M3-SLU: Evaluating Speaker-Attributed Reasoning in Multimodal Large Language Models

Yejin Kwon*, Taewoo Kang*, Hyunsoo Yoon†, and Changouk Kim†
Department of Industrial Engineering, Yonsei University
Seoul, Republic of Korea
{beckykwon, hs.yoon, kimco}@yonsei.ac.kr, gangtaeu02@gmail.com

Abstract

We present M3-SLU, a new multimodal large language model (MLLM) benchmark for evaluating multi-speaker, multi-turn spoken language understanding. While recent models show strong performance in speech and text comprehension, they still struggle with speaker-attributed reasoning, the ability to understand who said what and when in natural conversations. M3-SLU is built from four open corpora (CHiME-6, MELD, MultiDialog, and AMI) and comprises almost 12K validated instances with paired audio, transcripts, and metadata. It includes two tasks: (1) Speaker-Attributed Question Answering and (2) Speaker Attribution via Utterance Matching. We provide baseline results for both cascaded pipelines and end-to-end MLLMs, evaluated using an LLM-as-Judge and accuracy metrics. Results show that while models can capture what was said, they often fail to identify who said it, revealing a key gap in speaker-aware dialogue understanding. M3-SLU offers as a challenging benchmark to advance research in speaker-aware multimodal understanding. 🤗 M3-SLU-Task1 & 🤗 M3-SLU-Task2.

Keywords: Multi-Speaker Spoken Language Understanding, Speech-LLM Benchmark, Evaluation

1. Introduction

Multimodal Large Language Models (MLLMs) have begun to blur the boundary between modalities, that is, seeing, hearing, and reasoning. Advances in models such as Qwen3-Omni (Xu et al., 2025), Gemini 2.5 (Comanici et al., 2025), and GPT-5 (Wang et al., 2025b) have demonstrated how far AI systems can extend their understanding beyond text, seamlessly integrating visual, auditory, and linguistic cues to perceive the world in richer ways. Particularly, Audio-Language Models (ALMs) that integrate auditory representations into large language models, including Qwen2-Audio (Chu et al., 2024), Audio Flamingo (Goel et al., 2025), and Voxtral (Liu et al., 2025), have recently achieved remarkable progress in bridging speech and language understanding. These models enable machines not only to transcribe speech but also to understand intent, summarize dialogues, and generate coherent responses.

Despite their impressive multimodal capabilities, most existing MLLMs still assume single-speaker conditions, achieving strong performance in Automatic Speech Recognition (ASR) but leaving Speaker Diarization (SD) largely unaddressed (Yin et al., 2025). Yet, real-world conversations are far more complex than such single-speaker settings. Understanding “who spoke when and what” offers a more comprehensive and meaningful perspective on real-world conversations (Gao et al., 2025). In particular, in natural interactions, utterance sequences may be immediately continuous or may overlap temporally; overlapping talk often occurs around turn transitions and is characterized by backchannels, interruptions, and simultaneous first-starts (Levinson and Torreira, 2015; Knudsen et al., 2020; Schegloff, 2000).

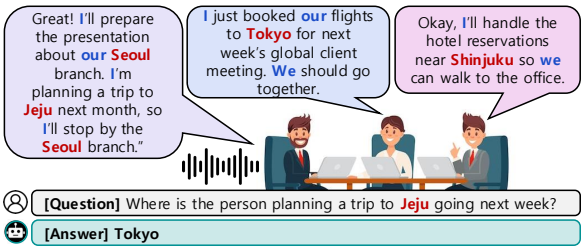

Figure 1: Overview of the M3-SLU Benchmark.

Understanding "Who spoke When and What" in multi-party conversations is a crucial step toward socially intelligent AI. However, most existing speech benchmarks (Chen et al., 2024; Yang et al., 2024; Sakshi et al., 2024) address speaker-related and general dialogue tasks together, without separately examining the distinct challenges of speaker-centric understanding in real conversations (Wang et al., 2025c). To close the gap between current MLLM evaluation and real-world conversational complexity, we introduce the **M3-SLU Benchmark** (**M**ulti-Speaker, **M**ulti-Turn, and **M**ulti-Modal **S**poken **L**anguage **U**nderstanding), as illustrated in Figure 1. Our key contributions:

- We constructed M3-SLU benchmark using four open multi-speaker corpora — **CHiME-6** (Watanabe et al., 2020), **MELD** (Poria et al., 2019), **MultiDialog** (Park et al., 2024), and **AMI** (Kraaij et al., 2005), reflecting diverse

*These authors contributed equally.
†Corresponding authors.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5722–5736
11–16 May 2026.

Benchmarks	SLURP	VoiceBench	MMSU	MMAU	AudioBench	MSU-Bench	**M3-SLU (Ours)**
Speaker-Oriented	X	X	X	X	X	O	**O**
Multi-speaker	X	X	X	X	X	O	**O**
Speaker-attributed Reasoning	X	X	X	X	X	△	**O**
Audio Source	TTS+RPC	TTS+RPC	RPC	RPC	RPC	RPC	**RPC**
Conversation Type	Monologue	Monologue	Dialogue	Dialogue	Dialogue	Dialogue	**Dialogue**
Conversation Length	Short	Short	Short	Short	Short	Short	**Long (Over 1 Min)**

Table 1: Comparison between M3-SLU and existing speech-language understanding benchmarks.

acoustic conditions and conversational patterns such as overlaps and rapid turns.

- We propose the M3-SLU benchmark and **evaluation framework for MLLMs**, designed to measure performance on **two simple yet challenging tasks that can only be solved by correctly identifying the speaker**.

2. Related Work

2.1. Speech Understanding Models

Speech understanding has advanced beyond mere transcription toward comprehension of spoken language – capturing both meaning and paralinguistic cues such as prosody and tone. Earlier pipeline systems that linked ASR to NLP models often lost acoustic detail and propagated recognition errors, prompting a shift toward end-to-end architectures that map raw audio directly to semantic representations. This architectural shift has enhanced robustness and enabled deeper speech understanding, as shown by recent models such as SpeechGPT (Zhang et al., 2023), Salmonn (Tang et al., 2023; Yu et al., 2025), GLM-4-voice (Zeng et al., 2024), and Gemini (Team et al., 2023). These models typically follow two main approaches: (1) using an audio adaptor, as in Audio Flamingo 3 (Goel et al., 2025) and Voxtral (Liu et al., 2025), or (2) directly combining an audio encoder with an LLM, as in Qwen2-Audio (Chu et al., 2024).

