



POLITECNICO
MILANO 1863

**SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE**

EXECUTIVE SUMMARY OF THE THESIS

AI Moderation in Multi-Party Voice Conversations: Architecture, Strategies, and Evaluation

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA INFORMATICA

Author: SALVATORE PIO CONTE

Advisor: PROF. FRANCA GARZOTTO

Co-advisors: PIETRO AMMATURO, THOMAS SCHOEGJE

Academic year: 2025-2026

1. Introduction and Motivation

Recent advances in automatic speech recognition, speech synthesis, and large language models (LLMs) have made voice-based interaction with conversational systems increasingly practical. Most existing systems, however, are still designed around a dyadic interaction model, where a single user interacts with a single artificial assistant. Many real-world conversations are instead multi-party: meetings, study groups, collaborative tasks, and decision-making processes involve several participants who alternate, overlap, and negotiate the floor while addressing different members of the group.

Extending conversational AI from one-to-one interaction to group voice conversation is therefore not a simple scaling problem. A multi-party system must manage who is allowed to speak, attribute each contribution to the correct participant, avoid uncontrolled overlap, and define the role of the artificial agent within the group. In this setting, the agent cannot be treated only as a respondent or information provider. Its behaviour affects the structure of the conversation itself.

This thesis investigates the role of an AI moderator in real-time multi-party voice conversa-

tions. The moderator is conceived as a neutral, process-oriented agent: it does not suggest solutions or take part in the content of the group decision, but supports the way in which the discussion unfolds. Its role is to help the group remain focused, encourage balanced participation, provide recaps on request, and intervene when the conversation stalls or drifts away from the task.

To study this problem, the thesis presents *Alutami*, a real-time speech-based system for AI-moderated group discussions, and makes three contributions. First, it formulates the AI moderator as a process-oriented agent for multi-party voice conversations, distinct from an assistant that contributes to the content of the task. Second, it presents Alutami as a working prototype architecture combining explicit turn management, contribution attribution, real-time speech processing, and controlled AI moderation. Third, it provides a within-subjects evaluation over ten groups, comparing discussions with and without the moderator and measuring its effects on decision quality, session efficiency, participation balance, and user satisfaction.

2. AIutami: System and Moderation Strategy

AIutami supports small groups in a collaborative voice discussion and integrates a moderator that regulates the conversational process rather than solving the task. Its design addresses the specific challenges of group voice interaction: floor management, contribution attribution, overlap avoidance, and deciding when the agent should intervene without becoming an additional decision-maker.

Turn-taking and attribution. AIutami adopts an explicit turn-taking mechanism. At any moment, only one participant may hold the floor, and taking the floor is a deliberate action. A participant presses a push-to-talk button to start speaking and releases the floor when the contribution is complete. While one participant is speaking, the others can reserve the next turn and claim it through a short priority window once the floor becomes available.

This design choice makes contribution attribution part of the interaction protocol. Instead of allowing free overlapping speech and later reconstructing who said what through speaker recognition or diarisation, the system knows by design who is authorised to speak at each moment. Each transcription can therefore be linked to the current floor holder without relying on biometric speaker identification. The same turn state also prevents human speech from overlapping with the moderator’s voice: when the moderator is speaking, human turns are temporarily blocked until the intervention is complete.

The moderator as a process-oriented agent. The moderator is designed as a neutral, process-oriented agent. It does not express opinions on the substance of the discussion, does not suggest rankings or solutions, and does not participate in the group decision. Its role is instead to support how the discussion unfolds. It can help the group remain focused, encourage quieter participants to contribute, provide recaps on request, and intervene when the conversation stalls, drifts off topic, becomes unbalanced, or violates task-specific ground rules.

The moderator is occasional rather than continuous: it does not fill every silence, but acts only when the system detects a reason to, and frames its interventions as procedural support rather than content-level guidance.

