The Imperative of Conversation Analysis in the Era of LLMs: A Survey of Tasks, Techniques, and Trends

Anonymous ACL submission

Abstract

In the era of large language models (LLMs), a vast amount of conversation logs will be accumulated thanks to the rapid development trend of language UI. Conversation Analysis (CA) strives to uncover and analyze critical information from conversation data, streamlining manual processes and supporting business insights and decision-making. The need for CA to extract actionable insights and drive empowerment is becoming increasingly prominent and attracting widespread attention. However, the lack of a clear scope for CA leads to a dispersion of various techniques, making it difficult to form a systematic technical synergy to empower business applications. In this paper, we perform a thorough review and systematize CA task to summarize the existing related work. Specifically, we formally define CA task to confront the fragmented and chaotic landscape in this field, and derive four key steps of CA from conversation scene reconstruction, to in-depth attribution analysis, and then to performing targeted training, finally generating conversations based on the targeted training for achieving the specific goals. In addition, we showcase the relevant benchmarks, discuss potential challenges and point out future directions in both industry and academia. In view of current advancements, it is evident that the majority of efforts are still concentrated on the analysis of shallow conversation elements, which presents a considerable gap between the research and business, and with the assist of LLMs, recent work has shown a trend towards research on causality and strategic tasks which are sophisticated and high-level. The analyzed experiences and insights will inevitably have broader application value in business operations that target conversation logs.

1 Introduction

With the development of large language models (LLMs), the next generation of system interactions is rapidly advancing towards natural language conversation interfaces (Language UI) (Wang et al., 2023a; Dong et al., 2023; Wu et al., 2024a), leading to a vast amount of natural language interaction logs. It is more valuable to extract, summarize, analyze, and reason from these conversation[1] logs, which produces numerous new applications such as system optimization (Xu et al., 2020; Liu et al., 2023a), customer operations (Wang et al., 2020; Zou et al., 2021), and demand insights (Deriu et al., 2021; Li et al., 2021; He et al., 2023). The goal of conversation analysis (CA) is to mine and analyze out critical information (such as customer profiles, purchase intentions, mood changes, sales skills, shortcomings and corresponding improvement proposals, etc.) from conversational data, which contributes to simplifying manual processes and assisting in business insights and decision-making.

In the era of small language models (SLMs),

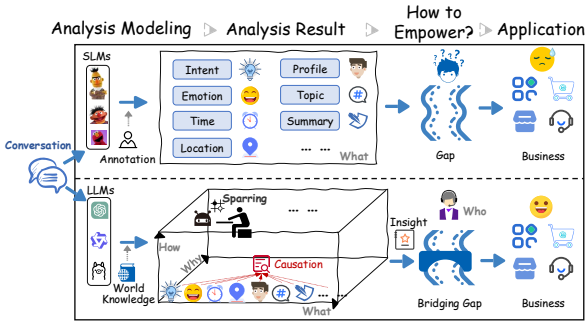

Figure 1: Conversation analysis results are usually shallow in the era of small language models (SLMs) and cannot well empower the real-world applications, while large language models (LLMs) possess world knowledge, and can contribute to in-depth attributions and provide constructive insights or experiences.

[1] In the literature, "conversation" and "dialogue" are both commonly used, referring to different communication types. CA aims to analyze any form and type of communications, such as human-human, human-machine, multi-party (Ganesh et al., 2023), etc. We refer to them collectively as "conversation" in this paper.

conversational data based analysis did not form a systematic task but was broken down into atomic tasks with strict output constraints, such as intent classification, slot filling (Qin et al., 2020; Louvan and Magnini, 2020; Jiang et al., 2023), and conversation summarization (Fabbri et al., 2021; Ouyang et al., 2023; Ramprasad et al., 2024) as shown in the upper part of Figure 1. Meanwhile, the analysis results generated by these atomic tasks are generally shallow, flat and there is a significant gap between the results and the genuine needs of the business, failing to address the industry's pain points. The core challenge was that small models could not model the world, thus they were unable to reconstruct the actual scene, such as conversation participants, scenarios, insightful rules from the "conversation". However, as shown in the lower part of Figure 1, with the rise of LLMs which fulfil a wealth of world knowledge (Zhang et al., 2023c; Yildirim and Paul, 2024; Zhao et al., 2024), it has become possible to infer detailed "real-world scene" from conversations and generate multifaceted and in-depth conversation analysis results, from "*What is the scene elements behind the conversation*", to "*Why do these elements exist*", then to "*How to deal with experience or problem*", finally answering "*Who should participant become*". The more detailed and accurate the inferred conversation information is, the more precise and valuable insights we can provide for business applications.

Despite achieving significant progress in the era of LLMs, CA continues to encounter certain difficulties: (1) **Definition**: there is a lack of a systematic definition of CA, which leads to a dispersion in the optimization objectives and technical aspects. (2) **Data**: there is a scarcity of CA datasets that encompass all essential elements of conversations, making it challenging to accurately model and evaluate the conversation background information, which affects the development of CA. (3) **Method**: Unlike flat text or documents, conversations inherently possess characteristics such as multi-turn interactions, strict contextual dependencies, implicit ambiguity, and colloquial language, which necessitate deeper modeling techniques. (4) **Application**: the current analysis yields rather shallow results when it comes to gauging emotions, opinions, and intentions, lacking a unified, profound, and constructive analytical perspective, which leads to the extensive gap between the realm of research and practical application.

However, there is currently a lack of technical surveys to showcase the research progress and to outline the research blueprint for CA. To bridge this gap, we make the first attempt to give the thorough review of existing research related to CA. According to the procedure of conversation analysis, the field of CA is segmented into four essential parts: (1) *Scene Reconstruction* for inferring the latent scene elements in conversation and relatively shallow analyses, (2) *Causality Analysis* for undertaking in-depth attributions, and (3) *Skill Enhancement* for training employees or models based on the attributions, and (4) *Conversation Generation* following the improved skills for generating conversations. Centered around the **goals** of system optimization (such as increasing customer retention rate), these four procedures seamlessly integrate from preliminary reconstruction, to in-depth attribution, and then to the targeted capability enhancement, ultimately re-generating the conversations and are continuously iterated in pursuit of more effectively achieving the goals. The major contributions of this survey are summarized as follows:

- **First technical review**: To our knowledge, we present the first technical CA survey from the viewpoint of goal-directed analysis with four critical procedures, integrating related techniques.
- **Systematic task definition**: Centered around the goal improvements, we give the clear task definition and optimization objectives of CA.
- **Abundant data resource**: We organize the existing data resources and evaluation criteria.
- **Insightful discussions**: We discuss the existing challenges and shed light on the future directions of CA in the era of LLMs.

2 Concept and Formulation

This section gives the definition of conversation analysis (CA), and then formally describes the critical procedures within CA.

Conversation analysis aims to identify critical information from human-human, human-machine, machine-machine, and multi-party conversations, derive the underlying causes, and develop the solutions to drive relevant improvements for more effective goal achievement continuously, such as *elevating customer experience*, *reducing complaint rate*. As shown in Figure 2, CA typically includes several components/procedures:

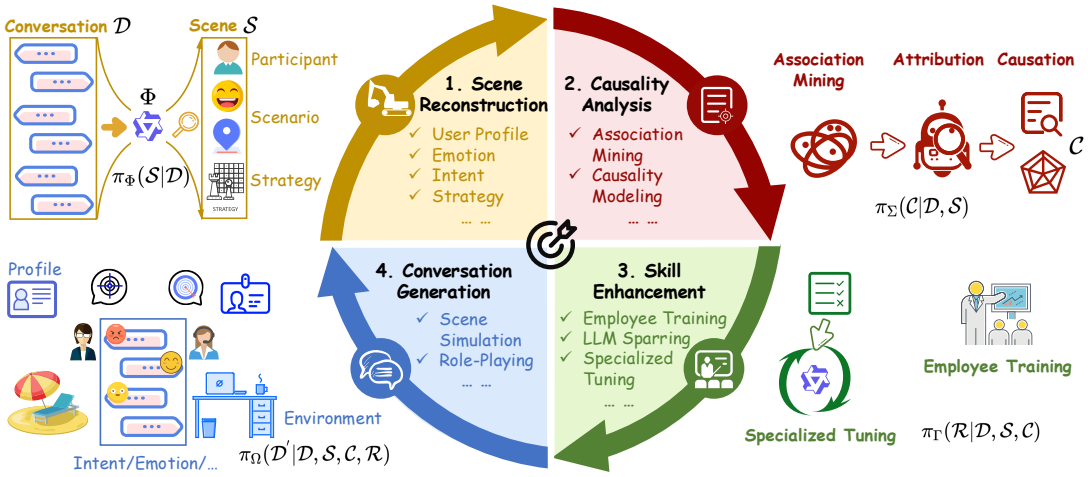

Figure 2: Overview of Conversation Analysis, which consists of 1) Scene Reconstruction, 2) Causality Analysis, 3) Skill Enhancement, and 4) Conversation Generation, centered around the goal-directed optimization.