This paradigm has also paved the way for the emergence of more specialized capabilities, such as the Speaker LM (Yin et al., 2025), which captures individual vocal signatures and uses them in reasoning, and the MT-LLM (Meng et al., 2025), designed to disentangle and process dialogue from concurrent speakers. Looking forward, the frontier of research is expanding into multimodal domains where models like GPT-5 (Wang et al., 2025b) and Qwen2.5-Omni (Xu et al., 2025) fuse auditory streams with visual data to achieve a contextually richer, more human-aligned interaction.

2.2. Speech Understanding Benchmarks

As the performance of Large Audio-Language Models (LALMs) has advanced, developing benchmarks that evaluate their complex speech understanding capabilities has become a major research focus. Early benchmarks in Table 1 such as SLURP (Bastianelli et al., 2020) and VoiceBench (Chen et al., 2024) primarily focused on single-turn, single-speaker intent classification using synthetic or short speech segments (TTS + RPC), laying the groundwork for fundamental speech-language understanding. Subsequent datasets including MMAU (Sakshi et al., 2024), MMSU (Wang et al., 2025a), and AudioBench (Wang et al., 2024) expanded their evaluation scope to multi-turn audio-based question and tasks such as intent classification and emotion interpretation, aiming to assess more nuanced aspects of speech understanding.

However, these benchmarks, while often built from long recordings, evaluated only short conversational segments (typically under 30 seconds) and were mainly oriented toward intent, emotion, or speaker recognition rather than context-grounded reasoning. Moreover, they did not address multi-speaker scenarios where understanding requires reasoning over speaker identities and interactions.

2.3. Multi-speaker Speech Understanding Benchmarks

As summarized in Table1, MSU-Bench (Wang et al., 2025c) marked the first benchmark specifically dedicated to the rigorous evaluation of multi-speaker understanding in realistic conversational scenarios. Yet, it still focused on short dialogues and failed to capture reasoning that spans multiple turns and speakers. Building on this trajectory, we propose the **M3-SLU** benchmark, designed to assess long-form conversational understanding in segments over one minute long. Unlike prior benchmarks, M3-SLU comprises real multi-speaker dialogues lasting between one and three minutes and features speaker-attributed question answering tasks that require identifying concrete nouns (e.g., objects, places, times, numbers, names) in-

TTS (Text-to-Speech): Speech audio is synthetically generated from written text using a text-to-speech engine.

RPC (Real/Recorded Speech Corpus): Speech audio is naturally recorded from human speakers in real environments.

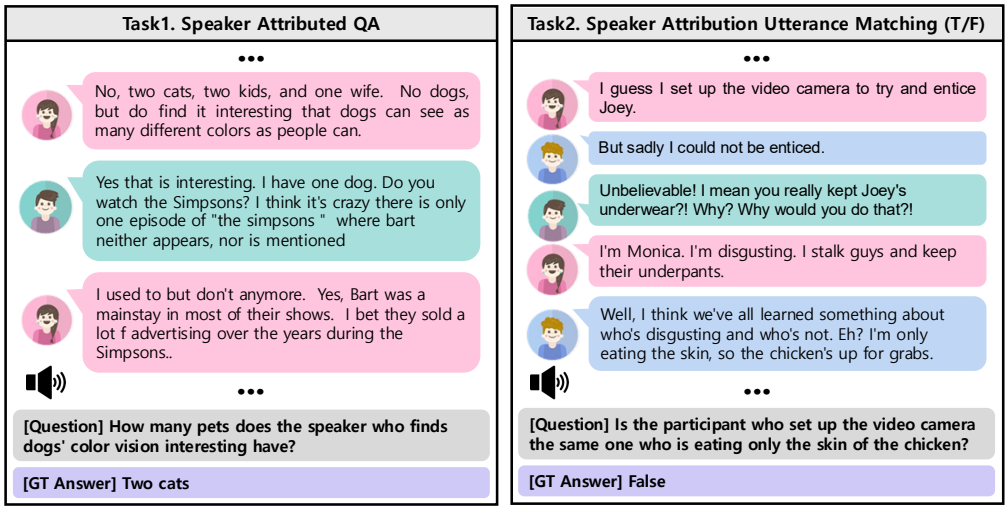

Figure 2: Example of M3-SLU Benchmark. Although they are presented in text form, they are all spoken conversation datasets. The generated Question and GT Answer are in text format.

stead of abstract intents or emotions.

3. M3-SLU Benchmark

3.1. Purpose of Benchmark

Our M3-SLU benchmark is designed to measure how accurately a model can understand complex multi-speaker conversations and provide appropriate answers to related questions. Therefore, **we designed tasks that require the MLLM models to listen to the conversation and distinguish between speakers in order to answer correctly**. As illustrated in Figure 2, we propose two tasks: **Task 1. Speaker-Attributed QA** and **Task 2. Speaker Attribution Utterance Matching (T/F)**.

3.2. Overview of Benchmarks

M3-SLU consists of two core evaluation tasks generated from four public multi-speaker dialogue datasets — CHiME-6 (Watanabe et al., 2020), MELD (Poria et al., 2019), MultiDialog (Park et al., 2024), and AMI (Kraaij et al., 2005). **CHiME-6** consists of real dinner-party recordings captured in noisy environments with overlapping speech and distant microphones, making it ideal for evaluating speech robustness and diarization accuracy. **MultiDialog** covers multi-topic conversations designed for contextual understanding across diverse domains. **MELD**, based on the Friends TV series, provides multimodal emotional dialogues that emphasize emotion recognition and sentiment analysis. Finally, **AMI** includes real business meeting recordings commonly used for summarization and decision-making tasks, capturing realistic multi-speaker interactions in professional settings. (Details and statistics for each dataset are provided in Appendix A.)

	CHIME-6	MELD	MultiDialog	AMI
Total (h)	19.76	2.74	70.04	38.56
Seg. Dur (s)	116.81	56.12	97.92	126.64
Utt./Seg	66.40	13.92	15.51	87.70
# of Speakers	2–4	2–8	2	2–5

Table 2: Segment-level statistics of datasets used in M3-SLU, including total audio duration, average segment length, average number of utterances per segment, and the number of speakers.