Intervention logic. The moderation strategy combines deterministic triggers with LLM-based judgement (Figure 1). Deterministic triggers are derived from rules and timers over the session state. They include the opening message, prolonged collective silence, inactivity of a single participant, and the approach of the session time limit. Participant inactivity is handled progressively: the system first issues a private text nudge, and only later, if the participant remains silent, produces a voiced prompt by name.

LLM-derived triggers are based on the analysis of the discussion. After each human turn, the model receives the running summary of the conversation, the latest contribution, and the current participation metrics. It returns a structured judgement containing an updated summary, a reason for intervention, a severity score, and a proposed message. The possible reasons include monopolisation, exclusion, off-topic drift, conflict, explicit requests, and task-specific ground-rule violations.

A key control principle governs this process: the LLM proposes, but the backend decides. The model never controls the conversational floor directly. A proposed intervention becomes speech only if it passes deterministic backend filters, including a minimum severity threshold, per-reason cooldowns, a warm-up period before judging participation imbalance, and phase guards that restrict moderation to the active discussion. Conflict and explicit requests are treated as exceptions, since a conflict may require immediate de-escalation and a direct request from the group should be answered.

Participation imbalance is also handled through backend-computed metrics rather than left entirely to the model. A participant is treated as over-participating when their speaking time exceeds twice the group average, and as under-participating when it falls below half of it. Following inclusive moderation principles, the moderator does not reprimand dominant participants directly; it instead invites quieter members to contribute and anchors the invitation to a concrete point raised in the discussion [3]. This keeps the intervention supportive rather than punitive, and helps preserve the moderator’s neutral role within the group.

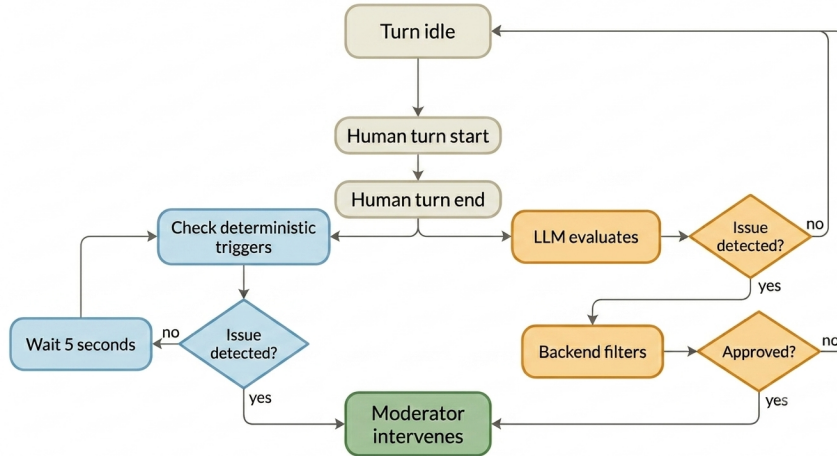


Figure 1: How the moderator decides to intervene. Deterministic triggers lead straight to an intervention, while the model’s proposals must clear the backend filters first; otherwise the floor returns to idle.

3. Architecture and Implementation

AIutami is a web-based real-time system: a frontend interface, an asynchronous backend, a shared runtime state, and external AI services for transcription, reasoning, and speech synthesis. The implementation enforces the design principles introduced above, keeping the conversational floor explicit, attributing each contribution to the correct participant, preventing overlaps between participants and the moderator, and letting the agent intervene only through a controlled backend pipeline. A task-agnostic core is separated from task-specific components, so the same moderation infrastructure serves different collaborative scenarios.

Communication and audio. The backend uses Django with Channels, which moves the application to the asynchronous ASGI model and adds WebSocket support. WebSockets carry control and session events (turn changes, updates, audio signalling), while audio travels separately over WebRTC, better suited to real-time browser voice. Through `aiortc` the server acts as a WebRTC peer: it receives the participants’ streams, forwards the current speaker’s audio to the others, and feeds the same stream into transcription. Since only one participant holds the floor at a time, a per-session audio hub forwards the floor holder’s frames without mixing. Transcription is gated by the floor state, so the recogniser listens only to the authorised speaker,

which keeps background voices out of the transcript and links each contribution to a known speaker.