1) *Scene Reconstruction*, 2) *Causality Analysis*, 3) *Skill Enhancement*, and 4) *Conversation Generation*. Specifically, CA system first analyzes the latent scene elements $\mathcal{S}$ such as participant profile, emotion, intent within the conversation, and then the factors behind these elements are uncovered systematically, drafting a comprehensive analytical report or making corresponding attributions $\mathcal{C}$. Furthermore, built upon these factors, the latent insights $\mathcal{R}$ of system can be improved via intelligent agent or employee training to facilitate the attainment of our goals $\mathcal{G}$. Finally, CA system generates the conversational content $\mathcal{D}$ following the improved insights for next-round iteration.

Scene Reconstruction. This procedure tends to attain the scene elements $\mathcal{S}$ (e.g., participants, scenarios including emotion, intent, environment, etc) from the conversational content $\mathcal{D}$. In fact, $\mathcal{D}$ arises within the specific scene $\mathcal{S}$, and this process aims to reconstruct the actual $\mathcal{S}$ according to $\mathcal{D}$ through tailor-designed modeling Φ. The scene reconstruction procedure can be formalized as $\pi_\Phi(\mathcal{S}|\mathcal{D})$. This procedure takes conversational data $\mathcal{D}$ as input and then restores the objective elements within it for contributing to insightful analyses in the subsequent steps.

Causality Analysis. This procedure seeks to delve deeply into the rationales $\mathcal{C}$ behind the components $\mathcal{S}$, such as the reasons for participants' attitude shifts, persuasive strategies. To achieve this, some causal modeling methods Σ can be developed. For instance, association mining can be performed within multi-group conversations and their corresponding scene components, and insightful reports and causation $\mathcal{C}$ are reached by condensing these items. The causality analysis procedure can be formalized as $\pi_\Sigma(\mathcal{C}|\mathcal{D},\mathcal{S})$. Causality analysis takes conversational content $\mathcal{D}$ and $\mathcal{S}$ reconstructed in §*Scene Reconstruction* as input and then derives more precise factors and induced reasons which contribute to the scene elements $\mathcal{S}$.

Skill Enhancement. This procedure strives to optimize the entire system towards the goal according to the causation $\mathcal{C}$. Concretely, for human participants such as call-center employees, there is a need for relevant departments to provide training Γ for human participants based on the feedback. As for AI agent, we need experts to optimize the agent model Γ for better insights $\mathcal{R}$ towards achieving the goals according to $\mathcal{C}$. The skill enhancement procedure can be formalized as $\pi_\Gamma(\mathcal{R}|\mathcal{D},\mathcal{S},\mathcal{C})$. Skill enhancement refers $\mathcal{D}$, $\mathcal{S}$, and $\mathcal{C}$ analyzed in §*Causality Analysis* to train the personnel or optimize the model for attaining the refined $\mathcal{R}$, which guides the direction towards achieving the goal.

Conversation Generation. This procedure aims to collect the conversational content $\mathcal{D}$ from real-world conversation data after employee training, such as online customer service and hotline data. In addition, we also drive the models to synthesize data based on the refined insights $\mathcal{R}$ (including scene elements, fundamental abilities, etc.) within the conversation via the techniques such as role-playing model Ω. The conversation generation procedure can be formalized as $\pi_\Omega(\mathcal{D}'|\mathcal{D},\mathcal{S},\mathcal{C},\mathcal{R})$. The conversation generation results in the collection of vast amounts of data, supporting thorough analysis of conversation and serving as the most

direct evidence of goal achievement.

According to multi-action reinforcement learning (Lei et al., 2022a; Li et al., 2022a; Wang et al., 2022b), the entire process of conversation analysis can be expressed formally as an Markov decision process (MDP) defined by a tuple ($\mathcal{D}$, $\mathcal{A}$, P, r). $\mathcal{D}$ is the state and represents the conversational content in CA. The action set $\mathcal{A} = \mathcal{S} \times \mathcal{C} \times \mathcal{R}$ is three-dimensional, in which $\mathcal{S}$ is the scene reconstruction sub-action set, $\mathcal{C}$ is the causality analysis sub-action set, and $\mathcal{R}$ is the skill enhancement sub-action set. P is the transition probability distribution and r is the reward function. The MDP with a multi-action space is referred to as multiple MDPs (MMDPs) in Lei et al. (2022a) and we defines the MMDPs over the states, actions, and reward in CA as follows:

- **States:** The global state is $\mathcal{D}_t$ at timestep t which denotes the conversational content generated subsequent to the t-th iteration. Following Wang et al. (2022b), the local states corresponding to the sub-action sets are enriched by a concatenation of the state $\mathcal{D}_t$ and the history of previous k-1 dimension sub-action choice.

- **Actions:** The actions at timestep t consist of scene reconstruction action $\mathcal{S}_t$, causality analysis action $\mathcal{C}_t$, and skill enhancement action $\mathcal{R}_t$, which correspond to analyzing the conversation scene, latent causation, and enhanced insights separately.

- **Policy:** The policy $\pi_\Theta(\mathcal{A}|\mathcal{D})$ consists of three stochastic sub-policies $\pi_\Phi(\mathcal{S}|\mathcal{D})$, $\pi_\Sigma(\mathcal{C}|\mathcal{D},\mathcal{S})$, and $\pi_\Gamma(\mathcal{R}|\mathcal{D},\mathcal{S},\mathcal{C})$, where Φ, Σ, and Γ are respectively models for scene reconstruction, causality analysis, and skill enhancement. $\Theta = [\Phi, \Sigma, \Gamma]$ is the entire set of all models. $\pi_\Theta(\mathcal{A}|\mathcal{D})$ selects a series of actions $\mathcal{A}$ (e.g., $\mathcal{S}$, $\mathcal{C}$, $\mathcal{R}$) in their respective sub-action space.

- **Transition:** At timestep t, given the sub actions $\mathcal{S}_t$, $\mathcal{C}_t$, $\mathcal{R}_t$, and state $\mathcal{D}_t$, the global state is updated as $\mathcal{D}_{t+1}$. The process happens to be the conversation generation $\pi_\Omega(\mathcal{D}^{'}|\mathcal{D},\mathcal{S},\mathcal{C},\mathcal{R})$, which can be reformulated as $P(\mathcal{D}_{t+1}|\mathcal{D}_t,\mathcal{S}_t,\mathcal{C}_t,\mathcal{R}_t)$.

- **Reward:** CA is to learn a whole policy to make the conversations achieve the goals such as *elevating customer experience*, *reducing complaint rate*, and its reward can be formalized as $\sum_t r(\mathcal{D}_t,\mathcal{S}_t,\mathcal{C}_t,\mathcal{R}_t)$.

3 Taxonomy of Conversation Analysis

This section mainly describes the research related to CA according to the taxonomy (shown in Fig. 3) of analysis procedure in brief, including ***scene reconstruction***, ***causality analysis***, ***skill enhancement***, and ***conversation generation***, centered around the ***goal-directed optimization***. The details are described in Appendix A.

Scene Reconstruction. A conversation contains conversational content, participants, scenarios (such as location, topic, atmosphere, emotion, intent) and so on. The elements other than the conversational content are collectively referred to as the scene. Scene reconstruction aims to mine these scene elements which are the background information of a conversation. Existing studies for scene elements include 1) *Participant* such as extracting personal attributes (Wang et al., 2022c; DeLucia et al., 2024), 2) *Scenario* such as intention detection (Sakurai and Miyao, 2024), conversation summarization (Tian et al., 2024), and 3) *Strategy* (Ou et al., 2024).

