Mistral Large 2, and AI21 Jamba-Instruct—to synthesize consensus proposals that align with participants' viewpoints. Unlike conventional techniques, the system integrates adaptive facilitation strategies, including clarifying misunderstandings, summarizing discussions, and proposing compromises, enabling the LLMs to refine consensus proposals based on user feedback iteratively. Experimental results indicate that ChatGPT 4.0 achieved the highest alignment with participant opinions and required fewer iterations to reach consensus. A one-way ANOVA confirmed that differences in performance between models were statistically significant. Moreover, descriptive analyses revealed nuanced differences in model behavior across various sustainability-focused discussion topics, including climate action, quality education, good health and well-being, and access to clean water and sanitation. These findings highlight the promise of LLM-driven facilitation for improving collective decision-making processes and underscore the need for further research into robust evaluation metrics, ethical considerations, and cross-cultural adaptability.

Keywords: consensus-building; artificial intelligence; collaborative decision-support systems; large language models; cosine similarity



Academic Editors: Diego Vergara and Pablo Fernández-Arias

Received: 7 August 2025

Revised: 4 September 2025

Accepted: 4 September 2025

Published: 6 September 2025

Citation: Triantafyllopoulos, L.; Kalles, D. From Divergence to Alignment: Evaluating the Role of Large Language Models in Facilitating Agreement Through Adaptive Strategies. *Future Internet* **2025**, *17*, 407. <https://doi.org/10.3390/fi17090407>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The concept of “achieving consensus” is often misinterpreted as complete unanimity among group members [1]. In fact, consensus does not require absolute agreement but the acceptance of a proposal that is considered satisfactory by all interested parties, considering the existing constraints [2]. Nevertheless, achieving this kind of agreement is a complex challenge, especially when participants have conflicting views [3,4]. Differences in opinion often arise from personal beliefs and priorities [5], while biases like attachment to initial

positions [6,7] and confirmation bias [8] can hinder consensus by limiting openness to alternative perspectives.

Given the importance of this issue in collective decision-making, particularly on matters of significant concern to humanity, various methods have been employed over time to facilitate the process. One common approach is including a facilitator, often an expert or knowledgeable individual on the topic under discussion [9,10]. The facilitator guides and structures the dialogue, helping participants navigate differing viewpoints, identify common ground, encourage critical thinking, and propose mutually acceptable ideas to reach consensus [11]. While this practice has proven effective in many cases, it has shown limitations when addressing particularly complex or contentious issues, especially in large-scale and fast-paced online discussions [12,13].

On this issue, the research community has increasingly focused on automating the role of the facilitator [14], with the use of artificial intelligence (AI) playing a pivotal role in advancing this effort in recent years [12]. Notably, a growing body of research highlights the increasing interest in the potential of AI-driven facilitators over their human counterparts [13,15]. This trend has further evolved with adopting large language models (LLMs), which are increasingly recognized as effective tools for facilitating complex discussions and consensus-building processes [16].

LLMs possess a unique ability to process and synthesize vast information [17]. Analyzing diverse perspectives, identifying common themes, and proposing neutral, balanced statements can serve as powerful tools for fostering consensus [18]. Additionally, as unbiased facilitators, these models can synthesize varied viewpoints, ensuring all group members have an equitable voice in the conversation [19].

Measures like cosine similarity are particularly useful for evaluating LLMs' success in achieving consensus [3,20,21]. Cosine similarity is effective in ranking alternatives and assessing how closely participants' views align during the consensus process [22], often providing more accurate results than Euclidean distance in such scenarios [23]. Additionally, cosine similarity allows for the assessment of alignment in high-dimensional spaces [24], which is common in LLM-based tasks where the complexity of the data often extends beyond simple scalar comparison.

In this framework, the current study aims to evaluate the effectiveness of three popular LLMs in facilitating consensus among individuals on critical issues. To accomplish this, a chat-based environment was developed from scratch, enabling synchronous communication between participating users and following a predefined discussion scenario where the LLM served as the facilitator. Its role was especially to synthesize consensus proposals based on user feedback, employing, in cases of disagreement, four well-known strategies from the literature for achieving consensus. After that, the study used cosine similarity to measure the alignment between users' initial positions and the final consensus, providing a precise metric to evaluate the effectiveness of these strategies and the overall ability of the LLMs to foster agreement.

The key contributions of this paper can be outlined as follows:

1. First, the development of a fully custom, synchronous multi-user chat platform that enables real-time LLM-facilitated discussions, going beyond existing static or asynchronous systems such as D-agree;
2. Second, the integration of an adaptive facilitation framework, allowing LLMs to dynamically select among five literature-based consensus strategies during live interactions—an innovation not previously tested with multiple LLMs in a controlled setting;
3. Third, the use of cosine similarity with sentence-level vectorization (via the Universal Sentence Encoder) to evaluate consensus quantitatively across iterative discussion

rounds, offering a novel and replicable method to benchmark alignment between participants' opinions and LLM-generated consensus proposals;

4. Finally, the comparative empirical assessment of three cutting-edge LLMs across a diverse set of sustainability-related questions yielded new insights into their behavior, convergence speed, and performance variability across discussion topics.

Beyond its empirical findings, this study contributes conceptually by operationalizing a replicable framework for AI-facilitated consensus evaluation. It bridges real-time interaction design with measurable alignment metrics, offering a structured methodology for future research across models, domains, and facilitation styles. Nevertheless, these contributions should be interpreted in the context of discussions involving only two participants, which inherently limits the generalizability of the results obtained.

The organization of this paper is structured as follows: Section 2 reviews relevant literature and outlines the research questions that guide this study. Section 3 details the methodology employed, including the consensus-building framework and tools used, and explains the selection criteria for the sample in the pilot test. Section 4 presents the empirical findings. Section 5 discusses these results, explores their implications, and addresses the study's limitations. Finally, Section 6 offers concluding remarks and highlights the key insights derived from the research.

2. Background

Within the context of the relevant literature, we focus on two critical issues. The first concerns the role of artificial intelligence (AI) in facilitating consensus building; the second addresses the various approaches used to measure AI's effectiveness in this area. After reviewing the pertinent information, we present the current study's approach as an advancement over the existing literature, highlighting the gaps it aims to fill and its potential contributions to the scientific community.

2.1. The Role of Artificial Intelligence in Facilitating Consensus Building

A review of the recent literature reveals numerous research efforts aimed at evaluating AI as a reliable tool for mediating and facilitating consensus-building. For this purpose, scholars often compare the outcomes of two scenarios: one with AI assistance and one without it throughout the process.

One popular approach discussed in the literature involves regularly integrating conversational agents into discussion platforms such as D-agree [15,25–28]. This online text-based platform allows participants to exchange messages with each other and with conversational AI. The AI is based on the Issue-Based Information System (IBIS) framework [29], which helps guide participants toward consensus by deconstructing dialogues into issues, ideas, pros, and cons. Additionally, these agents can summarize data, visualize discussion structures, and enhance participation with adaptive messages. Utilizing this framework, Haqbeen et al. conducted studies on achieving consensus regarding sustainability issues in Kabul, Afghanistan, and addressing solidarity in the Ukraine conflict [25,26]. Similarly, Sahab et al. [28] noted the effectiveness of AI agents in facilitating consensus in diverse environments across five Afghan cities, while Ito et al. [15] assessed the social influence of this platform in Japan. Indeed, Sahab et al. [30] later compared the results between these two countries to evaluate the effectiveness of AI agents across different cultural settings. Finally, Hadifi et al. [27] explored how conversational agents could act as facilitators to bolster women's participation in online debates in culturally conservative settings.

Despite the popularity of the D-agree platform and the use of IBIS for guiding the behavior of AI agents, some researchers consider the use of IBIS problematic in facilitating consensus achievement. Dong [31] cites two critical issues. Regarding the discussion

environment, he argues that it is not inclusive, as various characteristics of participants are ignored, leading to insufficient discussions. Additionally, he contends that the facilitation messages generated by IBIS-based AI agents are not sufficiently natural, resulting in reduced participant engagement and a lack of flow in topic discussions. As an ideal alternative, Dong et al. [13] propose using LLMs for guiding AI agents, presenting a series of additional functionalities that these models offer. They strengthen their position by comparing the results of the two approaches in their publication, distinguishing LLMs (tested using GPT-3.5) as more effective for assisting the consensus process. In a related development, Nomura et al. [32] have historically integrated LLMs with the IBIS framework to enhance brainstorming interactions. By employing GPT-3.5-turbo, they assigned AI agents roles that facilitated more dynamic and inclusive discussions, addressing potential concerns about engagement and the natural flow of dialogue. This approach leveraged the advanced capabilities of LLMs to improve the effectiveness of consensus-building processes, demonstrating the benefits of merging structured facilitation methods with the adaptability of modern AI technologies.

