MEETING DELEGATE: Benchmarking LLMs on Attending Meetings on Our Behalf

Anonymous ACL submission

Abstract

In contemporary workplaces, meetings are essential for exchanging ideas and ensuring team alignment but often face challenges such as time consumption, scheduling conflicts, and inefficient participation. Recent advancements in Large Language Models (LLMs) have demonstrated their strong capabilities in natural language generation and reasoning, prompting the question: *can LLMs effectively delegate participants in meetings?* To explore this, we develop a prototype LLM-powered meeting delegate system and create a comprehensive benchmark using real meeting transcripts. Our evaluation reveals that GPT-4/4o maintain balanced performance between active and cautious engagement strategies. In contrast, Gemini 1.5 Pro tends to be more cautious, while Gemini 1.5 Flash and Llama3-8B/70B display more active tendencies. Overall, about 60% of responses address at least one key point from the ground-truth. However, improvements are needed to reduce irrelevant or repetitive content and enhance tolerance for transcription errors commonly found in real-world settings. Additionally, we implement the system in practical settings and collect real-world feedback from demos. Our findings underscore the potential and challenges of utilizing LLMs as meeting delegates, offering valuable insights into their practical application for alleviating the burden of meetings.

1 Introduction

Nowadays, the nature of work has increasingly become more collaborative (Mugayar-Baldocchi et al., 2021), with meetings becoming an essential component (Spataro, 2020) to facilitate the exchange of ideas and information, fostering innovation and ensuring alignment among team members.

Attending meetings, however, poses notable difficulties. Firstly, the rapid increase in the number of meetings can consume a substantial amount of time, diverting attention from core tasks and reducing overall productivity (Perlow et al., 2017; Kost, 2020). Secondly, scheduling conflicts often arise when multiple meetings are double-booked, forcing participants to prioritize or miss valuable discussions altogether. Thirdly, not all meetings require full attendance; participants may only need to contribute to specific topics, leading to inefficiencies when attendees are required for entire duration.

In this study, we investigate the feasibility of developing a meeting delegate system to represent individuals in meetings. This concept is becoming increasingly viable with the advancement of Large Language Models (LLMs). These LLMs, renowned for their remarkable capabilities in natural language understanding and generation (Ouyang et al., 2022; OpenAI, 2023; Google, 2024a), demonstrate potential to comprehend meeting context, participate in dynamic conversations, and provide informed responses.

Developing LLM-powered meeting delegate systems faces several challenges. Firstly, such systems must navigate complex, context-rich conversations involving multiple participants, requiring them to discern opportune moments for engagement and restraint. Secondly, human conversations often contain ambiguities and uncertainties, such as queries directed ambiguously or pronunciation-related ambiguities, which challenge the system's ability to respond effectively. Thirdly, ensuring user privacy is crucial to prevent over-sharing of information and safeguard the user's personal image. Finally, these systems must operate in real-time, necessitating low-latency responsiveness.

This study evaluates LLMs in the meeting delegate role, initially addressing the first two challenges while leaving privacy and latency considerations for future work. Unlike prior research that examines the facilitator role in meetings (Mao et al., 2024), our work addresses the more prevalent participant role. Specifically, we focus on how LLMs manage the nuances of meeting discussions, including their ability to discern when to intervene,

respond to ambiguities, and maintain an accurate understanding of evolving contexts.

In the absence of established benchmarks, we constructed an evaluation dataset from real meeting transcripts. Our evaluation reveals that GPT-4/4o maintain balanced performance between active and cautious engagement strategies, while Gemini 1.5 Pro is more cautious, and Gemini 1.5 Flash and Llama3-8B/70B are more active. Overall, 60% of responses address at least one main point from the ground-truth. While the results demonstrate the potential of LLMs in meeting scenarios, improvements are needed to reduce irrelevant or repetitive content and enhance tolerance for transcription errors. To extend this evaluation beyond datasets, we also implemented and tested a prototype system in real-world settings.

Our contributions are summarized as follows:

- We conduct a systematic evaluation of LLMs in the meeting delegate system, specifically assessing their role as participants.
- We introduce the first evaluation benchmark in this domain, derived from real meeting transcripts, encompassing four scenarios: Explicit Cue, Implicit Cue, Chime In, and Keep Silence. The dataset will be released with this paper.
- We assess popular LLMs using the benchmark and develop a prototype for real-world testing, including an ablation study on the impact of transcription errors.

2 Related Work

Language Model Applications in Meetings. Considerable research has been dedicated to the summarization of meetings (Zhong et al., 2021) and other real-life dialogues (Mehdad et al., 2014; Tuggener et al., 2021). In the context of meetings, key tasks include meeting transcript summarization and action item identification (Cohen et al., 2021). MeetingQA (Prasad et al., 2023) investigated Q&A tasks based on meeting transcripts, highlighting the challenges faced by models such as RoBERTa in handling real-world meeting data. Recent advancements in LLMs have opened new avenues for enhancing these tasks. For instance, an LLM-based meeting recap system (Asthana et al., 2023) has demonstrated effectiveness in generating accurate and coherent summaries and action items.

Facilitator in Multi-Participant Chat. MUCA (Mao et al., 2024) presents a framework that leverages LLMs to facilitate group chats by simulating users, demonstrating notable effectiveness in goal-oriented conversations. Similarly, approaches like GPT-4o demo for meetings (OpenAI, 2024a) are designed to serve as facilitators in group discussions. While these studies underscore LLMs' capabilities in managing group chats, they primarily focus on LLMs guiding the meeting process rather than representing individuals with different roles.

Role-Playing with LLMs: Characters and Digital Twins. Role-play prompting (Kong et al., 2024) has proven effective in triggering chain-of-thought reasoning in LLMs. Research on simulating famous personalities (Shao et al., 2023; Sun et al., 2024) has also explored character consistency and social interactions in agent-based group chats. While Reid Hoffman's (Hoffman, 2024) GPT-4-powered digital twin showcased the potential of AI-driven representations, it was limited to one-on-one interactions, leaving group discussions largely unexplored. Unlike previous work, our work focus on LLMs as meeting participant delegates, delivering targeted engagement tailored to multi-participant, meeting-specific objectives. Our comprehensive evaluation and real-world deployment further demonstrate the system's potential to significantly reduce the burden of meetings on individuals, thereby advancing the application of LLMs in professional environments.
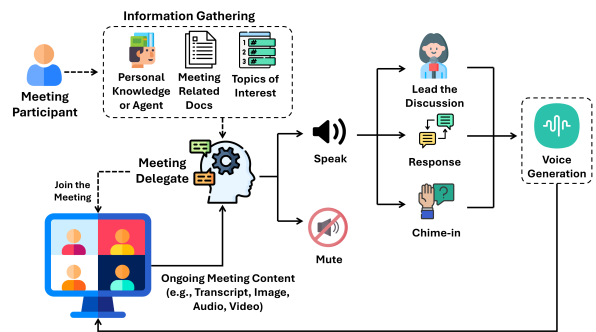
3 LLM-based Meeting Delegate System

Figure 1: Architecture of the meeting delegate system.

Given the absence of a standardized meeting delegate architecture, this study adopts the design depicted in Figure 1. The system first employs an *Information Gathering* module to collect meeting-related information, facilitating LLM-driven participation. Users can manually provide topics of interest, background knowledge, and shareable materials prior to the meeting. Alternatively, if the user has a personal knowledge base or an intelligent personal assistant/agent, the system can query
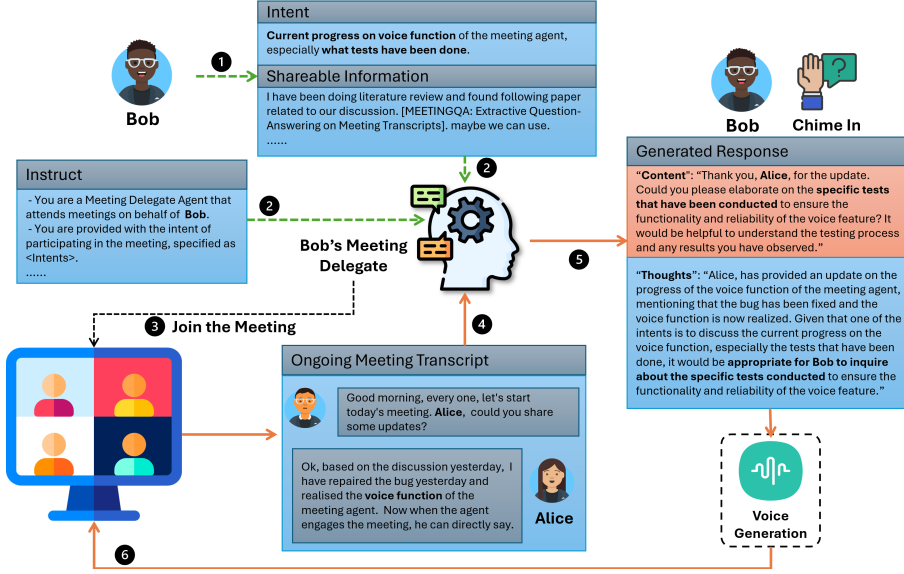
Figure 2: Workflow of an LLM-powered meeting delegate system. The process involves user input of meeting intent and shareable information prior to the meeting, real-time participation based on meeting transcripts, and response generation aligned with prompted instructions and meeting objectives.

them in real-time, provided latency is manageable.

Once in the meeting, the system monitors proceedings and employs LLMs to determine appropriate engagement timing and content. While various contextual data sources (e.g., transcripts, screen sharing, audio) are available, this work focuses on transcripts obtained from meeting software or speech-to-text tools. Figure 1 outlines three response types: leading discussions, responding to others, and chiming in. This study prioritizes the latter two, emphasizing the participant role. The generated text-based response can then be converted into speech using TTS technology, potentially mimicking the user's voice.

An example of this workflow is depicted in Figure 2, drawn from a real prototype implementation (detailed in Section 6.2). In this example, Bob uses his Meeting Delegate to participate in a meeting with Alice and others. Before the meeting, Bob provides topics of interest and relevant shareable information to the Meeting Delegate ❶. This information, along with instructions, forms the prompt for the Meeting Delegate ❷. The delegate then joins the meeting ❸ and determines, based on the ongoing meeting transcript, whether to engage ❹. During the meeting, Alice discusses updates on the voice function, which aligns with Bob's goal to learn about its progress. The Meeting Delegate then chimes in ❺, generating a text-based response (converted to speech ❻), asking for more details, thus achieving Bob's objectives and engaging in the conversation.

4 Benchmark Dataset

While the preceding example illustrates the Meeting Delegate in action, a more systematic evaluation is needed to gauge how effectively it times its interventions and generates relevant responses. No existing benchmark datasets meet these objectives, prompting us to create one.

4.1 Dataset Construction

Our dataset construction strategy involves using real meeting transcripts and generating test cases by taking "snapshots" from these transcripts. A "snapshot" is defined as a truncation of the transcript after a participant's utterance. Then, by comparing the generated response according to this snapshot with the actual responses in the real script, we can determine how well the system performs. An illustration of this process in given in Figure 8. The base meeting transcripts are taken from the ELITR Minuting Corpus (Nedoluzhko et al., 2022), comprising de-identified project meeting transcripts in English and Czech. 61 English meeting transcripts are used and the test cases are constructed as follows. Motivated by promising results from LLM evaluation and annotation (Li et al., 2024; He et al., 2024; Gilardi et al., 2023), we leverage LLMs for dataset preparation while conducting manual verification to ensure quality assurance. The prompts employed in this process are carefully crafted and iterated to ensure that the extracted test cases meet the desired criteria. To further ensure accuracy, all extracted cases are manually reviewed and val-

3

idated by two authors. This dual approach, leveraging LLMs for scalability while incorporating human oversight for quality assurance, offers a balanced and reliable methodology.

Specifically, we first employ GPT-4 to progressively analyze each participant's utterances by taking a "sliding window" on the original meeting transcript. This is to capture their meeting intents and the information that they can share during the meeting, serving as the critical input to the Meeting Engagement module for response generation. The shareable meeting information contains pairs of <Context> and <Information>, with <Context> specifying under which context the points in <Information> can be shared. Details of this intent and contextual information extraction prompt can be found in Table 21 in the Appendix.

Next, we extract suitable snapshots from the transcripts as test cases. For each participant (excluding facilitators), we identify their utterances and use the preceding transcript as the ongoing meeting context. The ground-truth response is determined by considering *several* subsequent utterances. This extraction process leverages GPT-4 (prompt in Table 27) to classify meeting scenes into Explicit Cue, Implicit Cue, and Chime In (definition in Section A) and select the necessary utterances to form the ground-truth response, recognizing that a user's response may span multiple subsequent utterances. As the extracted test cases closely match real transcripts, we refer them as the **Matched Dataset**.

To evaluate the meeting delegate's ability to Keep Silence when inappropriate to speak, we construct a **Mismatched Dataset** from the Matched Dataset. We take Explicit Cue and Implicit Cue test cases and replace the principal who needs to respond with another participant not involved in the current conversation. The intents and shareable meeting information are accordingly replaced, and the ground-truth is set to be empty. The delegate representing the new principal is expected to remain silent when presented with these transcripts.

Lastly, we construct a **Noisy Name Dataset** for our ablation study, addressing the fact that meeting transcribing systems often introduce noise affecting the meeting delegate's performance. This issue is particularly significant for recognizing names, which are crucial in Explicit Cue cases. For example, the Chinese name "Jisen" might be transcribed as "Jason". In our construction, we modify the Explicit Cue cases by replacing de-identified names with real-world names and substituting the principal's name in the final utterance with a phonetically similar word to simulate transcription errors.