Causality Analysis. The process of conversation interaction is sophisticated with the latent induced factors. Analyzing these factors has important guiding significance for gaining a deeper understanding of the causes and development during conversations. By performing *causality modeling* (Zhou et al., 2021; Zhao et al., 2023; Chen et al., 2024a), and *generating analysis reports* (Wang et al., 2023d; Chiang et al., 2024), causality analysis aims to delve into the causes which lead to the emergence of scene elements, such as emotional changes and user personality.

Skill Enhancement. After conducting attribution analysis, it is necessary to better utilize these feedback to provide targeted training for individuals such as call-center employees or AI agents. For human participants, some improvement suggestions or suggested wording can be provided, while prompts can be optimized along with some foundational capabilities such as consistency or emotion support for AI agents (*Automatic Prompt Optimization* (Pryzant et al., 2023; Cheng et al., 2024) and *Specialized Tuning* (Wang et al., 2024a; Zhong et al., 2024)). In addition, AI agent can serve as sparring partners to help human enhance skills (*LLM Sparring* (Xiao et al., 2024)).

Conversation Generation. Conversation generation focuses on generating concrete conversational content relying on the refined experiences/insights

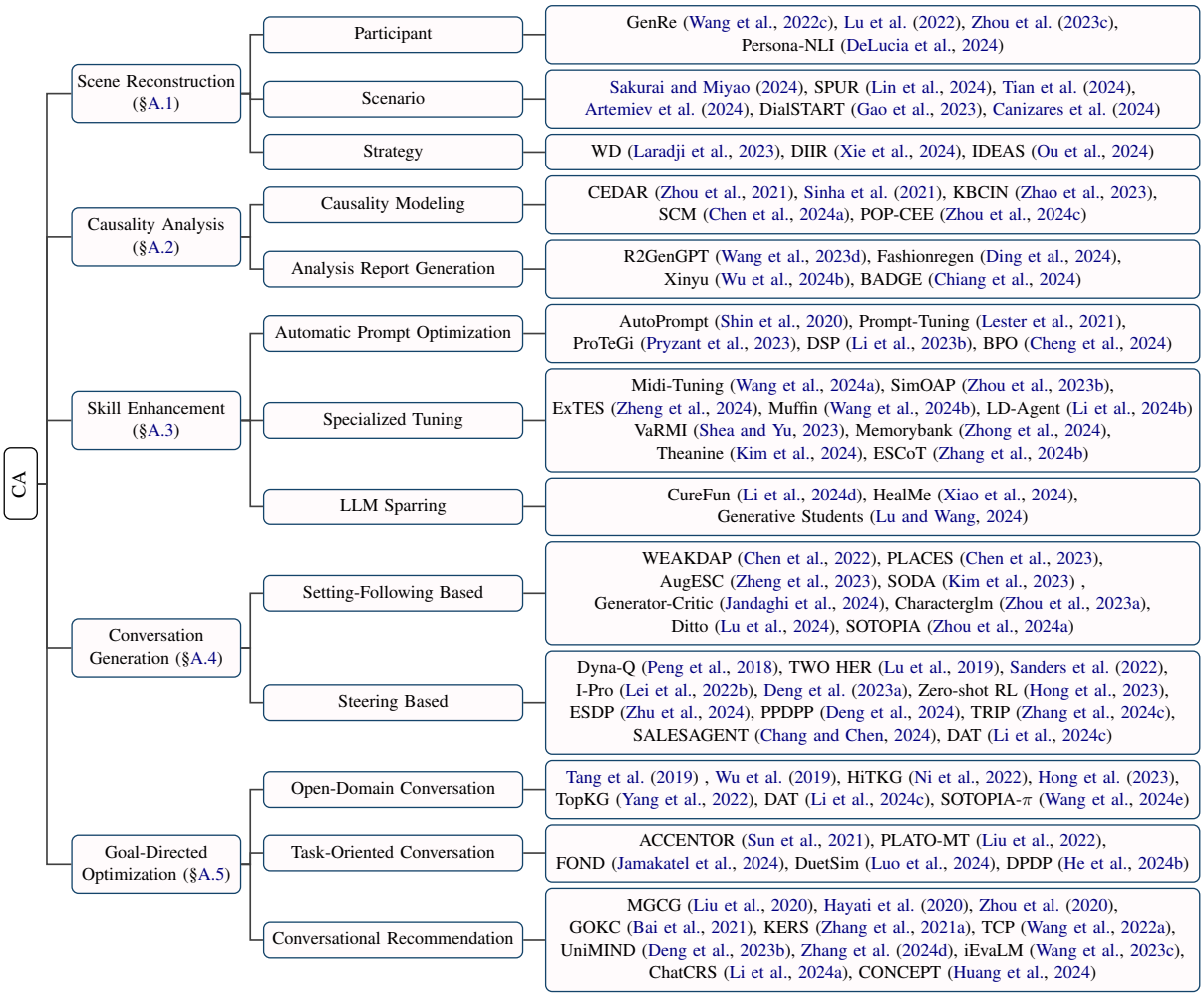

Figure 3: Taxonomy for Conversation Analysis (CA).

where the core issue is how to make the model follow these experience settings (*Setting-Following Based* (Jandaghi et al., 2024; Zhou et al., 2023a)) such as the refined scene elements and inject the analyzed insights into the model, steering agents to generate conversations towards specific goals (*Steering Based* (Chang and Chen, 2024)).

Goal-Directed Optimization. In fact, all conversations are arguably goal-directed (Searle, 1969; Austin, 1975), such as helpfulness, harmlessness in open-domain conversation, and completing booking of flight tickets in task-oriented conversation. The goal-directed modeling is gradually gaining attention in open-domain conversation (Tang et al., 2019; Li et al., 2024c), task-oriented conversation (Liu et al., 2022; Luo et al., 2024), and conversational recommendation (Li et al., 2024a). However, the goals studied in existing research are relatively shallow and narrow, and lack the universality, most of which aim to achieve a specific goal, such as confirming the harmlessness of the response. Differently, CA seeks to analyze out the common, representative and constructive information in conversations for elevating the system's effectiveness in achieving its goals which are generally implicit and grand.

4 Benchmark and Evaluation

This section assembles the comprehensive datasets pertinent to *scene reconstruction*, *causality analysis*, *skill enhancement*, *conversation generation*, and *goal-directed optimization*. Despite the availability of numerous relevant datasets, as shown in Appendix B, they only contain ample conversational content and lack detailed conversational scene elements, such as participants (user profile), scenarios, strategies, and goals, which hinders the comprehensive and insightful analyses.

Scene Reconstruction. Scene elements include participant such as user profile, scenario such as intent, emotion, and strategy. Existing conversation datasets contain some of these elements, but do not

include all of them simultaneously. These datasets either contain information about participants (Chen et al., 2023; Zhang et al., 2018; DeLucia et al., 2024), scenario elements (Chen and Yang, 2021; Sakurai and Miyao, 2024) such as intent and summary, or possess some strategy annotations (Chen et al., 2021; Wu et al., 2022). Due to the relatively objective nature of these elements, the evaluation metrics are mostly based on hard indicators, such as F1-score, Accuracy, BLEU, and ROUGE-L. Recently, researchers have begun to pay attention to LLM evaluators (Wang et al., 2024d; Zhang et al., 2023b; Zeng et al., 2024; Xia et al., 2024).

Causality Analysis. Currently, causality analysis related conversation datasets are not sufficient. The task definition is actually complex as the causes are more high-level, subjective, and open-ended. To make the task purer and more conducive to manual annotation, existing work tends to simplify the task format. For example, researchers define the emotion cause task, aiming to identify the utterances which stimulate the emotion of the target utterance in conversations (Zhao et al., 2023; Zhou et al., 2024c). Therefore, most of the evaluation criterion are discriminative metrics such as Accuracy and F1-score.

Skill Enhancement. For AI agents' skill enhancement, existing datasets focus on the improvement of fundamental abilities, such as conversational consistency (Urbanek et al., 2019; Wang et al., 2023b), long-term memory (Maharana et al., 2024; Xu et al., 2022a,b; Bae et al., 2022; Jang et al., 2023), and emotion support (Liu et al., 2021a; Zhang et al., 2024b). Currently, the benchmarks for automatic prompt optimization and LLM sparring in conversations are still scare. The evaluation metrics are some common criteria such as F1-score, Accuracy, BLEU, and ROUGE-L, and gradually introducing LLM evaluators.