Other researchers recognize the superiority of LLMs as tools for facilitating consensus in various ways. For example, Ding and Ito [3] developed the Self-Agreement framework, utilizing GPT-3 to autonomously generate and evaluate diverse opinions, facilitating consensus without human input. Using a different approach, Song et al. [33] demonstrated how multi-agent AI systems powered by LLMs can significantly influence social dynamics and steer opinions towards consensus by increasing social pressure among participants. Additionally, Tessler et al. [34] introduced the “Habermas Machine,” an AI designed to mediate discussions on divisive political topics. Their AI was trained to produce group statements that not only garnered broad agreement among UK groups discussing issues like Brexit and immigration but also did so more effectively than human mediators. The “Habermas Machine” helped create clearer, more logical, and informative statements, ensuring minority perspectives were not alienated, thus demonstrating another powerful application of LLMs in consensus-building within diverse groups.

Building on the capabilities of LLMs to facilitate consensus-building, Govers et al. [12] focused on the specific strategies that AI agents must use to mediate online debates effectively. Their research utilized the Thomas–Kilmann Conflict Mode Instrument (TKI) to implement conflict resolution strategies through LLMs. The study particularly emphasized high-cooperativeness strategies, which were most effective at depolarizing discussions and enhancing the perception of achievable consensus among participants. This approach demonstrated how tailored AI-driven strategies could significantly influence online interactions, guiding diverse groups toward agreement and more constructive dialogue.

2.2. Measuring the Effectiveness of AI in Consensus-Building

As noted in the previous section, various approaches illustrate how AI can be valuable and necessary for facilitating the consensus-building process. However, a subsequent step involves assessing the extent to which AI achieves this goal. In the literature, numerous metrics and methodologies have been proposed to measure AI’s impact on consensus formation, ranging from semantic similarity scores to perceptual and qualitative assessments. This diversity reflects the complexity of consensus-building, shaped by objective factors, such as agreement metrics, and subjective elements, such as participant satisfaction and perceived fairness.

One prominent line of research focuses on using semantic models to compare participant views and gauge overall agreement. Specifically, Ding and Ito [3] introduce a scoring function that evaluates candidate agreements generated by improved language models, relying on semantic similarity scores—such as cosine similarity computed with

BERT embeddings—to determine how well different viewpoints are unified. This logic is also applied by Pérez et al. [35], who draw on soft consensus measures and fuzzy logic to quantify partial agreements, using indices like Euclidean, Manhattan, and cosine distances, along with adaptive thresholds.

Another approach examines the broader social context influenced by AI. Song et al. [33] use a 6-point Likert scale to measure opinion shifts and polarization before and after AI-mediated discussions, shedding light on the social forces that foster or hinder consensus. Likewise, Hadfi and Ito [36] adopt “Augmented Democratic Deliberation,” using feedback loops in knowledge graphs to assess the depth and quality of group deliberations.

Additionally, exploring participants’ subjective experiences sheds further light on how AI can affect consensus outcomes. Govers et al. [12] introduce a Perceived Consensus Score that allows participants to indicate how realistic they believe potential agreements are, potentially revealing unspoken reservations or hidden divergences. This metric is complemented by qualitative insights from Pham et al. [5] and de Rooij et al. [37], who conduct interviews to capture richer details about participants’ perceptions of AI’s role in shaping group dynamics and decision-making.

Finally, integrating quantitative and qualitative data provides a comprehensive view of how AI can facilitate consensus. Kim et al. [38] and Sahab et al. [39] show that AI-facilitated platforms not only lead to measurable shifts in opinions but also improve perceptions of fairness and inclusivity. This dual impact underscores AI’s potential to enhance the structural efficiency of consensus-building—via semantic modeling and feedback loops—and the human experience of collaboration. AI interventions increase the likelihood of durable and broadly accepted agreements by fostering more productive, transparent, and balanced discussions.

3. A Layered Path to Opinion Convergence

Building on the relevant literature on using AI to facilitate consensus-building, particularly the potential of LLMs for this purpose, we developed a systematic approach to the process and the measurement of its effectiveness. Specifically, we designed a web application from the ground up to enable structured communication and message exchange among participants, as well as interactions with the LLM models under examination, ChatGPT 4.0, Mistral Large 2, and AI21 Jamba. These models were integrated into an organized discussion framework incorporating specific consensus-facilitating strategies, allowing each model to select appropriate actions based on conversation history and user feedback. The discussions were analyzed to evaluate the extent to which participants reached a consensus proposal aligned with their initial positions. The following section outlines the employed approach and its key components.

3.1. System Architecture and Workflow

The system architecture of the web application was designed to facilitate efficient interaction between participants and the LLMs, supporting real-time consensus-building processes. Figure 1 illustrates the consensus formation workflow, which begins with participants responding to a question posed during the discussion and expressing their opinions on the topic. The LLM synthesizes these opinions to formulate a proposal to achieve consensus among the participants.

Participants then indicate their acceptance or rejection of the consensus proposal. In cases of rejection, the system collects feedback from the participant(s) who disagreed, while those who accepted the proposal await the next iteration. The LLM then evaluates the accumulated conversation history and the newly provided feedback to autonomously determine which of the five predefined facilitation strategies to apply. This choice was not

governed by any fixed algorithm or manual intervention but emerged dynamically from the model’s real-time interpretation of the dialogue. Each chosen strategy guides the LLM through tailored prompts to generate a revised proposal. This iterative process continues until a consensus is reached or the discussion ends, with the LLM adapting its strategy in real time to the evolving dynamics of the dialogue.

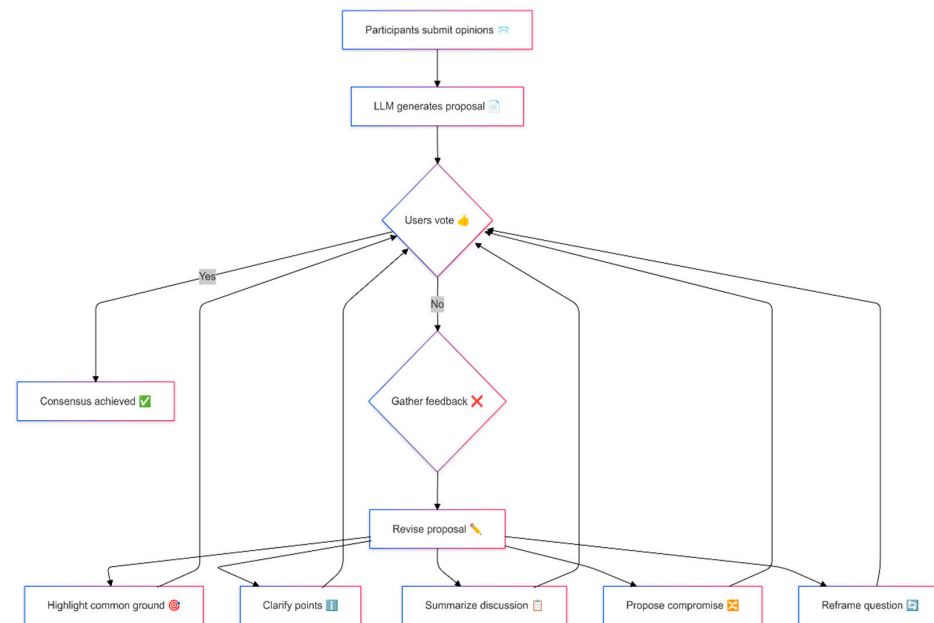


Figure 1. Consensus formation workflow.

The five strategies that the LLM could select from at each rejection point are summarized in Table 1. These strategies, derived from a literature review, aim to promote understanding, summarize key points, highlight agreements, propose compromises, and reframe questions when necessary. For instance, Bosch et al. [40] and Dworetzky et al. [41] emphasize the importance of the “ClarifyUnderstanding” strategy for eliminating ambiguities, while Chosokabe et al. [42] highlight the “SummarizeDiscussion” strategy for condensing complex conversations into their essential elements. Similarly, Nieto-Romero et al. [43] stress the value of the “HighlightCommonGround” approach, which effectively identifies shared agreements among diverse perspectives. Complementing these strategies is the “ProposeCompromise” method, emphasized by Opricovic [44] and Basuki [45], which seeks to offer balanced solutions that accommodate differing viewpoints. Additionally, Kettenburg et al. [46] advocate for the “ReframeQuestion” strategy, which involves skillfully rephrasing or redirecting research questions to foster new insights and facilitate consensus.

