Multi-Party Conversational Agents: A Survey

Sagar Sapkota
University of Central Florida
sagar.sapkota@ucf.edu

Mohammad Saqib Hasan
Stony Brook University
mdshasan@cs.stonybrook.edu

Mubarak Shah
University of Central Florida
shah@crcv.ucf.edu

Santu Karmaker
University of Central Florida
santu@ucf.edu

Abstract

Multi-party Conversational Agents (MPCAs) are systems designed to engage in dialogue with more than two participants simultaneously. Unlike traditional two-party agents, designing MPCAs faces additional challenges due to the need to interpret both utterance semantics and social dynamics. This survey explores recent progress in MPCAs by addressing three key questions: 1) Can agents model each participants' mental states? (*State of Mind Modeling*); 2) Can they properly understand the dialogue content? (*Semantic Understanding*); and 3) Can they reason about and predict future conversation flow? (*Agent Action Modeling*). We review methods ranging from classical machine learning to Large Language Models (LLMs) and multi-modal systems. Our analysis underscores *Theory of Mind* (ToM) as essential for building intelligent MPCAs and highlights multi-modal understanding as a promising yet underexplored direction. Finally, this survey offers guidance to future researchers on developing more capable MPCAs.

1 Introduction

Developing agents for conversational understanding and natural dialogue generation has been a long-standing research agenda within the NLP community (Wang et al., 2024a). Traditionally, most conversational agents have been designed for two-party interactions (Zheng et al., 2022). However, real-world conversations—such as group chats, team meetings, online gaming, and customer support forums—often involve more than two participants. As a result, there has been growing interest in developing agents capable of engaging with multiple participants simultaneously, commonly referred to as ***Multi-Party Conversational Agents*** (MPCAs).

Compared to traditional two-party systems, building MPCAs presents greater challenges due to the complex dynamics in multi-party interactions. Effective MPCAs must simultaneously infer

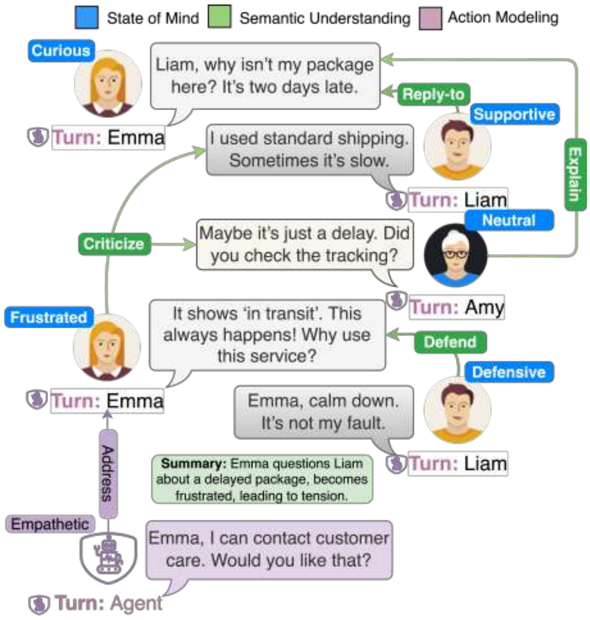

Figure 1: Example of a multi-party conversation demonstrating key challenges that MPCAs must handle. At each time step of the conversation, the MPCA must identify the states of mind of each participant (e.g., curiosity, frustration, etc.), have semantic understanding of the conversation (e.g., speaker actions like criticize and explain, dialog summary, etc.), and be able to take the appropriate action (e.g., response, turn-taking, identifying addressee, etc.). Combining these capabilities makes for a social and intelligent agent.

participants' mental states (Cohen et al., 2024), comprehend user content (Ma et al., 2023a), and anticipate future dialogue flow (Houde et al., 2025). Mastering these capabilities is crucial for enabling collaborative tools, socially adept robots, and intelligent assistants. For instance, Figure 1 illustrates a two-speaker interaction with an agent, highlighting core challenges such as emotional variance, evolving latent states, and context-sensitive response planning (Wei et al., 2023; Houde et al., 2025). While recent advances in conversational AI (AI, 2024; OpenAI, 2024; Qwen et al., 2025; Jiang et al., 2024a) have improved these components individually, integrating them into fully functional MPCAs still remains an open challenge (Zhao et al., 2023; Tan et al., 2023).

This research gap has recently attracted significant attention from the community, leading to a surge in efforts to design multi-party conversational systems (Ganesh et al., 2023). In synergy with with recent advances in generative AI, particularly Large Language Models (LLMs), most of these efforts have focused on leveraging LLMs for this domain (Wang et al., 2024c; Tan et al., 2023). As such, this survey aims to serve as a comprehensive resource for AI researchers interested in developing MPCAs. We review over **70** research articles relevant to multi-party conversations and categorize their contributions into three broad themes:

State of Mind Modeling: Works under this theme study the following question: "*Can models accurately infer each participant's mental and emotional states?*" Key tasks under this theme include emotion recognition, participants' engagement detection, personality identification, and recognition of user intents.

Semantic Understanding: Under this theme, works assess MPCAs' ability to properly understand utterances and the context of the multi-party dialogue. This covers tasks like dialogue summarization, conversation disentanglement, discourse structure analysis, and representation learning.

Agent Action Modeling: Tasks pertaining to this theme describe MPCAs' ability to predict future conversation flow. This involves tasks like turn detection, addressee selection, and response selection/generation.

For each of the aforementioned themes, we start off by discussing research works utilizing traditional ML algorithms (Bayser et al., 2019; Majumder et al., 2019; Ghosal et al., 2019) and NLP models (Shen et al., 2020; Wen et al., 2024; Li et al., 2021b). We then move towards Large Language Models (LLMs) as many MPC tasks are being modeled by leveraging LLMs either through fine-tuning or prompting (Wu et al., 2024; Li et al., 2025; Jiang et al., 2024b). Finally, since the proliferation of multi-modal data (e.g., meeting videos, group chats with audio-visual context) (Zhang et al., 2024a; Carletta et al., 2005; Sasu et al., 2025; Müller et al., 2024), we discuss works where MPC tasks have been studied using visual and auditory cues (Yoon et al., 2024; Lee and Deng, 2024; Jain et al., 2023). In addition, we discuss the major datasets associated with these tasks.

Based on our thematic analysis of MPCA research, we identify two promising future directions: integrating *Theory of Mind* (ToM) (Premack and Woodruff, 1978) into multi-party dialogue systems and advancing *Multi-modal Fusion and Grounding*. For ToM integration, we highlight gaps in the current literature and suggest tailoring models to infer mental states within task-specific contexts. For multi-modal grounding, we highlight that key tasks—such as conversation disentanglement, representation learning, addressee selection, and response generation—remain largely underexplored in multi-modal settings. Consequently, even state-of-the-art multi-modal MPCAs often struggle to simulate realistic multi-party interactions. We then discuss how incorporating open-ended multi-modal benchmarks (Tang et al., 2025) can enhance MPCA development. We hope that the insights drawn from our analysis will guide the community for future research toward more robust and socially aware MPCA systems.

2 A Thematic Taxonomy

To organize this survey paper, we create a taxonomy of existing research efforts on multi-party conversations (MPC) by categorizing them into three core themes: *State of Mind Modeling*, *Semantic Understanding*, and *Agent Action Modeling*. This taxonomy reflects essential capabilities for human-like social communication in group settings. Indeed, a socially intelligent agent must determine when to speak, whom to address, and what to say, collectively referred to as *actions*. According to Social Identity Theory (Tajfel and Turner, 1979), such actions are influenced by the agent's own and others' mental states, while reciprocal effects also exist, where actions can shape mental states (Hatfield et al., 1993). Semantic understanding (Kopp and Krämer, 2021) further mediates these dynamics by enabling agents to interpret social cues and context (Rijpma et al., 2023). As illustrated in Figure 1, these three components interact dynamically throughout the course of a conversation.

Next, we formalize our proposed taxonomy by mapping existing MPC tasks to these three themes as follows: tasks aligned with *State of Mind Modeling* include emotion recognition, personality detection, dialog act recognition, and engagement detection; tasks supporting *Semantic Understanding* include dialog summarization, conversation disentanglement, discourse analysis, and representation

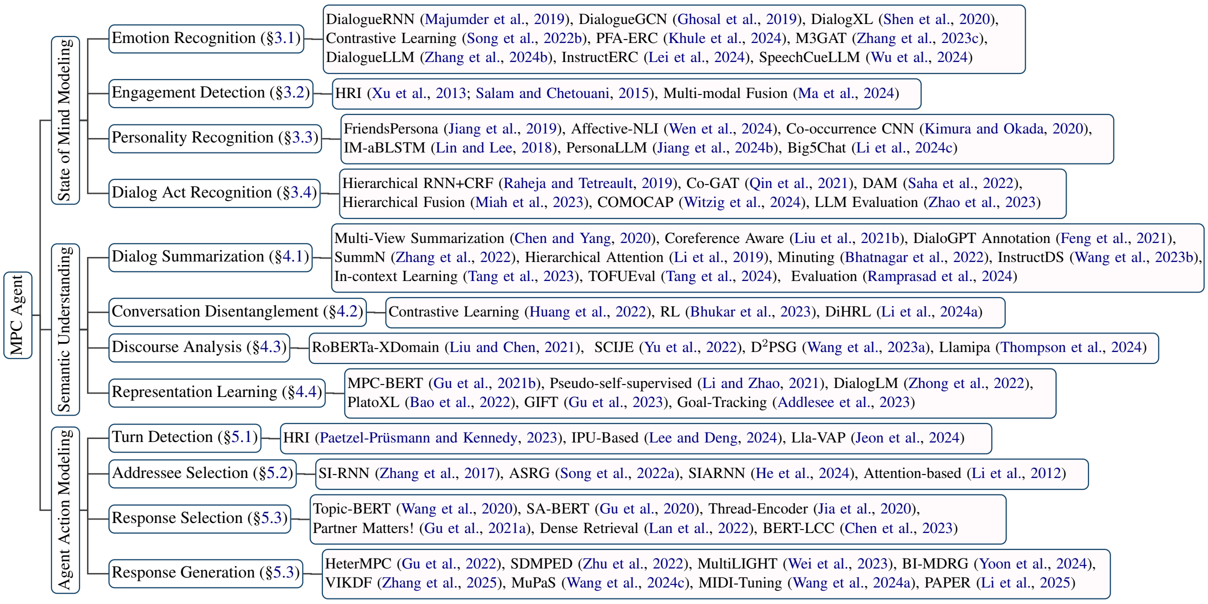

Figure 2: Thematic taxonomy of MPC tasks and recent works focusing on these tasks.

learning; and tasks contributing to *Agent Action Modeling* include turn detection, addressee selection, and response generation. A visual taxonomy of these skills and associated tasks is shown in Figure 2, with an extended taxonomy—including technique categorization (traditional, multi-modal, and LLM-based)—available in Appendix A.