Dataset	Task 1 (Q&A)	Task 2 (T/F)	Total
CHiME-6	532	1,006	1,538
MultiDialog	3,257	3,791	7,048
MELD	51	100	151
AMI	1,086	2,131	3,217
Total	**4,926**	**7,028**	**11,954**

Table 3: Composition of the M3-SLU Benchmark

Task 1 (QA) and Task 2 (T/F) were designed from approximately 8k multi-speaker dialogue segments, each longer than one minute. As shown in Tables 2 and 3, we finalized 11,954 challenging data instances (segments) for the benchmark, and each instance involves at least two speakers. And, every data instance must consist of at least two speakers. The two proposed tasks are as follows:

- **Task 1. Speaker-Attributed QA**: Evaluates a model's ability to extract concise noun-phrase answers from conversations, testing how well it links information to the correct speaker.
- **Task 2. Speaker Attribution Utterance**

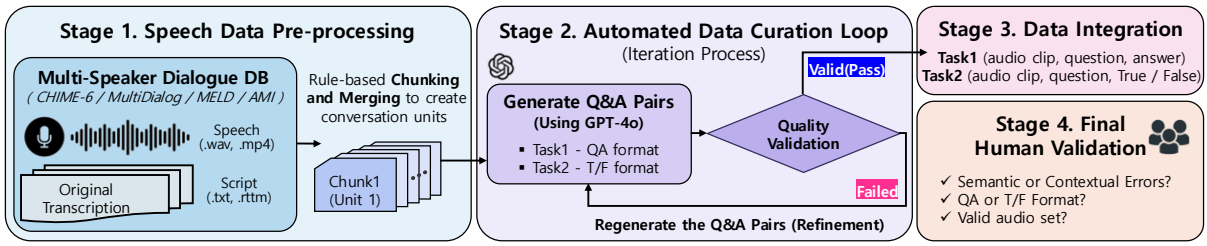

Figure 3: The 4-Stage Hybrid Pipeline for M3-SLU Benchmark Construction.

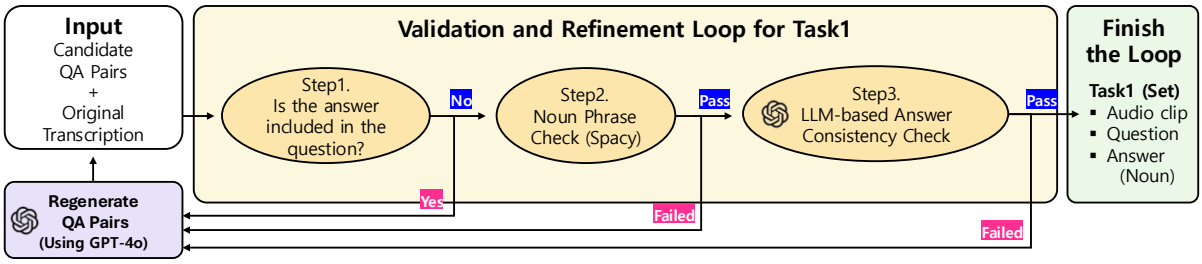

Figure 4: Detailed Pipeline for Validation and Refinement for Task 1 (QA)

Match (T/F): Evaluates reasoning about speaker identity by checking if two utterances or actions were made by the same person.

3.3. Benchmark Construction Pipeline

The 4-stage hybrid benchmark construction pipeline was implemented as shown in Figure 3, in which LLM-driven automatic screening and human review were jointly employed to maximize the reliability and accuracy of the generated dataset.

3.3.1. Stage 1. Speech Data Pre-processing

To ensure that the benchmark embodies its core characteristics of multi-speaker and multi-turn conversations, we carefully selected and collected high-quality public datasets (CHIME-6, MultiDialog, MELD, and AMI). Long dialogue recordings, mostly around or over one hour in length, were segmented into semantically coherent conversation units, each lasting between one and 3.5 minutes. Only samples containing two or more speakers were retained, resulting in **refined long multi-speaker conversation chunks** that serve as the raw materials for the second-stage processing engine.

3.3.2. Stage 2. Automated Data Curation Loop

This critical engine stage manages data quality systematically by cycling through **Generation → Validation → Refinement**. Preprocessed conversation units are passed to LLM (GPT-4o) to generate task-specific data, immediately followed by automated verification and refinement cycles.

Generation. For both Task 1 and Task 2, around 8,000 multi-speaker conversation chunks are preprocessed. The original transcripts (ground-truth scripts) of these chunks were provided to GPT-4o (Hurst et al., 2024), which automatically generated corresponding question and answer pairs: short noun-phrase answers for Task 1 and True/False statements for Task 2. For each task, a small set of manually created QA examples was also supplied as few-shot guidance.

Validation and Refinement in Task 1. After the initial generation stage, an iterative validation and refinement loop in Figure 4 was applied to ensure the factual accuracy and linguistic precision of the generated QA pairs. Each iteration involved feeding the original ground-truth scripts and generated QA outputs back into GPT-4o for self-evaluation and regeneration. Through this process, only samples that consistently produced clear, contextually relevant, and noun-phrase-based answers were retained across the entire dataset.

As shown in Figure 4, **Step 1** checks whether the answer is redundantly included in the question. **Step 2** verifies the linguistic validity of the answer as a noun phrase using the SpaCy library. **Step 3** conducts an LLM-based consistency check, in which GPT-4o evaluates whether the generated QA pair is contextually appropriate given the original transcript and whether the provided answer correctly responds to the question. If a candidate pair fails any step, new QA pairs are regenerated by GPT-4o, and the entire validation loop is repeated until all conditions are satisfied.

The iteration stopped after six cycles because the proportion of samples passing Step 2 stayed

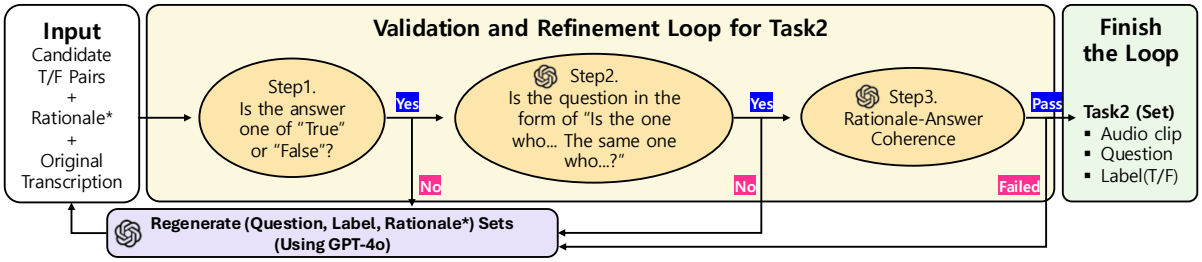

Figure 5: Detailed Pipeline for Validation and Refinement in Task 2(T/F)

around 78%, showing no further improvement. Conversation chunks that repeatedly failed Step 2 were deemed unsuitable for QA generation, as their original scripts lacked extractable noun-phrase answers. These samples were therefore excluded from the dataset. Consequently, the remaining 5,531 candidates advanced to human validation.