Table 1. Available strategies for selection by the LLM to achieve overall consensus.

Strategy	Prompt
ClarifyUnderstanding	Provide additional explanations, definitions, or examples to eliminate misunderstandings or ambiguities related to the research question or discussion points.
SummarizeDiscussion	Provide a concise summary of the discussion, highlighting key points of agreement and disagreement.
HighlightCommonGround	Identify and emphasize points of agreement among participants.
ProposeCompromise	Suggest potential compromises or middle-ground solutions.
ReframeQuestion	Rephrase or adjust the focus of the research question to make it more agreeable or clearer.

To support this architecture, we utilized several technologies, including HTML, JavaScript, PHP, CSS, and AJAX, to ensure the application’s user-friendliness and compati-

bility across devices. An open-source server environment built using Node.js facilitated real-time interaction with the LLMs. Critical for sustained discussions, conversation history was stored in a specially configured MySQL database. Finally, API technology was employed to integrate the LLMs into the platform, with necessary code modifications to align with the previously described workflow in Figure 1.

3.2. The Characteristics of the Evaluated LLMs

The three LLMs selected in this research are distinguished by their popularity and promising results in addressing complex tasks. Developed by three companies, each model possesses unique features while sharing common attributes.

3.2.1. The ChatGPT Model

ChatGPT, developed by OpenAI, has rapidly gained popularity due to its powerful capabilities [47]. Version 4.0, released in March 2023, has been extensively tested by the research community across various tasks, including its effectiveness in facilitating consensus among participants [33].

This success is partly due to its architecture, which follows the Transformer model ([48]; see Figure 2), an approach that excels at processing input sequences and generating outputs by predicting the next word in a sentence based on context [49]; unlike previous recurrent neural network models, which process text sequentially from left to right, the Transformer can simultaneously consider all words in a sentence, even those at the ends of the input. It achieves this by employing multiple self-attention mechanisms, which calculate a weighted sum of the input sequence. These weights are determined based on the similarity between each input token and the rest of the sequence [50].

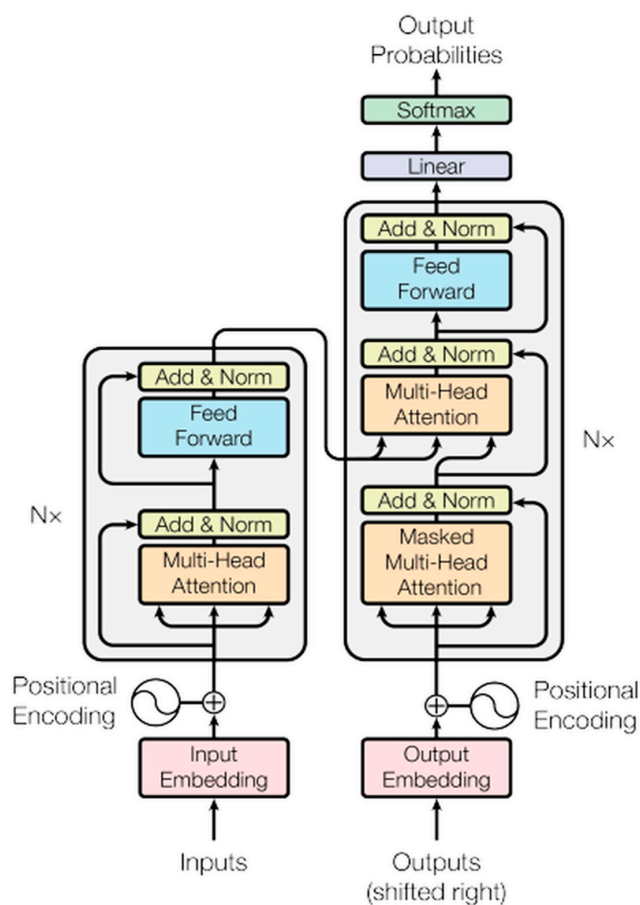


Figure 2. The Transformer—model architecture (Ashish [48]).

In addition, ChatGPT introduced some unprecedented advancements. Specifically, a unique training approach, based on unsupervised learning, was employed to process vast amounts of internet data autonomously. This approach involved incorporating approximately 1.76 trillion parameters [51] to capture intricate patterns and connections in textual information. Furthermore, reinforcement learning with human feedback was applied in a subsequent step to enhance the model's responses, refining its ability to generate contextually appropriate and accurate outputs [52].

3.2.2. The Mistral Large 2 Model

Mistral Large 2 is an LLM developed by Mistral AI and released in July 2024, significantly later than ChatGPT [53]. Unlike the previously mentioned LLM, this model was trained predominantly on French texts rather than English and utilizes a smaller number of parameters, approximately 123 billion [54]. Despite this difference, performance measurements indicate that it is an equally reliable choice, offering competitive results in general knowledge and reasoning tasks [55]. Furthermore, Mistral Large 2 supports dozens of foreign languages and more than 80 programming languages (e.g., Python, Java, C, C++, JavaScript, and Bash), making it highly capable in multilingual text processing and code generation [56,57].

Mistral Large 2 also builds upon the traditional Transformer architecture [55], introducing several improvements aimed at reducing computational complexity without significantly compromising the quality of its outputs. One notable enhancement is the Sliding Window Attention mechanism, which limits attention to a specific window of tokens. Additionally, the model incorporates mechanisms such as the Key-Value Cache with Rolling Buffer, enabling the storage and reuse of precomputed values during text generation. This reduces redundant calculations and improves the model's speed and efficiency [58]. Furthermore, as an open-source LLM, Mistral Large 2 allows the research community to make modifications, fostering further advancements and optimizations [53].

3.2.3. The AI21 Jamba-Instruct Model

The third and final model evaluated in this study is AI21's Jamba-Instruct, which was released in early May 2024 [59]. This particular LLM has been trained on a relatively minor number of parameters, approximately 52 billion [60]. Despite this, it achieves remarkable results across several performance indicators, making it a standout choice among open-source LLMs [61].

The architecture of the Jamba model series by AI21 Labs is particularly noteworthy. It employs an innovative hybrid approach that integrates Mamba layers beyond the traditional Transformer architecture. These layers enhance the processing of long-distance relationships in significant texts while reducing memory requirements. Additionally, the inclusion of Mixture of Experts (MoE) enables the use of multiple experts per layer, significantly increasing the model's capacity without a proportional increase in computational cost. It can handle content lengths of up to 256,000 tokens, with a KV cache memory footprint of only 4 GB—substantially lower than the 32 GB required by comparable LLMs. This efficiency enables Jamba to achieve top-tier performance in benchmarks such as HellaSwag and MMLU, all while being up to three times more efficient than other models of similar size [60,61].

3.3. Calculating Text Similarity as a Surrogate for Consensus

The cosine similarity index is a common and widely used method for comparing text similarity [62]. This index is efficient in computation [63] and its results are easy to interpret [64]. It quantifies the cosine of the angle between two non-zero vectors in a

multidimensional space [65]. Specifically, its calculation formula is defined by the following equation [66]:

$$\text{Cosine Similarity} : \frac{A * B}{\|A\| \|B\|} \quad (1)$$

where

- A and B represent the user opinion and the consensus sentence provided by LLM, respectively, as the two vectors;
- $A * B$ represents the dot product of the two vectors;
- $\|A\|$ and $\|B\|$ represent the magnitudes (or norms) of the vectors A and B.

The resolution of this relationship involves four distinct steps. First, the texts under consideration are converted into vector representations, followed by calculating the dot product between them. Next, the norms of these vectors are computed. Finally, the cosine similarity is determined by dividing the dot product by the product of the magnitudes of the two vectors.

In this study, this process was applied to analyze the cosine similarity between the initial opinions of the participants and the final proposal that was ultimately accepted in an effort to achieve consensus. It was also used to assess the differences between intermediate states, namely, participants' opinions regarding consensus proposals that were not accepted. The study required participants to provide feedback on the reasons for rejecting these proposals. This analysis helped to determine the extent to which participants maintained their original positions throughout the discussion or modified their stances by accepting proposals that did not align with their initial positions.