3 Research on State of Mind Modeling

Understanding the state of mind of participants is essential for developing socially intelligent multi-party conversational agents. But what exactly constitutes a participant's *state of mind*? Our survey identifies four key components that together provide a comprehensive understanding of participants' mental states during conversation: emotion recognition, engagement detection, personality recognition, and conversational intentions (i.e., dialog act recognition). These essential components of mental states are defined and modeled by legacy psychological frameworks, such as the OCEAN model for personality traits (Kasap et al., 2009) and the OCC model (Ortony et al., 1988) for emotions (Irfan et al., 2020; Gebhard, 2005). Below, we discuss these four key components pivotal to *State of Mind Modeling*.

3.1 Emotion Recognition

Emotion recognition (ER) is the task of classifying emotions as positive (e.g., joy) or negative (e.g., anger) (Ortony et al., 1988). It helps MPCAs to perceive and respond empathetically. Initial approaches did contextual modeling (Sun et al., 2021; Xie et al., 2024a) using models like recurrent neural network-based DialogueRNN (Majumder et al., 2019) and graph convolutional neural network-based DialogueGCN (Ghosal et al., 2019) that focus on the speaker's emotional context. Other works tried to solve different limitations of ER: Shen et al. (2020) used XLNet (Yang et al., 2020) to handle long-term emotional dependencies through integrating global and speaker attentions, while Li et al. (2022); Song et al. (2022b) used contrastive learning to help models differentiate similar emotional signals. Other methods include isolating emotion contexts by incorporating commonsense knowledge, speaker relationships (Xu and Li, 2023), and conversation disentanglement (Zhao et al., 2022); assessing emotional changes using personality (Wang et al., 2024d) and pseudo-future forecasting (Khule et al., 2024).

Recently, LLMs and multi-modal models are common for these tasks. Multi-modal methods work by fusing text, audio, and image via attention and/or loss functions (Chudasama et al., 2022; Rasendrasoa et al., 2022), showing how integrating facial expressions (Zheng et al., 2023; Zhang et al., 2023c) and topic modeling (Yuan et al., 2024) enhance ER performance. LLM-based methods like structured prompting and fine-tuning (Tan et al., 2023; Zhao et al., 2023) have been shown to improve ER performance significantly. Techniques like multi-step prompting and generative tasks better captured emotional dynamics (Hama et al., 2024; Lei et al., 2024), while adding vocal descriptions bridged modality gaps (Wu et al., 2024).

3.2 Engagement Detection

Engagement detection measures the interaction level of participants in conversations, essential for managing interactions and turns between participants. Despite being a difficult problem, recent multi-modal methods used visual and audio cues, enhancing performance by continuously tracking participants' behaviors (Bohus and Horvitz, 2009; Xu et al., 2013). Other methods increase accuracy by integrating group interaction and spontaneous conversation (Salam and Chetouani, 2015; Fedotov et al., 2018). Recently, multi-modal LLMs have shown promise, offering interpretability and flexibility for engagement detection (Ma et al., 2024). However, effectively using multi-modal frameworks and LLMs for this task remains a challenge.

3.3 Personality Recognition

Personality recognition (PR) helps MPCAs identify personality traits of participants from conversation modalities (video, text, etc.), commonly using the Big-Five personality model: extraversion, agreeableness, conscientiousness, neuroticism, and openness (Cunningham, 1977). Initial text-based methods analyzed and aggregated predictions over dialogues to predict personality (Ríssola et al., 2019). Later, transformers like BERT and RoBERTa enhanced PR through attention mechanisms (Jiang et al., 2019). Other approaches framed PR as a natural language inference task using emotion annotations (Wen et al., 2024).

Multi-modal methods for PR integrate vision, audio, and behavioral cues to solve the task. Initial methods clustered co-occurring non-verbal behaviors (Okada et al., 2015), while advanced techniques employ convolutional neural networks (CNNs) (Kimura and Okada, 2020) and attention-based models (Lin and Lee, 2018). Recent multi-modal advancements combine information from wearable devices and video data (Cabrera-Quiros et al., 2022). Facial expressions and speech overlap are also important multi-modality features for models to do effective PR (Song et al., 2023; Yu et al., 2019). With the widespread use of LLMs, recent progress involves shifting focus from doing PR to incorporating personalities in language models (Jiang et al., 2024b). In this paradigm, tuning LLMs on target personality datasets is more effective than using general prompt-based methods (Li et al., 2024c).

3.4 Dialog Act Recognition

Dialog Act Recognition (DAR) classifies participant utterances into certain actions, such as whether they are asking questions or making statements. Different methods have been applied to improve upon this task: hierarchical CNNs and RNNs, which model current utterance DAR based on previous predictions for better context (Liu et al., 2017); self-attention and Conditional Random Fields (CRFs) to capture utterance and conversation level context (Raheja and Tetreault, 2019); joint modeling of dialogue act and sentiment through co-attention and Graph Attention Networks (GAT) (Qin et al., 2020, 2021); and speaker transitions and pretrained embeddings to incorporate speaker information (He et al., 2021b).

Recently, multi-modal approaches integrating non-verbal cues, such as tone and facial expressions, via attention-based multi-modal transformers further improved DAR (Saha et al., 2020, 2022; Witzig et al., 2024). Hierarchical multi-modal fusion using early-stage Bi-LSTMs and late-stage CNN-based encoders was also shown to be effective (Miah et al., 2023). LLMs like ChatGPT, however, were less effective for DAR tasks (Zhao et al., 2023). Methods like dual-process masking, emphasizing informative tokens, showed improvements across LLMs like BERT, RoBERTa, and Llama (Kim et al., 2024). However, LLM-based DAR approaches are still an unsolved challenge.

4 Research on Semantic Understanding

Proper semantic understanding is essential for any conversational system. But what comprises *semantic understanding*? We identify four tasks that collectively define semantic understanding in MPC: *Conversation disentanglement* (Elsner and Charniak, 2008), *Dialogue summarization* (Zhu and Penn, 2006), *Discourse Structure Analysis* (Webber et al., 2011), and *Representation Learning* (Lowe et al., 2015). These tasks help MPCAs interpret the meaning and flow of conversation, which is challenging due to overlapping utterances (Sugawara, 2012), parallel threads of discussion (Mayfield et al., 2012), temporal dependencies (Xing and Tsang, 2023), and dynamic context shifts (Galley et al., 2003). Traditionally, graph-based and sequential models were used (Wang et al., 2021; Gu et al., 2023; Shi and Huang, 2019), but recent works shifted towards transformer-based architectures (Xie et al., 2024b; Ma et al., 2023a). In the following sections, we review existing works on each aspect of semantic understanding.

4.1 Dialog Summarization

Dialog summarization (DS) involves creating coherent summaries of conversations, highlighting the main topics and participant actions. This task can be challenging due to overlapping topics, informal language, and varied speaker roles. As a result, methods use multiple views like global structure, utterance-level details, topic segmentation, dialogue progression, and coreference resolution for coherent summaries (Chen and Yang, 2020; Liu et al., 2021b; Feng et al., 2021). For long dialogues, works propose using retrieve-then-summarize or recursive segmentation (split-then-summarize) or segment-wise summaries for better summarization (Zhang et al., 2021, 2022; Han et al., 2024; Yin et al., 2024).

In recent times, dialogue summaries have incorporated multi-modal methods, as leveraging audio and visual cues improves topic relevance and speaker importance estimation (Li et al., 2019). Technologies like automatic speech recognition, machine translation, perspective-aware encoding, and topic modeling have also been developed for this task (Bhatnagar et al., 2022; Jain et al., 2023). LLMs are also being used for dialogue summarization but are not as effective due to hallucination, as current models, including GPT-4, continue to generate plausible but unsupported details (Tang et al., 2023, 2024; Ramprasad et al., 2024). Works like Xu et al. (2022) mitigate this by formatting dialogues to pre-training formats, while Zhu et al. (2024) employ symbolic knowledge distillation and contrastive learning to enhance the factual consistency. LLMs are also being used to generate synthetic data for the summarization task (Mishra et al., 2023; Wang et al., 2023b).