Storage and runtime state. Persistent data (users, sessions, rankings, transcripts, reports) is stored in PostgreSQL, since it must survive the session and support the later analysis. Fast-changing state (current floor holder, reservations, moderation timers, transcript fragments) lives in Redis, which loses relevance once the session ends and also serves as the channel layer that publishes events to all participants of a session.

AI services. Three functions are delegated to OpenAI: reasoning with `gpt-4o-mini`, streaming transcription with `gpt-4o-mini-transcribe` (Realtime API, text returned while the participant is still speaking), and speech synthesis with `gpt-4o-mini-tts`. The mini models are an explicit latency/cost trade-off: transcription, turn-boundary evaluation, and synthesis must stay fast enough not to break the rhythm of the discussion.

Prompt construction. The moderator prompt is assembled modularly: a fixed core defines role, tone, intervention reasons, severity scale, and response format, while a task-specific block adds the scenario and ground rules. The model returns a structured output (updated summary, reason, severity score, message), with the message generated last so it is conditioned on the judgement already made. Task blocks include right/wrong examples, following a few-

shot prompting strategy [1], and critical role instructions are repeated to keep the moderator’s behaviour consistent across tasks.

4. Empirical Evaluation

The moderator was evaluated through a consensus-ranking task, a standard group decision-making paradigm in which participants receive an emergency scenario, rank fifteen items individually, and then discuss as a group to agree on a shared ranking. This paradigm was chosen because it produces sustained collaborative debate, naturally creates divergent opinions among participants, and provides an expert reference ranking against which the quality of the group decision can be measured. Two validated scenarios were used: NASA Moon Survival [2] and Lost at Sea [5]. Since each group performed two sessions, using two different tasks reduced the learning effect that would have been introduced by repeating the same scenario.

The study compared two conditions: with the AI moderator active (ON) and without the AI moderator (OFF). In both conditions, participants received the same task ground rules before the discussion; the difference was that in the ON condition the moderator monitored and enforced them during the conversation, while in the OFF condition they were only presented at the beginning. The study followed a within-subjects design: each group completed both conditions, so that every group served as its own control. Thirty adults took part, organised into ten groups of three, for a total of twenty sessions. Task order and condition order were counterbalanced across four configurations to reduce order effects. A preliminary pilot was conducted before the main study to validate the protocol and correct practical issues related to connectivity, repeated interventions, and user feedback during the moderation step.

The evaluation was organised around four research questions. RQ1 addressed effectiveness: whether the moderator influences the quality of the group decision, measured as the distance between the group ranking and the expert ranking. RQ2 addressed efficiency: whether the moderator changes the time needed by the group to reach consensus. RQ3 addressed group dynamics: whether the moderator influences the equity of participation among the three members of the

group. RQ4 addressed satisfaction: how participants perceive the system and whether they prefer the moderated or unmoderated condition. These dimensions follow the ISO 9241-11 axes of effectiveness, efficiency, and satisfaction [4], extended with a group-dynamics dimension specific to multi-party interaction.

For RQ1, the main metric was the *error_score*, computed as the sum of the absolute distances between the group ranking and the expert ranking across the fifteen items; lower values indicate better decisions. A secondary metric, *synergy_gain*, compared the group result with the participants’ individual rankings, measuring whether discussion improved the decision with respect to the initial individual answers. RQ2 was measured through discussion duration in minutes. RQ4 was measured through the System Usability Scale, five moderator-specific Likert items available only in the ON condition, comparative preference questions in the second session, and open-ended responses analysed qualitatively.