In the first step of this methodology, the examined texts were converted into vector representations using the Universal Sentence Encoder (USE) available from Tensor Flow Hub. This model was chosen over traditional approaches such as Term Frequency–Inverse Document Frequency (TF-IDF) and basic word embeddings, as it captures the semantic meaning of entire sentences and paragraphs more effectively [67]. Although more recent alternatives like Sentence-BERT (SBERT) offer state-of-the-art performance in semantic similarity tasks, USE was selected because it provided an optimal balance between accuracy, computational efficiency, and multilingual robustness. In particular, its proven performance across languages, including Greek, made it especially suitable for the present study, which required real-time processing of participant input during ongoing discussions. Furthermore, USE's lightweight integration with our development environment ensured consistent and fast embedding generation under the iterative consensus-building workflow. These features rendered USE not a matter of convenience but a methodologically appropriate and reliable choice for computing similarity scores across participant data. The entire vectorization process was implemented in Python using the Google Colab platform, which provided an accessible and efficient environment for embedding generation and data processing.

Next, we proceeded to calculate the dot product (numerator of Equation (1)), which concerns the sum of the vector products between each user's opinions and the consensus proposals from the LLMs. More specifically, we have Equation (2) as follows:

$$A * B = \sum_{i=0}^n A_i * B_i \quad (2)$$

where

- A_i and B_i are the elements of vectors A and B, respectively;
- n is the dimension of the vectors.

Then, we calculated the magnitudes (norms) of the vectors by taking the square root of the sum of the squares of their elements. These magnitudes represent the length of each vector in the multidimensional space. In this context, the following holds:

$$\|A\| = \sqrt{\sum_{i=1}^n A_i^2} \text{ and } \|B\| = \sqrt{\sum_{i=1}^n B_i^2} \quad (3)$$

Finally, we calculated the cosine similarity by dividing the dot product, as detailed in relation (2), by the product of the magnitudes of the two vectors from relation (3). This index ranges between -1 and 1 , where:

- A cosine similarity of 1 indicates that the vectors are identical, meaning there is complete agreement between the user's proposal and the consensus proposal;
- A cosine similarity of 0 indicates that the vectors are orthogonal, suggesting no alignment between the user's proposal and the consensus proposal generated by the LLM;
- A cosine similarity of -1 indicates that the vectors are diametrically opposed, meaning the user's proposal and the consensus proposal from the LLM are completely opposite.

It should be noted that cosine similarity operationalizes consensus regarding semantic proximity to initial views. While this allows for a precise and reproducible comparison, it may underestimate cases where participants converge on a novel compromise distant from their original positions.

4. Experimental Setup

This study aimed to evaluate the AI-assisted application by focusing on a group with diverse views on sustainability, making students the ideal choice for participation [68,69]. In August 2024, the Hellenic Open University launched a call for volunteers, targeting Informatics students from all years through its official website. The announcement provided a straightforward application process with clear instructions, a consent form, and a video guide to ensure participants could easily navigate the application. To uphold ethical standards and protect privacy, the University's Ethics and Research Committee approved the study, requiring participants only to create a personal username and password, with anonymity fully safeguarded throughout.

Additionally, implementing the sessions between participants and using the three LLMs via API required resources, which were provided by both OpenAI and Amazon. Specifically, OpenAI offered free API access through the "OpenAI Researcher Access Program". At the same time, the other two models were made available through Amazon Bedrock at no cost as part of the Open Cloud for Research Environments (OCRE) initiative.

In total, 30 students expressed interest in participating in the research following the university's invitation. Using the specially designed management platform, these students organized sessions among themselves, sending invitations, selecting their questions, and scheduling the day and time for each session. As a result, between 24 August and 30 September 2024, a total of 75 sessions were conducted, with each student participating in at least one session. The sessions were evenly distributed across the three LLM models under study, with the platform occasionally adjusting the question availability to ensure a balanced evaluation of each model.

Six questions were selected for discussion, covering several SDGs, such as good health and well-being (Goal 3) and quality education (Goal 4). The set of questions used is listed in Table 2 below.

Table 2. Questions for discussion and consensus-building.

#	Question
1	Should patients with a healthy lifestyle be prioritized in healthcare provision compared to those who choose a lifestyle that increases the risk of serious conditions?
2	Should international collaborations be strengthened to address pandemics like COVID-19, or would it be better to focus on national strategies to protect public health?
3	Should priority be given to the integration of digital technologies in education to better prepare students for the modern age, or should we focus more on strengthening basic skills in literacy and numeracy?
4	Should educational policy place greater emphasis on personalized learning and support for students with different learning needs, or is it more important to maintain uniform standards and approaches for all students?
5	Should global cooperation be strengthened to improve access to clean drinking water and sanitation services, or is it better to tailor policies to the specific needs of each region?
6	Should economically disadvantaged countries adopt stricter policies to reduce carbon dioxide emissions, or should greater emphasis be placed on adaptation and the resilience of societies to climate change?

5. Results and Analysis

This section presents various analyses using cosine similarity calculations to compare the three LLMs in terms of their ability to facilitate users' consensus during a controversial discussion. This study was performed at two levels: first, by examining the distance between the participants' initial and final positions and, second, through a more detailed graphical evaluation of the intermediate stages that evolved, ultimately leading to an acceptable consensus proposal.

5.1. Evaluation of the Distance Between Users' Initial and Final Positions

Regarding the first part of the results, Tables 3 and 4 are presented. Table 3 initially analyzes the degree of similarity between the users' initial and final positions at the model level, using the average cosine similarity as a metric. As shown, ChatGPT-4 achieved the highest mean cosine similarity ($M = 0.7015$, $SD = 0.1480$, 95% CI [0.6605, 0.7425]), followed by A21-Lab ($M = 0.6127$, $SD = 0.1485$, 95% CI [0.5628, 0.6626]) and Mistral Large 2 ($M = 0.5807$, $SD = 0.1807$, 95% CI [0.5247, 0.6367]). For the total of 124 calculated values, the overall average cosine similarity index was 0.6381.

Table 3. Average cosine similarity per examined LLM.

LLM	No. of Occasions	Average Cosine Similarity	Std. Dev.	95% CI (Lower—Upper)
ChatGPT 4.0	50	0.701	0.1480	[0.6605, 0.7425]
Mistral Large 2	40	0.5807	0.1807	[0.5247, 0.6367]
AI21 Jamba	34	0.6126	0.1485	[0.5628, 0.6626]
Total	124	0.6381		

A one-way ANOVA revealed a statistically significant difference among the three models ($F(2, 121) = 6.98$, $p = 0.00136$), indicating that the type of LLM significantly affected the degree of semantic alignment between user positions and the generated consensus proposals.

Table 4. Average cosine similarity per sustainable task and LLM.

Sustainable Task	LLM	No. of Occasions	Avg. Cosine Similarity	Std. Dev.	95% CI (Lower—Upper)
Good Health and Well-Being	ChatGPT 4.0	22	0.70499	0.1558	[0.6399, 0.7701]
	Mistral Large 2	16	0.5962	0.0987	[0.5479, 0.6446]
	AI21 Jamba	14	0.6574	0.1531	[0.5773, 0.7376]
	Total	52	0.658	0.1449	[0.6194, 0.6981]
Climate action	ChatGPT 4.0	2	0.8488	0.0021	[0.8459, 0.8517]
	Mistral Large 2	2	0.5574	0.1365	[0.3682, 0.7466]
	AI21 Jamba	4	0.4966	0.0399	[0.4575, 0.5357]
	Total	8	0.599	0.1663	[0.4846, 0.7151]
Quality Education	ChatGPT 4.0	20	0.7069	0.1071	[0.66, 0.7539]
	Mistral Large 2	16	0.5922	0.2482	[0.4706, 0.7138]
	AI21 Jamba	12	0.6356	0.1457	[0.5532, 0.718]
	Total	48	0.650	0.1784	[0.6004, 0.7013]
Clean water and sanitation	ChatGPT 4.0	6	0.6212	0.2287	[0.4381, 0.8042]
	Mistral Large 2	6	0.5167	0.1751	[0.3766, 0.6568]
	AI21 Jamba	4	0.5033	0.1272	[0.3788, 0.628]
	Total	16	0.552	0.1842	[0.4623, 0.6428]

Next, Table 4 examines the extent to which the LLMs facilitated consensus, this time considering the discussion topics, which were drawn from four of the 17 SDGs. As shown, higher average cosine similarity indices were observed in discussions concerning good health and well-being and quality education, both of which were supported by a larger number of cases in this study. Specifically, for the two questions related to SDG 3 (good health and well-being)—“Should patients with a healthy lifestyle be prioritized in health-care provision compared to those who choose a lifestyle that increases the risk of serious conditions?” and “Should international collaborations be strengthened to address pandemics like COVID-19, or would it be better to focus on national strategies to protect public health?”—the overall average cosine similarity was 0.658. Similarly, for the topics associated with SDG 4 and especially quality education—“Should priority be given to the integration of digital technologies in education to better prepare students for the modern age, or should we focus more on strengthening basic skills in literacy and numeracy?” and “Should educational policy place greater emphasis on personalized learning and support for students with different learning needs, or is it more important to maintain uniform standards and approaches for all students?”—the resulting average cosine similarity index reached 0.650. In contrast, a lower average cosine similarity score was observed for discussions on climate action (0.599), while the lowest average, 0.552, was recorded for the question addressing the global issue of clean water and sanitation. These descriptive results suggest that the alignment between participant views and consensus proposals may vary by topic. However, differences should be interpreted with caution due to differences in sample size across domains.