Conversation Generation. In order to evaluate and enhance the LLMs' ability to follow the role settings, some research efforts have been made in dataset construction. These datasets (Zhou et al., 2023a; Chen et al., 2023; Jandaghi et al., 2024; Lu et al., 2024) are built by the following ways: 1) recruiting crowd-sourcing workers for playing the setting, such as role or character, 2) prompting LLMs such as GPT-4 to generate synthetic data, 3) extracting conversations for specific characters from literary sources like novels. Since the quality of conversations, especially whether they adhere to the relevant settings, is highly subjective, most evaluations are conducted by pre-defining several evaluation dimensions or aspects (such as consistency, engagement) and then using powerful LLMs such as GPT-4 or human to evaluate. In addition, for steering conversation generation, most datasets are sourced from existing proactive conversation datasets (Deng et al., 2024; Chang and Chen, 2024). Besides the quality of conversations, goal completion is also considered.

Goal-Directed Optimization. In open-domain conversation, task-oriented conversation, and conversational recommendation, a substantial amount of relevant datasets have been accumulated with the task development. These datasets have either been derived from existing datasets (Tang et al., 2019; Yang et al., 2022) or specifically constructed through crowdsourcing (Moon et al., 2019; Budzianowski et al., 2018; Liu et al., 2022) such as human-to-human written conversations to meet the requirements of the task, where the goals are shallow and are a series of keywords, and the goals are considered met if these keywords are mentioned in the conversation. In addition, the evaluation metrics primarily consist of hard indicators, such as Accuracy, F1-score, Success Rate, BLEU, ROUGE-L, hits@K, PPL, and soft ones such as human and LLM evaluations. These metrics are tailor-designed and exploited for evaluating the achievement of the goals. For instance, success rate refers to the fraction of conversations that successfully accomplish the user's task (Luo et al., 2024).

5 Discussions

The above sections have provided an overview of the current landscape in the research field related to CA. Following this, the discussion will summarize the evolving trends in current research. Additionally, this section will outline potential avenues for future research, aiming to address gaps and explore new horizons in the field.

5.1 Research Trends

Current work is not systematic and the corresponding techniques are scattered and fragmented, but there is an overarching trend moving towards the following advancements:

Task Formulation. There is a shift from *requiring human intervention to decompose tasks into atomic ones with strict input-output formats* towards ***more natural language interactions with loose ones***. This implies that tasks are becom-

ing more flexible and necessitate the enhanced instruction-following ability of models. For example, a lot of work (Louvan and Magnini, 2020) in the past required a clear distinction between discriminative and generative tasks, with rigidly defined input and output formats, making it difficult to handle open-ended problems and resulting in poor scalability. However, LLMs adopt a unified instruction-following and generative paradigm, which enhances the flexibility and openness of task modeling (Wang et al., 2022c; Zhang et al., 2023a).

Task Complexity. We observe an evolution from *surface-level character reassembly* to ***deeper and implicit semantic restructuring***, which requires models to better understand the subtleties and implied meanings within conversations. In earlier studies, researchers focus on ASR error correction (Mani et al., 2020) or coreference resolution (Zhang et al., 2021b; Kim et al., 2021) in conversations, which requires a relatively superficial understanding to achieve promising performance. However, real-world conversations often contain many expressions that carry implications beyond the literal words, such as irony, which presents a significant challenge to comprehending the conversations and receives more attention in recent years (Li et al., 2023a; Ruis et al., 2023; Yue et al., 2024; Qian et al., 2024).

Task Modeling. There is a transition from *third-person point of view* to ***first-person interactive simulation modeling***. This necessitates greater generalization, transferability, and deductive reasoning capabilities within the models. In previous studies, the model directly takes the conversations as input to obtain the representations or generate chain-of-thought style rationales for specific tasks (Lin et al., 2022; Feng et al., 2023; Zhang et al., 2024b) under the third-person perspective. With the development of LLMs, some research efforts try to situate the model in a specific task context for task solution under the first-person perception paradigm, such as by simulating the process of data generation and outputting the relevant analysis results by the way (Liu et al., 2023a; Wang et al., 2024f; Niu et al., 2024), which have a stronger sense of immersion.

5.2 Future Directions

This subsection will explore emerging frontiers in CA research, with the aim of inspiring future investigations in this field.

LLM Conversation Simulator. With the development of LLMs which gradually gain powerful instruction-following abilities, researchers start to explore simulation modeling of LLMs. The simulation is mainly divided into two parts: 1) conversational content simulator where the model serves as a conversational AI agent (Li et al., 2022b), and 2) conversational scene simulator where the model acts as a conversational analyzer (Wang et al., 2024f). The conversational content simulator generates conversation data according to relevant settings, which have attracted a huge amount of interest in the era of LLMs, such as role-playing, conversation data augmentation. The conversational scene simulator mines useful conversational elements such as user profile, emotion, strategy, summarization by simulating conversations instead of providing chain-of-thought style rationales. The simulation process integrates the model into specific conversational contexts, thereby enhancing its situational awareness and in-depth analyses. Therefore, how to construct the conversational scene simulator for conversation analysis is worth exploring.

Fine-Grained Conversation Benchmark. Although tasks involving conversations have been extensively studied, the accumulated datasets are only applicable to one or a few specific atomic tasks (DeLucia et al., 2024) and pay more attention to conversation generation (Zhou et al., 2024a). There is a lack of conversation analysis datasets that contain comprehensive and fine-grained conversational scene elements (participant profile, topic, scenario, strategy, etc.), which hinders the evaluation and modeling of conversation analysis, such as the ability to understand conversational scenes and perform the causality analysis. Therefore, the research communities urgently need high-quality and comprehensive conversation analysis benchmark.

Long-Context Conversation Modeling. Conversation analysis takes the conversational content as input and also faces the long-context challenge, such as the loss of key information, confusion between the parties (attributing one participant's viewpoint to another) in conversations. Although the long-context modeling has been extensively studied in documents (Wang et al., 2024c; He et al., 2024a; Chen et al., 2024b), LLMs encounter the significantly different challenges in long-context conversational content from the documents, such as contextual response dependency and inconsistency as the participants may change their minds. Therefore, it is necessary to conduct specialized research on long-context modeling in conversations.

In-Depth Conversation Reasoning & Attribu-

tion. Most research related to conversation analysis tends to uncover the relatively shallow elements, such as emotion (Lin et al., 2024), intent (Song et al., 2022), and summarization (Tian et al., 2024), which leads to the significant gap between the analysis results and business application. The business personnel are not clear on how to utilize these superficial results for empowering the business innovation. What they really need are the underlying reasons behind these shallow results, along with some constructive suggestions. For example, in customer service conversations, while models can detect the user's angry emotion, the causes of this frustration and how to prevent these occurrences in subsequent services remain under-explored. However, it is extremely challenging to perform these in-depth attribution explorations in the era of small language models due to their limited modeling ability, while the breakthrough in the large language models brings a ray of hope for in-depth reasoning thanks to their powerful instruction-following ability and abundant world knowledge. Building an ecosystem for evaluating and modeling in-depth attribution is vital, such as introducing real-world tasks with corresponding benchmarks.

Goal-Directed Conversation Optimization & Evaluation. Although goal-directed conversation modeling has been extensively explored in recent years, most of these goals are generally simple and restricted to a certain scenario, such as mentioning specific keywords (Tang et al., 2019) or persuading users to accept recommendations (Wang et al., 2022a; Li et al., 2024a) within a conversation. In fact, real-world business goals are often very complex and abstract within multiple conversations, such as improving the customer's experience and increasing user utilization, which raises higher requirements for the inductive and deductive abilities of models. How to bridge the gap between these goals and the analysis results from models is worth further exploring. In addition, these goals are usually abstract and high-level, so traditional metrics such as F1-score, BLEU cannot accurately measure the achievement of goals, and LLM evaluators also face significant challenges due to the abstractness. Therefore, the ability to evaluate goal achievement for LLM evaluators is of great importance and requires additional investigation.

Cross-Session Conversation KV Cache. The reuse of key-value cache has always been one of the key technologies for reducing inference costs (Liu et al., 2023c; Chen et al., 2024c). From inter-sentence eviction to the recent cross-layer reuse, it can effectively reduce the cost of a single inference. However, the number of users using LLM-based conversation systems is gradually increasing, and the conversation content between different users and LLMs is not completely isolated; there are many similar topics. Efficiently reusing cross-session key-value cache and effectively storing the conversation history cache of different types of users will be key to improving the efficiency of conversation systems and reducing costs.