Validation and Refinement in Task 2. Unlike Task 1, Task 2 additionally required GPT-4o to produce a rationale explaining the reasoning behind each True/False(boolean type) answer. This rationale was used temporarily during the validation process to assess the consistency of each candidate pair, though it was not included in the final benchmark.

The verification process for Task 2 (True/False reasoning) followed the validation and refinement loop illustrated in Figure 5. Each candidate set (question, Boolean label, and rationale) underwent a three-step validation loop. **Step 1** verified whether the label was a valid Boolean ("True" or "False"). **Step 2** checked whether the question conformed to the required speaker-attributed form (e.g., "Is the one who … the same one who … ?"), ensuring that the reasoning explicitly involved speaker identity. **Step 3** evaluated the coherence between the rationale and the answer, filtering out logically inconsistent or semantically irrelevant cases.

Samples that failed any step entered the regeneration step, in which GPT-4o regenerated the candidate set. Each regenerated set was re-validated through the same three steps. After repeating the loop three times, we observed no further increase in the proportion of valid samples. Thus, we obtained a final set of 8,020 True/False labeled question pairs, each aligned with its corresponding audio clip and ready for subsequent human verification.

3.3.3. Stage 3. Data Integration

Text-based data pairs (QA, T/F) that pass the automated stages are matched to corresponding original audio clips, packaging both text and speech information into complete multimodal datasets for human annotators.

3.3.4. Stage 4. Final Human Validation

The final and most critical stage ensured the overall integrity and contextual quality of both tasks. In this phase, human annotators manually inspected every remaining instance to identify subtle semantic or contextual errors that automated processes could not capture. For **Task 1**, annotators carefully examined each question–answer pair to remove cases where the generated noun-phrase answers were vague or linguistically invalid (e.g., phrases like "that thing" or "something"), as well as instances where the question already contained the answer, rendering the question meaningless. For **Task 2**, reviewers focused on verifying speaker attribution and contextual alignment, discarding samples in which the True/False reasoning relied on incorrect or ambiguous speaker identification.

4. Evaluation of M3-SLU Benchmark

Since Task 1 checks noun-phrase matching (QA) and Task 2 judges whether two utterances or actions were made by the same speaker (True/False), we propose distinct evaluation methods for assessing an MLLM's speaker-attributed reasoning ability.

Task 1 (QA) Evaluation Traditional QA evaluation metrics typically rely on Exact Match (EM) and token-level F1 scores. The EM score is the percentage of predictions that match any one of the ground truth answers exactly. The F1 score measures the average overlap between the prediction and ground truth answer (Rajpurkar et al., 2016). However, these metrics assume textual inputs and do not account for variations arising from speech or pronunciation differences.

Therefore, a new evaluation metric was required for Task 1, which involves listening to speech and identifying the words that appear in the conversation. To address this, we adopted an LLM-as-a-Judge approach, in which GPT-4o evaluated model outputs by considering semantic similarity and phonetic plausibility, a strategy inspired by the evaluation framework in AudioBench (Wang et al., 2024). As shown in Table 5, this allowed for mi-

Model (SD + ASR)	CHiME-6		MELD		MultiDialog		AMI	
	WER	cpWER	WER	cpWER	WER	cpWER	WER	cpWER
Pyannote + Whisper-Medium	0.631	0.712	0.601	0.712	0.356	0.335	0.472	0.451
Pyannote + Whisper-Large	0.635	0.713	0.604	0.707	0.391	0.354	0.487	0.478
DiariZen + Whisper-Medium	0.631	0.601	0.600	0.581	0.355	0.162	0.472	0.377
(Closed SDR) AssemblyAI	0.532	0.631	0.509	0.678	0.236	0.157	0.531	0.472
(Closed SDR) Google STT*	0.710	0.720	0.545	0.662	0.394	0.542	0.552	0.450

Table 4: Comparison of WER and cpWER across four dialogue audio sets. Both metrics indicate better performance with lower scores. Experiments marked with an asterisk (*) were conducted on 500 randomly sampled instances from each dataset due to research resource limitations.

nor pronunciation- or transcription-related variations—such as "NITE XML" vs. "Night XML"—to be accepted as correct, ensuring a more human-aligned and speech-aware assessment of answer quality. We conducted an LLM-as-a-Judge evaluation using prompts similar to the prompt of (Badshah and Sajjad, 2024). The final evaluation score was defined as the proportion of samples that GPT-4o (LLM-as-Judge) evaluated as correct among all instances.

$$\text{Score}_{\text{Task1}} = \frac{N_{\text{Correct}}^{\text{GPT-4o}}}{N_{\text{Total}}} \quad (1)$$

Case	EM	F1 Score	LLM-as-Judge
(GT) *NITE XML*	-	-	-
Nite xml	Incorrect	Partial	Correct
Night xml	Incorrect	Partial	Correct
Nite x-m-l	Incorrect	Incorrect	Correct

Table 5: Comparison of Evaluation Metrics

Task 2 (T/F) Evaluation For Task 2, each prediction was evaluated using the standard **Accuracy** metric, which measures the proportion of correctly classified True/False labels.

5. Experiments and Results

First, we conducted a Speaker Diarization and Recognition (SDR) Test to verify whether the audio clips in M3-SLU can be accurately transcribed with correct speaker attribution by existing models. Then, we evaluated the M3-SLU benchmark to assess the capability of current E2E MLLMs and cascaded SD + ASR + LLM pipelines in performing speaker-attributed reasoning across multi-speaker dialogues.