At the LLM level, ChatGPT-4 showed higher average cosine similarity scores across all four discussion topics included in this study (see Table 4). For example, ChatGPT-4 recorded a mean cosine similarity of 0.8487 in the context of climate action, compared to 0.5570 for Mistral Large 2 and 0.4966 for A21-Lab. In discussions related to good health and well-being, ChatGPT-4 achieved a similarity of 0.7050, while A21-Lab and Mistral Large 2 scored 0.6570 and 0.5962, respectively. The same pattern was observed in quality education, with ChatGPT-4 reaching 0.7069, compared to 0.6355 for A21-Lab and 0.5921

for Mistral Large 2. Even in the topic of clean water and sanitation, where scores were generally lower, ChatGPT-4 had a higher similarity (0.6211) than A21-Lab (0.5033) and Mistral Large 2 (0.5166).

While these results suggest a consistent trend in higher alignment scores for ChatGPT-4—meaning that its proposals remained more closely aligned with participants’ initial positions—these observations should be interpreted descriptively rather than as evidence of broader facilitation success.

5.2. Assessment of the Intermediate State Until Users Reach Consensus

Figures 3–5 are presented below to help us better understand the consistency of the examined LLMs in generating proposals that align with the participants’ feedback and the conversation history throughout the interaction until consensus is reached.

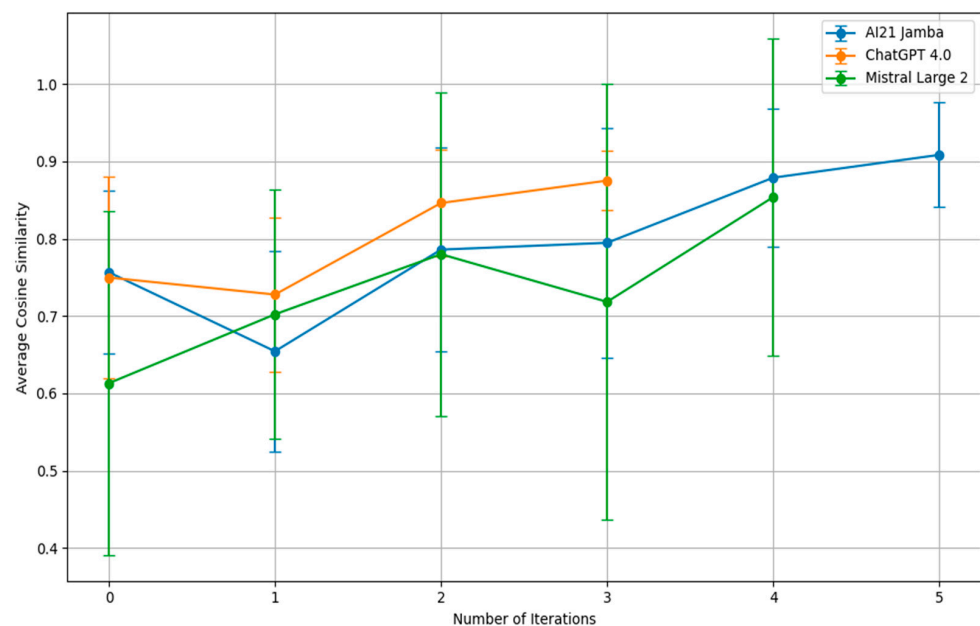


Figure 3. Average cosine similarity across iterations.

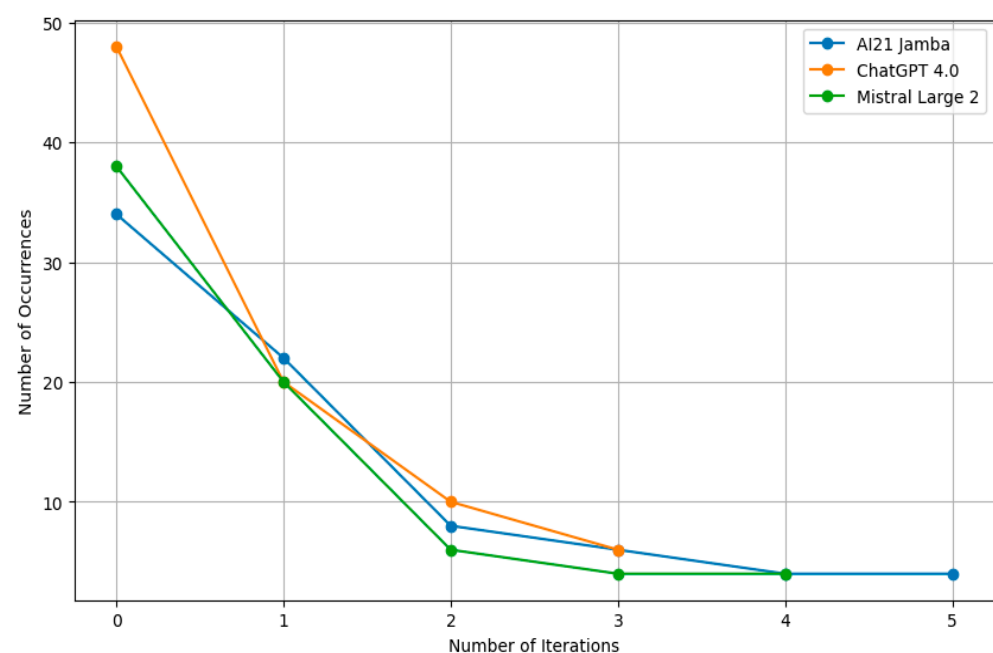


Figure 4. Number of cases per iteration for achieving consensus.

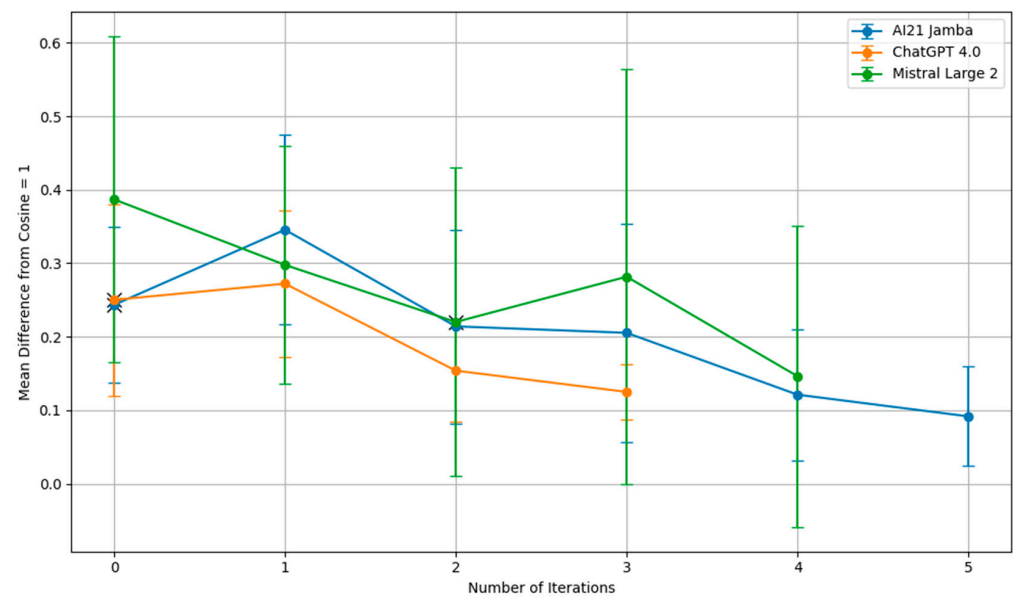


Figure 5. Mean difference in cosine similarity per iteration.