Conversation Security. LLMs has made the conversation a predominant medium of interaction, both among people and between humans and machines. The security of conversations should receive increasing attention. As shown in Figure 2, CA is the analysis process in continuous and iterative manners, centered around specific goals such as improving user satisfaction or discovering violations. As CA systems continuously analyze out critical information and give insights for reducing violations, the evolution of non-compliant behavior to resist regulation will be ongoing. Therefore, the constant evolution of the CA system in such an adversarial context is worth exploring, such as in the self-play manner (Zhang et al., 2024a). In addition, the extreme imbalance between non-compliant and normal samples in conversations poses serious challenges for modeling. Fortunately, LLMs make it possible to simulate the non-compliant data, enhancing the modeling capability for non-compliant samples. The exploration of how to jailbreak LLMs (Wei et al., 2023; Zhou et al., 2024b) to generate these positive examples that are beneficial for modeling has a wide range of needs.

6 Conclusion

This paper presents the first endeavor to outline conversation analysis (CA) and perform a relevant literature review from a technical viewpoint. CA is framed by four principal steps centered around the goal achievement: (1) scene reconstruction, (2) causality analysis, (3) skill enhancement, and (4) conversation generation, which reveals the enormous application value of CA in the era of rapid development of language UI under the backdrop of LLMs. We have also gathered benchmarks related to CA and provided some discussions on development trends and future directions, hoping to kick-start and accelerate the rapid development and business application of CA.

Limitations

We provide a comprehensive review on recent techniques related to CA, and make a first attempt to define the CA task from the viewpoint of technique, including four critical procedures. We have also reorganized the existing data resources and pointed out the potential value of CA in the context of business goal achievement. Nevertheless, we only make the high-level descriptions on existing work related to scene reconstruction, causality analysis, skill enhancement, conversation generation, and goal-directed optimization. In the next version, we will give detailed analyses in each procedure and build an arena to conduct in-depth comparisons of various techniques.

References

Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. 2024. Leveraging summarization for unsupervised dialogue topic segmentation. In *Findings of NAACL*, pages 4697–4704.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of EMNLP*, pages 3769–3787.

Jiaqi Bai, Ze Yang, Xinnian Liang, Wei Wang, and Zhoujun Li. 2021. Learning to copy coherent knowledge for response generation. In *Proceedings of AAAI*, volume 35, pages 12535–12543.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*, pages 5016–5026.

Pablo C Canizares, Jose María López-Morales, Sara Pérez-Soler, Esther Guerra, and Juan de Lara. 2024. Measuring and clustering heterogeneous chatbot designs. *ACM Transactions on Software Engineering and Methodology*, 33(4):1–43.

Wen Chang and Yun-Nung Chen. 2024. Injecting salesperson's dialogue strategies in large language models with chain-of-thought reasoning. In *Findings of ACL*, pages 3798–3812.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of NAACL*, pages 3002–3017.

Hang Chen, Bingyu Liao, Jing Luo, Wenjing Zhu, and Xinyu Yang. 2024a. Learning a structural causal model for intuition reasoning in conversation. *IEEE Trans. on Knowl. and Data Eng.*, 36(7):3210–3223.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of NAACL: Human Language Technologies*, pages 1380–1391.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024b. Long context is not long at all: A prospector of long-dependency data for large language models. In *Proceedings of ACL*, pages 8222–8234. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of EACL*, pages 844–868.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. In *Proceedings of NeurIPS Workshop*.

Yilong Chen, Guoxia Wang, Junyuan Shang, Shiyao Cui, Zhenyu Zhang, Tingwen Liu, Shuohuan Wang, Yu Sun, Dianhai Yu, and Hua Wu. 2024c. NACL: A general and effective KV cache eviction framework for LLM at inference time. In *Proceedings of ACL*, pages 7913–7926. Association for Computational Linguistics.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of ACL*, pages 3201–3219.

Shang-Hsuan Chiang, Lin-Wei Chao, Kuang-Da Wang, Chih-Chuan Wang, and Wen-Chih Peng. 2024. Badge: Badminton report generation and evaluation with llm. *arXiv preprint arXiv:2406.18116*.

Ananlada Chotimongkol. 2008. *Learning the structure of task-oriented conversations from the corpus of in-domain dialogs*. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Alexandra DeLucia, Mengjie Zhao, Yoshinori Maeda, Makoto Yoda, Keiichi Yamada, and Hiromi Wakaki. 2024. Using natural language inference to improve persona extraction from dialogue in a new domain. *arXiv preprint arXiv:2401.06742*.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of EMNLP*, pages 10602–10621.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *Proceedings of ICLR*.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023b. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems*, 41(3):1–25.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat-Seng Chua, and Qing Li. 2024. Fashionregen: Llm-empowered fashion report generation. In *Companion Proceedings of the ACM Web Conference 2024*, page 991–994.

Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of SIGKDD*. Association for Computing Machinery.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428. European Language Resources Association.

Aciel Eshky. 2014. *Generative probabilistic models of goal-directed users in task-oriented dialogs*. Ph.D. thesis, The University of Edinburgh.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of ACL*, pages 6866–6880.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Yue Feng, Yunlong Jiao, Animesh Prasad, Nikolaos Aletras, Emine Yilmaz, and Gabriella Kazai. 2023. Schema-guided user satisfaction modeling for task-oriented dialogues. In *Proceedings of ACL*, pages 2079–2091.

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.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware contrastive learning. In *Proceedings of SIGIR*, page 2481–2485.

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.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of EMNLP*, pages 8142–8152.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024a. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of ACL*, pages 13628–13642. Association for Computational Linguistics.

Ming He, Jiwen Wang, Tianyu Ding, and Tong Shen. 2023. Conversation and recommendation: knowledge-enhanced personalized dialog system. *Knowledge and Information Systems*, 65(1):261–279.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024b. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of ACL*, pages 4768–4791. Association for Computational Linguistics.

Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via RL on imagined conversations. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversation recommender systems with system-and user-centric factors. *arXiv preprint arXiv:2404.03304*.

Prakash Jamakatel, Rebecca De Venezia, Christian Muise, and Jane Jean Kiam. 2024. A goal-directed dialogue system for assistance in safety-critical application. In *Proceedings of IJCAI*, pages 7859–7867. International Joint Conferences on Artificial Intelligence Organization.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Findings of ACL*, pages 15245–15270.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of EMNLP*, pages 13584–13606.

Sheng Jiang, Su Zhu, Ruisheng Cao, Qingliang Miao, and Kai Yu. 2023. SPM: A split-parsing method for joint multi-intent detection and slot filling. In *Proceedings of ACL: Industry Track*, pages 668–675. Association for Computational Linguistics.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of ACL*, pages 6130–6141.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of EMNLP*, pages 12930–12949.

Seo Hyun Kim, Kai Tzu iunn Ong, Taeyoon Kwon, Namyoung Kim, Keummin Ka, SeongHyeon Bae, Yohan Jo, Seung won Hwang, Dongha Lee, and Jinyoung Yeo. 2024. Theanine: Revisiting memory management in long-term conversations with timeline-augmented response generation. *arXiv preprint arXiv:2406.10996*.

Issam H Laradji, Stefania Raimondo, David Vazquez, Pau Rodriguez, Christopher Pal, et al. 2023. Workflow discovery from dialogues in the low data regime. *Transactions on Machine Learning Research*.

Kun Lei, Peng Guo, Wenchao Zhao, Yi Wang, Linmao Qian, Xiangyin Meng, and Liansheng Tang. 2022a. A multi-action deep reinforcement learning framework for flexible job-shop scheduling problem. *Expert Systems with Applications*, 205:117796.

Wenqiang Lei, Yao Zhang, Feifan Song, Hongru Liang, Jiaxin Mao, Jiancheng Lv, Zhenglu Yang, and Tat-Seng Chua. 2022b. Interacting with non-cooperative user: A new paradigm for proactive dialogue policy. In *Proceedings of SIGIR*, page 212–222.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pages 3045–3059.

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2024a. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *arXiv preprint arXiv:2405.01868*.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024b. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023a. Diplomat: A dialogue dataset for situated pragmatic reasoning. In *Proceedings of NeurIPS*, volume 36, pages 46856–46884. Curran Associates, Inc.