5.1. Experiment Setting for SDR Test

Following prior work (Yin et al., 2025), we evaluated cascade SD+ASR pipelines on our audio clips in M3-SLU. In particular, we used **Pyannote 3.1** (Bredin et al., 2020) and **DiariZen** (Han et al., 2025) as speaker diarization (SD) modules, both widely recognized for their strong performance on English conversational audio. These were combined with **Whisper** (Radford et al., 2023) models of varying sizes (Medium and Large) for ASR. To further compare with end-to-end commercial systems, we also included two proprietary SDR models, **AssemblyAI** and **Google STT**.

We also measured two metrics, average of **WER** and average of **cpWER** for audio clips, to assess whether our benchmark achieves proper SDR performance with existing models. WER (Word Error Rate), commonly used in ASR assessment, measures the proportion of word errors between reference and predicted transcripts (Morris et al., 2004). And cpWER (concatenated minimum-permutation WER) adapts WER for multi-speaker data by optimally permuting speakers' transcriptions before scoring (Watanabe et al., 2020).

5.2. Results of SDR Performance

Table 4 summarizes the SDR performance of different models across four multi-speaker dialogue audio sets in M3-SLU. Among the open cascade SD + ASR pipelines, **Diarizen + Whisper-Medium** consistently achieved the most balanced WER and cpWER. Based on this observation, we adopted this Diarizen + Whisper-Medium configuration as our default SD + ASR setting and integrated it with LLMs to conduct the final experiments of our benchmark. In addition, closed SDR models such as AssemblyAI achieved further reductions in both WER and cpWER on MultiDialog audio sets.

Since M3-SLU consists of long segments, rather than the short clips (within 30 seconds) used in previous studies (Watanabe et al., 2020; Yin et al., 2025), the reported WER and cpWER are naturally higher. This is because longer segments inherently increase the chance of accumulated transcription errors over time, leading to higher WER, while the frequent speaker transitions in extended dialogues also raise cpWER for each instance.

M3-SLU Benchmark	Task1	Task2
SD + ASR + LLM		
GT Script(Gold)+Llama3.1(8B)	0.9577	0.5787
GT Script(Gold)+Mistral(7B)	0.8717	0.5409
Diarizen+whisper+Llama3.1(8B)	0.7863	0.5620
AssemblyAI+Llama3.1(8B)	0.9192	0.5452
Diarizen+whisper+Mistral(7B)	0.7665	0.5490
Diarizen+whisper+Mistral(24B)	0.8068	0.6544
E2E Speech LLM		
Qwen2-Audio(7B)	0.0602	0.4960
MistralAI-Voxtral(24B)	0.8375	0.5169
E2E Multimodal LLM		
Qwen2.5-Omni(7B)	0.6883	0.5071
Qwen3-Omni(30B)	0.7762	0.5760

Table 6: M3-SLU benchmark results comparing cascaded (SD+ASR+LLM) and E2E Speech/Multimodal LLM models. For Task 1, the evaluation scores were obtained using the LLM-as-Judge approach, while for Task 2, the values represent accuracy based on the correctness of True/False judgments. Both scores are higher-the-better metrics.

5.3. Experiment Setting for M3-SLU Benchmark Evaluation

To evaluate current models' speaker-attributed reasoning ability on the M3-SLU benchmark, we adopted both cascaded (SD + ASR + LLM) and end-to-end (E2E) MLLM methodologies, following prior approaches in spoken language understanding research (Yin et al., 2025; Wang et al., 2025c).

In the cascade setting, we first combined speaker diarization (SD) and automatic speech recognition (ASR) models before passing the transcribed text to a large language model (LLM) for question answering. Based on the SDR Test results in Table 4, the **Diarizen + Whisper-Medium** combination exhibited consistently balanced multi-speaker transcription performance across our audio datasets, therefore we adopted this combination as the default SD + ASR configuration. The transcribed text was then passed to LLMs such as **Llama3.1-8B** and **Mistral-7B/24B** to investigate the impact of LLM scale. We also included commercial **AssemblyAI**'s transcription results to analyze the difference in transcription text quality.

In parallel, we evaluated end-to-end Speech-LLMs and Multimodal LLMs, which directly process raw audio inputs without intermediate transcription. For the Speech-LLM evaluation, we tested **Qwen2-Audio-7B** and **Voxtral-Small-24B**, while the Multimodal LLM evaluation included **Qwen2.5-Omni-7B** and **Qwen3-Omni-30B**, which jointly handle audio and textual reasoning in a unified framework.

5.4. Results of M3-SLU Benchmark

The top two scores in Table 6 correspond to the GT Script (Gold) configuration, where the model was provided with the original ground-truth transcription that had been perfectly segmented by speakers.

Evaluation on Task 1 As expected, when the GT Script(Gold) was provided to the LLM, this configuration achieved the highest score, since the model received perfectly transcribed and speaker-segmented text, effectively removing any noise or cascading errors during the SD and ASR stages. In addition, the cascade setting (SD + ASR + LLM) also demonstrated reasonably strong QA performance, indicating that despite inevitable errors from the SD + ASR modules, the overall pipeline was still able to preserve a substantial amount of semantic and speaker-related information.

In the upper part of Table 6, the results demonstrate how the quality of the SD + ASR pipeline directly influences the final QA performance in the cascade setting. Although both settings use the same LLM (Llama 3.1-8B), AssemblyAI + Llama 3.1-8B achieves 0.9192, substantially outperforming Diarizen + Whisper + Llama 3.1-8B (0.7863). As shown in Table 4, AssemblyAI exhibits remarkably higher transcription accuracy, particularly on the MultiDialog Audio dataset, showing a clear margin over the Diarizen + Whisper-Medium results. That is, **more precise transcriptions and speaker labels allow the LLMs to better understand who said what**, thereby yielding QA results nearly comparable to those obtained with gold transcriptions (GT Script + LLM). Also, the performance difference between Mistral-7B (0.7665) and Mistral-24B (0.8068) shows that increasing the model size leads to a moderate improvement in Task 1 results, suggesting that **larger LLM better leverage the transcribed and diarized input for understanding speaker-attributed content**.

In the lower part of Table 6, E2E Speech-LLM and Multimodal LLM models show relatively lower performance on Task 1, compared to the cascade setting. Only the larger models, such as MistralAI-Voxtral (24B) and Qwen3-Omni (30B), achieved scores approaching 80%, indicating that **our M3-SLU Task 1 remains a challenging problem for current E2E Speech-LLMs and MLLMs**.