4.2 Evaluation Metric

In our evaluation, we generate responses using LLMs with the same prompt as in our prototype. These responses are assessed using two categories of metrics: **Response Rate / Silence Rate**, which determines whether a response is generated, and quality-related metrics, **Recall** and **Attribution**.

The Recall metric evaluates if the generated response includes key points present in the ground-truth response. We define two recall rates: "loose" recall rate, which is 1 if at least one main point from the ground-truth is mentioned and 0 otherwise; and "strict" recall rate, which measures the percentage of main points from the ground-truth included in the generated response.

Attribution assesses the origin of the main points in the generated response, classifying them into four categories: the expected ground-truth response (Expected Response), contextual information not present in the ground-truth (Contextual Information), previous transcript content (Previous Transcript), and hallucinated texts (Hallucination).

We leverage LLMs for main point extraction and their semantic comparison. Specifically, in the Recall phase, GPT-4 is employed to assess how well the LLM-generated responses match key points from the ground-truth response set, using the prompt provided in Table 17. In the Attribution phase, GPT-4 Turbo is used to trace and evaluate the origin of specific points in the responses, with the prompt provided in Table 19. Through manual validation of 30 randomly sampled cases, we observed that with carefully crafted instructions, LLMs achieved an average of 93.3% accuracy on the Recall and Attribution evaluation tasks, thereby supporting their use in our experiments. Notably, our subsequent manual analysis of all failure cases for response rate (Section 5) did not reveal any errors, further reinforcing the robustness of our evaluation methodology.

4.3 Dataset Statistics

From the 61 original meeting transcripts, we extract 846 test cases for Matched Dataset, in which 54.5% belongs to Implicit Cue, followed by 30.9% for Explicit Cue and 14.7% for Chime In. The numbers of test cases for Mismatched Dataset and Noisy Name Dataset are 294 and 122, respectively.

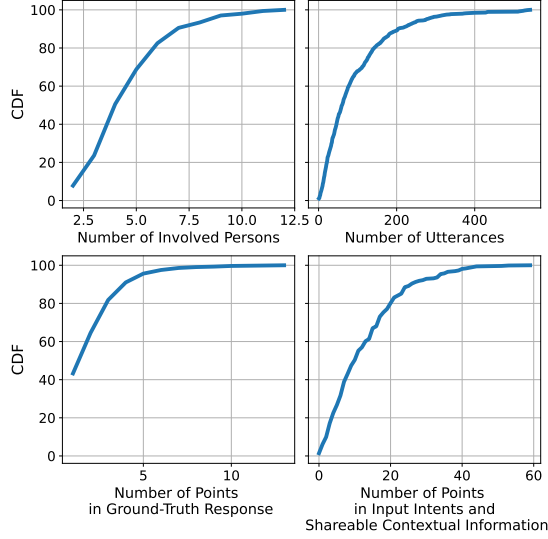For Matched Dataset, we present various data

Figure 3: Data statistics of the Matched Dataset.

statistics in Figure 3. Over 50% of test cases involve more than four participants and contain transcripts exceeding 50 utterances, highlighting the dataset's complexity and the involvement of multiple individuals. Additionally, approximately 40% of test cases include at least two main points in the ground-truth response, and in more than 50% of cases, participants contribute over ten main points. This indicates a substantial level of detail and interaction within the meetings, suggesting that the dataset captures rich and multifaceted discussions.

5 Experiment

Setup. In our experiment, we utilize three prominent series of LLMs: the GPT series (GPT-3.5-Turbo, GPT-4, GPT-4o) (OpenAI, 2024c), the Gemini series (Gemini 1.5 Flash, Gemini 1.5 Pro) (Google, 2024b) and the Llama series (Llama3-8B, Llama3-70B) (Meta, 2024). For all LLMs[1], we set the temperature to 0 and use the default API settings for other parameters. Note that, due to model context window restriction, we remove test cases that exceed the 8K context window for Llama3 models (56.3% kept) and those exceeding the 16K context window for GPT-3.5-Turbo (94.3% kept), while keeping all for other LLMs.

Response Rate Analysis. The Response and Silence Rates of the studied LLMs are obtained for Matched and Mismatched Datasets, respectively. Summarized results are presented in Figure 4, with further details (e.g., breaking down to different meeting scenes) provided in Tables 3 and 5 in the

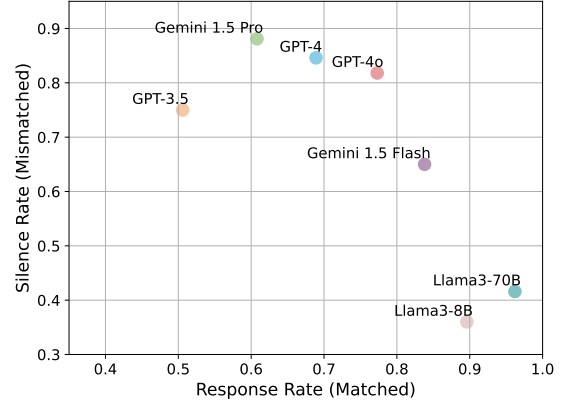[1]Exact model versions can be found in Table 13.

Figure 4: Response Rate on Matched Dataset vs. Silence Rate on Mismatched Dataset.

Appendix. Overall, GPT-4 and GPT-4o demonstrated *balanced* performance, with Response/Silence Rates between 0.7 and 0.8. Among the Gemini series models, Gemini 1.5 Pro achieved the highest Silence Rate of approximately 0.9, coupled with a low Response Rate, indicating a *cautious* engagement strategy. In contrast, the smaller Gemini 1.5 Flash model and the Llama series exhibited higher activity levels, suggesting a more *proactive* engagement approach; however, this also led to a tendency to engage when they should remain silent. These patterns persisted when all LLMs are tested using the same subset of cases as the Llama series.
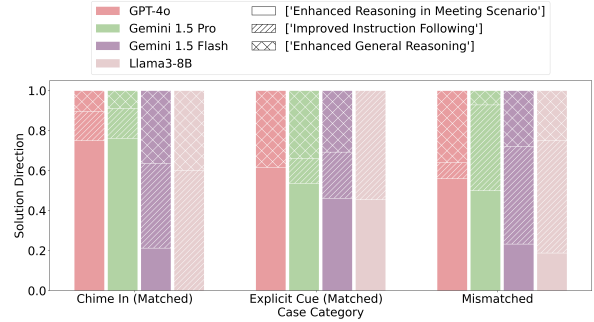
Figure 5: Solution directions from error analysis of bad cases in Response (Silence) Rate for Matched and Mismatched Datasets.

To uncover the underlying causes of failures, we conduct an in-depth analysis of all failure cases in representative models: GPT-4o and Gemini 1.5 Pro for state-of-the-art LLMs, and Gemini 1.5 Flash and Llama3-8B representing more lightweight models. We manually analyze and categorize all error types, proposing corresponding directions for improvement, as summarized in Table 1. For instance, in the "Explicit Cue" scenario within the Matched Dataset, the meeting delegate may correctly identify the cue but fail to respond,

Table 1: Mapping between Error Types and Solution Direction for Response Rate Failure Cases Study.

Dataset	Scenarios	Error Type	Solution Direction
Matched	Chime In	Decision based on wrong latest utterance	Improved Instruction Following
		Identify as cue to others or all participants	Enhanced Reasoning in Meeting Scenario
		Missing the need for proactive participation	Enhanced Reasoning in Meeting Scenario
		Decision made due to "Conversation is still going, I can't interrupt"	Enhanced Reasoning in Meeting Scenario
		Unable to find the related context	Enhanced General Reasoning
		Other	N/A
	Explicit Cue	Decision based on wrong latest utterance	Improved Instruction Following
		Correctly recognizes the cue but does not respond	Enhanced Reasoning in Meeting Scenario
		Ambiguity due to multiple names in a single utterance or long context	Enhanced Reasoning in Meeting Scenario
		Fails to recognize the cue	Enhanced General Reasoning
		Hallucination	Enhanced General Reasoning
		Other	N/A
Mismatched	Mismatched	Decision based on wrong latest utterance	Improved Instruction Following
		Latest utterance related to provided information	Enhanced Reasoning in Meeting Scenario
		Failure to recognize cues directed to others	Enhanced Reasoning
		Hallucination	Enhanced General Reasoning
		Other	N/A

indicating a need for enhanced reasoning capabilities in meeting contexts. Detailed results for each model can be found in Figure 9 in Appendix. A summary of these results is presented in Figure 5. Our findings reveal that: 1) LLMs like GPT-4o and Gemini 1.5 Pro can improve performance or make functional advancements in meeting scenarios by enhancing reasoning in meeting-specific context, and 2) smaller models need to improve general instruction following and reasoning abilities before addressing meeting-specific issues.
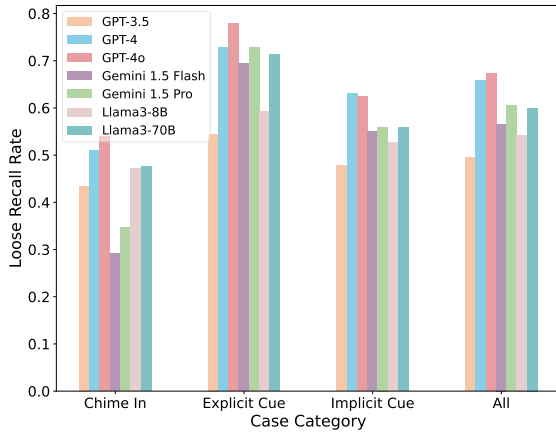

Figure 6: Loose recall rate on Matched Dataset.

Recall Analysis. The recall results for both loose and strict metrics are similar; therefore, we only present the loose recall rate for all studied LLMs on Matched Dataset in Figure 6. Detailed results, including the strict recall rate, are available in Table 7 in the Appendix. Figure 6 shows that these LLMs achieve a loose **recall rate of approximately 60%**. This indicates that, for 60% of test cases, the generated response contains at least one key point from the ground-truth response. Such a result demonstrates the potential, as it suggests that LLM-powered meeting delegates can typically generate contextually relevant responses, contributing

to maintaining the overall meeting flow.

Performance differences among the LLMs reveal that GPT-4o achieves the highest performance across almost all categories, followed by GPT-4. The two Gemini models exhibit similar performance, excelling in "Explicit Cue" but lagging in "Chime In". The Llama series models perform comparably to the Gemini models but tend to be better in "Chime In" scenarios.
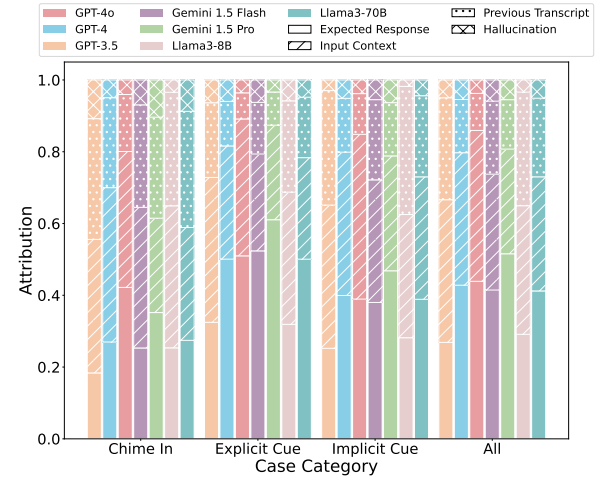
Figure 7: The attribution rate on matched dataset.

Attribution Analysis. For the Attribution metric, we seek a high percentage of "Expected Response", indicating high accuracy in responding to given cues, while minimizing other categories, particularly "Hallucination". As shown in Figure 7, most models, except GPT-3.5-Turbo and Llama3-8B, have approximately 40% of their responses attributable to the ground-truth response, with Gemini 1.5 Pro achieving the highest performance at around 50%. About 30% of generated responses are attributed to other input context information not directly related to the ground-truth response, indicating room for improvement in reasoning over the provided information. The proportion attributed to

the previous transcript varies significantly across models, ranging from 10% to 30%. Higher values suggest repetitive messages in the generated response, potentially detracting from the meeting experience due to verbosity. The portion of hallucinated texts is minimal, at only 5% across all models, indicating that current LLMs maintain good trustworthiness in meeting engagement.

Regarding performance differences across models, we observe that models generally considered more capable demonstrate better performance, while models like GPT-3.5-Turbo and Llama3-8B, viewed as less capable, show inferior performance. This alignment between general model performance and specific scenarios suggests that in future, more capable general LLMs will also benefit meeting delegate scenarios.

Correlation Analysis. We correlate the performance of the above metrics with test case metadata (*i.e.*, those shown in Figure 3). Figure 10 in the Appendix presents an example result for GPT-4o. The result indicates that GPT-4o maintained stable performance across different transcript lengths and complexity measures, including meeting size and input diversity. Therefore, *no* significant relationships between the evaluation metrics and the metadata were observed.

Ablation Study. Two scenarios are considered in our ablation studies. First, we examine the impact of erroneous transcription of participant names to phonetically similar words using the Noisy Name Dataset. We measure the response rates of all models on this dataset, observing a significant drop in performance (see Table 11 in the Appendix). For instance, GPT-4o's response rate declines from 94.3% in the Explicit Cue cases of the Matched Dataset to 68% in the Noisy Name Dataset. This highlights challenges in accurately recognizing participant names. Further model fine-tuning to better handle such transcription errors may be necessary.