4.2 Conversation Disentanglement

Conversation disentanglement is the task of classifying overlapping utterances into coherent threads. Traditionally, hierarchical neural models and clustering algorithms were used but failed to capture global context and required ample data (Jiang et al., 2018; Li et al., 2021b). This problem was mitigated by contrastive learning via clustering related utterances without the need for large datasets (Huang et al., 2022; Gao et al., 2023), unsupervised learning via self-generated pseudo labels (Liu et al., 2021a), and reinforcement learning by directly optimizing on coherence using reward functions (Bhukar et al., 2023). Recently, heterogeneous discourse graphs combined with Graph Convolutional Networks (GCN) have outperformed GPT-4, showing that adding structured discourse can help with this task (Li et al., 2024a).

4.3 Discourse Structure Analysis

Discourse-structure analysis predicts dependencies among elementary discourse units (EDUs), such as speakers and utterances, aiding MPCAs in contextual understanding. Major approaches for discourse structure analysis include: sequential decision models that are prone to long-distance errors and error-propagation (Shi and Huang, 2019); graph-based methods with EDU graphs and gated message passing (Wang et al., 2021); utilizing structure distillation and speaker-aware masks to refine interaction parsing (Yu et al., 2022; Ma et al., 2023a); dialogue-based pretraining for improved robustness (Liu and Chen, 2021); and joint optimization with auxiliary tasks like QA (He et al., 2021a).

Recent advances also incorporate multi-modal inputs (vision, audio, text) via cross-modal attention (Gong et al., 2024) or reframe parsing as a Seq2Seq task using LLMs (Wang et al., 2023a). LLMs also enable incremental discourse graph generation (Thompson et al., 2024) and explanation-driven supervision to train smaller models in distinguishing correct structures for better accuracy (Liu et al., 2025).

4.4 Representation Learning

Representation learning aims to encode elements of the conversation, like utterances and speaker roles, into a vector latent space. Most works employ label-agnostic self-supervised methods with custom objectives like reply-to prediction, speaker identification, and key-utterance detection for representation learning (Gu et al., 2021b; Li and Zhao, 2021; Zhong et al., 2022) with improvements like role embeddings or speaker-specific attention (Bao et al., 2022; Ma et al., 2023a). Another label-agnostic method is inferring hidden dialogue structures (e.g., speaker roles, utterance connections) using expectation-maximization (Li and Zhao, 2023; Li et al., 2023). Also, methods where dialogues are modeled as graphs tend to enhance transformer-based representation (Gu et al., 2023) and adapt to tasks like emotion recognition or dialogue act (Xie et al., 2024b).

Despite advances in dialogue encoding, multi-modal representation learning is still behind, mainly due to data scarcity. LLM techniques like zero-shot and few-shot prompting fail without fine-tuning in complex dialogues, lacking explicit modeling of discourse structures or speaker roles, limiting LLMs' effectiveness as encoders (Addlesee et al., 2023).

5 Research on Agent Action Modeling

Agent action modeling predicts the next move of participants, such as deciding turn-taking, whom to address, and selecting and generating appropriate responses, for a smooth and coherent conversation. To do this, MPCAs must decide when (turn detection), to whom (addressee selection), and what to speak (agent response). These abilities make MPCAs natural and effective for interaction, which are discussed further below.

5.1 Turn Detection

Turn detection involves knowing when a speaker ends their utterance and who should start next and is essential for ensuring smooth speaker transitions. Simple techniques like maximum likelihood estimation, support vector machines, and convolutional neural networks failed in this task due to subtle turn shifts (Bayser et al., 2019), so annotations with syntax and prosodic features were introduced (Enomoto et al., 2020). Sequence-based neural models like GRUs and transformers further advanced turn prediction (Lee and Deng, 2024).

Overcoming text-only methods, multi-modality integrated speech pauses, gaze, and listener behaviors (Żarkowski, 2019; Paetzel-Prüsmann and Kennedy, 2023; Lee et al., 2023). LLM-based methods, like multi-task instruction tuning and prompting, enabled more flexible cue interpretation for detecting turns and backchannels (Wang et al., 2024b; Pinto and Belpaeme, 2024). Recent innovations also include full-duplex models to handle overlapping speech and real-time feedback (Veluri et al., 2024), strategies using social reasoning (Nonomura and Mori, 2024), and hybrid LLM-VAP (voice activity detection) methods effective across fluent dialogue (Jeon et al., 2024).

5.2 Addressee Selection

Addressee selection identifies intended listeners. Early methods used static embeddings or recurrent neural networks (RNNs) to model roles and context (Ouchi and Tsuboi, 2016). Later approaches dynamically updated embeddings by jointly optimizing addressee and response selection for coherence (Zhang et al., 2017). For low-resource languages, multilingual embeddings, adversarial learning, and transfer learning were used to enhance performance (Sato et al., 2018). Recent models combine addressee selection with response generation using attention mechanisms and role-aware transformers (Song et al., 2022a; He et al., 2024). Further, multi-modal data is used to assess engagement through gaze, posture, and lip movements to compute attention scores (Li et al., 2012).

5.3 Agent Response

Agent response skills help MPCAs deliver contextually appropriate and coherent utterances. There are two ways MPCAs can respond: response selection and response generation. Response selection involves selecting the best candidate out of a set of predefined responses. There are different methods for effective response selection: discourse parsing to separate mixed conversations for improved context matching (Wang et al., 2020; Jia et al., 2020; Zhang et al., 2023b); using speaker identity and roles (Gu et al., 2020); persona-aware models using learned persona embeddings (Gu et al., 2021a; Li et al., 2021a); and graphs linking persona with dialogue elements for responses relevant to participants' preferences (Ju et al., 2022). Furthermore, there are works implementing contrastive learning (Li et al., 2021c; Feng et al., 2023b), dense retrieval (Lan et al., 2022), and local context predictions (Chen et al., 2023; Su et al., 2024) for selecting better context-response pairs.

Response generation, on the other hand, synthesizes the response, ensuring coherence across speakers and conversation threads. Traditional approaches used graphs capturing semantics between dialog and speaker to generate responses (Gu et al., 2022; Ju et al., 2022). Adding emotional dynamics (Zhu et al., 2022), speaker persona (Chauhan et al., 2023), and discourse relations (Li et al., 2024b; Fan et al., 2024) within these graphs also enhanced empathy, engagement, and conversational flow, leading to better responses. Multi-modal methods introduced visual grounding (Chen et al., 2021; Sun et al., 2022; Luo et al., 2023) and employed cross-modal attention with contrastive learning to synchronize multi-modal information (Zhang et al., 2023a; Yoon et al., 2024; Zhang et al., 2025) for improved generation of responses. Alternatively, LLMs brought new opportunities for

Dataset	Modality	Model	Task	Eval Metric	Eval Result
EmoryNLP (Zahiri and Choi, 2017)	Text	PFA-ERC (Khule et al., 2024)	Emotion Recognition	F1 Score	42.94
MELD (Poria et al., 2019)	Audio, Video, Text	DialogueLLM (Zhang et al., 2024b)	Emotion Recognition	F1 Score	71.9
EmoryNLP (Zahiri and Choi, 2017)	Text	Affect-NLI (Wen et al., 2024)	Personality Recognition	Accuracy	61.3
DailyDialog (Li et al., 2017)	Text	COMOCAP (Witzig et al., 2024)	Dialog Act	F1 Score	79.66
EMOTyDA (Saha et al., 2020)	Audio, Video, Text	DAM (Text+Video) (Saha et al., 2022)	Dialog Act	F1 Score	57.01
SAMSum (Gliwa et al., 2019)	Text	InstructDS (Wang et al., 2023b)	Dialog Summary	ROUGE-1	58.4
Ubuntu IRC (Kummerfeld et al., 2019)	Text	Bi-Level CL (Huang et al., 2022)	Disentanglement	F1 Score	70.4
STAC (Asher et al., 2016)	Text	D2PSG (Wang et al., 2023a)	Discourse Parsing	F1 Score	62.77
CCPE (Radlinski et al., 2019)	Text	LLa-VAP (Jeon et al., 2024)	Turn Detection	F1 Score	83.13
Ubuntu IRC (Kummerfeld et al., 2019)	Text	ASRG (Song et al., 2022a)	Addressee Selection	Accuracy	84.65
Ubuntu IRC (Kummerfeld et al., 2019)	Text	Dense Retrieval (Lan et al., 2022)	Response Selection	Recall	91
TopDial (Wang et al., 2023c)	Text	MIDI-Tuning (Wang et al., 2024a)	Response Generation	BLEU	39.6

Table 1: Benchmark datasets for multi-party conversations and performance of the state-of-the-art models on them.

response generation by leveraging their pretraining to learn speaker roles, turn-taking, conversational threads, and persona dynamics (Wang et al., 2024c,a; Li et al., 2025; T et al., 2024). Some research also extends LLM response generation to multilingual settings (Gunson et al., 2024; Hu et al., 2025).

6 Evaluation and State-of-the-Art

So far, we have discussed a range of tasks within our thematic taxonomy, each associated with multiple benchmark datasets for evaluation. Table 1 summarizes the most widely used benchmarks along with state-of-the-art (SOTA) model performance. Further details are provided in Appendixes B and C.

These datasets vary in size, domain, and modality and are typically designed to evaluate a single MPCA skill, such as emotion recognition (Poria et al., 2019), personality detection (Sanchez-Cortes et al., 2012), or dialogue summarization (Zhong et al., 2021), which may limit their generalizability. Domain specificity also presents a challenge. For example, the AMI Meeting Corpus (Carletta et al., 2005), despite offering annotations for multiple tasks, is underutilized due to its narrow focus on formal product design meetings (see Table 3).