Evaluation on Task 2 As shown on the right side of Table 6, the Task 2 results reveal that no model configurations exceeded 70%, a surprisingly low performance considering that the Task 2 is a binary (True/False) classification. This suggests that **accurate speaker-attributed utterance matching is impossible for both cascade and E2E models under the current framework**. Even

in the cascade setting with gold transcripts provided, the models failed to accurately distinguish speakers. The best result was obtained with the Diarizen + Whisper + Mistral (24B) combination, which achieved only 0.6544 on Task 2.

As shown in the comparison between Task 1 and Task 2 results in Table 6, models demonstrated a **noticeable gap between understanding what was said and who said it**. While Task 1 could be partially solved by leveraging contextual cues without explicit speaker distinction, Task 2 inherently required precise speaker identification to match utterances correctly. This means that **although current models can comprehend the content of conversations, they remain largely incapable of reasoning about speaker attribution**, highlighting the persistent gap toward true multi-speaker understanding. Therefore, advancing and evaluating future MLLMs will benefit from benchmarks such as M3-SLU as a foundation for genuine multi-speaker understanding.

Evaluation with Closed Models As shown in Table 7 below, current commercial models such as GPT-4o-Audio and Gemini-2.5-Flash-Audio completely fail to perform multi-speaker understanding, indicating that they are still unable to distinguish and reason over different speakers in conversational audio.

M3-SLU Benchmark	Task1	Task2
Closed Models		
GPT-4o-Audio*	0.32	0.51
Gemini-2.5-Flash-Audio*	0.41	0.54

Table 7: M3-SLU benchmark results using closed commercial models. Due to limited research resources, both models were evaluated on a randomly selected 100 samples.

5.5. Human Verification of LLM-as-Judge Evaluation

Since M3-SLU Task 1 involves predicting noun phrases from audio inputs, we adopted an LLM-as-a-Judge evaluation method using GPT-4o. To verify its reliability, we manually compared GPT-4o's judgments with human judgments on 200 randomly selected samples. Specifically, we used GT noun answers and predictions from the Diarizen + Whisper-Medium + LLaMA 3.1 (8B) experiment. The results showed 96.5% agreement between GPT-4o and human evaluators, demonstrating that our LLM-as-Judge evaluation is consistent with human judgment and not arbitrarily biased.

6. Conclusion

We introduced M3-SLU, a benchmark that reveals a key limitation of current MLLMs—the inability to comprehend "who spoke when and what" in long multi-speaker dialogues. Through two targeted tasks, Speaker-Attributed QA and Utterance Matching, M3-SLU isolates the challenge of speaker reasoning beyond simple transcription. Our experiments show that while existing cascaded pipelines and MLLMs can capture what was said, they consistently fail to track who said it, even with accurate transcripts. This underscores a critical gap in speaker attribution and multi-speaker reasoning. Building on real, naturally occurring conversations with speaker-attributed annotations, M3-SLU offers a structured evaluation setting that addresses the limitations of synthetic or short-turn benchmarks. The consistently low performance of current state-of-the-art models across both tasks reflects the complexity of speaker-grounded reasoning, which is unlikely to be resolved through scaling alone. M3-SLU offers a practical testbed for studying how multimodal language models handle multi-speaker conversations in realistic settings. We anticipate that it will guide the development of modeling strategies, evaluation methods, and training practices that explicitly incorporate speaker roles, turn-taking, and conversational structure, all of which are essential to dialogue comprehension.

7. Limitations

Our current benchmark is primarily focused on English conversational data. Future work could expand M3-SLU to include a wider range of languages and even more complex, overlapping speech scenarios to further probe the robustness of MLLMs. In addition, our evaluation framework currently relies on GPT-4o as an LLM-as-Judge to assess model outputs. While this approach enables flexible and semantic-level evaluation, it may overlook subtle variations that arise from speech input, such as minor pronunciation or transcription differences. We are actively exploring more refined evaluation strategies that can account for these speech-induced variations while maintaining fairness and consistency across models.

8. Ethical Consideration

All audio data used in this study are sourced from publicly available corpora (CHiME-6, MELD, MultiDialog, and AMI) that provide appropriate research licenses. No private or personally identifiable information (PII) was included. The dataset construction and experiments fully comply with the ethical use policies of the original sources.

9. Bibliographical References

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ming Gao, Shilong Wu, Hang Chen, Jun Du, Chin-Hui Lee, Shinji Watanabe, Jingdong Chen, Siniscalchi Sabato Marco, and Odette Scharenborg. 2025. The multimodal information based speech processing (misp) 2025 challenge: Audio-visual diarization and recognition. *arXiv preprint arXiv:2505.13971*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Birgit Knudsen, Ava Creemers, and Antje S Meyer. 2020. Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11:593671.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, et al. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2025. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16348.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party

dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025a. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025b. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. 2025c. Msu-bench: Towards understanding the conversational multi-talker scenarios. *arXiv preprint arXiv:2508.08155*.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2025. Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. *arXiv preprint arXiv:2508.06372*.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2025. Salmonn-omni: A standalone speech llm without codec injection for full-duplex conversation. *arXiv preprint arXiv:2505.17060*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

10. Language Resource References

Wessel Kraaij and Thomas Hain and Mike Lincoln and Wilfried Post. 2005. *The AMI Meeting Corpus*. Augmented Multiparty Interaction (AMI) Project. Distributed via ELRA, ELRA-ID: ELRA-S0940, ISLRN 776-138-369-477-5. PID https://groups.inf.ed.ac.uk/ami/download/. A multi-modal corpus of meeting recordings with transcriptions and annotations.

Se Park and Chae Kim and Hyeongseop Rha and Minsu Kim and Joanna Hong and Jeonghun Yeo and Yong Ro. 2024. *Let's Go Real Talk: A Multimodal Spoken Dialogue Dataset for Face-to-Face Conversations*. KAIST MISLAB (Multimodal Intelligence Systems Laboratory), Multimodal Spoken Dialogue Resources, 1.0. PID https://huggingface.co/datasets/IVLLab/MultiDialog. Multimodal dataset containing synchronized audio, video, and text dialogues for spoken conversation modeling.

Soujanya Poria and Devamanyu Hazarika and Navonil Majumder and others. 2019. *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*. Nanyang Technological University, University of Michigan. Available at http://affective-meld.github.io, ISLRN 190-235-172-195-7. Multimodal emotional dialogue dataset based on TV series Friends.