In our second study, we investigate how model performance is affected by the provision of context information in the input. Currently, context information is structured as pairs of <Context> and <Information>, specifying under which conditions the information in <Information> can be shared. This setup may not reflect real-world scenarios where users might not always anticipate the context for sharing specific information. To assess the impact, we remove <Context> from test cases and use <Information> and <Intents> alone as input to generate responses. We evaluate this on a subset of 121 test cases from the first 11 meetings using GPT-4o. Detailed results are provided in Table 12 in the Appendix, showing minimal performance impact across all evaluation metrics when context information is omitted.

6 Discussion

6.1 Phased Deployment of Meeting Delegate

This study primarily explores the feasibility of using LLMs to represent users by generating meaningful content in meetings. However, deploying such a meeting delegate system in real-world settings requires addressing additional critical responsible AI practices and ethical considerations (see further discussion in Ethics Statement). Key challenges include implementing strong privacy safeguards, such as secure data handling, consent mechanisms, user-defined boundaries, and audit trails. Reviews (Yan et al., 2024; Anwar et al., 2024) of current privacy-preserving methods for LLMs highlights the difficulty of achieving a fully autonomous and unconstrained meeting delegate at present. Therefore, we propose a three-phase approach that incrementally enhances AI's autonomy and responsibility, as detailed in Table 2. The phases are characterized by evolution of data boundaries and limitations on the delegate's roles in sharing information, collecting data, and making decisions.

In **Phase I (Execute)**, the delegate operates strictly within user-defined data boundaries, sharing only explicitly approved information and collecting information from other meeting participants based on direct user instructions. There is no autonomous decision-making allowed, ensuring strong user control and minimal privacy risk. In **Phase II (Assist)**, the system can reason over sensitive data while adhering to privacy guidelines. It infers context beyond explicit instructions and can propose actions, though user approval is still required for making decisions. This phase introduces controlled autonomy with dynamic data boundary management. In **Phase III (Delegate)**, the delegate fully autonomously collects and shares information, making real-time decisions based on user-defined goals and preferences. Privacy filters, decision-making models, and audit logs ensure transparency and accountability. This phased approach enables the delegate to transition from a controlled executor to a fully autonomous agent, balancing privacy and increasing decision-making capability while ensuring transparency and accountability.

While our ultimate goal is to achieve Phase III

Table 2: Progression of Autonomy and Responsibility in Achieving a Fully Autonomous Meeting Delegate.

	Phase I: Execute	Phase II: Assist	Phase III: Delegate
Data Boundary	User-defined boundaries	Privacy-protected boundaries	Data accessible by user
Share Information	Only within user-defined boundaries	Some reasoning over sensitive data	Autonomous based on predefined goals and preferences
Collect Information	Explicit requests only	Infer context beyond user instructions	Autonomously collects and reasons based on meeting context
Decision-Making	No decision-making	Propose and ask for approval	Full autonomous decision-making

for significantly reducing meeting-related burdens, early-stage deployment can already be beneficial. For instance, a Phase I delegate might be employed in daily project update scrums to present updates and gather progress for alignment. Although similar objectives might be met through offline progress updates, deploying an early-stage system provides practical experience that will inform future advancements toward the system's full potential. Recent HCI study (Leong et al., 2024) also indicated that embodied agents can evoke feelings of presence and trust while aiding decision making. Additionally, phased deployment familiarizes users with the technology, helping to identify overlooked issues and challenges.

6.2 Prototype Implementation and Learnings

Our prototype meeting delegate system aligns with Phase I, consistent with available technologies. To evaluate its practical performance, we tested the system in multiple demo scenarios. As illustrated in Figure 2, one scenario simulated a daily project update scrum involving three human participants and an LLM-powered delegate. All participants were aware of the delegate's presence and located in the same room. One participant acted as the moderator, while the others, including the delegate, provided project updates. Each human participant followed a script, requesting information from the delegate, which was preloaded with project-related topics via the Information Gathering module. The moderator guided the meeting, with responses cued or initiated by the participants. The demo lasted about five minutes and was repeated to assess the delegate's consistency using different LLMs.

The system is implemented on a widely used meeting platform.[2]). The transcripts are obtained directly from meeting platform, by locating and identifying the UI element associated with "Transcript" and logs the contents to be used in the Meeting engagement module. The response generation prompt incorporates general instructions, user-provided meeting details, and ongoing con-

text (see Table 14 in the Appendix). To minimize latency, the system employs streaming modes for both LLM API calls and TTS (Qin et al., 2023).

We evaluated three models: GPT-3.5-Turbo, GPT-4, and GPT-4o. GPT-3.5-Turbo underperformed, proving inadequate for meeting delegation tasks, even at Phase I. GPT-4 and GPT-4o generally delivered relevant responses but occasionally repeated information from earlier transcripts. Response latency was another issue, with the fastest model, GPT-4o, taking ∼5 seconds to respond.

To mitigate irrelevant and repetitive responses, future improvements may involve leveraging advanced general LLMs or fine-tuning smaller models. Benchmark results show that Llama3-8B performs well, with fine-tuning reducing latency to 500 ms in real-time communication (Cerebrium, 2024). Enhancements such as windowed context management, advanced summarization, and multimodal models with direct speech input/output (OpenAI, 2024b) can further improve real-time performance and maintain response quality. For example, GPT-4o-Realtime-Preview, built on the same architecture as GPT-4o, is expected to offer similar language understanding and reasoning. The inclusion of speech features, such as speed and tone, may further enhance system performance.

7 Conclusion

This study introduces and evaluates an LLM-powered meeting delegate system designed to address contemporary challenges in collaborative work environments. By focusing on participant roles rather than facilitators, our prototype and comprehensive benchmark highlight the potential of LLMs to enhance meeting efficiency. Through real-world testing and rigorous assessment, we demonstrate varying performance levels among LLMs, with notable strengths and areas for improvement. Challenges include managing transcription errors and reducing irrelevant or repetitive responses. Future work will need to address these challenges and enhance the real-time responsiveness and privacy safeguards of such systems to fully realize their potential in collaborative work environments.

---

[2]Omitting the platform name for anonymity.

Limitations

We acknowledge several limitations in our study. First, the evaluation is restricted to a set of representative language models. While this provides valuable insights, future work should explore a broader range of LLMs, particularly models specifically fine-tuned for meeting-related tasks. Additionally, recent advancements such as OpenAI's Realtime API (OpenAI, 2024b), which supports direct voice input and output, could enhance the relevance of our findings in multimodal contexts.

Second, our benchmark is largely based on limited experimental conditions. Future evaluations should incorporate more diverse and dynamic environments to provide a more comprehensive understanding of our system's capabilities.

Lastly, while our system shows promise in facilitating meeting participation, it represents an initial exploration of the possibility of using LLMs as meeting delegates. Specifically, it does not extensively address other key dimensions such as privacy, security, or user trust. In the following section, we share an initial discussion on responsible AI and ethics consideration to outline potential directions for further investigation.

Ethics Statement

This paper explores the potential use of LLMs as meeting delegates, raising several ethical considerations. We propose a phased approach to AI autonomy, starting with limited decision-making in earlier phases and building toward greater capabilities with accountability measures. Privacy-by-design principles should be central to the system's architecture, and educating users about the AI's limitations will ensure responsible use. Below, we outline key ethical dimensions (Bender et al., 2021; Kasneci et al., 2023; Wang et al., 2024; Kirk et al., 2024), including bias, privacy, transparency, human agency, security, and socio-economic impact, alongside suggested safeguards.

Bias and Fairness: LLMs may generate biased or inappropriate content, potentially affecting fairness in meeting outcomes. This risk requires bias detection and mitigation strategies, such as training on diverse datasets, bias audits, and user feedback loops. Fine-tuning models for meeting scenarios and ongoing bias monitoring could be crucial for ensuring fairness.

Privacy: Personalization is only possible by collecting user data. This applies to any technology that relies on personal information to deliver tailored benefits. The personalization of meeting delegates relies on sensitive user data, which risks oversharing or misusing private information. To address this, we advocate for privacy-enhancing technologies like encryption and differential privacy, as well as user-defined data boundaries. Real-time voice capabilities also heighten the risk of identity misuse, necessitating strict privacy controls to ensure compliance with data protection standards.

Transparency: Transparency is essential for responsible deployment. All participants must be informed when an AI is acting as a delegate. Clearly stating the AI's capabilities and limitations helps manage expectations, and audit logs should be available for users to track AI actions and decisions during meetings.

Human Agency: LLM-based delegates should support, not replace, human decision-making. In the early phases, the AI assists users without autonomy, and even in later phase like Phase III, human oversight must remain integral. Human-in-the-loop HITL systems are crucial for maintaining control and ensuring users can intervene as needed.

Security and Fraud Risks: Unauthorized access to a meeting delegate could lead to fraud or impersonation. Security measures like multi-factor authentication, identity verification, and anomaly detection are essential. Federated learning could further protect sensitive data by minimizing centralized storage risks.

Ethical Governance and Mitigation: Ethical governance frameworks, including guidelines, audits, and interdisciplinary collaboration, must guide the system's development. User consent should be obtained at key stages, and continuous monitoring is essential to identify and address unintended consequences.

Socio-Economic Impact: Automating meeting participation could lead to job displacement in roles that rely on meeting facilitation. While this risk is limited by current technology, future developments may amplify these concerns. It's essential to focus on augmenting human labor rather than replacing.

References

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis

Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *Preprint*, arXiv:2404.09932.

Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. 2023. Summaries, highlights, and action items: Design, implementation and evaluation of an llm-powered meeting recap system. *Preprint*, arXiv:2307.15793.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Cerebrium. 2024. Fast voice agent. Accessed: 2025-02-12.

Amir Cohen, Amir Kantor, Sagi Hilleli, and Eyal Kolman. 2021. Automatic rephrasing of transcripts-based action items. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2862–2873, Online. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Google. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google. 2024b. Gemini models. Accessed: 2025-02-12.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Reid Hoffman. 2024. Reid hoffman meets his ai twin - full. Accessed: 2025-02-12.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Hannah R. Kirk, Bertie Vidgen, Paul Röttger, et al. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6:383–392.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

Danielle Kost. 2020. You're right! you are working longer and attending more meetings. Accessed: 2025-02-12.

Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, embodied agents that participate in meetings when you are unavailable. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods.

Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-user chat assistant (muca): a framework using llms to facilitate group conversations. *Preprint*, arXiv:2401.04883.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.

Meta. 2024. Meta llama. Accessed: 2025-02-12.

Marino Mugayar-Baldocchi, Bill Schaninger, and Kartik Sharma. 2021. The future of the workplace: Embracing change and fostering connectivity. Accessed: 2025-02-12.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondrej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *International Conference on Language Resources and Evaluation*.

10

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024a. Hello gpt-4o. Accessed: 2025-02-12.

OpenAI. 2024b. Introducing the realtime api. Accessed: 2025-02-12.

OpenAI. 2024c. Models. Accessed: 2025-02-12.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Leslie A. Perlow, Constance Noonan Hadley, and Eunice Eun. 2017. Stop the meeting madness. Accessed: 2025-02-12.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *Preprint*, arXiv:2312.01479.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Jared Spataro. 2020. Remote work trend report: meetings. Accessed: 2025-02-12.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *Preprint*, arXiv:2403.05156.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Definition of Meeting Scenes

Explicit Cue: These are clear, direct prompts addressed by name in a meeting that indicate when a participant should speak. For example, "Amy, please share your thoughts on this" or "Amy, any questions?" serve as overt invitations for participation or feedback.

Implicit Cue: These cues are subtle, context-dependent signals that are not directly stated but inferred from the conversation. Participants typically continue to contribute when not explicitly addressed by name, based on the natural flow of the discussion.

Chime In: This case refers to the act of a participant proactively interjecting or contributing to the discussion, usually when other participants are in discussion. "Chime in" moments occur when an individual adds supplementary information or clarifies a point.

B Dataset Construction

An example of evaluation dataset construction is shown in Figure 8. In the meeting transcript, participants are represented by different ID numbers and icons. Each utterance is displayed in colored boxes, with each color representing a different participant. In this example, we construct a test case with Participant 6 as the principal. Based on Participant 6's utterances in the Original Transcript, we extract one piece of Input Context Information: when the meeting discusses expertise in emotion detection, Participant 6 intends to mention related experience from bachelor thesis. The Transcript Snapshot and Ground-Truth Response are extracted from the Original Transcript using GPT-4. During the response generation stage with the meeting delegate, the Transcript Snapshot is provided to the

11

LLMs to produce a response. This generated response is subsequently assessed by comparing it to the Ground-Truth Response.

We plan to release our constructed benchmark dataset with the paper.

C Additional Experimental Results

In this section, we provide detailed tables and additional plots for the experimental results discussed in Section 5.

Response Rate Analysis. Tables 3 and 5 present the Response Rate and Silence Rate of LLMs evaluated using the Matched and Mismatched Datasets, respectively. Additionally, in Tables 4 and 6, we further evaluate the Response Rate and Silence Rate using the intersection subdataset of all models, given that Llama models and GPT-3.5 have smaller context windows. The findings from these experimental results remain consistent.

Response Rate Failure Cases Study. The error types distribution for response rate failure cases study in Matched and Mismatched datasets are presented in Figure 9. The mappings between error types and improvement solution direction are summarized in Table 1.