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. 2024c. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint arXiv:2406.11978*.

Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. 2021. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–29.

Wenhao Li, Xiangfeng Wang, Bo Jin, Dijun Luo, and Hongyuan Zha. 2022a. Structured cooperative reinforcement learning with time-varying composite action space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8618–8634.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024d. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022b. Controllable dialogue simulation with in-context learning. In *Findings of EMNLP*, pages 4330–4347. Association for Computational Linguistics.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023b. Guiding large language models via directional stimulus prompting. In *NeurIPS*, volume 36, pages 62630–62656.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of EMNLP*, pages 4436–4451. Association for Computational Linguistics.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of ACL*, pages 2545–2558.

Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. In *Proceedings of ACL*, pages 11100–11115.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proceedings of ACL*, pages 3469–3483. Association for Computational Linguistics.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023a. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of ACL*, pages 1–21.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021b. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of EMNLP*, pages 4335–4347.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of ACL*, pages 1036–1049.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of ACL*, pages 1024–1034.

Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy, and Nancy Chen. 2023b. Joint dialogue topic segmentation and categorization: A case study on clinical spoken conversations. In *Proceedings of EMNLP: Industry Track*, pages 185–193.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023c. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. In *Proceedings of NeurIPS*, volume 36, pages 52342–52364. Curran Associates, Inc.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of COLING*, pages 480–496.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proceedings of NAACL: Human Language Technologies*, pages 5200–5212.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of ACL*, pages 7828–7840.

Keting Lu, Shiqi Zhang, and Xiaoping Chen. 2019. Goal-oriented dialogue policy learning from failures. In *Proceedings of AAAI*, volume 33, pages 2596–2603.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, volume 8 of *L@S '24*, page 16–27.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. DuetSim: Building user simulator with dual large language models for task-oriented dialogues. In *Proceedings of LREC-COLING*, pages 5414–5424.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of ACL*, pages 13851–13870.

Anirudh Mani, Shruti Palaskar, and Sandeep Konam. 2020. Towards understanding asr error correction for medical conversations. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 7–11.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of ACL*, pages 845–854.

Jinjie Ni, Vlad Pandelea, Tom Young, Haicang Zhou, and Erik Cambria. 2022. Hitkg: Towards goal-oriented conversations via multi-hierarchy learning. In *Proceedings of AAAI*, page 11112–11120.

Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. 2024. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of ACL*, pages 8724–8741. Association for Computational Linguistics.

Jiao Ou, Jiayu Wu, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Inductive-deductive strategy reuse for multi-turn instructional dialogues. *Preprint*, arXiv:2404.11095.

Siru Ouyang, Jiaao Chen, Jiawei Han, and Diyi Yang. 2023. Compositional data augmentation for abstractive conversation summarization. In *Proceedings of ACL*, pages 1471–1488. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of ACL*, pages 2182–2192.

Olivier Pietquin. 2006. Consistent goal-directed user model for realisitc man-machine task-oriented spoken dialogue simulation. In *2006 IEEE International Conference on Multimedia and Expo*, pages 425–428. IEEE.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing

emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of EMNLP*, pages 7957–7968.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! towards implicit user intention understanding of language model driven agents. In *Proceedings of ACL*, pages 1088–1113. Association for Computational Linguistics.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of EMNLP*, pages 1807–1816. Association for Computational Linguistics.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Proceedings of ACL*, pages 12549–12561. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of AAAI*, volume 34, pages 8689–8696.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In *Proceedings of NeurIPS*, volume 36, pages 20827–20905. Curran Associates, Inc.

Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of ACL*, pages 1635–1657.

Abraham Sanders, Tomek Strzalkowski, Mei Si, Albert Chang, Deepanshu Dey, Jonas Braasch, and Dakuo Wang. 2022. Towards a progression-aware autonomous dialogue agent. In *Proceedings of NAACL: Human Language Technologies*, pages 1194–1212.

John R Searle. 1969. Speech acts: An essay in the philosophy of language. *Cambridge UP*.

Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser. 2021. OTTers: One-turn topic transitions for open-domain dialogue. In *Proceedings of ACL*, pages 2492–2504.

Ryan Shea and Zhou Yu. 2023. Building persona consistent dialogue agents with offline reinforcement learning. In *Proceedings of EMNLP*, pages 1778–1795.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of EMNLP*, pages 4222–4235.

Priyanka Sinha, Pabitra Mitra, Antonio Anastasio Bruto da Costa, and Nikolaos Kekatos. 2021. Explaining outcomes of multi-party dialogues using causal learning. *arXiv preprint arXiv:2105.00944*.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proceedings of EMNLP*, pages 7967–7977. Association for Computational Linguistics.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of COLING*, pages 717–729.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of NAACL*, pages 1570–1583.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of ACL*, pages 5624–5634.

Yuanhe Tian, Fei Xia, and Yan Song. 2024. Dialogue summarization with mixture of experts based on large language models. In *Proceedings of ACL*, pages 7143–7155.

Anna Tigunova. 2020. Extracting personal information from conversations. In *Companion Proceedings of the Web Conference 2020*, pages 284–288.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of EMNLP*, pages 673–683.

Bryan Wang, Gang Li, and Yang Li. 2023a. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023b. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of EMNLP*, pages 1132–1143. Association for Computational Linguistics.

Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. 2024a. Instruct

once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *Proceedings of ACL*, pages 3993–4010.

Jian Wang, Dongding Lin, and Wenjie Li. 2022a. Follow me: Conversation planning for target-driven recommendation dialogue systems. *arXiv preprint arXiv:2208.03516*.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multitask learning. In *Proceedings of AAAI*, volume 34, pages 9177–9184.

Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024b. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. In *Findings of ACL*, pages 567–585.

Keqin Wang, Alison Bartsch, and Amir Barati Farimani. 2022b. Man: Multi-action networks learning. *arXiv preprint arXiv:2209.09329*.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024c. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Preprint*, arXiv:2406.17419.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024d. Large language models are not fair evaluators. In *Proceedings of ACL*, pages 9440–9450. Association for Computational Linguistics.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024e. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of ACL*, pages 12912–12940.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023c. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of EMNLP*, pages 10052–10065.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024f. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In *Proceedings of ACL*, pages 15880–15893. Association for Computational Linguistics.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023d. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Zhilin Wang, Xuhui Zhou, Rik Koncel-Kedziorski, Alex Marin, and Fei Xia. 2022c. Extracting and inferring personal attributes from dialogue. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 58–69. Association for Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? In *Proceedings of NeurIPS*, volume 36, pages 80079–80110. Curran Associates, Inc.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024a. Autogen: Enabling next-gen LLM applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of ACL*, pages 3794–3804.

Yiquan Wu, Bo Tang, Chenyang Xi, Yu Yu, Pengyu Wang, Yifei Liu, Kun Kuang, Haiying Deng, Zhiyu Li, Feiyu Xiong, et al. 2024b. Xinyu: An efficient llm-based system for commentary generation. In *Proceedings of SIGKDD*, pages 6003–6014.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP*, pages 6177–6181. IEEE.

Tingyu Xia, Bowen Yu, Yuan Wu, Yi Chang, and Chang Zhou. 2024. Language models can evaluate themselves via probability discrepancy. In *Findings of ACL*, pages 4889–4901. Association for Computational Linguistics.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of ACL*, pages 1707–1725.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of ACL*, pages 13207–13219. Association for Computational Linguistics.

Haotian Xu, Haiyun Peng, Haoran Xie, Erik Cambria, Liuyang Zhou, and Weiguo Zheng. 2020. End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web*, 23(3):1989–2002.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of ACL*, pages 5180–5197. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of ACL*, pages 2639–2650.

Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. TopKG: Target-oriented dialog via global planning on knowledge graph. In *Proceedings of COLING*, pages 745–755.

Ilker Yildirim and LA Paul. 2024. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature–a case study with a chinese sitcom. *arXiv preprint arXiv:2404.19509*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *Proceedings of ICLR*.

Jun Zhang, Yan Yang, Chencai Chen, Liang He, and Zhou Yu. 2021a. KERS: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals. In *Findings of EMNLP*, pages 1092–1101.

Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, Wenliang Chen, and Dong Yu. 2023a. Safeconv: explaining and correcting conversational unsafe behavior. In *Proceedings of ACL*, pages 22–35.