Shinji Watanabe and Michael Mandel and Jon Barker and others. 2020. *CHiME-6: Tackling Multispeaker Speech Recognition for Unsegmented Recordings*. Carnegie Mellon University, University of Sheffield, INRIA, and CHiME Challenge Organization. Available at https://chimechallenge.github.io/chime6/. PID https://openslr.org/150/. Dataset and challenge for distant multi-speaker speech recognition.

11. Appendices

11.1. *Appendix A. Source Datasets Used in M3-SLU benchmark*

In this appendix, we provide detailed descriptions of the four publicly available multi-speaker dialogue datasets used to construct the M3-SLU benchmark: **CHiME-6, MELD, MultiDialog, and AMI**. Each dataset contributes distinct conversational properties, acoustic conditions, and interaction patterns that collectively enable a comprehensive evaluation of multi-speaker, multi-turn spoken language understanding.

CHiME-6 The CHiME-6 dataset is a large-scale, real-world multi-channel conversational speech corpus designed to advance research on robust speech processing in challenging acoustic environments. The data were collected in realistic dinner party scenarios held in everyday home environments, where multiple participants engage in spontaneous conversations. As a result, the recordings exhibit natural turn-taking, frequent speaker overlap, background noise such as clattering dishes and music, and acoustic distortions including reverberation and distance attenuation.

A defining characteristic of CHiME-6 is that the audio is recorded in a non-segmented manner using distant microphone arrays, without prior separation of speakers or utterances. This setting closely reflects real deployment conditions and poses significant challenges for speech enhancement, speaker diarization, and automatic speech recognition. The dataset includes both close-talking microphones and far-field multi-channel microphone arrays, enabling comparative studies across recording conditions.

The dataset is divided into training, development, and evaluation splits, with a total size exceeding 120 GB of audio data. In addition to raw audio, CHiME-6 provides detailed annotations in JSON format, including time-aligned transcriptions, speaker identities, and overlap information. Furthermore, floorplan annotations describing microphone and speaker locations are included, allowing for spatially informed analysis and modeling.

CHiME-6 was introduced as the first community-scale challenge to systematically address non-segmented multi-speaker automatic speech recognition, and is accompanied by fully open-source baseline systems covering speech enhancement, speaker diarization, and recognition modules. Owing to its realism, scale, and comprehensive annotations, CHiME-6 has become a foundational benchmark for developing end-to-end speech processing systems under realistic far-field, multi-speaker conditions. The dataset is publicly available via the OpenSLR repository at `https://openslr.org/150/`.

MELD The MELD (Multimodal EmotionLines Dataset) is a multimodal conversational corpus designed for emotion and sentiment recognition in multi-party dialogue settings. It is constructed from scenes of the popular TV series *Friends*. MELD provides aligned textual, acoustic, and visual modalities for each utterance, enabling comprehensive analysis of emotional expressions in conversational contexts involving multiple speakers.

The dataset contains over 1,400 dialogues and more than 13,000 utterances, each drawn from scripted yet naturalistic conversations in the TV series. Owing to its dialogue-centric structure, MELD captures rich contextual dependencies across utterances, including turn-taking, speaker interactions, and emotion flow over time. Each dialogue involves multiple participants, reflecting realistic multi-speaker conversational dynamics rather than isolated utterances.

For each utterance, MELD provides textual transcriptions, corresponding audio signals, and visual data capturing facial expressions and body language. All utterances are annotated with one of

seven categorical emotion labels—anger, disgust, sadness, joy, neutral, surprise, and fear—as well as a sentiment polarity label indicating positive, negative, or neutral sentiment. This dual annotation scheme supports both fine-grained emotion recognition and coarse-grained sentiment analysis within conversational settings.

The availability of synchronized text, speech, and visual cues makes MELD a representative benchmark for multimodal affective computing and dialogue emotion recognition. Its rich emotional variability, combined with explicit conversational context, has led to its widespread adoption in research on multimodal dialogue understanding, emotion-aware conversational agents, and context-sensitive affective modeling. The MELD dataset is publicly available through its official website at https://affective-meld.github.io/.

MultiDialog The MultiDialog dataset is a large-scale multi-party conversational corpus designed for modeling face-to-face spoken dialogue under realistic and noisy conditions. It was introduced in the work *"Let's Go Real Talk: Spoken Dialogue Model for Face-to-Face Conversation"* (ACL 2024, Oral) by the Integrated Vision Language Lab (IVLLab), and is publicly available under the CC BY-SA 4.0 license. The dataset consists of English multi-speaker conversations recorded in realistic social environments.

MultiDialog contains 8,733 dialogues comprising a total of 187,859 utterances, amounting to approximately 340 hours of audio. Each dialogue spans on average about 2.3 minutes and includes roughly 21 utterances, with an average utterance duration of 6.5 seconds. Each utterance is paired with time-aligned audio recorded at a sampling rate of 16 kHz, along with its corresponding textual transcription, speaker identifier, and emotion label. The emotion annotation scheme includes seven categories(neutral, happy, fear, angry, disgusting, surprising, and sad).

Beyond emotion analysis, MultiDialog is explicitly designed to support complex multi-domain dialogue understanding. Within a single conversation, speakers may transition across multiple tasks or intents, such as booking accommodation followed by requesting restaurant recommendations. This characteristic makes the dataset well suited for studying long-context dialogue modeling, multi-intent reasoning, and compositional task execution.

Audio, transcription, and emotion annotations are distributed via Hugging Face, while the corresponding video recordings are provided separately, enabling research on multimodal dialogue systems that jointly reason over speech, language, and visual cues. The MultiDialog dataset is publicly available at https://huggingface.co/datasets/IVLLab/MultiDialog.

AMI Meeting Corpus The AMI (Augmented Multi-party Interaction) Corpus is a large-scale multimodal meeting dataset designed to support research on multi-party interaction, meeting understanding, and conversational analysis. The corpus comprises approximately 100 hours of recorded meetings, including both scenario-driven meetings and naturally occurring business meetings. Scenario-based meetings account for roughly two-thirds of the dataset and are structured as role-playing design team sessions that follow a complete project lifecycle from initial briefing to final presentation, while the remaining portion consists of naturally occurring meetings covering a variety of domains.