The central metric for RQ3 was the *gini_index* of speaking time. In AIutami, speaking time is operationalised through the turn-management logs: it corresponds to the time interval during which a participant holds the conversational floor, from taking the turn to releasing it. It is therefore a floor-holding measure, rather than a voice-activity-based estimate of pure acoustic speech production. The index was computed using the canonical Gini coefficient (Equation 1):

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |t_i - t_j|}{2n \sum_{i=1}^n t_i}, \quad (1)$$

where t_i is the total floor-holding time of the i -th participant and $n = 3$ is the number of participants in each group. The value ranges between 0 and 1: 0 corresponds to perfectly equal participation, while higher values indicate increasingly unbalanced participation. Since the design was paired, the analysis compared ON and OFF sessions within the same groups, using paired statistical tests and reporting effect sizes alongside significance values.

5. Results and Discussion

Table 1 summarises the main quantitative results across the four research questions. The results show a clear separation between decision

outcome and conversational process. The moderator did not produce a statistically detectable improvement in the quality of the group decision, but it substantially changed the way the discussion unfolded: moderated sessions lasted longer, participation became more balanced, and participants clearly preferred the moderated condition.

Table 1: Summary of results by research question, comparing sessions with the moderator active (ON) and without the moderator (OFF). Effect-size magnitude follows conventional thresholds.

Metric	ON	OFF	Effect
error_score (RQ1)	45.6	49.6	negligible
synergy_gain (RQ1)	7.5	9.4	negligible
Duration, min. (RQ2)	24.0	14.8	large
gini_index (RQ3)	0.128	0.263	large
SUS (RQ4)	83.3	80.5	small

For RQ1, the moderator did not produce a detectable effect on decision quality. The mean *error_score* was slightly lower in the ON condition than in the OFF condition (45.6 vs. 49.6), but the paired test showed no significant difference, $t(9) = -0.44$, $p = 0.67$, with a negligible effect size ($d = -0.14$). The same pattern emerged for *synergy_gain*, where the difference between ON and OFF was again negligible. This indicates that groups converged to rankings of similar quality with and without the moderator. The result should not be read as proof that moderation can never affect decision quality, but as no measurable effect in this study and with this sample size. An exploratory analysis by task suggests that the moderator’s effect on decision quality may depend on how competence is distributed in the group: balancing participation appears to help when relevant knowledge is spread across the members, so the discussion can aggregate it, and to hurt when a single member already holds the best answer, whose view is then diluted, or when none does. This pattern (a moderator $\times$ task interaction on *synergy_gain*, $p = .043$) rests on small, unbalanced per-task cells and remains exploratory.

For RQ2, the moderator clearly increased the time required to reach consensus. Mean discussion duration was 24.0 minutes in the ON condition and 14.8 minutes in the OFF condition.

Because an outlier was detected, the Wilcoxon signed-rank test was used; the result was significant, $W = 1.0$, $p = 0.004$, with a large effect size ($r = 0.91$). Nine groups out of ten took longer with the moderator, with a median increase of about ten minutes. This shows that the moderator changed the rhythm of the discussion and introduced a cost in terms of efficiency.

For RQ3, the same process change appears in a positive form. The mean *gini_index* of floor-holding time was 0.128 in the ON condition and 0.263 in the OFF condition. The Wilcoxon test returned $W = 5.0$, $p = 0.02$, with a large effect size ($r = -0.74$). Participation imbalance was therefore roughly halved when the moderator was active, and all groups except one moved in the expected direction (Figure 2). This is the strongest objective result of the study: the moderator achieved the main process-oriented goal for which it was designed, namely making participation more evenly distributed across group members.