A notable limitation is the scarcity of robust *multi-modal* datasets that integrate non-verbal cues such as facial expressions, gaze, or vocal tone—essential for capturing participants' mental states and enriching semantic context.

Analysis of SOTA models across benchmarks reveals significant performance disparities. Tasks involving agent action modeling (e.g., turn detection, addressee and response selection) often exceed 80% accuracy, while most other tasks average around 60%. Although certain models achieve near-perfect scores on specific benchmarks (e.g., Dense Retrieval models attaining 91% recall on Ubuntu IRC (Lan et al., 2022; Kummerfeld et al., 2019)), such datasets are often confined to niche domains, limiting their real-world applicability and transferability. Moreover, SOTA models for multi-modal benchmarks remain scarce, despite the critical role of cross-modal information in enhancing MPCA performance.

Overall, these observations underscore the limitations of current benchmarking practices and highlight the need for more comprehensive, diverse, and multi-modal datasets to advance future research.

7 Future Directions

Building intelligent MPCAs still remains an open challenge. While our survey highlights significant progress, there remain large gaps in performance. We recommend focusing on three major research thrusts as follows to tackle these challenges.

Theory of Mind: Theory of Mind (ToM) refers to the ability to understand, reason, and build beliefs about what participants might be thinking or feeling at each step of a conversation, even if those thoughts and feelings are not explicitly expressed (Premack and Woodruff, 1978). ToM enables agents to reason and infer the beliefs and intentions of their own and other participants.

ToM is an essential skill for MPCAs as it falls under the umbrella of *State of Mind Modeling* (Vanderlyn et al., 2025). Techniques like MindDial (Qiu et al., 2024) and benchmarks like MuMA-TOM (Shi et al., 2025) and KokoMind (Shi et al., 2023) show promise in modeling of ToM in conversations. However, the benchmarking is done on synthetic datasets, thereby not emulating realistic multi-party settings for ToM. There are works probing LLMs for ToM reasoning capabilities (Chen et al., 2024; Hou et al., 2024), but studies show that scaling up LLMs does not always improve ToM reasoning (Sclar et al., 2023). On standard ToM tasks, some LLMs achieve near-human performance (Ma et al., 2023b), suggesting the ability to track beliefs and intentions. Nevertheless, there is an ongoing debate as to whether these models truly reason or simply memorize patterns about mental states (Ullman, 2023; Kosinski, 2024). Latest benchmarks, ToMATO (Shinoda et al., 2025), reveal gaps for strong models like GPT-4.

Also, there is little work in integrating ToM capabilities into other MPCA tasks. Our intuition is that such integration will improve MPCAs performance. Table 2 is a small experiment where we show this by finetuning two open-source models, LLaMA3.1 8B and Mistral v0.3 7B, on a popular dialogue act recognition dataset, EMOTyDA (Saha et al., 2020), which also contains emotion tags of participants. We model two scenarios, one where the models predict the action based solely on the conversation till now (Act) and one where the emotions of participants are provided as additional context. The table shows that the performance of small language models improves when trained to perform dialogue act classification with emotion recognition (one capability identified in ToM) as additional context. If belief tracking (a core part of ToM) could also be integrated in a similar manner, it could lead to more intelligent MPCAs.

Model	Task	F1-Score
LLaMA3.1 8B	Act	55.01%
	Act w. Emotion	**55.80%**
Mistral v0.3 7B	Act	43.34%
	Act w. Emotion	**55.94%**

Table 2: Results of open-source LLMs finetuned (1 epoch, temperature=0.01) on the EMOTyDA dataset for the dialogue act recognition task without and with participant emotion as additional context. We see that emotion as context improves performance in both cases, highlighting that state of mind modeling, and thereby ToM, is heavily intertwined with other MPCA tasks.

To address all the gaps highlighted till now, future research should explore transforming traditional state-of-mind modeling tasks to scale ToM modeling for MPCAs, integrating emotion, persona, engagement, and intention reasoning along with belief tracking.

Multi-Modal Fusion and Grounding: From Section 6, we see that multi-modal modeling is an underexplored area in key MPC tasks like addressee selection, conversation disentanglement, and agent action modeling. Conversation disentanglement, in particular, could benefit from integrating modalities like spatial audio and/or face tracking, but most works are still text-based. In some cases, multi-modal dialogue models with video context fail to consistently outperform text-only baselines on tasks like dialogue act recognition and response generation (Wang et al., 2024e). Additionally, there is a lack of multi-modal ToM benchmarking despite promising results on synthetic datasets (Shi et al., 2025). Future research directions should develop methods to fuse different modalities effectively to address these identified gaps to create more robust MPCAs.

Evaluation Benchmarks: xAs per Section 6, we observe three limitations of current MPCA evaluation benchmarks: (i) each dataset focuses on a single specific skill of MPCAs; (ii) lack of benchmarks for multi-modal settings; and (iii) lack of realistic metrics/simulations for evaluation. There have been works addressing the lack of multi-modal benchmarks (Mahajan and Shaikh, 2021) through developing datasets like Friends-MMC (Wang et al., 2024e) for response prediction, and MuMA-TOM (Shi et al., 2025) for mental-state Q&A. However, they are mostly synthetic data covering individual tasks, thereby not solving the first. Furthermore, most evaluation is done using static metrics like BLEU or accuracy, which do not fully capture the performance of MPCAs' interactions. Future research must create real-world multi-modal benchmarks with simulation-based metrics where an MPCA is inserted into multi-party settings to interact in real time (Gunson et al., 2024).

8 Conclusion

The development of multi-party conversation agents has come a long way from text-based machine learning to multi-modal and LLM-based approaches. In this survey paper, we discussed the recent literature on MPCAs by defining a thematic taxonomy and categorizing works/tasks according to it. Our survey provides details on how different methods solve each problem of MPCAs. It also provides insights into the limitations and challenges faced when developing MPCAs. Finally, we discuss future directions for this field of research, specifically how Theory of Mind, multi-modal fusion, and grounding, and better benchmarks can help us develop better MPCAs. Addressing these aspects will advance MPCAs toward genuinely engaging and socially aware interactions, bridging the gap between current systems and human-like conversational intelligence.

Limitations

Our survey primarily focuses on research published after 2021, which may exclude some foundational or emerging contributions outside this period. Given the extensive body of related work, we acknowledge the possibility of overlooking equally significant studies. The future directions we propose are based on our interpretation of current research trends and supported by cursory experiments, which may not capture all perspectives. Additionally, many works we reviewed rely on custom datasets and models, some of which do not align with recent standard benchmarks. This reliance may introduce biases in our analysis. We recommend further, more comprehensive research to cross-verify our findings.

References

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party Goal Tracking with LLMs: Comparing Pre-training, Fine-tuning, and Prompt Engineering. *arXiv preprint*. ArXiv:2308.15231.

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zheng-Yu Niu. 2022. PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation. *arXiv preprint*. ArXiv:2109.09519.

Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning Multi-Party Turn-Taking Models from Dialogue Logs. *arXiv preprint*. ArXiv:1907.02090.

Aakash Bhatnagar, Nidhir Bhavsar, Muskaan Singh, and Petr Motlicek. 2022. An End-to-End Multilingual System for Automatic Minuting of Multi-Party Dialogues. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 582–589, Manila, Philippines. Association for Computational Linguistics.

Karan Bhukar, Harshit Kumar, Dinesh Raghu, and Ajay Gupta. 2023. End-to-End Deep Reinforcement Learning for Conversation Disentanglement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12571–12579. Number: 11.

Dan Bohus and Eric Horvitz. 2009. Models for Multiparty Engagement in Open-World Dialog. In *Proceedings of the SIGDIAL 2009 Conference*, pages 225–234, London, UK. Association for Computational Linguistics.

Laura Cabrera-Quiros, Ekin Gedik, and Hayley Hung. 2022. Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras. *IEEE Transactions on Affective Computing*, 13(1):46–59.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: a pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.

Kent Chang, Danica Chen, and David Bamman. 2023. Dramatic conversation disentanglement. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4020–4046, Toronto, Canada. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. *MHaDiG*: A Multilingual Humor-aided Multiparty Dialogue Generation in multimodal conversational setting. *Knowledge-Based Systems*, 278:110840.

Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou. 2021. Multimodal Incremental Transformer with Visual Grounding for Visual Dialogue Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 436–446, Online. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Zelin Chen, Lvmin Liu, Yujie Wan, Yuzhong Chen, Chen Dong, Weiwei Li, and Yuhang Lin. 2023. Improving BERT with local context comprehension for multi-turn response selection in retrieval-based dialogue systems. *Computer Speech & Language*, 82:101525.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang.

2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Chernyavskiy and Dmitry Ilvovsky. 2023. Transformer-based Multi-Party Conversation Generation using Dialogue Discourse Acts Planning. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 519–529, Prague, Czechia. Association for Computational Linguistics.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4651–4660, Los Alamitos, CA, USA. IEEE Computer Society.

Philip R. Cohen, Lucian Galescu, and Maayan Shvo. 2024. An explainable collaborative dialogue system using a theory of mind. *Preprint*, arXiv:2302.09646.

M. R. Cunningham. 1977. Personality and the structure of the nonverbal communication of emotion. *Journal of Personality*, 45:564–584.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842.

Mika Enomoto, Yasuharu Den, and Yuichi Ishimoto. 2020. A Conversation-Analytic Annotation of Turn-Taking Behavior in Japanese Multi-Party Conversation and its Preliminary Analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 644–652, Marseille, France. European Language Resources Association.

Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2024. Improving Multi-party Dialogue Generation via Topic and Rhetorical Coherence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3240–3253, Miami, Florida, USA. Association for Computational Linguistics.

Dmitrii Fedotov, Olga Perepelkina, Evdokia Kazimirova, Maria Konstantinova, and Wolfgang Minker. 2018. Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, MCPMD '18, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023a. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Jiazhan Feng, Chongyang Tao, Xueliang Zhao, and Dongyan Zhao. 2023b. Learning Multi-turn Response Selection in Grounded Dialogues with Reinforced Knowledge and Context Distillation. *ACM Trans. Inf. Syst.*, 41(4):115:1–115:27.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 259–273, Kyoto, Japan. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. *arXiv preprint*. ArXiv:2105.12544 [cs].

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154, Toronto, Canada. Association for Computational Linguistics.

Jingsheng Gao, Zeyu Li, Suncheng Xiang, Ting Liu, and Yuzhuo Fu. 2023. CluCDD:Contrastive Dialogue Disentanglement via Clustering. *arXiv preprint*. ArXiv:2302.08146 [cs].

Patrick Gebhard. 2005. Alma: a layered model of affect. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '05, page 29–36, New York, NY, USA. Association for Computing Machinery.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Chen Gong, DeXin Kong, Suxian Zhao, Xingyu Li, and Guohong Fu. 2024. MODDP: A Multi-modal Open-domain Chinese Dataset for Dialogue Discourse Parsing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10561–10573, Bangkok, Thailand. Association for Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2041–2044, New York, NY, USA. Association for Computing Machinery.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Cong Liu, and Guoping Hu. 2023. GIFT: Graph-induced fine-tuning for multi-party conversation understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11645–11658, Toronto, Canada. Association for Computational Linguistics.

Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021a. Partner Matters! An Empirical Study on Fusing Personas for Personalized Response Selection in Retrieval-Based Chatbots. *arXiv preprint*. ArXiv:2105.09050 [cs].

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097, Dublin, Ireland. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021b. MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

Nancie Gunson, Angus Addlessee, Daniel Hernández García, Marta Romeo, Christian Dondrup, and Oliver Lemon. 2024. A holistic evaluation methodology for multi-party spoken conversational agents. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, IVA '24, New York, NY, USA. Association for Computing Machinery.

Kenta Hama, Atsushi Otsuka, and Ryo Ishii. 2024. Emotion recognition in conversation with multi-step prompting using large language model. In *Social Computing and Social Media*, pages 338–346, Cham. Springer Nature Switzerland.

Qinyu Han, Zhihao Yang, Hongfei Lin, and Tian Qin. 2024. Let Topic Flow: A Unified Topic-Guided Segment-Wise Dialogue Summarization Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2021–2032. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. 1993. Emotional contagion. *Current Directions in Psychological Science*, 2(3):96–100. Original work published 1993.

Xiaoming He, Yunzhe Jiang, Huajun Cui, and Yinqiu Liu. 2024. Addressee Selection for Multi-Party Conversations in Cyber-Physical-Social Systems. In *2024 IEEE Cyber Science and Technology Congress (CyberSciTech)*, pages 180–186.

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021a. Multi-tasking Dialogue Comprehension with Discourse Parsing. *arXiv preprint*. ArXiv:2110.03269 [cs].

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021b. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2150–2157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11532–11547, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling ai agent participation in group conversations: A human-centered approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 390–408, New York, NY, USA. Association for Computing Machinery.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotion-Lines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhongtian Hu, Yiwen Cui, Ronghan Li, Meng Zhao, and Lifang Wang. 2025. Can xLLMs Understand the Structure of Dialog? Exploring Multilingual Response Generation in Complex Scenarios. *arXiv preprint*. ArXiv:2501.11269 [cs].

Chengyu Huang, Zheng Zhang, Hao Fei, and Lizi Liao. 2022. Conversation Disentanglement with Bi-Level Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2985–2996, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bahar Irfan, Anika Narayanan, and James Kennedy. 2020. Dynamic Emotional Language Adaptation in Multiparty Interactions with Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Raghav Jain, Tulika Saha, Jhagrut Lalwani, and Sriparna Saha. 2023. Can you Summarize my learnings? Towards Perspective-based Educational Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3158–3173, Singapore. Association for Computational Linguistics.

Hyunbae Jeon, Frederic Guintu, and Rayvant Sahni. 2024. Lla-VAP: LSTM Ensemble of Llama and VAP for Turn-Taking Prediction. *arXiv preprint*. ArXiv:2412.18061 [cs].

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn Response Selection using Dialogue Dependency Relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. 2019. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. *Preprint*, arXiv:1911.09304.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Dongshi Ju, Shi Feng, Pengcheng Lv, Daling Wang, and Yifei Zhang. 2022. Learning to Improve Persona Consistency in Multi-party Dialogue Generation via Text Knowledge Enhancement. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 298–309, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zerrin Kasap, Maher Ben Moussa, Parag Chaudhuri, and Nadia Magnenat-Thalmann. 2009. Making them remember–emotional virtual characters with memory. *IEEE computer graphics and applications*, 29(2):20–29.

Tanmay Khule, Rishabh Agrawal, and Apurva Narayan. 2024. PFA-ERC: Psuedo-future augmented dynamic emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16196–16207, Miami, Florida, USA. Association for Computational Linguistics.

Yeo Jin Kim, Halim Acosta, Wookhee Min, Jonathan Rowe, Bradford Mott, Snigdha Chaturvedi, and James Lester. 2024. Dual process masking for dialogue act recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15270–15283, Miami, Florida, USA. Association for Computational Linguistics.

Ryo Kimura and Shogo Okada. 2020. Personality trait classification based on co-occurrence pattern modeling with convolutional neural network. In *HCI International 2020 – Late Breaking Papers: Interaction, Knowledge and Social Media: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, page 359–370, Berlin, Heidelberg. Springer-Verlag.

Stefan Kopp and Nicole Krämer. 2021. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, Volume 12 - 2021.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Heyan Huang, and Xian-Ling Mao. 2022. Exploring Dense Retrieval for Dialogue Response Selection. *arXiv preprint*. ArXiv:2110.06612 [cs].

Meng-Chen Lee and Zhigang Deng. 2024. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*, ICMI '24, pages 57–65, New York, NY, USA. Association for Computing Machinery.

Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, pages 436–444, New York, NY, USA. Association for Computing Machinery.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *Preprint*, arXiv:2309.11911.

Aaron W. Li, Veronica Jiang, Steven Y. Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2020a. Aloha: Artificial learning of human attributes for dialogue agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8155–8163.

Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-seng Chua, and Donghong Ji. 2024a. Revisiting conversation discourse for dialogue disentanglement. *ACM Trans. Inf. Syst.*, 43(1).

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020b. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingyang Li, Shengli Song, Yixin Li, Hanxiao Zhang, and Guangneng Hu. 2024b. ChatMDG: A discourse parsing graph fusion based approach for multi-party dialogue generation. *Information Fusion*, 110:102469.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021a. Dialogue History Matters! Personalized Response Selection in Multi-Turn Retrieval-Based Chatbots. *ACM Trans. Inf. Syst.*, 39(4):45:1–45:25.

Liyuan Li, Qianli Xu, and Yeow Kee Tan. 2012. Attention-based addressee selection for service and social robots to interact with multiple persons. In *Proceedings of the Workshop at SIGGRAPH Asia*, WASA '12, pages 131–136, New York, NY, USA. Association for Computing Machinery.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11002–11010.

Tianda Li, Jia-Chen Gu, Xiaodan Zhu, Quan Liu, Zhen-Hua Ling, Zhiming Su, and Si Wei. 2021b. DialBERT: A Hierarchical Pre-Trained Model for Conversation Disentanglement. *arXiv preprint*. ArXiv:2004.03760 [cs].

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024c. Big5-chat: Shaping llm personalities through training on human-grounded data. *Preprint*, arXiv:2410.16491.

Yameng Li, Shi Feng, Daling Wang, Yifei Zhang, and Xiaocui Yang. 2025. PAPER: A Persona-Aware Chain-of-Thought Learning Framework for Personalized Dialogue Response Generation. In *Natural Language Processing and Chinese Computing*, pages 215–227, Singapore. Springer Nature.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yiyang Li, Xinting Huang, Wei Bi, and Hai Zhao. 2023. Pre-training Multi-party Dialogue Models with Latent Discourse Inference. *arXiv preprint*. ArXiv:2305.15175.

Yiyang Li and Hai Zhao. 2021. Self- and Pseudo-self-supervised Prediction of Speaker and Key-utterance for Multi-party Dialogue Reading Comprehension. *arXiv preprint*. ArXiv:2109.03772.

Yiyang Li and Hai Zhao. 2023. EM Pre-training for Multi-party Dialogue Response Generation. *arXiv preprint*. ArXiv:2305.12412.

Yuntao Li, Can Xu, Huang Hu, Lei Sha, Yan Zhang, and Daxin Jiang. 2021c. Small Changes Make Big Differences: Improving Multi-turn Response Selection in Dialogue Systems via Fine-Grained Contrastive Learning. *arXiv preprint*. ArXiv:2111.10154 [cs].