In Figure 3, the variation in the average cosine similarity per LLM is observed as the discussion progresses, particularly in cases where no acceptable consensus proposal exists among the participants. Even in the initial phase of interaction between the AI and users, ChatGPT 4.0 demonstrates a cosine similarity index close to 0.75. This value represents the similarity between the consensus proposal it generates and the initial opinions of the participants. The similarity increases significantly after the second and third iterations, as the model applies strategies to assist the discussion. Notably, in all discussions involving this LLM, no more than three iterations were required for the sample to reach a consensus. Similarly, AI21 Jamba achieves a high initial cosine similarity of 0.75. However, its behavior diverges as the process continues. According to Figure 3, its index drops significantly to 0.65 after the first round of user feedback but subsequently improves through better integration of user input. Finally, Mistral Large 2 starts with a lower initial cosine similarity index, slightly above 0.60, but shows improvement across iterations, ultimately reaching a peak of 0.85 after the fourth round. Standard deviation error bars have been included to illustrate the variability of similarity scores at each iteration step, offering a clearer view of performance consistency.

Next, Figure 4 illustrates the distribution of cases across different phases of the consensus-building process. ChatGPT 4.0 demonstrates faster convergence toward consensus compared to the other examined LLMs. Specifically, it achieves consensus more frequently within the first round of proposal generation, even without user feedback, reaching 50 cases. This surpasses the number of immediate unanimous cases observed with Mistral Large 2 and AI21 Jamba, which remain below 40. When user feedback is incorporated, the differences among the three LLMs diminish over the first three iterations. Mistral Large 2 and AI21 Jamba require additional iterations to reach consensus in some cases (up to the fourth and fifth rounds), whereas ChatGPT 4.0 achieves consensus earlier. As this figure displays frequency counts, error bars are not applicable.

Lastly, aiming to understand how the similarity between states evolved throughout the discussions, an additional analysis focused on identifying the elbow point [70] for each LLM by examining the average difference in cosine similarity from a value of 1 (entirely identical) across iterations.

Figure 5 shows no clear elbow point was detected for any model. ChatGPT 4.0 showed a steady decline in cosine difference over time, while Mistral Large 2 and AI21 Jamba

exhibited fluctuations—indicating that more complex discussions may require additional iterations for convergence. To support interpretability, standard deviation error bars have also been included in this figure, providing insight into the variability of semantic drift across the interaction timeline.

6. Discussion

This study evaluated the potential of large language models (LLMs) to facilitate consensus during contentious discussions by developing a novel, real-time conversation platform. The platform employed cosine similarity as a reliable metric to assess the alignment between participants' views and proposed consensus solutions. Unlike previous studies that utilized AI tools with fixed facilitation functions, this research introduced a key innovation, dynamic and adaptive strategy selection by the AI itself, based on the discussion history. The performance of three popular LLMs—ChatGPT 4.0, Mistral Large 2, and AI21 Jamba—was compared in contemporary discussions on sustainability issues.

The analysis of all conducted conversations focused on two key areas: (1) the degree of differentiation between participants' initial positions and the consensus proposals ultimately accepted and (2) the effort required—measured by the number of iterative rounds—to reach agreement. Regarding the first axis, notable differences were observed in how each model formulated consensus proposals. ChatGPT 4.0 tended to generate proposals more closely aligned with participants' initial positions, as reflected in higher cosine similarity scores. In contrast, AI21 Jamba and Mistral Large 2 produced proposals that deviated more from participants' original views, which resulted in lower similarity scores under our operationalization. It should be emphasized that such deviations do not necessarily indicate weaker facilitation quality, as effective consensus-building may also involve movement away from initial views toward a shared compromise. This trend was consistent across all discussions in the pilot study. Specifically, ChatGPT 4.0 showed higher alignment in climate-change-related discussions, whereas AI21 Jamba adopted the most divergent approach in formulating consensus proposals on the same topic. As for the second research issue, ChatGPT 4.0 demonstrated a higher percentage of discussions that reached agreement more quickly, often requiring only one or two iterations of the scenario application process. In contrast, Mistral Large 2 and AI21 Jamba generally required more iterations—typically four to five—before producing proposals acceptable to all participants. This translated into increased time and effort under our study design.

Despite the promising results of this study, several limitations must be acknowledged, as they may have influenced the findings and affected the generalizability of the results. First, all participants were Computer Science students, a homogeneous group with high digital literacy and familiarity with structured reasoning and technological tools. While this choice reduced variability for pilot study, it may also have made participants unusually adept at interacting with LLMs, potentially inflating their apparent performance and limiting the generalizability of the findings. Second, the discussions took place in dyads, simplifying the consensus-building process by reducing the complexity of negotiation dynamics. In larger groups, where a wider range of perspectives and potential conflicts must be addressed, LLMs may require different facilitation strategies and more iterations to achieve consensus. Third, all conversations were conducted in Greek, which is not a primary training or instruction language for the evaluated LLMs. This linguistic mismatch may have impacted the models' ability to interpret nuance, produce syntactically correct responses, and engage effectively with participant input—particularly for models less optimized for multilingual tasks. Fourth, while cosine similarity is a widely used and effective metric for assessing textual alignment, it primarily measures the semantic proximity between the final consensus proposal and participants' initial views. This operationalization may

underestimate successful facilitation in cases where participants reach consensus by moving away from their initial positions toward a novel compromise, as such outcomes would register lower similarity scores. In addition, cosine similarity mainly captures surface-level or lexical overlap and may overlook deeper aspects of consensus, such as shared reasoning, emotional alignment, perceived fairness, or mutual understanding. It also does not account for participants' satisfaction with the facilitation process or the perceived legitimacy of the outcome. Given that consensus involves both semantic and interpersonal dimensions, reliance on this metric alone may present an incomplete picture of facilitation effectiveness. Fifth, the study did not include a direct comparison with human facilitators, which could have offered valuable insights into the respective strengths and limitations of AI-mediated versus human-guided consensus building. However, it is essential to note that such a comparison was beyond the scope of this pilot study, as the platform was specifically designed for real-time facilitation using LLMs. Incorporating human mediators would have required a fundamentally different technical and procedural setup—including adjustments to interaction dynamics, facilitator training, and participant anonymity protocols—with any small-scale or isolated comparison under such divergent conditions potentially introducing biases that would undermine the interpretability and generalizability of the results. Finally, this study's thematic focus on sustainability—though timely and significant—may limit the transferability of findings to other domains or organizational contexts, where the nature and dynamics of discussion may differ substantially.

Given these limitations, future research could expand on the present study by involving larger groups of participants in the discussions. Engaging more individuals would introduce complex group dynamics and surface real-world challenges such as power imbalances, dominance of certain voices, or resistance to consensus. In such contexts, exploring additional facilitation strategies—such as voting mechanisms or structured debate formats—may become necessary to understand better how different interaction practices influence the consensus-building process. Moreover, incorporating participants with diverse demographic and cultural backgrounds could provide valuable insights into how these variables shape engagement with AI facilitators. Conducting discussions in the primary training languages of the LLMs would also help clarify the impact of language on facilitation quality, model comprehension, and overall user experience. Finally, expanding the thematic scope of the discussion scenarios beyond sustainability—into areas such as ethics, healthcare, or public policy—could shed light on the adaptability of LLMs in contexts marked by substantial social, moral, or ideological differences. Such extensions would be instrumental in assessing the broader applicability of AI-facilitated consensus-building in diverse real-world settings.

From a methodological perspective, an auspicious direction for future research involves conducting a comparative study between AI-driven and human-led facilitation, using a matched evaluation framework. Such a study would allow for direct assessment of differences in effectiveness, efficiency, and participant experience, offering a clearer understanding of each approach's unique contributions and limitations. Additionally, complementing cosine similarity with alternative computational metrics—such as Jaccard similarity—and incorporating qualitative indicators of agreement could offer a more nuanced understanding of conceptual convergence. This would help capture lexical overlap and the depth and substance of shared meaning among participants. To further enrich the analysis, participant-centered measures such as satisfaction ratings, perceptions of fairness, and post-discussion reflections could be integrated. These metrics would help assess whether the consensus reached was computationally valid and experienced as acceptable, equitable, and meaningful from the participants' perspectives.

Finally, an important avenue for future research involves the development of transparency and control mechanisms to safeguard ethical conduct among participants throughout the discussion process. Addressing this issue is essential for the design of AI facilitation systems that are not only effective but also inclusive, interpretable, and responsive to the ethical and social needs of diverse communities. Ensuring transparency in facilitation decisions and offering participants a degree of agency or oversight will be critical to fostering trust and legitimacy in AI-mediated deliberation.