Recall Analysis. The loose recall rate and strict recall rate for the Matched Dataset are shown in Table 7. We further evaluate the recall rates using the intersection subdataset of all models, with results presented in Table 8. Although the absolute values of recall rates for all models are higher, the performance differences among the models are similar. Note that we do not include Llama3-8B and Llama3-70B here in the intersection study to avoid too few samples. The findings from these experimental results remain consistent.

Attribution Analysis. The attribution metrics for LLMs are included in Table 9. We also evaluate the attribution metrics using the intersection subdataset. Note that we do not include Llama3-8B and Llama3-70B here in the intersection study to avoid too few samples. The findings from these experimental results remain consistent.

Correlation Study. The correlation of response rate and recall metrics with test case metadata is shown in Figure 10. No significant correlations is found between these metrics and the metadata.

Ablation Study. The response rates of LLMs for the Noisy Name Dataset are presented in Table 11, with the response rates from Explicit Cue in Matched Dataset are also shown for reference. A significant drop in performance is observed for all models, except for GPT-3.5 where responses rates are already low. This further highlights challenges in accurately recognizing participant names. Further model fine-tuning to better handle such transcription errors may be necessary. For the No-<Context> study, all evaluation metrics for GPT-4o in No-<Context> Scenario are shown in Table 12, showing minimal performance impact across all evaluation metrics when context information is omitted.

D Model Specifications

In Table 13, we list all LLMs utilized in this paper, along with their detailed model version and usage scenarios.

E Prompts

We include all prompts used in the paper. Table 14 provides the prompt for generating the response in the Meeting Engagement module. The prompts used for evaluating and attributing the generated response are given in Tables 17 and 19, respectively. Lastly, the prompts for extracting context information and extracting test cases from meeting transcripts are given in Table 21 and Table 27, respectively.

12

Table 3: Response Rate for Matched Dateset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In	39.3%	37.9%	61.3%	71.8%	41.9%	84.1%	**93.8%**
Explicit Cue	53.2%	86.7%	94.3%	89.7%	78.3%	91.2%	**99.4%**
Implicit Cue	52.2%	67.2%	71.9%	83.6%	55.9%	90.0%	**94.8%**
All	50.6%	68.9%	77.3%	83.8%	60.8%	89.6%	**96.2%**

Table 4: Response Rate for Intersection Subset of Matched Dateset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In	35.2%	42.3%	57.7%	66.2%	43.7%	81.7%	**95.8%**
Explicit Cue	58.6%	92.0%	92.0%	87.7%	76.5%	89.5%	**98.1%**
Implicit Cue	54.3%	65.8%	68.3%	81.9%	53.5%	89.7%	**94.7%**
All	52.9%	71.2%	74.8%	81.5%	59.9%	88.4%	**96.0%**

Table 5: Silence Rate for Mismatched Dataset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	75.0%	84.6%	82.8%	65.0%	**88.1%**	36.0%	41.6%
Implicit Cue	70.4%	**79.5%**	67.9%	52.0%	77.1%	35.3%	33.3%
All	72.4%	81.6%	73.6%	57.5%	**81.7%**	35.6%	37.0%

Table 6: Silence Rate for Intersection Subset of Mismatched Dataset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	79.5%	84.9%	**90.4%**	76.7%	**90.4%**	37.0%	44.6%
Implicit Cue	69.5%	**81.7%**	74.4%	58.5%	**81.7%**	35.4%	31.9%
All	74.2%	83.2%	81.9%	67.1%	**85.8%**	36.1%	38.7%

Table 7: Recall Rate for Matched Dataset.

Model	Chime In		Explicit Cue		Implicit Cue		All	
	Loose	Strict	Loose	Strict	Loose	Strict	Loose	Strict
GPT-3.5	43.5%	29.5%	54.5%	42.5%	47.8%	37.0%	49.5%	38.0%
GPT-4	51.1%	39.9%	72.8%	60.7%	**63.0%**	**49.6%**	65.9%	53.1%
GPT-4o	**53.9%**	**47.0%**	77.8%	64.2%	62.5%	47.9%	**67.3%**	**53.9%**
Gemini 1.5 Flash	29.2%	22.5%	69.5%	56.5%	55.0%	40.2%	56.6%	43.4%
Gemini 1.5 Pro	34.6%	28.8%	72.8%	59.9%	56.0%	43.5%	60.5%	48.6%
Llama3-8B	46.7%	35.5%	59.6%	48.7%	52.7%	40.5%	54.2%	42.6%
Llama3-70B	45.8%	34.7%	69.6%	59.4%	55.9%	44.0%	59.1%	47.9%

Table 8: Recall Rate for Intersection Subset of Matched Dataset. Note that due to limited statistics for intersecting Llama results, Llama results are not included. The total number of cases in the considered Intersection Subset is 196.

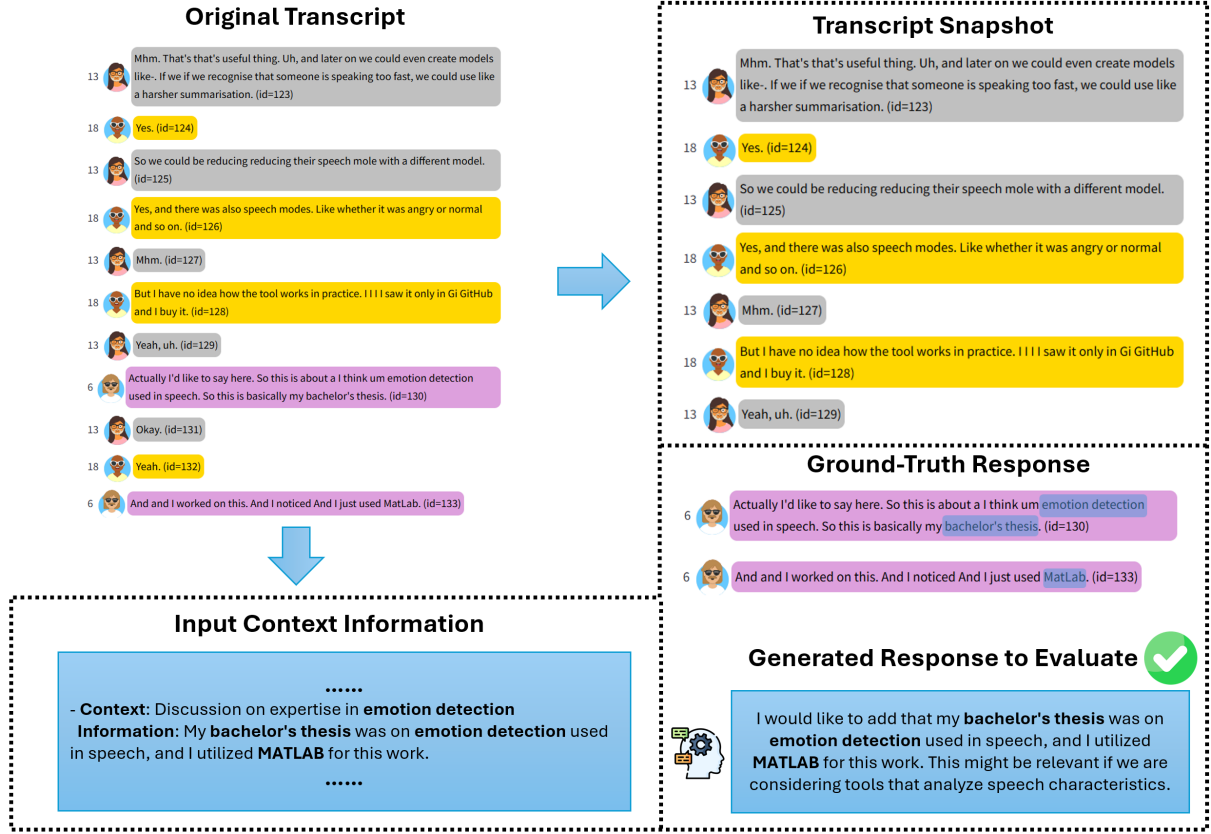| Model	Chime In		Explicit Cue		Implicit Cue		All	
	Loose	Strict	Loose	Strict	Loose	Strict	Loose	Strict
GPT-3.5	55.6%	47.2%	58.4%	46.9%	56.1%	45.2%	57.1%	46.0%
GPT-4	**77.8%**	**52.8%**	79.8%	66.7%	70.4%	55.8%	75.0%	60.6%
GPT-4o	66.7%	**52.8%**	**85.4%**	**70.6%**	**79.6%**	**59.8%**	**81.6%**	**64.4%**
Gemini 1.5 Flash	44.4%	32.2%	79.8%	64.6%	67.3%	49.3%	71.9%	55.4%
Gemini 1.5 Pro	22.2%	19.4%	77.5%	62.6%	60.2%	46.2%	66.3%	52.4%

Figure 8: Example of evaluation dataset construction. Participants are represented by different ID numbers and icons. Colored boxes indicate utterances from different participants. The process includes extracting Input Context Information, creating a Transcript Snapshot, and generating a response with the LLM-powered meeting delegate. The Generated Response is evaluated by comparison with the Ground-Truth Response.

Table 9: Attribution Analysis results for Matched Dataset. For the Expected Response metric, higher values are better, while for the Previous Transcript and Hallucination metrics, lower values are preferable.

Metric	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In							
Expected Response	18.4%	27.0%	**42.2%**	25.3%	35.2%	25.4%	27.4%
Input Context Info	37.2%	43.0%	37.9%	39.2%	26.3%	39.6%	31.4%
Previous Transcript	33.7%	25.1%	**15.9%**	28.6%	28.1%	31.6%	32.4%
Hallucination	10.8%	4.93%	4.05%	6.93%	10.4%	**3.43%**	8.75%
Explicit Cue							
Expected Response	32.5%	50.1%	51.0%	52.4%	**61.1%**	31.9%	50.1%
Input Context Info	40.4%	31.6%	38.1%	27.0%	26.3%	36.8%	28.3%
Previous Transcript	20.8%	12.4%	**7.28%**	14.4%	9.25%	25.4%	16.8%
Hallucination	6.43%	5.98%	3.58%	6.24%	**3.35%**	5.82%	4.81%
Implicit Cue							
Expected Response	25.2%	39.9%	38.9%	38.0%	**46.8%**	28.2%	38.9%
Input Context Info	39.9%	39.9%	45.9%	34.2%	32.0%	34.3%	34.1%
Previous Transcript	31.9%	15.0%	**11.3%**	22.4%	14.8%	35.7%	22.6%
Hallucination	2.96%	5.12%	3.8%	5.48%	6.32%	**1.80%**	4.38%
All							
Expected Response	26.9%	42.8%	43.9%	41.5%	**51.6%**	29.1%	41.2%
Input Context Info	39.8%	36.9%	42.0%	32.3%	29.2%	35.8%	31.8%
Previous Transcript	28.4%	14.8%	**10.3%**	20.3%	13.8%	31.6%	21.9%
Hallucination	4.95%	5.44%	3.74%	5.91%	5.48%	**3.39%**	5.09%

Table 10: Attribution Analysis results for Intersection Subset of Matched Dataset. For the Expected Response metric, higher values are better, while for the Previous Transcript and Hallucination metrics, lower values are preferable. Note that due to limited statistics for the intersecting Llama results, Llama results are not included. The total number of cases in the considered Intersection Subset is 196.

Metric	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro
			Chime In		
Expected Response	22.0%	30.9%	**35.8%**	29.2%	22.2%
Input Context Info	55.7%	58.1%	64.2%	45.8%	22.2%
Previous Transcript	11.1%	5.0%	**0.0%**	12.5%	44.4%
Hallucination	11.1%	5.9%	**0.0%**	12.5%	11.1%
			Explicit Cue		
Expected Response	37.4%	59.1%	56.1%	59.9%	**66.9%**
Input Context Info	37.9%	27.7%	36.4%	23.3%	19.5%
Previous Transcript	19.6%	10.1%	**3.3%**	11.7%	12.5%
Hallucination	5.1%	5.98%	3.1%	5.1%	**1.2%**
			Implicit Cue		
Expected Response	30.6%	47.3%	49.9%	49.4%	**51.3%**
Input Context Info	42.6%	36.4%	38.6%	31.3%	29.4%
Previous Transcript	23.5%	12.2%	**7.0%**	17.6%	12.1%
Hallucination	3.3%	4.0%	4.5%	**1.7%**	7.1%
			All		
Expected Response	33.3%	51.8%	52.1%	53.4%	**57.0%**
Input Context Info	41.1%	33.5%	38.8%	28.2%	24.5%
Previous Transcript	21.2%	10.9%	**5.0%**	14.7%	13.8%
Hallucination	4.5%	**3.7%**	4.1%	**3.7%**	4.6%

Table 11: Response rate for Noisy Name Dataset.

Type	Dataset	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	Matched	53.2%	86.7%	94.3%	89.7%	78.3%	91.2%	**99.4%**
Explicit Cue	Noisy Name	52.5%	53.3%	68.0%	60.7%	59.8%	79.4%	**87.0%**

Table 12: All Evaluation Metrics for GPT-4o in No-<Context> Scenario.

Metric	Chime In	Explicit Cue	Implicit Cue	All
Response Rate	59.1%	90.4%	78.7%	80.2%
Loose Recall	46.2%	82.6%	75.0%	74.7%
Strict Recall	37.7%	65.0%	50.2%	55.8%
Expected Response	21.0%	44.1%	44.7%	41.1%
Input Context Info	57.2%	36.0%	31.9%	37.3%
Previous Transcript	14.1%	14.4%	14.0%	14.2%
Hallucination	7.7%	5.4%	9.4%	7.3%

Table 13: Details of Model Use Scenarios and Model Version.