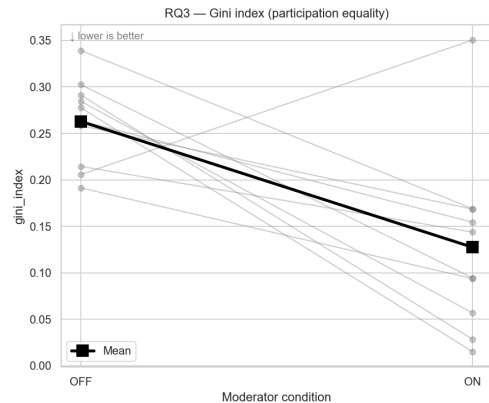


Figure 2: Paired plot of the Gini index of speaking time across the ten groups. Each segment connects the ON and OFF session of the same group.

For RQ4, the general usability of the system remained high in both conditions. The SUS score was 83.3 in ON and 80.5 in OFF, with no detectable difference between the two conditions. Both scores are above the usual threshold for excellent usability, suggesting that the moderator did not make the system harder to use. The moderator-specific Likert items were consistently positive, with a median score of 6 out of 7 across the evaluated dimensions. The comparative questions gave the clearest subjective result: after experiencing both conditions, 94%

of participants placed themselves above the neutral midpoint in preferring the moderated version, with a preference mean of 6.4 and a median of 7.

The qualitative side helps explain this preference. The 65 open-ended responses were examined through a thematic analysis, with two independent coders reaching total agreement on the final coding. The most recurrent positive theme was BALANCE: participants perceived the moderator as useful because it encouraged quieter members to speak and helped prevent the discussion from being dominated by a single participant. This qualitative result is consistent with the quantitative reduction of the Gini index. The main criticism was instead concentrated in the theme MISUNDERSTANDING: in some cases, the moderator classified as off-topic contributions that were actually related to the task. This points to a specific calibration issue in the classifier prompt, rather than to a failure of the moderation concept as a whole.

Taken together, the four research questions tell a coherent story. The AI moderator did not substantially change what the group decided, but it changed how the group decided and how participants experienced the discussion.

6. Conclusion and Future Work

This thesis investigated whether an artificial agent can act as a moderator in real-time multi-party voice conversations. The work addressed the problem from three complementary perspectives: it formulated the moderator as a neutral, process-oriented agent; it implemented AIutami as a working multi-party speech-based system with explicit floor management and controlled AI interventions; and it evaluated the moderator through a within-subjects study comparing discussions with and without moderation.

The empirical results show that the value of AI moderation in this setting lies less in changing what the group decides and more in shaping how the group discusses. The moderator did not produce a measurable improvement in decision quality, but it made participation significantly more balanced and was clearly preferred by participants. At the same time, moderated sessions lasted longer, revealing a trade-off between efficiency and inclusiveness. Within the limits of the study, AIutami should therefore be

read not as a definitive solution to multi-party AI moderation, but as a working prototype and an empirical step toward conversational systems that take part in group interaction as process-oriented moderators rather than as additional speakers.

Future work should address three main directions. First, interaction should move beyond push-to-talk toward more natural free voice conversation, combining predictive turn-taking with automatic speaker attribution and allowing participation to be measured as actual speaking time rather than floor-holding time. Second, the moderator should become more sensitive to the state of the group by incorporating audio-native or multimodal cues such as tone, pace, intensity, and possibly visual signals. Third, the system should be evaluated beyond survival-ranking tasks, including other decision-making settings, open-ended discussions, creative collaboration, and social contexts in which moderation may play a different role.

References

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language Models are Few-Shot Learners, 2020.
- [2] Jay Hall and W. H. Watson. The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations*, 23(4):299–317, 1970.
- [3] Mo Houtti, Moyan Zhou, Loren Terveen, and Stevie Chancellor. Observe, Ask, Intervene: Designing AI Agents for More Inclusive Meetings. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.
- [4] International Organization for Standardization. ISO 9241-11:2018, Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. <https://www.iso.org/standard/63500.html>, 2018.
- [5] Paul M. Nemiroff and William A. Pasmore. Lost at Sea: A Consensus-Seeking Task. In *The 1975 Annual Handbook for Group Facilitators*, pages 28–34. Pfeiffer, J. William, 1975.