7. Conclusions

The present study explored the use of LLMs in supporting consensus-building processes within structured online discussions. Among the three evaluated models, ChatGPT 4.0 achieved higher alignment with participants' initial views and required fewer iterations to reach agreement compared to Mistral Large 2 and AI21 Jamba based on cosine similarity measures. These findings suggest that LLMs can function as mediators in collaborative discussions and highlight important areas for further investigation, including the impact of language, cultural context, and participant diversity. Future research could build on these results by involving larger and more heterogeneous groups, testing additional facilitation strategies, and integrating qualitative assessments to capture the deeper dynamics of agreement formation. Such efforts may contribute to the continued refinement of AI-assisted decision-making frameworks in various domains.

Author Contributions: Conceptualization, L.T. and D.K.; methodology, L.T. and D.K.; software, L.T.; validation, L.T.; formal analysis, L.T.; investigation, L.T.; resources, L.T.; data curation, L.T.; original draft preparation, L.T.; writing—review and editing, D.K.; supervision, D.K.; project administration, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study is only contained in the article itself.

Acknowledgments: The authors would like to thank the students from the undergraduate IT department at the Hellenic Open University for their participation in the sessions conducted using the pilot version of the application.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hanson, M.P. *Clues to Achieving Consensus: A Leader's Guide to Navigating Collaborative Problem Solving*; Rowman & Littlefield Education: Lanham, MD, USA, 2005.
2. Neves, M.P. Consensus. In *Encyclopedia of Global Bioethics*; Springer: Cham, Switzerland, 2016; pp. 740–747.
3. Ding, S.; Ito, T. Self-agreement: A framework for fine-tuning language models to find agreement among diverse opinions. In *Pacific Rim International Conference on Artificial Intelligence*; Springer Nature Singapore: Singapore, 2023; pp. 298–309.
4. Foroughi, M.; de Andrade, B.; Roders, A.P.; Wang, T. Public participation and consensus-building in urban planning from the lens of heritage planning: A systematic literature review. *Cities* **2023**, *135*, 104235. [[CrossRef](#)]
5. Pham, T.V.; Weisswange, T.H.; Hassenzahl, M. Embodied Mediation in Group Ideation—A Gestural Robot Can Facilitate Consensus-Building. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, Copenhagen, Denmark, 1–5 July 2024; pp. 2611–2632.
6. Kahneman, D.; Slovic, P.; Tversky, A. (Eds.) *Judgment Under Uncertainty: Heuristics and Biases*; Cambridge University Press: Cambridge, UK, 1982; pp. 3–20.
7. Liu, W.; Xiao, S.; Browne, J.T.; Yang, M.; Dow, S.P. ConsensusUs: Supporting multi-criteria group decisions by visualizing points of disagreement. *ACM Trans. Soc. Comput.* **2018**, *1*, 1–26. [[CrossRef](#)]
8. Toma, C.; Butera, F. Hidden profiles and concealed information: Strategic information sharing and use in group decision making. *Personal. Soc. Psychol. Bull.* **2009**, *35*, 793–806. [[CrossRef](#)] [[PubMed](#)]
9. Straus, D. *How to Make Collaboration Work: Powerful Ways to Build Consensus, Solve Problems, and Make Decisions*; Berrett-Koehler Publishers: Oakland, CA, USA, 2002.

10. Pound, D. Designing and facilitating consensus-building—keys to success. In *Conflicts in Conservation: Navigating Towards Solutions*; Redpath, S.M., Gutierrez, R.J., Wood, K.A., Young, J.C., Eds.; Cambridge University Press: Cambridge, UK, 2015; pp. 240–256.
11. Cowell, S.; Begg, K.; Clift, R. Support for sustainable development policy decisions a case study from highway maintenance. *Int. J. Life Cycle Assess.* **2006**, *11*, 29–39. [\[CrossRef\]](#)
12. Govers, J.; Velloso, E.; Kostakos, V.; Goncalves, J. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2024; pp. 1–18.
13. Dong, Y.; Ding, S.; Ito, T. An Automated Multi-Phase Facilitation Agent Based on LLM. *IEICE Trans. Inf. Syst.* **2024**, *107*, 426–433. [\[CrossRef\]](#)
14. Filip, F.G.; Zamfirescu, C.B.; Ciurea, C. Collaborative activities and methods. In *Computer-Supported Collaborative Decision-Making*; Springer International Publishing: Cham, Switzerland, 2017; pp. 71–120.
15. Ito, T.; Hadfi, R.; Suzuki, S. An agent that facilitates crowd discussion: A crowd discussion support system based on an automated facilitation agent. *Group. Decis. Negot.* **2022**, *31*, 621–647. [\[CrossRef\]](#)
16. Zhou, L.; Farag, Y.; Vlachos, A. An LLM Feature-based Framework for Dialogue Constructiveness Assessment. *arXiv* **2024**, arXiv:2406.14760.
17. Burton, J.W.; Lopez-Lopez, E.; Hechtlinger, S.; Rahwan, Z.; Aeschbach, S.; Bakker, M.A.; Becker, J.A.; Berditchevskaia, A.; Berger, J.; Brinkmann, L.; et al. How large language models can reshape collective intelligence. *Nat. Hum. Behav.* **2024**, *8*, 1643–1655. [\[CrossRef\]](#)
18. Abdelnabi, S.; Goma, A.; Sivaprasad, S.; Schönherr, L.; Fritz, M. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 83548–83599.
19. Westermann, H.; Savelka, J.; Benyekhlef, K. Llmmediator: Gpt-4 assisted online dispute resolution. *arXiv* **2023**, arXiv:2307.16732. [\[CrossRef\]](#)
20. Xu, S.; Wu, Z.; Zhao, H.; Shu, P.; Liu, Z.; Liao, W.; Li, S.; Sikora, A.; Liu, T.; Li, X. Reasoning before comparison: LLM-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv* **2024**, arXiv:2402.11398. [\[CrossRef\]](#)
21. Zhang, L.; Jijo, K.; Setty, S.; Chung, E.; Javid, F.; Vidra, N.; Clifford, T. Enhancing large language model performance to answer questions and extract information more accurately. *arXiv* **2024**, arXiv:2402.01722. [\[CrossRef\]](#)
22. Chiclana, F.; García, J.T.; del Moral, M.J.; Herrera-Viedma, E. Analyzing consensus measures in group decision making. *Procedia Comput. Sci.* **2015**, *55*, 1000–1008. [\[CrossRef\]](#)
23. Bakhshi, F.; Ashtiani, M. An approach for reaching consensus in large-scale group decision-making focusing on dimension reduction. *Complex. Intell. Syst.* **2024**, *10*, 4223–4251. [\[CrossRef\]](#)
24. Ingram, S.; Munzner, T. Dimensionality reduction for documents with nearest neighbor queries. *Neurocomputing* **2015**, *150*, 557–569. [\[CrossRef\]](#)
25. Haqbeen, J.; Takayuki, I.; Hadifi, R.; Nishida, T.; Sahab, Z.; Sahab, S.; Roghmal, S.; Amiryar, R. Promoting discussion with AI-based facilitation: Urban dialogue with Kabul city. In *Proceedings of the 8th ACM Collective Intelligence, ACM Collective Intelligence Conference Series, Boston (Virtual Conference), South Padre Island, TX, USA, 18–19 June 2020*.
26. Haqbeen, J.; Sahab, S.; Ito, T. A Case Study on the Comparison of AI-facilitated Threaded Conversation versus Threaded Conversation. *IIAI Lett. Inform. Interdiscip. Res.* **2023**, *4*. [\[CrossRef\]](#)
27. Hadfi, R.; Okuhara, S.; Haqbeen, J.; Sahab, S.; Ohnuma, S.; Ito, T. Conversational agents enhance women’s contribution in online debates. *Sci. Rep.* **2023**, *13*, 14534. [\[CrossRef\]](#)
28. Sahab, S.; Haqbeen, J.; Ito, T. Conversational ai as a facilitator improves participant engagement and problem-solving in online discussion: Sharing evidence from five cities in afghanistan. *IEICE Trans. Inf. Syst.* **2024**, *107*, 434–442. [\[CrossRef\]](#)
29. Kunz, W.; Rittel, H.W.J. *Issues as Elements of Information Systems*; IURD Working Paper Series; Institute of Urban and Regional Development, University of California: Berkeley, CA, USA, 1 July 1970.
30. Sahab, S.; Haqbeen, J.A.; Ito, T. Comparative analysis of AI facilitator impact in online discussions: A cross-cultural study. In *INFORMATIK 2024*; Gesellschaft für Informatik eV: Berlin, Germany, 2024; pp. 715–725.
31. Dong, Y. The multi-agent system based on llm for online discussions. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 6 May–10 May 2024*; pp. 2731–2733.
32. Nomura, M.; Ito, T.; Ding, S. Towards Collaborative Brain-storming among Humans and AI Agents: An Implementation of the IBIS-based Brainstorming Support System with Multiple AI Agents. In *Proceedings of the ACM Collective Intelligence Conference*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–9.
33. Song, T.; Tan, Y.; Zhu, Z.; Feng, Y.; Lee, Y.C. Multi-Agents are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions. *arXiv* **2024**, arXiv:2411.04578.
34. Tessler, M.H.; Bakker, M.A.; Jarrett, D.; Sheahan, H.; Chadwick, M.J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D.C.; et al. AI can help humans find common ground in democratic deliberation. *Science* **2024**, *386*, eadq2852. [\[CrossRef\]](#) [\[PubMed\]](#)