Yun-Shao Lin and Chi-Chun Lee. 2018. Using interlocutor-modulated attention BLSTM to predict personality traits in small group interaction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*, pages 163–169. ACM.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Austin, Texas. Association for Computational Linguistics.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021a. Unsupervised Conversation Disentanglement through Co-Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2345–2356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shannan Liu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Enhancing Multi-party Dialogue Discourse Parsing with Explanation Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1531–1544, Abu Dhabi, UAE. Association for Computational Linguistics.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengyuan Liu and Nancy F. Chen. 2021. Improving Multi-Party Dialogue Discourse Parsing via Domain Integration. *arXiv preprint*. ArXiv:2110.04526.

Zhengyuan Liu, Ke Shi, and Nancy F. Chen. 2021b. Coreference-Aware Dialogue Summarization. *arXiv preprint*. ArXiv:2106.08556 [cs].

Zhipeng Liu, Xiaoming Zhang, Litian Zhang, and Zelong Yu. 2024. MDS: A fine-grained dataset for multi-modal dialogue summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11123–11137, Torino, Italia. ELRA and ICCL.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Ge Luo, Manman Zhang, Yuchen Ma, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. 2023. Forward Creation, Reverse Selection: Achieving Highly Pertinent Multimodal Responses in Dialogue Contexts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 1687–1696, New York, NY, USA. Association for Computing Machinery.

Cheng Charles Ma, Kevin Hyekang Joo, Alexandria K. Vail, Sunreeta Bhattacharya, Álvaro Fernández García, Kailana Baker-Matsuoka, Sheryl Mathew, Lori L. Holt, and Fernando De la Torre. 2024. Multimodal Fusion with LLMs for Engagement Prediction in Natural Conversation. *arXiv preprint*. ArXiv:2409.09135 [cs].

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2023a. Enhanced speaker-aware multi-party multi-turn dialogue comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2410–2423.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023b. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Elijah Mayfield, David Adamson, and Carolyn Rose. 2012. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 60–69.

Md Messal Monem Miah, Adarsh Pyarelal, and Ruihong Huang. 2023. Hierarchical fusion for online multimodal dialog act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7532–7545, Singapore. Association for Computational Linguistics.

Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna, and Issam Laradji. 2023. LLM aided semi-supervision for efficient Extractive Dialog Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10002–10009, Singapore. Association for Computational Linguistics.

Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2024. Multimediate'24: Multi-domain engagement estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11377–11382, New York, NY, USA. Association for Computing Machinery.

Ryota Nonomura and Hiroki Mori. 2024. Who Speaks Next? Multi-party AI Discussion Leveraging the Systematics of Turn-taking in Murder Mystery Games. *arXiv preprint*. ArXiv:2412.04937 [cs] version: 1.

Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 15–22, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and Response Selection for Multi-Party Conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143, Austin, Texas. Association for Computational Linguistics.

Maike Paetzel-Prüsmann and James Kennedy. 2023. Improving a Robot's Turn-Taking Behavior in Dynamic Multiparty Interactions. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 411–415, New York, NY, USA. Association for Computing Machinery.

Maria J. Pinto and Tony Belpaeme. 2024. Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1733–1738. ISSN: 1944-9437.

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.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8665–8672.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13709–13717.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Sanjana Ramprasad, Elisa Ferracane, and Zachary C. Lipton. 2024. Analyzing LLM Behavior in Dialogue Summarization: Unveiling Circumstantial Hallucination Trends. *arXiv preprint*. ArXiv:2406.03487 [cs].

Sandratra Rasendrasoa, Alexandre Pauchet, Julien Saunier, and Sébastien Adam. 2022. Real-time multimodal emotion recognition in conversation for multiparty interactions. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 395–403, New York, NY, USA. Association for Computing Machinery.

M. G. Rijpma, M. Montembeault, S. Shdo, J. H. Kramer, B. L. Miller, and K. P. Rankin. 2023. Semantic knowledge of social interactions is mediated by the hedonic evaluation system in the brain. *Cortex; a journal devoted to the study of the nervous system and behavior*, 161:26–37.

Esteban A. Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Personality recognition in conversations using capsule neural networks. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 180–187.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Meta-learning based deferred optimisation for sentiment and emotion aware multi-modal dialogue act classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 978–990, Online only. Association for Computational Linguistics.

Hanan Salam and Mohamed Chetouani. 2015. Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 341–347. ISSN: 2156-8111.

David Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14(3-2):816–832.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. Akan cinematic emotions (ace): A multimodal multi-party dataset for emotion recognition in movie dialogues. *Preprint*, arXiv:2502.10973.

Motoki Sato, Hiroki Ouchi, and Yuta Tsuboi. 2018. Addressee and Response Selection for Multilingual Conversation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3631–3644, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Preprint*, arXiv:2012.08695.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *arXiv preprint*. ArXiv:2408.12574 [cs].

Weiyan Shi, Liang Qiu, Dehong Xu, Pengwei Sui, Pan Lu, and Zhou Yu. 2023. Kokomind: Can large language models understand social interactions?

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, volume 33 of *AAAI'19/IAAI'19/EAAI'19*, pages 7007–7014, Honolulu, Hawaii, USA. AAAI Press.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. *Preprint*, arXiv:2501.08838.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Qi Song, Sheng Li, Ping Wei, Ge Luo, Xinpeng Zhang, and Zhenxing Qian. 2022a. Joint Learning for Addressee Selection and Response Generation in Multi-Party Conversation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6587–6591. ISSN: 2379-190X.

Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. 2023. Learning person-specific cognition from facial reactions for automatic personality recognition. *IEEE Transactions on Affective Computing*, 14(4):3048–3065.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022b. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jindian Su, Shanshan Yu, Xiaobin Ye, and Dandan Ma. 2024. BERT-KRS: A BERT-Based Model for Knowledge-Grounded Response Selection in Retrieval-Based Chatbots. In *Applied Intelligence*, pages 310–321, Singapore. Springer Nature.

Kazuyoshi Sugawara. 2012. Interactive significance of simultaneous discourse or overlap in everyday conversations amongl gui former foragers. *Journal of Pragmatics*, 44(5):577–618.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Joshua T, Arnav Agarwal, Shriya Sanjay, Yash Sarda, John Sahaya Rani Alex, Saurav Gupta, Sushant Kumar, and Vishwanath Kamath. 2024. Thread Detection and Response Generation Using Transformers with Prompt Optimisation. In *2024 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. ISSN: 2474-915X.

H. Tajfel and J.C. Turner. 1979. An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 33(47):74.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. 2023. Is ChatGPT a good multi-party conversation solver? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4905–4915, Singapore. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W. Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. *arXiv preprint*. ArXiv:2402.13249 [cs].

Yiming Tang, Abrar Anwar, and Jesse Thomason. 2025. M3pt: A transformer for multimodal, multi-party social signal prediction with person-aware blockwise attention. *Preprint*, arXiv:2501.13416.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An Incremental Discourse Parser. *arXiv preprint*. ArXiv:2406.18256 [cs] version: 3.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Preprint*, arXiv:2302.08399.

Lindsey Morgan Vanderlyn, Dirk Väth, and Thang Vu. 2025. Understanding the role of mental models in user interaction with an adaptive dialog agent. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 989–1015, Albuquerque, New Mexico. Association for Computational Linguistics.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA. Association for Computational Linguistics.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A Structure Self-Aware Model for Discourse Parsing on Multi-Party Dialogues. In *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21*, volume 4, pages 3943–3949. ISSN: 1045-0823.

Ante Wang, Linfeng Song, Lifeng Jin, Junfeng Yao, Haitao Mi, Chen Lin, Jinsong Su, and Dong Yu. 2023a. D^2PSG: Multi-Party Dialogue Discourse Parsing as Sequence Generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:4004–4013. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023b. Instructive Dialogue Summarization with Query Aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023c. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143, Singapore. Association for Computational Linguistics.

Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiao-Yong Wei. 2024a. Instruct Once, Chat Consistently in Multiple Rounds: An Efficient Tuning Framework for Dialogue. *arXiv preprint*. ArXiv:2402.06967 [cs].

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024b. Turn-Taking and Backchannel Prediction with Acoustic and Large Language Model Fusion. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125. ISSN: 2379-190X.

Weishi Wang, Shafiq Joty, and Steven C. H. Hoi. 2020. Response Selection for Multi-Party Conversations with Dynamic Topic Tracking. *arXiv preprint*. ArXiv:2010.07785.

Xiaoyu Wang, Ningyuan Xi, Teng Chen, Qingqing Gu, Yue Zhao, Xiaokai Chen, Zhonglin Jiang, Yong Chen, and Luo Ji. 2024c. Multi-Party Supervised Fine-tuning of Language Models for Multi-Party Dialogue Generation. *arXiv preprint*. ArXiv:2412.05342 [cs].

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024d. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Qun Liu, and Dongyan Zhao. 2024e. Friends-mmc: A dataset for multi-modal multi-party conversation understanding. *Preprint*, arXiv:2412.17295.

B. Webber, M. Egg, and V. Kordoni. 2011. Discourse structure and language technology. *Nat. Lang. Eng.*, 18(4):437–490.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-Party Chat: Conversational Agents in Group Settings with Humans and Models. *arXiv preprint*. ArXiv:2304.13835.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Ruosong Yang, and Shuaiqi Liu. 2024. Affective- nli: Towards accurate and interpretable personality recognition in conversation. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 184–193.

Philine Witzig, Rares Constantin, Nikola Kovacevic, and Rafael Wampfler. 2024. Multimodal dialog act classification for digital character conversations. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *Preprint*, arXiv:2407.21315.