Ruize Zhang, Zelai Xu, Chengdong Ma, Chao Yu, Wei-Wei Tu, Shiyu Huang, Deheng Ye, Wenbo Ding, Yaodong Yang, and Yu Wang. 2024a. A survey on self-play methods in reinforcement learning. *arXiv preprint arXiv:2408.01072*.

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 ACL*, pages 2204–2213. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of ACL*, pages 13395–13412.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024c. Strength lies in differences! towards effective non-collaborative dialogues via tailored strategy planning. *arXiv preprint arXiv:2403.06769*.

Weinan Zhang, Yue Zhang, Hanlin Tang, Zhengyu Zhao, Caihai Zhu, and Ting Liu. 2021b. What did you refer to? evaluating co-references in dialogue. In *Findings of ACL*, pages 5075–5084.

Xiaotong Zhang, Xuefang Jia, Han Liu, Xinyue Liu, and Xianchao Zhang. 2024d. A goal interaction graph planning framework for conversational recommendation. In *Proceedings of AAAI*, volume 38, pages 19578–19587.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023b. Wider and deeper llm networks are fairer llm evaluators. *Preprint*, arXiv:2308.01862.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023c. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of EMNLP*, pages 8289–8311.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of AAAI*, volume 38, pages 19632–19642.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. *Proceedings of AAAI*, 37(11):14020–14028.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of ACL*, pages 1552–1568.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of ACL*, pages 11325–11345.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. *Proceedings of AAAI*, 38(17):19724–19731.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023a. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023b. SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In *Proceedings of ACL*, pages 9945–9959.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of COLING*, pages 4128–4139.

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021. Probing commonsense explanation in dialogue response generation. In *Findings of EMNLP*, pages 4132–4146.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2023c. Learning to predict persona information for dialogue personalization without explicit persona description. In *Findings of ACL*, pages 2979–2991.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2024a. Sotopia: Interactive evaluation for social intelligence in language agents. In *Proceedings of ICLR*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024b. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. *Preprint*, arXiv:2406.05644.

Zhihan Zhou, Xue Gu, Yujie Zhao, and Hao Xu. 2024c. POP-CEE: Position-oriented prompt-tuning model for causal emotion entailment. In *Findings of ACL*, pages 4199–4210. Association for Computational Linguistics.

Hui Zhu, Xv Wang, Zhenyu Wang, and Kai Xv. 2024. An emotion-sensitive dialogue policy for task-oriented dialogue system. *Scientific Reports*, 14(1):19759.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In *Proceedings of AAAI*, volume 35, pages 14665–14673.

A Taxonomy of Conversation Analysis

This section mainly describes the research related to CA according to the taxonomy of analysis procedure in detail, including scene reconstruction (§A.1), causality analysis (§A.2), skill enhancement (§A.3), and conversation generation (§A.4) for goal-directed optimization (§A.5).

A.1 Scene Reconstruction

A conversation contains conversational content, participants, scenarios (such as location, topic, atmosphere, emotion, intent) and so on. The elements other than the conversational content are collectively referred to as the scene. Scene reconstruction aims to mine these scene elements which are the background information of a conversation.

Participant. Tigunova (2020) investigated what salient attributes can be extracted from conversational texts and which tools may be applicable for that. Wang et al. (2022c) introduced the tasks of extracting and inferring personal attributes from conversations, such as hobbies, pets, family, likes and dislikes. They also utilized constrained attribute generation based on autoregressive language model with the discriminative reranker. Lu et al. (2022) used automatic partner personas generation to enhance the response generation. Zhou et al. (2023c) proposed a method which learnt to predict persona information based on the conversation history. DeLucia et al. (2024) post-hoc adapted a trained persona extraction model to a new domain by a natural language inference method.

Scenario. Sakurai and Miyao (2024) modified the existing datasets of persuasive conversation and created datasets using a multiple-choice paradigm to evaluate LLMs' intention detection capability in conversation. Lin et al. (2024) introduced SPUR to estimate user satisfaction, which developed an iterative prompting framework that uses an LLM to firstly extract signals of satisfaction in a labeled training set, then summarize the reasons into rubrics, and lastly apply the rubrics to predict satisfaction labels. Tian et al. (2024) proposed an LLM-based approach with role-oriented routing and fusion generation to utilize mixture of experts for conversation summarization. Due to the variety in segmentation granularity, as well as the lack of diverse annotated corpora, there is no universal models can be easily applied to domain-specific applications. Liu et al. (2023b) firstly introduced a joint model for conversation segmentation and topic classification, conducting a case study on healthcare follow-up calls for diabetes management. DialSTART (Gao et al., 2023) learned topic-aware utterance representations from unlabeled conversation data through adjacent utterance matching and pseudo-segmentation. These topic-aware utterance representations are then combined with the conversation coherence to perform unsupervised segmentation. Artemiev et al. (2024) proposed a novel conversation summarization-based approach, focusing on transcribed spoken conversations, achieving significant improvements in unsupervised topic segmentation. Feder et al. (2022) adopted the clustering method to associate toxic and non-toxic groupings among the formerly suspended users, tackling challenges for causal assumption. Canizares et al. (2024) developed the clustering methods to help in grouping chatbots along their conversation topics and design features.

Strategy. Laradji et al. (2023) introduced a new task called the workflow discovery which aims to discover the set of actions that have been taken to solve a specific requirement. Xie et al. (2024)

proposed DIIR framework to learn and apply conversation strategies in the form of natural language inductive rules from expert demonstrations, which generates and verifies a strategy description given a context and its corresponding gold response. Ou et al. (2024) derived instructional strategies from diverse real instruction conversations, which were then deductively exploited to new conversation scenarios, contributing to high-quality instructions.

A.2 Causality Analysis

The process of conversation interaction is sophisticated with the latent induced factors. Analyzing these factors has important guiding significance for gaining a deeper understanding of the causes and development during conversations. By performing causality modeling, and generating analysis reports, causality analysis aims to delve into the causes which lead to the emergence of scene elements, such as emotional changes and user personality.

Causality Modeling. Zhou et al. (2021) explored why models respond as they do by probing their understanding of commonsense reasoning that elicits proper responses. The authors framed commonsense as a latent variable to formally define the problem which aimed to generate commonsense explanations for responses. Sinha et al. (2021) analyzed why a conversation ends with a particular sentiment from the point of view of conflict analysis and future collaboration design. Zhao et al. (2023) concentrated on detecting causal utterances that lead to the target utterance with a non-neutral emotional tone in conversations, and developed the knowledge-bridged causal interaction network (KBCIN) with commonsense knowledge as bridges. Chen et al. (2024a) constructed a conversation cognitive model based on intuition theory to explain how each utterance engages and activates information channels in a recursive manner. The authors also synthesize datasets which incorporate implicit causes and complete cause labels. Zhou et al. (2024c) developed the position-oriented prompt-tuning method called POP-CEE to solve the causal emotion entailment task in conversations.

Analysis Report Generation. Wang et al. (2023d) explored the potential of LLMs for radiology report generation, and proposed R2GenGPT for generating radiology report. Ding et al. (2024) used LLMs to appropriately select and combine multimodal information, integrating text, charts, and images into a coherent, complete, and insightful report. Based on this report, they conducted higher-level understanding and analysis of specific domains. Wu et al. (2024b) developed a LLM-based system Xinyu to assist commentators in generating Chinese commentaries. Chiang et al. (2024) focused on badminton report generation, including details such as player names, game scores, and ball types for providing audiences with a comprehensive view, and proposed BADGE based on LLMs.

A.3 Skill Enhancement

After conducting attribution analysis, it is necessary to better utilize these insights to provide targeted training for individuals such as call-center employees or AI agents. For human participants, some improvement suggestions or suggested wording can be provided, while prompts can be optimized along with some foundational capabilities such as consistency or emotion support for AI agents. In addition, AI agent can serve as sparring partners to help human enhance skills.

Automatic Prompt Optimization. Directly Optimizing LLMs for specific tasks may be costly, leading researchers to focus on automatic prompt optimization. AutoPrompt (Shin et al., 2020) and Prompt-Tuning (Lester et al., 2021) firstly attempted to optimize prompts to enhance task performance without training LMs. Pryzant et al. (2023) proposed Text Gradient Prompt Optimization (ProTeGi), a general non-parametric algorithm that uses LLM feedback instead of differentiation and LLM editing instead of backpropagation to directly make discrete improvements to prompts. Li et al. (2023b) introduced a new component called directional stimulus into the prompt, guiding black-box LLMs toward specific desired outputs. Cheng et al. (2024) created an automatic prompt optimizer that rewrites disorganized or unclear human prompts into LLM-preferred prompts for better conveying human intent.