A distinctive feature of the AMI dataset is its rich multimodal recording setup. Each meeting is captured using multiple synchronized modalities, including close-talking microphones worn by individual participants, far-field microphones recording the entire room, and several video cameras that provide both individual participant views and a global room perspective. Additional contextual signals such as whiteboard activity, projected slides, and pen-based interactions are also recorded. All sensing devices are tightly synchronized, enabling precise temporal alignment across audio, visual, and auxiliary data streams.

In addition to raw recordings, the AMI Corpus provides extensive manual annotations that support a wide range of research tasks. These include orthographic transcriptions of speech, dialogue act labels (e.g., statements, questions, suggestions, agreements, and backchannels), and various behavioral and interactional annotations such as head movements and participant actions. The availability of structured annotations, combined with accurate time stamps, makes AMI particularly well suited for studies on meeting summarization, topic segmentation, dialogue act recognition, multimodal interaction modeling, and long-context conversational understanding.

Owing to its realistic meeting scenarios, comprehensive multimodal coverage, and high-quality annotations, the AMI Corpus has become a foundational benchmark for research on multi-speaker and multimodal conversational systems. The dataset is publicly available under a Creative Commons Attribution 4.0 license and can be accessed via the official AMI repository at https://groups.inf.ed.ac.uk/ami/download/.

11.2. Appendix B. Prompts

11.2.1. Prompt for Task 1 (Speaker-Attributed QA) Generation

Prompt for Task 1: Speaker-Attributed QA Generation

You are given a short conversation formatted as `<spkN>: <utterance>`.

First, count how many distinct speaker tags (e.g., `spk1`, `spk2`, ...) appear in the conversation.

Task

- If there are **two or more different speaker tags**:
  - Generate **exactly one question** whose solution requires knowing *which speaker* said a particular line.
  - The question must rely on speaker attribution, but the **answer must be the content or information** provided by that speaker, **never the speaker tag itself**.
  - **Forbidden**: Questions of the form *"Who said . . . ?"*, which would force the speaker identity as the answer.
  - **Allowed**: Questions such as *"What size did the speaker who described the tiny house give?"*, where the answer is the described content (e.g., `10×30`), not the speaker label.
- If there is **only one speaker tag**:
  - Generate **exactly one question** about the conversational content itself.

Answer Requirements (Strict)

- The answer must be a **keyword-level noun phrase**, headed by a noun.
- Do **not** use full sentences or clauses.
- Do **not** include finite verbs or auxiliaries (e.g., *is*, *are*, *was*, *were*).
- Do **not** include pronouns (e.g., *I*, *you*, *he*, *she*).
- Normalize numbers and units when appropriate (e.g., *"ten by thirty"* → `10×30`).
- Do **not** include trailing punctuation (., ?, !).
- Do **not** include speaker tags.
- Keep the answer short: **at most 7 words**.

Output Format
Return **exactly** the following JSON object (no markdown, no additional keys):

```
{
  "question": "<one single-sentence question>",
  "answer": "<short noun phrase>"
}
```

11.2.2. Prompt for Task 2 (Speaker Attribution Utterance Match (T/F)) Generation

Prompt for Task 2: Speaker Attribution Utterance Match (T/F) Generation

You are an expert in discourse analysis. Your goal is to create a single, sophisticated True/False question that tests a listener's ability to track who is who in a conversation based on what they say and how they say it. **Core Task:** Create a question based on the pattern: *"Is the participant who did/said X the same one who did/said Y?"* **Strict Rules**

- The question **must** connect two distinct pieces of information (actions, statements, opinions, roles) and ask if they belong to the same anonymous participant.
- The answer must be `True` or `False` based *only* on the provided text.
- The output must be a single JSON object, with no extra text or markdown.

Examples

Example 1 — Conversation:

```
<spk1>: I think we should start with the appetizers first.
        I brought that fancy cheese from my trip to France.
<spk2>: Oh, cheese! Great idea. So, are we cooking the steaks
        now or later? I'm getting hungry.
<spk1>: Let's do the steaks after everyone has had some cheese
        and wine. Patience, my friend.
```

Output:

```
{
  "question": "Is the participant who brought the cheese from
              France the same one who asked when the steaks
              would be cooked?",
  "answer": "False",
  "rationale": "The participant who brought the cheese (spk1)
               suggested waiting on the steaks, while another
               participant (spk2) asked about cooking them."
}
```

Example 2 — Conversation:

```
<spk3>: Honestly, this project is a mess. The deadline is too
        tight. I feel like I'm the sheriff in a lawless town.
<spk4>: I agree, it's chaotic. What's your plan for the next
        two days, then?
<spk3>: My plan is to delegate the testing phase to Mark's
        team immediately. It's the only way.
```

Output:

```
{
  "question": "Did the participant who described themselves as
              a 'sheriff' also propose a specific plan to
              delegate the testing phase?",
  "answer": "True",
  "rationale": "The same participant (spk3) first used the
               'sheriff' analogy and then outlined their plan
               to delegate the testing phase."
}
```

Output Format

Based on the input conversation, generate a new JSON output following the exact same pattern as the examples above (no markdown, no additional keys):

```
{
  "question": "<one single-sentence True/False question>",
  "answer": "<True or False>",
  "rationale": "<brief explanation>"
}
```

11.2.3. Prompt for Task 1 LLM-as-Judge Evaluation

Prompt for LLM-as-Judge Evaluation

You are an expert evaluator assessing the quality of answers to questions about spoken conversations. Your task is to determine whether a model answer is semantically equivalent to the ground-truth answer, even if the surface forms differ. **Evaluation Criteria**

- Focus on **semantic equivalence**, not exact string matching.
- Minor variations in the following should be judged as `Correct`:
  - Capitalization (e.g., *nite xml* vs. *NITE XML*)
  - Spacing or hyphenation (e.g., *x-m-l* vs. *xml*)
  - Common abbreviations or shorthand (e.g., *w/* vs. *with*)
  - Phonetic or informal spelling (e.g., *nite* vs. *night*)
  - Minor word order differences that preserve meaning
- Judge as `Incorrect` only when the model answer refers to a meaningfully different entity, fact, or concept than the ground truth.

Input

```
- Question:            {question}
- Ground-Truth Answer: {ground_truth}
- Model Answer:        {model_answer}
```

Output Format
Return a single JSON object with no markdown or extra text:

```
{
  "judgment":  "<Correct or Incorrect>",
  "rationale": "<one sentence explaining your judgment>"
}
```