Yunhe Xie, Chengjie Sun, Bingquan Liu, and Zhenzhou Ji. 2024a. Multi-view contrastive parsing network for emotion recognition in multi-party conversations. *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yunhe Xie, Chengjie Sun, Yifan Liu, Zhenzhou Ji, and Bingquan Liu. 2024b. UniMPC: Towards a Unified Framework for Multi-Party Conversations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 2639–2649, New York, NY, USA. Association for Computing Machinery.

Bowen Xing and Ivor W Tsang. 2023. Relational temporal graph reasoning for dual-task dialogue language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13170–13184.

Feifei Xu and Guangzhen Li. 2023. Multiple information extraction and interaction for emotion recognition in multi-party conversation. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 977–982.

Qianli Xu, Liyuan Li, and Gang Wang. 2013. Designing engagement-aware agents for multiparty conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2233–2242, New York, NY, USA. Association for Computing Machinery.

Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Narrate Dialogues for Better Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3565–3575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ze Yang, Wei Wu, Huang Hu, Can Xu, Wei Wang, and Zhoujun Li. 2021. Open domain dialogue generation with latent images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14239–14247.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

Yuan-Jhe Yin, Bo-Yu Chen, and Berlin Chen. 2024. A Novel LLM-based Two-stage Summarization Approach for Long Dialogues. *arXiv preprint*. ArXiv:2410.06520 [cs].

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Kang Zhang, Yu-Jung Heo, Du-Seong Chang, and Chang D. Yoo. 2024. BI-MDRG: Bridging Image History in Multimodal Dialogue Response Generation. *arXiv preprint*. ArXiv:2408.05926 [cs].

Mingzhi Yu, Emer Gilmartin, and Diane Litman. 2019. Identifying personality traits using overlap dynamics in multiparty dialogue. *Preprint*, arXiv:1909.00876.

Nan Yu, Guohong Fu, and Min Zhang. 2022. Speaker-Aware Discourse Parsing on Multi-Party Dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5372–5382, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peicong Yuan, Guoyong Cai, Ming Chen, and Xiaolv Tang. 2024. Topics guided multimodal fusion network for conversational emotion recognition. In *Advanced Intelligent Computing Technology and Applications*, pages 250–262, Singapore. Springer Nature Singapore.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *Preprint*, arXiv:1708.04299.

Bo Zhang, Hui Ma, Jian Ding, Jian Wang, Bo Xu, and Hongfei Lin. 2025. Distilling Implicit Multimodal Knowledge into Large Language Models for Zero-Resource Dialogue Generation. *Information Fusion*, 118:102985. ArXiv:2405.10121 [cs].

Bo Zhang, Jian Wang, Hui Ma, Bo Xu, and Hongfei Lin. 2023a. ZRIGF: An Innovative Multimodal Framework for Zero-Resource Image-Grounded Dialogue Generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pages 5464–5473, New York, NY, USA. Association for Computing Machinery.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024a. Mintrec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *Preprint*, arXiv:2403.10943.

Rongjunchen Zhang, Tingmin Wu, Sheng Wen, Surya Nepal, Cecile Paris, and Yang Xiang. 2023b. SAM: Multi-turn Response Selection Based on Semantic Awareness Matching. *ACM Trans. Internet Technol.*, 23(1):3:1–3:18.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2017. Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. *arXiv preprint*. ArXiv:1709.04005 [cs].

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. 2023c. M3gat: A multi-modal, multi-task interactive graph attention network for conversational sentiment analysis and emotion recognition. *ACM Trans. Inf. Syst.*, 42(1).

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2024b. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Preprint*, arXiv:2310.11374.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H. Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. *arXiv preprint*. ArXiv:2110.10150 [cs].

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *Preprint*, arXiv:2304.09582.

Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022. MuCDN: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. Ux research on conversational human-ai interaction: A literature review of the acm digital library. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, page 1–24. ACM.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multitask learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, Toronto, Canada. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773. Number: 10.

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.

Ling.Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307, Dublin, Ireland. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2024. Factual Dialogue Summarization via Learning from Large Language Models. *arXiv preprint*. ArXiv:2406.14709 [cs].

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Interspeech*.

Mateusz Żarkowski. 2019. Multi-party Turn-Taking in Repeated Human–Robot Interactions: An Interdisciplinary Evaluation. *International Journal of Social Robotics*, 11(5):693–707.

A MPC Thematic Taxonomy

To complement the brief taxonomy presented in Figure 2, we provide a comprehensive breakdown of the multi-party conversational studies in Figure 3. We expand each of the three themes—State of Mind Modeling, Semantic Understanding, and Agent Action Modeling—by listing all associated tasks and surveying representative works under each category. Further, we organize the literature based on the methodology adopted: traditional approaches, multi-modal systems, and LLM-based techniques. This detailed taxonomy serves as a valuable reference for understanding the full scope of MPC research and identifying key trends across different modeling strategies.

B MPC Datasets

In this section, we provide a comprehensive list of multi-party conversational (MPC) datasets we reviewed through our survey. We expand the datasets in Table 3 to include a broader range of datasets spanning diverse domains, modalities, and annotated task types. This detailed compilation supports further exploration and benchmarking across the different MPCA skill tasks discussed in this paper.

C Results

Table 4 provides a comprehensive overview of model performances across a wide range of MPCA tasks and datasets. This table captures the broader landscape of research efforts, including emerging baselines, multi-modal models, and task-specific strategies. Notably, we observe significant performance gaps across tasks like engagement detection, discourse parsing, and multi-modal response generation, underscoring persistent modeling challenges. Additionally, it reveals increased interest in multi-modal approaches, yet most still underperform compared to unimodal text-based models, highlighting the need for more robust multi-modal integration in future MPCA studies.

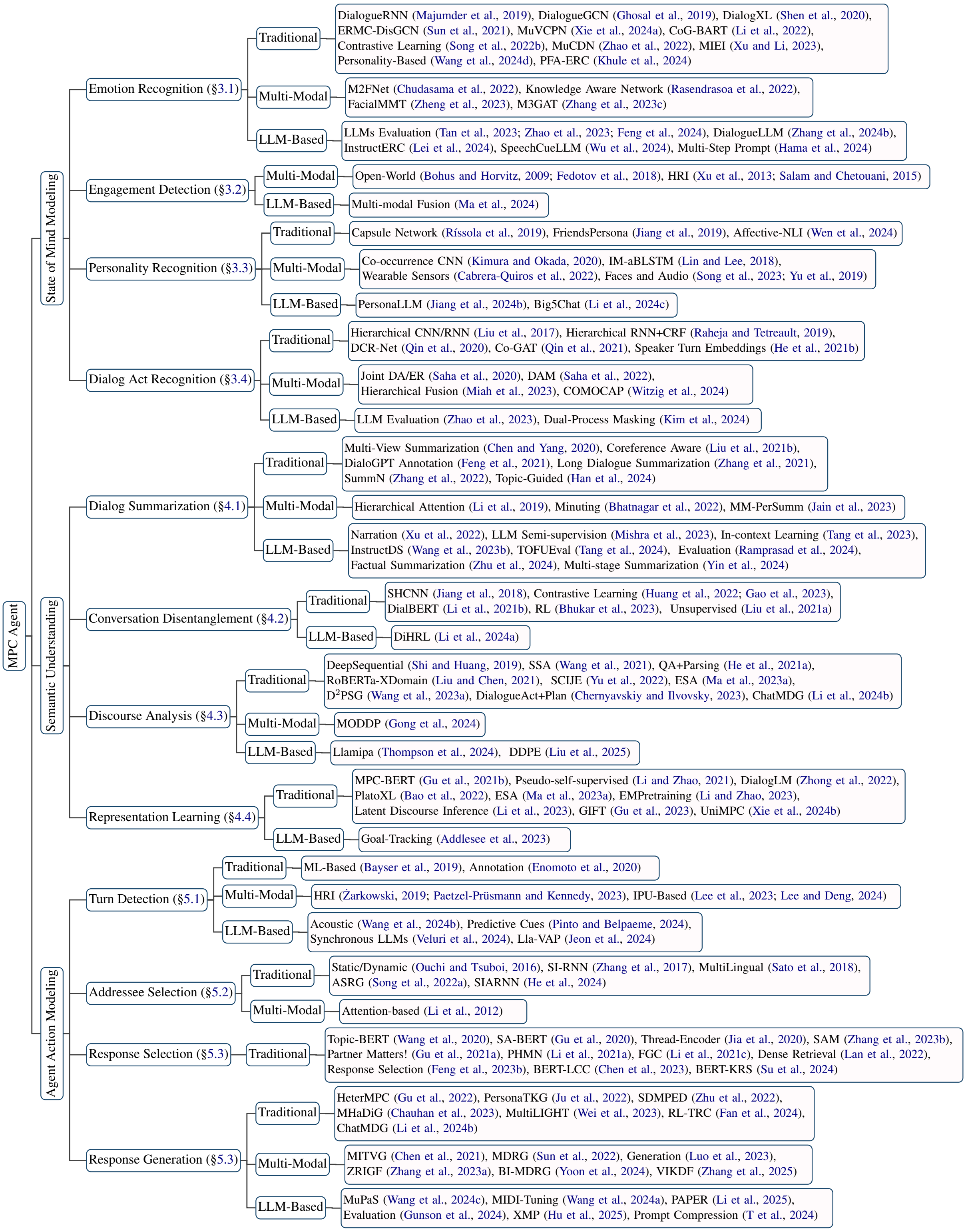

Figure 3: Full taxonomy of MPC tasks and recent work under them.