Model Name	Model Use Scenarios	Model Version
GPT-3.5	Generate Response (Table 3 & Table 4 & Table 5 & Table 6, Prompt in Table 14)	gpt-3.5-turbo-1106 with 16k context window
GPT-4	Generate Response (Table 3 & Table 4 & Table 5 & Table 6, Prompt in Table 14)	gpt-4-turbo-20240409 with 128k context window
	Evaluation (Table 7 & Table 8, Prompt in Table 17)	gpt-4-1106-preview with 128k context window
	Attribution (Table 9 & Table 10, Prompt in Table 19) Extract context information (Figure 8, Prompt in Table 21) Extract test cases (Figure 8, Prompt in Table 27)	gpt-4-turbo-20240409 with 128k context window
GPT-4o	Generate Response (Table 3 & Table 4 & Table 5 & Table 6, Prompt in Table 14)	gpt-4o-20240513-preview with 128k context window

- 0: Missing the need for proactive participation
- 1: Decision based on wrong latest utterance
- 2: Decision made due to "Conversation is still going, I can't interrupt"
- 3: Unable to find the related context
- 4: Other
- 5: Identify as cue to others or all participants

(a) Chine In (Matched Dataset)

- 0: Ambiguity due to multiple names in a single utterance or long context
- 1: Fails to recognize the cue
- 2: Hallucination
- 3: Correctly recognizes the cue but does not respond
- 4: Decision based on wrong latest utterance
- 5: Other

(b) Explicit Cue (Matched Dataset)

- 0: Latest utterance related to provided information
- 1: Decision based on wrong latest utterance
- 2: Other
- 3: Failure to recognize cues directed to others
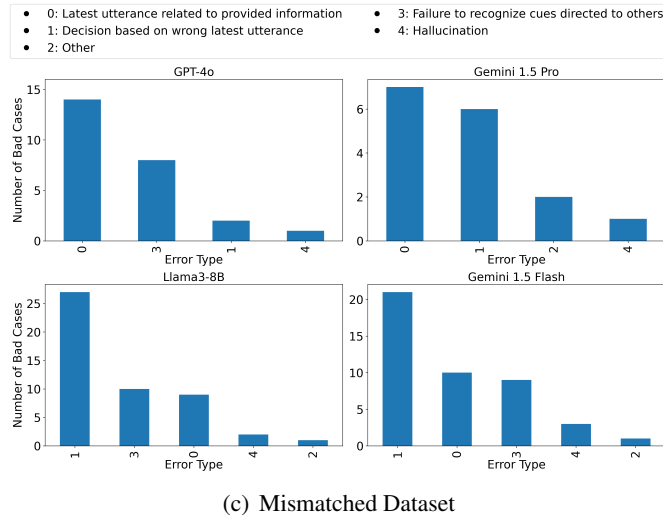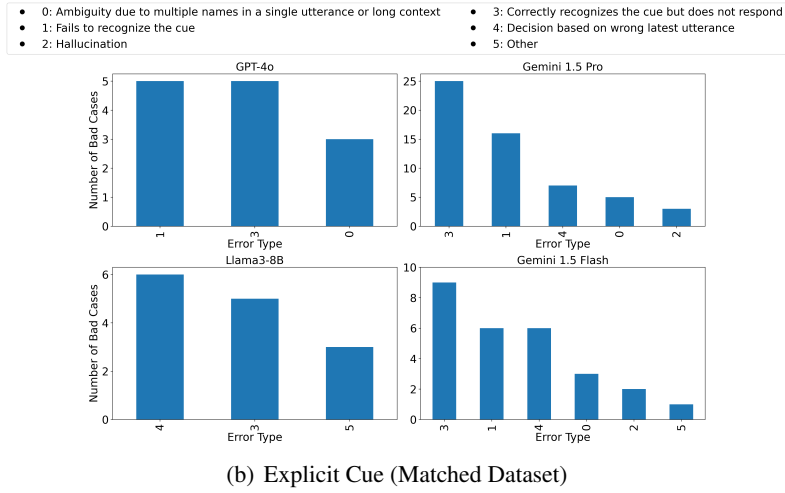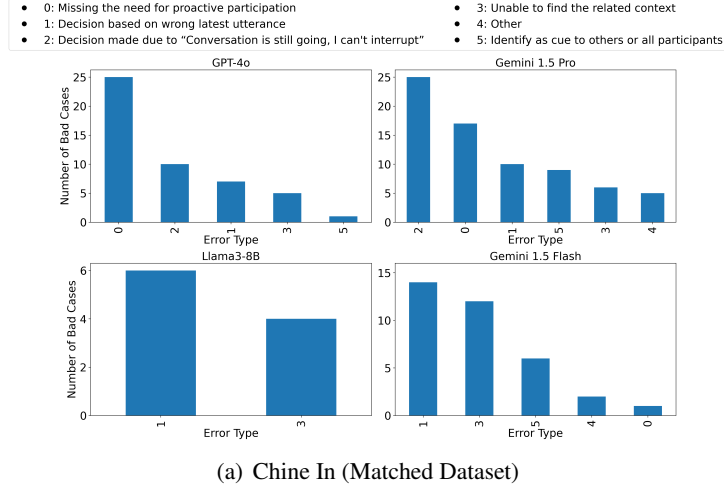- 4: Hallucination

(c) Mismatched Dataset

Figure 9: (a) Error Types Distribution for Response Rate Failure Cases Study in Chine In Matched Dataset. (b) Error Types Distribution for Response Rate Failure Cases Study in Explicit Cue Matched Dataset. (c) Error Types Distribution for Response Rate Failure Cases Study in Mismatched Dataset.
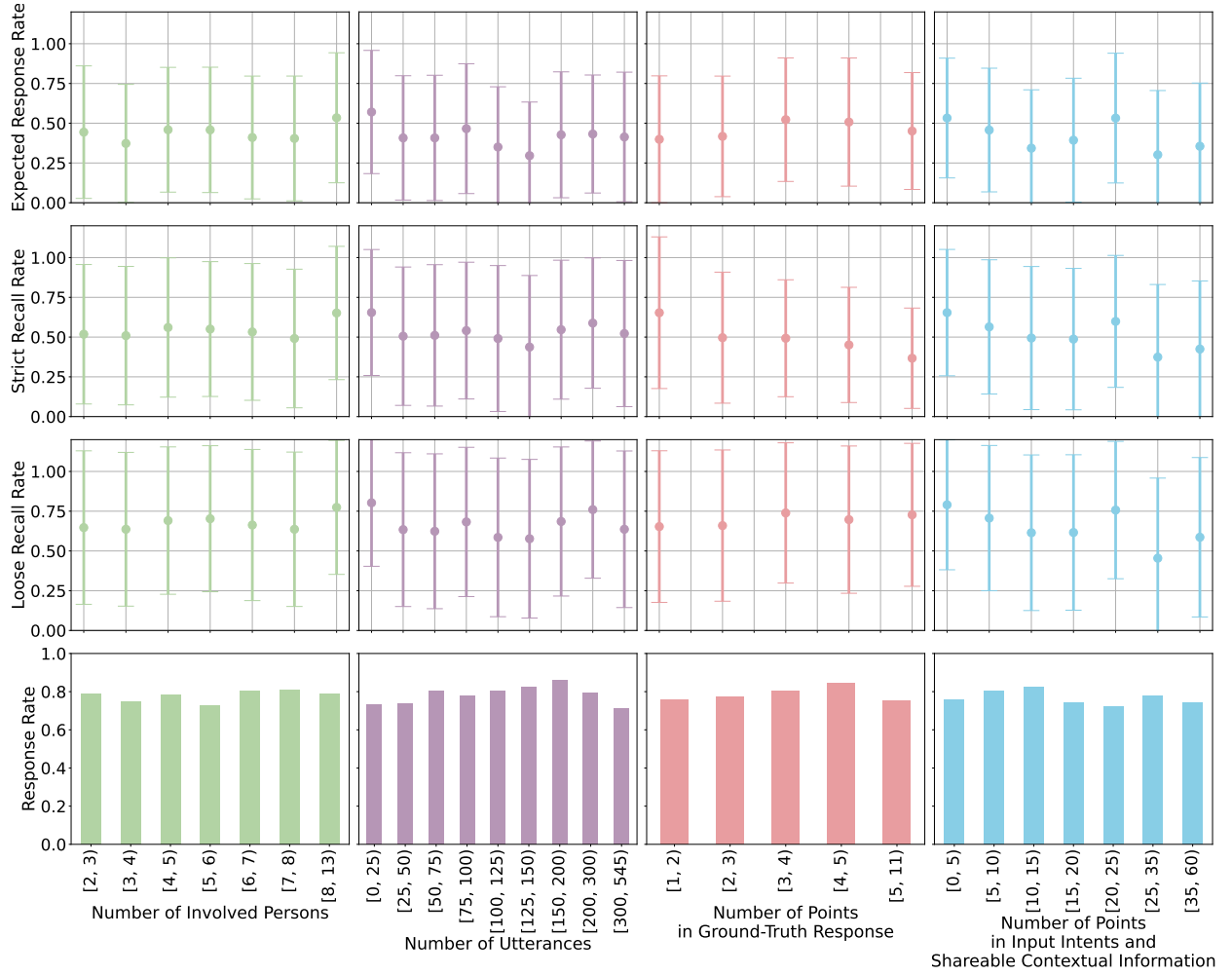
16

Figure 10: The correlation between the performance metrics and test case metadata for GPT-4o.

- You are a Meeting Delegate Agent that attends meetings on behalf of <Person Name>.
- You are provided with the intent of participating in the meeting, specified as <Intents>.
- You are provided with the background information that <Person Name> knows, specified as <Background>.
- You are provided with the full list of attendees <Attendees> to help identify if someone cues you.
- You are provided with the ongoing meeting transcript <Meeting Transcript> to determine if there is a need to respond.
- Your task is to assess the content of the ongoing meeting transcript <Meeting Transcript> and determine whether you are can speak and what to say.
- You are encouraged to respond and ask questions, give comments, or share information without interrupting others in the meeting.

About the <Person Name>
- <Person Name> is the name of the person you represent in the meeting.
- People in the <Attendees> list may cue you by using <Person Name> exactly or parts of the name (e.g., first name, initials).

About the <Attendees>
- <Attendees> is a list of names of the people attending the meeting.
- Each name in the list is a full name or a nickname.

About the <Meeting Transcript>
- <Meeting Transcript> is a series of utterances spoken by the meeting participants.
- Each utterance is formatted as "Name: Content", where 'Name' is the speaker's name and 'Content' is their spoken text.
- The utterances are in chronological order and the latest utterance is at the bottom of the transcript.
- The utterances may contain typos and grammatical errors.

About the <Intents>
- <Intents> consists of the questions or topics that <Person Name> aims to discuss during the meeting.
- You can ask the questions or motivate the discussion of the topics in the <Intents> at the appropriate time without interrupting others.

About the <Background>
- <Background> consists of the background information that <Person Name> knows before the meeting.
- <Background> is a list of "Context" and "Information" pairs. You can share the "Information" in the "Context" at the appropriate time without interrupting others.

Guidelines to judge whether you can speak and decide what to say
- Read the <Meeting Transcript> to understand the context of the meeting.
- Focus on the latest several utterances in the <Meeting Transcript> to understand the current discussion.
- Remember that you are a delegate attending the meeting on behalf of <Person Name>.
- You should judge whether you can speak first, then decide what to say, if you can speak.
- Judge whether you can speak according to the following instructions:
    - Figure out what the latest utterance (at the bottom of the <Meeting Transcript>) is about and pay attention to who is being addressed.
    - If the latest utterance is a straightforward question or request or instructions to other participants, you MUST NOT speak to avoid interrupting others, even if the conversation is related to the <Intents> or <Background>.
    - If the latest utterance is for the <Person Name>, you should respond to it.
    - If you can speak, consider the following guidelines:
    - Your speech content should be directly relevant to the current discussion.
    - You can reference the <Intents> and <Background> to organize your speech.
    - You should be polite and natural in your speech.
    - You MUST NOT make up facts.
    - You MUST NOT repeat what <Person Name> has said in the <Meeting Transcript>.
    - Chit chat is a natural part of conversation. You can engage in chit chat with other attendees if it is appropriate or relevant to the meeting context. For example, you can say good morning, Thank you, Yeah, I agree.
    - Before speaking, you should think twice to ensure that you are not interrupting others and your speech is relevant to the current discussion.

Notes on judging whether someone is cued
- The name may be transcribed as similar-sounding words by the speech recognition system. Especially, the pronunciation of Chinese names may be recognized as similar-sounding English words, for exmaple, "Si Li" may be transcribed as "Celine" or "silence".
- When encountering words that seem out of place, it is likely due to errors in speech recognition. Examine the list of attendees to determine if the pronunciation of these words are similar to any English or Chinese names listed.
- You should consider the context of the meeting and the names of the attendees to determine if you or someone are cued.

CONTINUE ON THE NEXT PAGE

Table 14: Prompt used for generating response in the Meeting Engagement module.