35. Pérez, I.J.; Cabrerizo, F.J.; Alonso, S.; Dong, Y.C.; Chiclana, F.; Herrera-Viedma, E. On dynamic consensus processes in group decision making problems. *Inf. Sci.* **2018**, *459*, 20–35. [\[CrossRef\]](#)
36. Hadfi, R.; Ito, T. Augmented Democratic Deliberation: Can Conversational Agents Boost Deliberation in Social Media? In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, New Zealand, 9–13 May 2022; pp. 1794–1798.
37. de Rooij, A.; Broek, S.V.D.; Bouw, M.; de Wit, J. Co-Creating with a Robot Facilitator: Robot Expressions Cause Mood Contagion Enhancing Collaboration, Satisfaction, and Performance. *Int. J. Soc. Robot.* **2024**, *16*, 2133–2152. [\[CrossRef\]](#)
38. Kim, S.; Eun, J.; Seering, J.; Lee, J. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum. Comput. Interact.* **2021**, *5*, 1–26. [\[CrossRef\]](#)
39. Sahab, S.; Haqbeen, J.; Ito, T. Facilitating Collaborative Consensus Building in Web-based Discussion through Collective Task-Based Roles: A Case Study. *IIAI Lett. Inform. Interdiscip. Res.* **2023**, *4*. [\[CrossRef\]](#)
40. BOBosch, D.; Pease, J.; Wolfe, M.L.; Zobel, C.; Osorio, J.; Cobb, T.D.; Evanylo, G. Community DECISIONS: Stakeholder focused watershed planning. *J. Environ. Manag.* **2012**, *112*, 226–232. [\[CrossRef\]](#)
41. Dworetzky, B.; Paladino, M.J.; Cruze, E.; Fox, D.; Roy, S.; Bakewell, T.; Coleman, C. Family engagement in systems change: Use of a new assessment tool in quality improvement. *Pediatrics* **2024**, *153* (Suppl. 1), e2023063424D. [\[CrossRef\]](#)
42. Chosokabe, M.; Koie, S.; Oyamada, Y. Examining the Effect of ChatGPT on Small Group Ideation Discussions. In *International Conference on Group Decision and Negotiation*; Springer Nature: Cham, Switzerland, 2024; pp. 142–153.
43. Nieto-Romero, M.; Milcu, A.; Leventon, J.; Mikulcak, F.; Fischer, J. The role of scenarios in fostering collective action for sustainable development: Lessons from central Romania. *Land Use Policy* **2016**, *50*, 156–168. [\[CrossRef\]](#)
44. Opricovic, S. A compromise solution in water resources planning. *Water Resour. Manag.* **2009**, *23*, 1549–1561. [\[CrossRef\]](#)
45. Basuki, A. Sustainable strategies selection in SMEs using MCDM approach. In *MATEC Web of Conferences*; EDP Sciences: Les Ulis, France, 2016; p. 02007.
46. Kettenburg, A.J.; Hanspach, J.; Abson, D.J.; Fischer, J. From disagreements to dialogue: Unpacking the Golden Rice debate. *Sustain. Sci.* **2018**, *13*, 1469–1482. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Easin, A.M.; Sourav, S.; Tamás, O. An intelligent llm-powered personalized assistant for digital banking using langgraph and chain of thoughts. In Proceedings of the 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), Pula, Croatia, 19–21 September 2024; pp. 625–630.
48. Ashish, V. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
49. Hariri, W. Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv* **2023**, arXiv:2304.02017.
50. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [\[CrossRef\]](#)
51. Hao, X.; Demir, E.; Eysers, D. Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction. *Technol. Soc.* **2024**, *78*, 102662. [\[CrossRef\]](#)
52. Roumeliotis, K.I.; Tselikas, N.D. Chatgpt and open-ai models: A preliminary review. *Future Internet* **2023**, *15*, 192. [\[CrossRef\]](#)
53. Mistral AI. Mistral Large: Introducing Our Advanced Model. Available online: <https://mistral.ai/news/mistral-large-2407> (accessed on 7 January 2025).
54. Da Silva, V.T.; Rademaker, A.; Lioni, K.; Giro, R.; Lima, G.; Fiorini, S.; Archanjo, M.; Carvalho, B.W.; Neumann, R.; Souza, A.; et al. Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature. *arXiv* **2024**, arXiv:2411.03484. [\[CrossRef\]](#)
55. Pawitan, Y.; Holmes, C. Confidence in the reasoning of large language models. *arXiv* **2024**, arXiv:2412.15296. [\[CrossRef\]](#)
56. Kanepajs, A.; Ivanov, V.; Moulange, R. Towards safe multilingual frontier ai. *arXiv* **2024**, arXiv:2409.13708.
57. Qiu, Z.; Liu, W.; Feng, H.; Liu, Z.; Xiao, T.Z.; Collins, K.M.; Tenenbaum, J.B.; Weller, A.; Black, M.J.; Schölkopf, B. Can Large Language Models Understand Symbolic Graphics Programs? *arXiv* **2024**, arXiv:2408.08313. [\[CrossRef\]](#)
58. Jiang, F. Identifying and Mitigating Vulnerabilities in LLM-Integrated Applications. Master’s Thesis, University of Washington ProQuest Dissertations & Theses, Washington, DC, USA, 2024.
59. AI21 Labs. Announcing Jamba-Instruct. Available online: <https://www.ai21.com/blog/announcing-jamba-instruct> (accessed on 7 January 2025).
60. Team, J.; Lenz, B.; Arazi, A.; Bergman, A.; Manevich, A.; Peleg, B.; Aviram, B.; Almagor, C.; Fridman, C.; Padnos, D.; et al. Jamba-1.5: Hybrid transformer-mamba models at scale. *arXiv* **2024**, arXiv:2408.12570. [\[CrossRef\]](#)
61. Lieber, O.; Lenz, B.; Bata, H.; Cohen, G.; Osin, J.; Dalmedigos, I.; Safahi, E.; Meirom, S.; Belinkov, Y.; Shalev-Shwartz, S.; et al. Jamba: A hybrid transformer-mamba language model. *arXiv* **2024**, arXiv:2403.19887. [\[CrossRef\]](#)

62. Shahmirzadi, O.; Lugowski, A.; Younge, K. Text similarity in vector space models: A comparative study. In Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; IEEE: New York, NY, USA, 2019; pp. 659–666.
63. Huang, A. Similarity measures for text document clustering. In Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 14–18 April 2008; pp. 9–56.
64. Li, B.; Han, L. Distance weighted cosine similarity measure for text classification. In Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2013), Hefei, China, 20–23 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 611–618.
65. Kalhori, H.; Alamdari, M.M.; Ye, L. Automated algorithm for impact force identification using cosine similarity searching. *Measurement* **2018**, *122*, 648–657. [[CrossRef](#)]
66. Mishra, H.; Soundarajan, S. Balancedqr: A framework for balanced query recommendation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer Nature: Cham, Switzerland, 2023; pp. 420–435.
67. TensorFlow. Semantic Similarity with TF Hub Universal Encoder. Available online: https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder (accessed on 7 January 2025).
68. Sakamoto, M.; Yamaguchi, E.; Yamamoto, T.; Wakabayashi, K. An intervention study on students' decision-making towards consensus building on socio-scientific issues. *Int. J. Sci. Educ.* **2021**, *43*, 1965–1983. [[CrossRef](#)]
69. Levett-Jones, T.; Catling, C.; Cheer, S.; Fields, L.; Foster, A.; Maguire, J.; McIntyre, E.; Tracey Moroney, O.A.M.; Pich, J.; Pitt, V.; et al. Achieving consensus on the essential knowledge and skills needed by nursing students to promote planetary health and sustainable healthcare: A Delphi study. *J. Adv. Nurs.* **2024**, *00*, 1–19. [[CrossRef](#)] [[PubMed](#)]
70. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.