Dataset	**Task**	**Modality**	**Description**
MELD (Poria et al., 2019)	Emotion Recognition	Audio, Video, Text	1,400 dialogues and approximately 13,000 utterances extracted from the TV series Friends
EmotionLines (Hsu et al., 2018)	Emotion Recognition	Text	2,000 dialogues with a total of 29,245 utterances
Akan Cinematic Emotions (ACE) (Sasu et al., 2025)	Emotion Recognition	Audio, Video, Text	385 dialogues and 6,162 utterances sourced from 21 Akan-language movies
EmoryNLP (Zahiri and Choi, 2017)	Emotion Recognition, Personality Recognition	Text	897 conversations, 12606 utterances from the TV series Friends
EMOTyDA (Saha et al., 2020)	Emotion Recognition, Dialog Act Recognition	Audio, Video, Text	Annotated subset of MELD and IEMOCAP
DailyDialog (Li et al., 2017)	Emotion Recognition, Dialog Act Recognition	Text	13118 conversations, 102979 utterances crawled from English learning websites
AMI Meeting Corpus (Carletta et al., 2005)	Emotion Recognition, Personality Recognition, Dialog Act Recognition, Dialogue Summarization	Audio, Video, Text	100 hours of meeting recordings
MPIIGroupInteraction (Müller et al., 2024)	Engagement Detection	Audio, Video	22 group discussions each with 3 to 4 participants exceeding 440 minutes of total recordings
Emergent LEAder (ELEA) (Sanchez-Cortes et al., 2012)	Personality Recognition	Audio, Video	27 meeting recordings each with 3 to 4 participants performing winter survival task
Team Corpus (Litman et al., 2016)	Personality Recognition	Audio	47 hours of multiparty interaction from 62 teams (3 to 4 person) playing the collaborative board game Forbidden Island
MRDA Corpus (Shriberg et al., 2004)	Dialog Act Recognition	Audio, Text	75 meetings with an average of about six participants per meeting
MIntRec2.0 (Zhang et al., 2024a)	Dialog Act Recognition	Audio, Video, Text	1,245 conversations, 15,040 utterances from TV series: Superstore, The Big Bang Theory, and Friends
TopDial (Wang et al., 2023c)	Dialog Act Recognition	Text	18,000 dialogues with 12.3 average utterances per dialogue
QMSum (Zhong et al., 2021)	Dialogue Summarization	Text	1,808 query-summary pairs derived from 232 meetings across various domains
SAMSum (Gliwa et al., 2019)	Dialogue Summarization	Text	16,000 messenger-like conversations
MDS (Liu et al., 2024)	Dialogue Summarization	Audio, Video, Text	16,000 minutes of video clips, with dialogues from TV series such as "The Big Bang Theory" and TikTok videos
Ubuntu IRC (Kummerfeld et al., 2019)	Conversation Disentanglement, Addressee Selection	Text	19,560 conversations collected from Ubuntu Internet Relay Chat (IRC) logs
Movie Dialogue Dataset (Chang et al., 2023)	Conversation Disentanglement	Text	2209 conversational threads derived from scripts of 831 movies and TV series
Multi-Domain Chat Disentanglement Dataset (Li et al., 2024a)	Conversation Disentanglement	Text	875 conversations across 4 distinct IRC channels
STAC (Asher et al., 2016)	Discourse Parsing	Text	1200 dialogues from from an online game (Settlers of Catan)
Molweni (Li et al., 2020b)	Discourse Parsing	Text	10,000 dialogues built upon the multi-party subset of the Ubuntu Dialogue Corpus
MODDP (Gong et al., 2024)	Discourse Parsing	Audio, Video, Text	864 dialogues, 18,114 utterances Chinese dataset with 12.7 hours of video clips

Table 3: Extended list of multi-party conversational datasets

Model	Task	Dataset	Eval Metric	Eval Result
DialogueLLM (Zhang et al., 2024b)	Emotion Recognition	EmoryNLP (Zahiri and Choi, 2017)	F1 Score	40.05
DialogueLLM (Zhang et al., 2024b)	Emotion Recognition	MELD (Poria et al., 2019)	F1 Score	71.9
InstructERC (Lei et al., 2024)	Emotion Recognition	EmoryNLP (Zahiri and Choi, 2017)	F1 Score	41.37
InstructERC (Lei et al., 2024)	Emotion Recognition	MELD (Poria et al., 2019)	F1 Score	69.15
M3GAT (Zhang et al., 2023c)	Emotion Recognition	MELD (Poria et al., 2019)	F1 Score	40.53
PFA-ERC (Khule et al., 2024)	Emotion Recognition	EmoryNLP (Zahiri and Choi, 2017)	F1 Score	42.94
PFA-ERC (Khule et al., 2024)	Emotion Recognition	MELD (Poria et al., 2019)	F1 Score	68.73
SpeechCueLLM (Wu et al., 2024)	Emotion Recognition	MELD (Poria et al., 2019)	F1 Score	67.6
Multi-Modal Fusion (Ma et al., 2024)	Engagement Detection	Custom	RMSE	1.338
Affect-NLI (Wen et al., 2024)	Personality Recognition	EmoryNLP (Zahiri and Choi, 2017)	Accuracy	61.3
ChatGPT (Zhao et al., 2023)	Dialog Act	DailyDialog (Li et al., 2017)	F1 Score	70
COMOCAP (Witzig et al., 2024)	Dialog Act	DailyDialog (Li et al., 2017)	F1 Score	79.66
DAM (Text+Video) (Saha et al., 2022)	Dialog Act	EMOTyDA (Saha et al., 2020)	F1 Score	57.01
Hierarchical Fusion (Miah et al., 2023)	Dialog Act	EMOTyDA (Saha et al., 2020)	F1 Score	50.3
Hierarchical Fusion (Miah et al., 2023)	Dialog Act	MRDA Corpus (Shriberg et al., 2004)	F1 Score	29.11
In-Context Learning (Tang et al., 2023)	Dialog Summary	SAMSum (Gliwa et al., 2019)	ROUGE-1	34.6
InstructDS (Wang et al., 2023b)	Dialog Summary	SAMSum (Gliwa et al., 2019)	ROUGE-1	58.4
Minuting (Bhatnagar et al., 2022)	Dialog Summary	SAMSum (Gliwa et al., 2019)	ROUGE-1	45
Bi-Level CL (Huang et al., 2022)	Disentanglement	Ubuntu IRC (Kummerfeld et al., 2019)	F1 Score	70.4
CLUCDD (Gao et al., 2023)	Disentanglement	Ubuntu IRC (Kummerfeld et al., 2019)	F1 Score	58.42
DiHRL (Li et al., 2024a)	Disentanglement	Ubuntu IRC (Kummerfeld et al., 2019)	F1 Score	47.63
RL (Bhukar et al., 2023)	Disentanglement	Ubuntu IRC (Kummerfeld et al., 2019)	F1 Score	51.9
D2PSG (Wang et al., 2023a)	Discourse Parsing	STAC (Asher et al., 2016)	F1 Score	62.77
Deep Sequential (Shi and Huang, 2019)	Discourse Parsing	STAC (Asher et al., 2016)	F1 Score	55.7
Llamipa (Thompson et al., 2024)	Discourse Parsing	STAC (Asher et al., 2016)	F1 Score	60.7
RoBERTa-XDomain (Liu and Chen, 2021)	Discourse Parsing	STAC (Asher et al., 2016)	F1 Score	57.1
SCIJE (Yu et al., 2022)	Discourse Parsing	STAC (Asher et al., 2016)	F1 Score	57.4
IPU-Based (Lee and Deng, 2024)	Turn Detection	Custom	F1 Score	87.3
LLa-VAP (Jeon et al., 2024)	Turn Detection	CCPE (Radlinski et al., 2019)	F1 Score	83.13
Predictive Cues (Pinto and Belpaeme, 2024)	Turn Detection	DailyDialog (Li et al., 2017)	F1 Score	90.3
ASRG (Song et al., 2022a)	Addressee Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Accuracy	84.65
SI-RNN (Zhang et al., 2017)	Addressee Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Accuracy	80.47
BERT-KRS (Su et al., 2024)	Response Selection	Persona-Chat (Zhang et al., 2018)	Recall	82
BERT-LCC (Chen et al., 2023)	Response Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Recall	88.9
Dense Retrieval (Lan et al., 2022)	Response Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Recall	91
FGC (Li et al., 2021c)	Response Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Recall	88.6
Partner Matters! (Gu et al., 2021a)	Response Selection	Persona-Chat (Zhang et al., 2018)	Recall	70.7
PLMKS+PLMCS+PLMRanker (Feng et al., 2023b)	Response Selection	WoW (Dinan et al., 2019)	Recall	92.7
SA-BERT (Gu et al., 2020)	Response Selection	Ubuntu IRC (Kummerfeld et al., 2019)	Recall	85.5
BI-MDRG (Yoon et al., 2024)	Response Generation	MMDialog (Feng et al., 2023a)	BLEU	27.6
HeterMPC (Gu et al., 2022)	Response Generation	Ubuntu IRC (Kummerfeld et al., 2019)	BLEU	12.61
MIDI-Tuning (Wang et al., 2024a)	Response Generation	TopDial (Wang et al., 2023c)	BLEU	39.6
MuPaS (Wang et al., 2024c)	Response Generation	Custom	Human	8.16/10
PAPER (Li et al., 2025)	Response Generation	HLA-Chat++ (Li et al., 2020a)	BLEU	44.16
VIKDF (Zhang et al., 2025)	Response Generation	Reddit (Yang et al., 2021)	BLEU	16.47
ZRIGF (Zhang et al., 2023a)	Response Generation	Reddit (Dziri et al., 2019)	BLEU	16.06

Table 4: Performance of MPCA models on various tasks