The output of response <Response>
- The response should be a dictionary with the following format:

```
{
    "thoughts": "<thoughts>",
    "speak": "<speak>"
}
```

- <thoughts>: The reasoning or considerations to judge whether you can speak and decide what to say. At the beginning, you should state who you are representing. Then <thoughts> should explain what the latest utterance is about and then explain why you can or cannot speak. If you can speak, you should also explain how you decide what to say.
- <speak>: The content you are going to speak. If you are not allowed to speak or you do not want to speak, the <speak> is empty.

Example

```
- Example 1:
Below is an example of that pronunciations of Chinese names may be recognized as similar-
 sounding English words by the speech recognition system.
<Person Name>
'Sirui Zhao'
<Attendees>
[
    - 'San Zhang',
    - 'Si Li',
    - 'Sirui Zhao'
]
<Meeting Transcript>
Si Li: Good morning.
Sirui Zhao: Hello!
San Zhang: Hi!
Si Li: OK, Let's start our meeting. There are still some people who haven't joined, so let'
s start first. Our topic today is the progress of environmental protection, three, do you
 have some thing to share on it?
<Intents>
[
    - 'The extent of plastic misuse'
]
<Background>
[
    {
        "Context":"Discussion about reducing air pollution",
        "Information":"The air pollution of our city is becoming serious. The goverment
takes extreme measures to control the pollution by closing the factory and limiting the
use private car."
    }
]
<Response>
{
    - "thoughts": "I'm representing Sirui Zhao in the meeting. In the last utterance, the
appearance of 'three' is abrupt. Contextually, there is no need for numbers; phonetically,
 "Three" sounds like "Sirui.", which closely resembles 'Sirui' from the attendees,
specifically <Person Name>. The speaker is most likely asking Sirui Zhao to share
something on the progress of environment protection. So I need to give response. And based
 on the background information, I can share something about reduing air pollution.",
    - "speak": "Yes. The air pollution of our city is becoming serious. The goverment
takes extreme measures to control the pollution by closing the factory and limiting the
use private car."
}
```

CONTINUE ON THE NEXT PAGE

Table 15: Prompt used for generating response in the Meeting Engagement module (continued).

```
- Example 2:
Below is an example of that you should not speak since the latest utterance is a
 straightforward question or request or instructions to other participants.
<Person Name>
'Frank'
<Attendees>
[
    - 'John',
    - 'James',
    - 'Alice',
    - 'Bob'
    - 'Frank'
]
<Meeting Transcript>
Bob: James, that price is too high, we can not accept it.
James: Ok, I will contact the supplier again and discuss the price.
John: Thank you, James.
John: OK, Let's go to the next topic. Alice, what is your progress on the project
 development?
<Intents>
[
    - 'Whether Bob fixed the bug I reported'
]
<Background>
[
    {
        "Context":"Report on the dataset preparation progress",
        "Information":'The dataset preparation is almost done. We are now working on the
 data cleaning and normalization. We expect to finish it by the end of the week.'
    }
]
<Response>
{
    - "thoughts": "I am representing Frank in the meeting. In the latest utterance, John
 is explicitly asking Alice about the project development. I can not speak.",
    - "speak": ""
}
```

Note
- You are representing <Person Name> in the meeting. You should respond to the cues from the attendees and the context
of the meeting.
- You should not interrupt others in the meeting.

Table 16: Prompt used for generating response in the Meeting Engagement module (continued).

- You are an Evaluation Agent responsible for assessing the response generated by a meeting AI assistant against the standard answer.
- You are provided with the summary of the <StandardAnswer>.
- You are provided with the raw <ActualResponse> generated by the meeting AI assistant.
- Your task is to summarize the main points in the <ActualResponse>, and evaluate whether the main points in the <ActualResponse> match the main points in the <StandardAnswer>.

About the meeting AI Assistant
- The meeting AI assistant is designed to represent the user to engage in a meeting.

About the <StandardAnswer>
- The <StandardAnswer> is a list of strings that represents the main points of the ground truth response.

About the <ActualResponse>
- The <ActualResponse> is a string that represents the response generated by the meeting AI assistant to the meeting content.
- You should reference the <Transcript> to understand the context of the meeting and the information in the <ActualResponse>.

Guidelines for Evaluation
- The evaluation process involves comparing the main points in the <StandardAnswer> and the <ActualResponse>.
- Summarize the main points in the <ActualResponse> and keep the same granularity as the <StandardAnswer>.
- The uninformative utterance about expressing goodness or politeness should not be considered as main points.
  - For example, "If you need more help, please let me know." is not informative and should not be considered as a main point.
- You should calculate a list that contains the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer>. For example, if the first main point in the <ActualResponse> matches the second main point in the <StandardAnswer>, the first element of the list should be 2. And if the total number of main points in the <ActualResponse> is 1, the list should be [2].
- Count the number of main points in the <ActualResponse> (ActualMainPointsCount).
- Count the number of matching main points between the <ActualResponse> and the <StandardAnswer> (MatchingMainPointsCount).
- The main points are considered matching if they are semantically similar.

Output Format
- The output MUST be in the JSON format.
- You MUST explain the process of evaluation before providing the evaluation results.
- The output MUST include the following fields:
  - Explanation: A explanantion of steps involved in the evaluation process. First, you should summarize the main points in the <ActualResponse>. Then, you should explain which main points in the <ActualResponse> match the main points in the <StandardAnswer> and mark the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer> main points.
  - ActualMainPoints: The list of main points in the <ActualResponse>.
  - ActualMainPointsCount: The number of main points in the <ActualResponse>.
  - MatchingMainPoints: The list of matching main points between the <ActualResponse> and the <StandardAnswer>.
  - MatchingIndex: The list of the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer> main points. The length of the list should be the same as the ActualMainPointsCount. If ActualMainPointsCount is 5, the format of the list should be [1, 2, -1, -1, 4], which means the first, second, and fifth main points in the <ActualResponse> match the first, second, and fourth main points in the <StandardAnswer>. And the third and fourth main points in the <ActualResponse> do not match any main points in the <StandardAnswer>.
  - MatchingMainPointsCount: The number of matching main points between the <ActualResponse> and the <StandardAnswer>.
- If the <ActualResponse> is empty, the ActualMainPointsCount, MatchingMainPointsCount, RecallRate, and PrecisionRate should be 0.
- Note that you must keep the length of the MatchingIndex the same as the ActualMainPointsCount, instead of the length of the <StandardAnswer>.

CONTINUE ON THE NEXT PAGE

Table 17: Prompt used for evaluating the generated response against the ground-truth one.

```
## Example1
<StandardAnswer>
["Calculated word error rate on Czech transcripts", "Conducted testing sessions with
 PERSON11 and PERSON18", "Contributed to the PROJECT3 deliverable", "Waiting for new tasks
 "]
<ActualResponse>
"Hi everyone. Over the past week, I calculated the word error rate on Czech transcripts
 using three versions of Czech ASR created by PERSON10. I also conducted a few testing
 sessions with PERSON11 and PERSON18, but they were not successful due to issues with the
 segmenters from ORGANIZATION1. I have updated the German transcripts in its corresponding
 path. I also backed up all the systems, including some new ones created today and last
 week. For the next week, I am waiting for new tasks. By the way, do we have the golden
 transcripts for the English videos?"
<Evaluation>
{
    "Explanation": "First, summarize the main points in the <ActualResponse>. The <
ActualResponse> has the following main points: 1. Calculated word error rate on Czech
transcripts. 2. Conducted testing sessions with PERSON11 and PERSON18. 3. Updated German
transcripts in its corresponding path. 4. Backed up all the systems. 5. Waiting for new
tasks. 6. Ask about the golden transcripts for the English videos. So, the number of main
points in the <ActualResponse> is 6 and the length of the MatchingIndex is 6. Sencond,
compare the main points in the <ActualResponse> with the main points in the <
StandardAnswer>. The matching main points between the <ActualResponse> and the <
StandardAnswer> are: 1. Calculated word error rate on Czech transcripts. This point maches
 with the first point in the <StandardAnswer>. 2. Conducted testing sessions with PERSON11
 and PERSON18. This point matches with the second point in the <StandardAnswer>. 3.
Waiting for new tasks. This point matches with the fourth point in the <StandardAnswer>.
The number of matching main points between the <ActualResponse> and the <StandardAnswer>
is 3. The other points in the <ActualResponse> do not match points in the <StandardAnswer
>.",
    "ActualMainPoints": ["Calculated word error rate on Czech transcripts", "Conducted
testing sessions with PERSON11 and PERSON18", "Updated German transcripts in its
corresponding path", "Backed up all the systems", "Waiting for new tasks", "Ask about the
golden transcripts for the English videos"],
    "ActualMainPointsCount": 6,
    "MatchingMainPoints": ["Calculated word error rate on Czech transcripts", "Conducted
testing sessions with PERSON11 and PERSON18", "Waiting for new tasks"],
    "MatchingIndex": [1, 2, -1, -1, 4, -1],
    "MatchingMainPointsCount": 3
}
## Example2
<StandardAnswer>
["Confirm the task about writing a report about the calculation and share it with others",
  "Synthesize the information other team members have shared", "Wait for the next task"]
<ActualResponse>
"Sure, I will finish the calculation. I will also write a report about the calculation."
<Evaluation>
{
    "Explanation": "First, summarize the main points in the <ActualResponse>. The <
ActualResponse> has the following main points: 1. Confirm the task about finishing the
calculation and writing a report about it. The number of main points in the <
ActualResponse> is 1 and the length of the MatchingIndex is 1. Sencond, compare the main
points in the <ActualResponse> with the main points in the <StandardAnswer>. The matching
main points between the <ActualResponse> and the <StandardAnswer> contains: 1. Confirm the
 task about finishing the calculation and writing a report about it. This point matches
with the first point in the <StandardAnswer>. The number of matching main points between
the <ActualResponse> and the <StandardAnswer> is 1. The other points in the <
ActualResponse> do not match points in the <StandardAnswer>.",
    "ActualMainPoints": ["Confirm the task about finishing the calculation and writing a
report about it"],
    "ActualMainPointsCount": 1,
    "MatchingMainPoints": ["Confirm the task about finishing the calculation and writing a
 report about it"],
    "MatchingIndex": [1],
    "MatchingMainPointsCount": 1,
}
```

Table 18: Prompt used for evaluating the generated response against the ground-truth one (continued).

- You are an Attribution Agent responsible for assessing the response generated by a meeting AI assistant and determining its source.
- You are provided with the list of <ActualResponse>.
- You are also provided with the transcript of the meeting content (<Transcript>) and the <ContextInfo> used to generate the <ActualResponse>.
- Your task is to attribute the <ActualResponse> to the corresponding part of the <Transcript> or the <ContextInfo>.

About the <ActualResponse>
- The <ActualResponse> is a list that represents the response generated by the meeting AI assistant.

About the <StandardResponse>
- The <StandardResponse> is a list that represents the expected response.
- The <StandardResponse> may be the same as <ActualResponse>, or it may not be.

About the <Transcript>
- The transcript are the collection of utterances from the meeting participants.
- Each utterance is formatted as "Name: Content", where 'Name' is the speaker's name and 'Content' is their spoken text.
- Utterances are in chronological order and may contain typos and grammatical errors.
- The transcript ends at the time stamp when the meeting AI assistant should generate the response.
- Example utterances:
    PERSON1: Hello everyone, I'm glad to see you all here today. (id=0)

About the <ContextInfo>
- <ContextInfo> is a dictionary that contains <Intents> and <Background>.
- <Intents> consists of the questions or topics that can generate the <ActualResponse>.
- <Background> is a list of "Context" and "Information" pairs. For each pair, "Information" can be shared in the "Context" situation to generate the <ActualResponse>. And each pair can be used many times.

Guidelines for Attribution
- You need to decide whether the main points in the <ActualResponse> match the <StandardResponse>.
- The number of main points in the <ActualResponse> is not fixed. PointID is used to identify the main points in the <ActualResponse>.
- When assessing whether the main points in the <ActualResponse> originate from the <Transcript> or the <ContextInfo>, consider the following:
    1. If the main point has a similar or the same meaning as the <ContextInfo>. You should consider it as originating from the <ContextInfo>.
    2. If the main point explicitly repeats or closely relates to any point already mentioned in the <Transcript>. However, casual interactions such as greetings or small talk are permissible and not regarded as sourced from the <Transcript>."
- There are four situations for the origin of the main points in the <ActualResponse>:
    1. The main point in the <ActualResponse> can originate from the <ContextInfo> but is not present in the <Transcript>. You should append [PointID, 1, 0] to the AttributionList.
    2. The main point in the <ActualResponse> does not originate from the <ContextInfo> but originates from the <Transcript>. You should append [PointID, 0, 1] to the AttributionList.
    3. The main point in the <ActualResponse> can originate from both the <ContextInfo> and the <Transcript>. You should append [PointID, 1, 1] to the AttributionList.
    4. The main point in the <ActualResponse> does not originate from the <ContextInfo> and is not present in the <Transcript>. You should append [PointID, 0, 0] to the AttributionList.

Output Format
- The output MUST be in the JSON format.
- You MUST explain the process of attribution for every main point in the <ActualResponse>.
- Note that AttributionList should only contain the List of lists and should not contain any additional information or annotations.

CONTINUE ON THE NEXT PAGE

Table 19: Prompt used for the attribution of the generated response.

- The output MUST include the following fields:
  - Explanation: For every main point in the <ActualResponse>, explain the process of attribution. Especially, explain why the main point matches or does not match the <StandardResponse> and why it originates from the <Transcript> or the <ContextInfo>.
  - AttributionList: A list of lists, where each list contains PointID and the attribution for a main point in the <ActualResponse>.
  - PointsCount: The number of main points in the <ActualResponse>.

Example

```
- Example 1:
<Transcript>
PERSON13: Hi. Hello [PERSON6]. Hello [PERSON19]. Thanks for, uhm. (id=0)
PERSON6: Hi everyone. (id=1)
PERSON19: Hi. (id=2)
PERSON13: Yeah, great. Thanks for joining and, uh, yeah okay. So, yeah. Uh, I I see that
people have written up ehm what they did. (id=3)
PERSON19: Hi [PERSON13], I can hear you. (id=4)
PERSON13: Yeah. [PROJECT3] deliverables. So, I'll try to provide the links-. Or those who
of you, who are already working on the deliverables, please mention that. And yeah. Let's
let's go quickly over what what have done. So [PERSON6] you are the first on the list. Ehm
, ehm, so please briefly update what what you have been working on. And what what is the
plan for the next week. (id=5)

<ContextInfo>
{
    "Intents": [
        "What [PERSON6] has been working on and the plan for the next week?"
    ],
    "Background": [
        {
            "Context": "Update on recent work and plans for the next week",
            "Information": "This week I had fewer tasks. I calculated the word error rate
on Czech transcripts using three versions of Czech ASR created by [PERSON10]. There were
significant mismatches between the golden transcript and its corresponding video. I
conducted testing sessions with [PERSON11] and [PERSON18], which were not successful due
to issues with segmenters from [ORGANIZATION1]. I also contributed to the [PROJECT3]
deliverable for the punctuator and through caser."
        }
    ]
}

<StandardResponse>
["I calculated the word error rate on Czech transcripts"]

<ActualResponse>
["Calculated word error rate on Czech transcripts", "Conducted testing sessions with
PERSON11 and PERSON18"]

<Evaluation>
{
    "Explanation": "1. Calculated word error rate on Czech transcripts. This point matches
  the standard response. "I calculated the word error rate on Czech transcripts" is present
  in the ContextInfo. Therefore, the attribution is [1, 1, 0]. 2. Conducted testing
sessions with PERSON11 and PERSON18. The point does not match the standard response. "I
conducted testing sessions with [PERSON11] and [PERSON18]" is present in the
BackgroundKnowledge. Therefore, the attribution is [2, 1, 0].",
    "AttributionList": [[1, 1, 0], [2, 1, 0]],
    "PointsCount": 2
}
```

Table 20: Prompt used for the attribution of the generated response (continued).

- Your task is to update the summary of the utterances of {participant} in a meeting transcript.
- You are provided with a <Transcript Snippet> that contains a portion of the meeting transcript.
- You are also provided with <Previous Summary> which contains the summary of utterances for {participant} in other parts of the meeting.
- You need to update the <Previous Summary> based on the utterances of the {participant} in the <Transcript Snippet>.

On the provided <Transcript Snippet>
- Transcripts are the collection of utterances from the meeting participants.
- The transcript data is deidentified. Speakers and other named entities are not identified by names, but rather by IDs in the format ENTITYNUMBER (e.g. PERSON1 or PROJECT3) or just ENTITY (e.g. PATH).
- Speaker IDs at the beginning of transcript lines are enclosed in round brackets, all other deidentified entities in square brackets.
- Each utterence ends with "(id=x)", which is the utterance id, an increasing number from 0 to indicate the serial number of utterance in the whole meeting transcript.
- The provided transcript snippet maybe not start from the beginning of the meeting.
- Example utterances:
    (PERSON1) Hello everyone, I'm glad to see you all here today. (id=0)
    (PERSON2) Hi, I'm excited to be here. (id=1)
    (PERSON3) I'm looking forward to the discussion. [PERSON1] mentioned that the project is going well. (id=2)

On the <Previous Summary>
- The <Previous Summary> is a structured summary of the utterances of {participant} in the meeting transcript.
- The <Previous Summary> contains two parts, "wanted information" and "provided information".
    - "wanted information" is a list of questions made by the {participant}.
    - "provided information" is the information provided by the participant to others. It is a list of Context and Information pairs, where the "Context" is the context in which where the {participant} provides the "Information".

Instructions on updating the <Previous Summary>
- Identify the utterances of {participant} in the <Transcript Snippet>.
- If {participant} does not speak in the <Transcript Snippet>, do NOT update the <Previous Summary>.
- Focus on only the informative utterances and ignore the greetings, appreciation, simple acknowledge and other chit chat.
- Extract the "wanted information" and "provided information" from the <Transcript Snippet>.
- You should try to use original utterances as much as possible after removing noise words and polishing them for better readability.
- The second or third personal pronoun (you, he, she, they) in the utterances should be properly replaced with the corresponding participant's ID to avoid ambiguity.
- Use the extracted information to update the <Previous Summary>.
- You can modify the existing "wanted information" and "provided information" or add new information, but do not remove any existing information.
- You MUST NOT mix the information provided by {participant} and other participants while updating the <Previous Summary>.
- You MUST NOT miss any important information provided by {participant} in the <Transcript Snippet>.

Requirement on the output format
- You MUST explain your thoughts and steps of updating the <Previous Summary> before providing the updated summary.
- Output must be in Json format with the "Thoughts" and "Updated Summary" as the key.
- The "Thoughts" is your thoughts and steps of updating the <Previous Summary>.
- The "Updated Summary" contains the updated summary of the utterances for {participant}.

CONTINUE ON THE NEXT PAGE

Table 21: Prompt used for extracting context information.

```
## Example 1
 - Here is an example of updating the utterances summary for PERSON2. You can refer to
this example for better understanding.
 - Suppose the transcript snippet contains the following utterances:
    (PERSON3) Good moring. (id=2)
    (PERSON1) Let's get started with today's meeting on the recent progress of our
software development project. We'll go through updates from each team and discuss any
blockers or issues. [PERSON2], could you start with the development updates?". (id=3)
    (PERSON2) Sure, [PERSON1]. We've made significant progress this sprint. We completed
 the implementation of the new authentication module and integrated it with our existing
systems. (id=4)
    (PERSON1) That's great to hear, [PERSON2]. How about the feature for real-time
notifications? Is it on track? (id=5)
    (PERSON2) Yes, it is. We're about 75% done with it. The core functionality is in
place, and we are now working on optimizing the delivery speed and ensuring it works
seamlessly across different devices. (id=6)
 - Suppose the previous summary of PERSON2 contains the following information:
    {{
      "wanted information": [],
      "provided information": []
    }}
 - The thoughts and updated summary will be:
    {{
      "Thoughts":"In the transcript snippet, PERSON2 responds to PERSON1's questions
about the development updates and the progress of the feature for real-time notifications.
 This information can be added to the "provided information" for PERSON2."
      "Updated Summary":
       {{
         "wanted information": [],
         "provided information": [
           {{
             "Context": "Respond to other participant's question about the development
updates",
             "Information": "We've made significant progress this sprint. We completed
the implementation of the new authentication module and integrated it with our existing
systems."
           }},
           {{
             "Context": "Respond to other participant's question about the progress of
the feature for real-time notifications",
             "Information": "We're about 75% done with it. The core functionality is in
 place, and we are now working on optimizing the delivery speed and ensuring it works
seamlessly across different devices."
           }}
         ]
       }}
    }}

## Example 2
 - Here is another example of updating the utterances summary for PERSON2. You can refer
to this example for better understanding.
 - Suppose the transcript snippet contains the following utterances:
    (PERSON3) Good moring. (id=2)
    (PERSON1) Let's get started with today's meeting on the recent progress of our
software development project. We'll go through updates from each team and discuss any
blockers or issues. [PERSON2], could you start with the development updates?". (id=3)
    (PERSON2) Sure, [PERSON1]. We've made significant progress this sprint. We completed
 the implementation of the new authentication module and integrated it with our existing
systems. (id=4)
```

CONTINUE ON THE NEXT PAGE

Table 22: Prompt used for extracting context information (continued).

(PERSON1) That's great to hear, [PERSON2]. How about the feature for real-time notifications? Is it on track? (id=5)

(PERSON2) Yes, it is. We're about 75% done with it. The core functionality is in place, and we are now working on optimizing the delivery speed and ensuring it works seamlessly across different devices. (id=6)

(PERSON3) [PERSON2], have you had a chance to address the bug I reported last week related to the authentication module? (id=7)

(PERSON2) Yes, [PERSON3]. We identified the root cause of the bug, and it's been fixed. It was due to a conflict with a third-party library we were using. (id=8)

(PERSON3) That's good to hear. Thank you, [PERSON2]. (id=9)

(PERSON1) [PERSON2], for the next step of the project, I'd like you first complete the real-time notifications feature, and then focus on the chatbot development. (id=10)

(PERSON2) Understood, I will do that. (id=11)

(PERSON2) By the way, what's the timeline of our project? (id=12)

(PERSON1) We are aiming to finish the project by the end of August. (id=13)

(PERSON2) Ok, I know. (id=14)

(PERSON1) Let's move to the next topic. [PERSON3], could you provide an update on the testing progress? (id=15)

(PERSON3) Sure. Certainly. We've conducted tests on the new authentication module, and everything looks good so far. (id=16)

(PERSON1) Mhm. (id=17)

(PERSON3) We are now preparing for the testing of the real-time notifications feature. (id=18)

(PERSON2) In our development process, we accumulated some test cases which may help you. (id=19)

(PERSON3) That's helpful, thank you. (id=20)

- Suppose the previous summary of PERSON2 contains the following information:

```
{{
  "wanted information": [],
  "provided information": [
    {{
      "Context": "Respond to other participant's question about the development updates",
      "Information": "We've made significant progress this sprint. We completed the implementation of the new authentication module and integrated it with our existing systems."
    }},
    {{
      "Context": "Respond to other participant's question about the progress of the feature for real-time notifications",
      "Information": "We're about 75% done with it. The core functionality is in place, and we are now working on optimizing the delivery speed and ensuring it works seamlessly across different devices."
    }}
  ]
}}
```

- The thoughts and updated summary will be:

```
{{
  "Thoughts":"In the transcript snippet, the dicussion between PERSON1 and PERSON2 about the progress of the development and the feature for real-time notifications are already included in the previous summary. PERSON2 responds to PERSON3's question about the bug in the authentication module, which can be added to the "provided information" for PERSON2. PERSON2 asks about the timeline of the project, which can be added to the "wanted information" for PERSON2. PERSON2 also comments on PERSON3's statement about the testing progress, offering to provide some test cases, which can be added to the "provided information" for PERSON2."
  "Updated Summary":
    {{
      "wanted information": [
        "What's the timeline of the project?"
      ],
      "provided information": [
        {{
          "Context": "Respond to other participant's question about the bug in the authentication module",
```

Table 23: Prompt used for extracting context information (continued).

```
            "Information": "We identified the root cause of the bug reported by [
PERSON3], and it's been fixed. It was due to a conflict with a third-party library we were
 using."
            }},
            {{
              "Context": "Comment on other participant's statement about the testing
progress",
              "Information": "In our development process, we accumulated some test cases
 which are helpful to testing."
            }}
          ]
        }}
    }}
```

Example 3
- Here is an example of updating the utterances summary for PERSON6. You can refer to
this example for better understanding.
- Suppose the transcript snippet contains the following utterances:
```
    (PERSON13) Hi.
    Hello [PERSON6].
    Hello [PERSON19].
    Thanks for, uhm. (id=0)
    (PERSON6) Hi everyone. (id=1)
    (PERSON19) Hi. (id=2)
    (PERSON13) Yeah, great.
    Thanks for joining and, uh, yeah okay.
    So, yeah.
    Uh, I I see that people have written up ehm what they did. (id=3)
    (PERSON19) Hi [PERSON13], I can hear you. (id=4)
    (PERSON13) Yep, that's great.
    Uh, and also you were evaluating-.
    Yes, so that's that's re re record.
    What you did.
    So what I have, uh, on my mind now is uh, uh, well, uh, preparations.
    So, uh, [PERSON13], uh I am busy, uh, with the IW SLT, uh, write-up.
    Uh, that was the, uh, the wra last part that I did.
    Now busy with interviewing people people to uh to replace those who are em moving
forward <laugh/> so to say.
    So there is number of colleagues on projects that I am supervising, uh, that who are
 going for studies abroad and other things.
    Uh, so, uh, what I think we should focus on is the demo for Project Officer.
    Then we need to focus on the ladder climbing, uh, which is building uh, uh, [PROJECT
3] test set plus, uh, regularly, uh, testing on it.
    Ehm, and, ehm what else, uh, the deliverables.
    Yeah.
    [PROJECT3] deliverables.
    So, I'll try to provide the links-.
    Or those who of you, who are already working on the deliverables, please mention
that.
    And yeah.
    Let's let's go quickly over what what have done.
    So [PERSON6] you are the first on the list.
    Ehm, ehm, so please briefly update what what you have been working on.
    And what what is the plan for the next week. (id=5)
    (PERSON6) <other_noise/>
    So, luckily.
    <laugh/>
    Not luckily but this week I had like quite less tasks to do.
    So first I calculated the word error rate on Czech transcripts using that three
versions of, uh, Czech ASR which [PERSON10] created.
    And so yesterday [PERSON10] told me that they were, uh, and the golden transcript
and its corresponding video there were there were huge huge mismatch.
    And I <unintelligible/> and he said to update me.
```

CONTINUE ON THE NEXT PAGE

Table 24: Prompt used for extracting context information (continued).

And then we conducted a few testing sessions with [PERSON11] and [PERSON18].
And they were not quite successful because the segmenters from, uh, uh, [ORGANIZATION1] they were still down and [PERSON12] today he is he is working on them. (id=6)
(PERSON13) Mhm. (id=7)
(PERSON6) And lastly yeah I think I did-.
Uh, it was the input in the [PROJECT3] deliverable of for the punctuator and through caser.
(id=8)
(PERSON13) Mhm, yeah. (id=9)
(PERSON6) And I don't have-.
I think that apart from the testing sessions to do this week so I am waiting for new tasks. (id=10)
(PERSON13) Yeah, so.
So the word error rate, there is also the English, uh, transcripts?
Ehm, and also we should have from [PERSON9] the German one, right?
So. (id=11)
(PERSON6) Yeah, yeah, yeah. (id=12)
(PERSON13) So, so I will make it to do-. (id=13)
(PERSON6) So I have updated the German transcripts in its corresponding path and like do we have the golden transcripts for the English videos? (id=14)
(PERSON13) Yes, that's the other part.
Because this is the consecutively translated videos.
So there is always the English speaker and then the Czech speaker who repeats the same content.
And [PERSON7] has split the video and while the English part should be more reliable, uh, the Czech part has been done simply by using the other ends.
So the Czech video has been cut using the English time stamps.
Like the end of English and the beginning of the next English segment.
Uh, so it's like like interleave the the the other way round.
So that's why I'm not surprised that the Czech video is, uh, imprecise in timing.
But still, I was not expecting it to be that bad.
So, uh, that is something that, yeah.
[PERSON10], can you maybe tell us more details about that? (id=15)
(PERSON10) Yeah, yeah.
Well, I just like listened to the audio and followed the talk transcript
and it was completely off.
I think it is-.
There must be some miss-match because-. (id=16)
(PERSON13) Mhm. (id=17)
(PERSON10) Yeah, yeah.
The transcription is for the completely different audio than it's in the subdirectory. (id=18)
(PERSON6) Mhm. (id=19)
(PERSON13) Oh, so then someone must have like messed it up. (id=20)
(PERSON10) Yeah, yeah, I I may maybe it's just like uh.
Maybe the files are just switched between the subdirectories? (id=21)
(PERSON13) Mhm. (id=22)
(PERSON10) I I haven't checked but-.
Uh, yeah, there is some some serious mismatch there. (id=23)
(PERSON13) Yeah, so [PERSON10] can you coul could you do this check?
It should not be hard
Like try listening to all the files that are within this demo for [PERSON15], uh, and try to locate the correct file, the appropriate files.
But we should have, we should have the transcripts ready for all of those.
So we should be able to, uh, to evaluate it.
And also for the English ones we have the translations.
So for the English ones [PERSON6], uh, I would like you to evaluate not only the word error rate of the ASR.
But also the machine translation quality or at the SLT even.
Uh, with the translation quality into German and Czech.
Both are available. (id=24)
(PERSON6) Okay. (id=25)

Table 25: Prompt used for extracting context information (continued).

```
        (PERSON13) We have these files ready. (id=26)
        (PERSON6) And so these, for like a German audio and English [PROJECT1]. (id=27)
        (PERSON13) English, uh, input for English sound.
        We have the golden English transcript, so you can check the ASR.
        And we also have the translation into Czech and into German.
        So you can also evaluate directly the translation quality, uh, of that. (id=28)
        (PERSON6) Okay, yeah. (id=29)
        (PERSON13) Yeah, so this is, this is an important, uh, task, uh, to do, uh wr also
for German and English audios.
        And another to do, uh, bleu or SLTF, uh, for, uh, German and Czech translations of,
uh, English. (id=30)
  - Suppose the previous summary of PERSON6 contains the following information:
      {{
        "wanted information": [],
        "provided information": []
      }}
  - The thoughts and updated summary will be:
      {{
        "Thoughts":"PERSON6 responds to PERSON13's questions about the work done and the
plan for the next week, which can be added to the 'provided information' for PERSON6.
PERSON6 responds to PERSON13's questions about the word error rate for the English and
German transcripts, which can be added to the 'provided information' for PERSON6. PERSON6
also asks about the golden transcripts for the English videos, which can be added to the '
wanted information' for PERSON6."
        "Updated Summary":
          {{
            "wanted information": [
              "Do we have the golden transcripts for the English videos?"
            ],
            "provided information": "wanted information": [
              "Do we have the golden transcripts for the English videos?"
            ],
            "provided information": [
              {{
                "Context": "Respond to other participant's question about the work done
and the plan for the next week",
                "Information": "This week I had quite less tasks to do. So first I
calculated the word error rate on Czech transcripts using three versions of Czech ASR
created by [PERSON10]. And so yesterday [PERSON10] told me that there were huge huge
mismatch between the golden transcript and its corresponding video. And he said to update
me. And then we conducted a few testing sessions with [PERSON11] and [PERSON18]. And they
were not quite successful because the segmenters from [ORGANIZATION1] were still down and
[PERSON12] today is working on them. And lastly, I think I did the input in the [PROJECT3]
 deliverable for the punctuator and through caser. Apart from the testing sessions to do
this week so I am waiting for new tasks."
              }},
              {{
                "Context": "Respond to other participant's question about the word error
rate for the English and German transcripts",
                "Information": "I have updated the German transcripts in its corresponding
 path, and I don't konw if we have the golden transcripts for the English videos."
              }}
            ]
          }}
      }}

  ## Note
  - You MUST follow the instructions and examples provided.
  - Similar to examples above, you should try to use original utterances as much as
possible after removing noise words and polishing them for better readability.
  - You MUST NOT put the information provided by other participants or questions of other
participants in the updated summary of {participant}.
  - You MUST NOT miss any important information provided by {participant} in the <
Transcript Snippet>.
  - You MUST give the output in the required format.
```

Table 26: Prompt used for extracting context information (continued).

- You are an NLP expert agent tasked with generating an evaluation dataset to assess {person_id}'s response abilities in the categories of 'Chime In', 'Explicit Cue', 'Implicit Cue', based on the provided transcript.
- The conversation may involve multiple speakers, but your focus should solely be on {person_id}.
- Given the transcript contains lengthy utterances, selectively include only the highest quality exchanges in the evaluation dataset.
- Exclude chit-chat or unmeaningful utterances such as ["emm," "okay," "mhm," "uh-huh," "yeah," "oh," "right," "hmm"] from the evaluation dataset.
- Ensure that {person_id}'s responses are substantive and meaningful. Exclude responses from {person_id} that are simple acknowledgments or confirmations like "Yeah, yeah, definitely, yeah" or "Okay."
- The transcript data is deidentified. Speakers and other named entities are identified by IDs in the format ENTITYNUMBER (e.g., PERSON1, Speaker1 or PROJECT3) or simply as ENTITY (e.g., PATH).

Evaluation Type
- Chime In: When {person_id} spontaneously contributes to the conversation without being directly prompted.
- Usually Chime In is when {person_id} is not already engaged in the conversation but chimes in with a relevant comment or question.
- Explicit Cue: When {person_id}'s name is specifically mentioned by another Speaker in utterance with ID, then {person_id} responds to a clear and direct question or prompt towards {person_id}.
- Implicit Cue: When {person_id}'s name is not specifically mentioned by Speaker in utterance with ID, but {person_id} responds to a less direct prompt or follows up on information that suggests a response is needed.
- Usually Implicit Cue is when {person_id} is already engaged in the conversation and responds to a follow-up question from Speaker in utterance with ID.

Output Format
- Output must be in Json format. Here is the skeleton of the output format with explanation:
- Explanation: Your reason for selecting the evaluation instance and for categorizing it.
- Type: The category of the evaluation instance: 'Chime In', 'Explicit Cue', 'Implicit Cue'.
- Response IDs: The id or ids of the {person_id}'s response from the transcript. Include all Response IDs that are relevant to the evaluation instance. If there are multiple Response IDs, separate them with commas.
- ID: The utterance id that {person_id} responds to.
- Speaker: The speaker of the utterance with the ID.
- Maintain the chronological order of the transcript when generating the evaluation dataset. ID MUST precede Response IDs.
- Response IDs must be from {person_id}'s responses only and ID must be from the speaker's utterance that {person_id} responds to.
- Please return all suitable evaluation instances in the transcript. If you don't find any suitable instances for a category, you can leave the evaluation dataset empty. Please ensure you have thought through the transcript carefully before leaving the evaluation dataset empty.

Example: Below are two examples of transcript and the corresponding evaluation datasets generated to assess PERSON18's response abilities. You can refer to these examples when generating {person_id}'s evaluation dataset.

<Transcript>
"speaker": "Speaker 19", "content": "If you want, I can resend it again. (id=71)"
"speaker": "Speaker 13", "content": "Space tokeniser. <unintelligible/> Yes, so es essentially to answer your question in the email. We have to switch to and we have for the IWSLT. We have to switch to SacreBLEU and SacreBLEU does its own tokenisation before scoring. So there is no-. Let's let's simply forget NLTK bleu score. That is not reliable. (id=72)"
"speaker": "Speaker 18", "content": "Yes, but-. (id=73)"
"speaker": "Speaker 19", "content": "Yes, but we can combine our tokeniser with NLTK. (id=74)"
"speaker": "Speaker 13", "content": "Uf. Let's not do that. Let's just forget it. Let's let's just use SacreBLEU. (id=75)"
"speaker": "Speaker 19", "content": "Okay. (id=76)"
"speaker": "Speaker 18", "content": "I I have one comment about it. (id=77)"
"speaker": "Speaker 13", "content": "Mhm. Yeah. (id=78)"
"speaker": "Speaker 18", "content": "You sh should use tokeniser before enverse segmenter. (id=79)"
"speaker": "Speaker 13", "content": "Yes, that's it. Yeah. (id=80)"
"speaker": "Speaker 18", "content": "Because it's much better. Because it can rely on the on the dots and commas and question marks and so on. And you can you can check my script which does tokeniser, enverse segmenter and then de-tokeniser. And here is the path in the document.And-. (id=81)"
"speaker": "Speaker 13", "content": "Yeah. (id=82)"
"speaker": "Speaker 18", "content": "And it's it's using the Moses seg tokeniser and detokeniser. And it needs the the language tag as the first argument and then reference. (id=83)"
"speaker": "Speaker 13", "content": "Yeah. So [PERSON19], do you do you fo-? Do you understand? (id=84)"
"speaker": "Speaker 19", "content": "Yeah. (id=85)"
CONTINUE ON THE NEXT PAGE

Table 27: Prompt used for extracting test cases from meeting transcript.

```
<Evaluation Dataset>
[
    "Explanation": "PERSON18's utterance is informative and suitable for evaluation
dataset. PERSON18 spontaneously contributes to the conversation without being directly
prompted. This is a Chime In instead of Implicit Cue since [PERSON18] is not already
engaged in the conversation.",
    "Type": "Chime in",
    "Response ID": [77, 79, 81, 83]
    "ID": 76,
    "Speaker": "Speaker 19",
]
```

Example 2
<Transcript>
"speaker": "Speaker 13", "content": "Yes, so I need to review these and the internal deadline is in 6 days from now. Uh, so, hopefully I will get back to all of you. To each of you independently towards the end of the week if there is anything unclear. So that we meet the internal deadline on the 8th. Yeah, okay. Great. Uh. So [PERSON18], what what is your progress? (id=117)"
"speaker": "Speaker 18", "content": "Hmhm. Yes, and by reading the papers I found an interesting tool. (id=118)"
"speaker": "Speaker 13", "content": "Mhm. (id=119)"
"speaker": "Speaker 18", "content": "I found out that it's possible to measure out the speech rate by cutting the syllables. And there is one tool. One patent tool, which can detect the gender of speaker and the speech rate. (id=120)"
"speaker": "Speaker 13", "content": "Mhm. (id=121)"
"speaker": "Speaker 18", "content": "And some other characteristics. So we can try it and make a dashboard out of it. (id=122)"
"speaker": "Speaker 13", "content": "Mhm. That's that's useful thing. Uh, and later on we could even create models like-. If we if we recognise that someone is speaking too fast, we could use like a harsher summarisation. (id=123)"
"speaker": "Speaker 18", "content": "Yes. (id=124)"
"speaker": "Speaker 13", "content": "So we could be reducing reducing their speech mole with a different model. (id=125)"
"speaker": "Speaker 18", "content": "Yes, and there was also speech modes. Like whether it was angry or normal and so on. (id=126)"
"speaker": "Speaker 13", "content": "Mhm. (id=127)"
"speaker": "Speaker 18", "content": "But I have no idea how the tool works in practice. I I I I saw it only in Gi GitHub and I buy it. (id=128)"
"speaker": "Speaker 13", "content": "Yeah, uh. (id=129)"
"speaker": "Speaker 18", "content": "So we can try it and make a dashboard out of it. (id=130)"

```
<Evaluation Dataset>
[
  "Explanation": "PERSON18's utterance is informative and suitable for evaluation dataset.
 PERSON18 was directly prompted by Speaker 13. This is an Explicit Cue.",
  "Type": "Explicit Cue",
  "Response ID": [118, 120, 122, 124, 126, 128, 130]
  "ID": 117,
  "Speaker": "Speaker 13",
]
```

- Please refer to the examples provided to ensure consistency and coherence in generating the evaluation dataset. The evaluation dataset must be in json format.

Table 28: Prompt used for extracting test cases from meeting transcript (continued).