Specialized Tuning. 1) *Conversation Consistency Capability*: Wang et al. (2024a) improved conversation consistency by separately modeling the speaker roles of agent and user, enabling the agent to adhere to its role consistently. Zhou et al. (2023b) proposed a two-stage strategy comprising over-sampling and post-evaluation to enhance the consistency of role-based conversation generation. Shea and Yu (2023) applied offline reinforcement learning to a role consistency task, demonstrating its ability to enhance role consistency and conversation quality over a system trained solely through imitation learning.

2) *Emotion Support Capability*: Wang et al. (2024b) introduced a novel multifaceted AI feedback mechanism to alleviate the helplessness support framework, aiming to reduce *"unhelpful messages"* in emotional support conversations. Inspired by the human emotional support generation process of identifying, understanding, and regulating emotions, Zhang et al. (2024b) proposed an interpretable emotional support response generation scheme called ESCoT. Based on this, they constructed the first CoT emotional support conversation dataset, ESDCoT. Zheng et al. (2024) utilized LLMs as "Counseling Teacher" to enhance smaller models' emotion support response abilities, significantly reducing the necessity of scaling up model size.

3) *Long-Term Conversation Capability*: Zhong et al. (2024) enabled LLMs to summon relevant memories, continually evolve through continuous memory updates, comprehend, and adapt to a user's personality over time by synthesizing information from previous interactions. Kim et al. (2024) proposed a novel timeline-augmented framework, THEANINE, which utilizes a series of memories demonstrating the development and causality of relevant past events to improve long-term conversation capability. To meet the real-world demands for long-term companionship and personalized interactions with chatbots, Li et al. (2024b) introduced a model-agnostic agent framework that simultaneously handles event summary and persona management, enabling the inference of appropriate long-term conversation responses.

LLM Sparring. Li et al. (2024d) leveraged the potential of LLMs in clinical medical education, facilitating natural conversations between students and LLM-Simulated Patients, evaluated their conversation, and provided suggestions to enhance students' clinical inquiry skills. Lu and Wang (2024) found that LLM-simulated students produced logical and believable responses that were aligned with their profiles in multiple-choice question tasks. They exhibit strong consistency with real students' response to the same set of questions. Teachers can use the feedback from LLM-simulated students to improve the quality of the examination. Xiao et al. (2024) proposed HealMe, a specialized cognitive reframing therapy model that emulates a complete psychotherapeutic procedure, helping clients to better understand their issues, accept new interpretations, and move toward constructive solutions.

A.4 Conversation Generation

Conversation generation focuses on generating concrete conversational content relying on the refined experiences where the core issue is how to make the model follow these experience settings such as the refined scene elements and inject the analyzed insights into the model, steering agents to generate conversations towards specific goals.

Setting-Following Based. Chen et al. (2022) explored few-shot data augmentation for conversation understanding by prompting LLMs and presented a novel approach that iterates on augmentation quality by applying weakly-supervised filters. Chen et al. (2023) used a small set of expert-written conversations as in-context examples to synthesize a social conversation dataset using prompting. Zheng et al. (2023) constructed AugESC, an augmented dataset for the ESC task. They prompted a fine-tuned language model to complete full conversations from available conversation posts of various topics, which are then postprocessed based on heuristics. Kim et al. (2023) designed CO_3 distillation framework and distilled SODA, the first publicly million-scale high-quality social conversation dataset, from the LLM. Jandaghi et al. (2024) proposed Generator-Critic architecture framework to create the conversational dataset. The Generator is an LLM prompted to output conversations. The Critic consists of a mixture of expert LLMs that control the quality of the generated conversations.

Various works also used the role-playing abilities of LLMs to generate conversation data. These tasks design conversation elements such as character traits and scenarios, allowing LLM to play specific roles and simulate conversation interactions. Zhou et al. (2023a) designed reasonable attributes and behaviors for AI Characters and adopted a few-shot approach by prompting GPT-4 to generate dialog data. Lu et al. (2024) introduced Ditto to generate conversation data by self-alignment way, which simulated role-play conversations as a variant of reading comprehension. Wang et al. (2024e) proposed SOTOPIA-π method to prompt LLMs generating social interaction (mainly conversations) data based on SOTOPIA Zhou et al. (2024a), then leveraging behavior cloning and self-reinforcement to improve social intelligence ability of LLM.

Steering Based. Peng et al. (2018) introduced Deep Dyna-Q, a deep RL framework for task-completion conversation policy learning. They integrated a world model into the conversation agent

to simulate user responses and experiences. This model is continuously updated with real user data to mimic user behavior better, optimizing the conversation agent with both real and simulated experiences. Lu et al. (2019) argued that hindsight experience replay (HER) enables learning from failures, but the vanilla HER is inapplicable to conversation learning due to the implicit goals. They developed two complex HER methods providing different tradeoffs between complexity and performance and enabled HER-based conversation policy learning. Strategy learning and goal planning attach great importance in proactive conversation systems. Sanders et al. (2022) conducted a comprehensive evaluation on how LLM-based conversation systems exhibit proactivity, including aspects such as clarification, goal-oriented guidance, and handling of non-cooperative conversations. Lei et al. (2022b) focused on proactive conversation policy when users exhibit non-cooperative behavior in interactive environments, balancing between reaching the goal topic and ensuring user satisfaction. Deng et al. (2023a) introduced the concept of global conversation state and proposed a framework in which conversation agents can understand the trajectory of the ongoing conversation, assess the likelihood of successful outcomes, and evaluate how their own response decisions impact the overall direction of the conversation.

Hong et al. (2023) explored a new method for adapting LLMs with RL for goal-directed conversation, which collected conversation data through sampling diverse synthetic rollouts of hypothetical in-domain human-human interactions. Zhu et al. (2024) proposed an ESDP, which incorporated user emotions information into conversation policy and selected the optimal action by combining top-k actions with the user emotions. In each turn, the user's emotional information is used as an immediate reward for the current conversation state to solve sparse rewards and the dependency on termination. Deng et al. (2024) introduced PPDPP, a conversation policy planning paradigm that uses a tunable language model plug-in for goal-directed conversation. It offers a training framework for supervised fine-tuning and reinforcement learning from self-play feedback. Chang and Chen (2024) injected conversation strategies and incorporated chain-of-thought reasoning for training LLMs, where it is equipped with the integrated capabilities of intent detection, policy selection, and response generation. Li et al. (2024c) presented DAT that adapts language model agents to plan goal-directed conversations. The core idea was to treat each utterance as an action, freeze a pre-trained language model, and train a small planner model that predicts a continuous action vector used for controlled generation in each round. Zhang et al. (2024c) investigated the limitations of current LLM-based conversation agents in strategic planning, with a key challenge being the agents' awareness of various non-cooperative user behaviors. To address these challenges, they proposed TRIP, which manipulates the experiences of conversation agents, enabling them to recognize the importance of tailoring strategies for individuals.

A.5 Goal-Directed Optimization

In fact, all conversations are arguably goal-directed (Searle, 1969; Austin, 1975), such as helpfulness, harmlessness in open-domain conversation, and completing booking of flight tickets in task-oriented conversation. The goal-directed modeling is gradually gaining attention in open-domain conversation, task-oriented conversation, and conversational recommendation.

Open-Domain Conversation. Tang et al. (2019) studied the conversational goals in open-domain conversation, and expected a conversational system to steer the chat toward a designated target subject by introducing coarse-grained keywords to control the intended content of system responses. Wu et al. (2019) wanted to endow a conversational agent with the ability of proactively leading the conversation, and introduced a new dataset named Konv which consists of a conversation leader and the followers. The leader sequentially changes the discussion topics, following the given conversation goal. Ni et al. (2022) studied how to plan short-term and long-term goal in open-domain conversations, and proposed the hierarchical model to learn goal planning in a hierarchical learning framework. Yang et al. (2022) proposed a global reinforcement learning with planned paths towards the global target.

Recently, goal awareness of open-domain conversation is evoking widespread interest in the era of large language models (LLMs). Hong et al. (2023) argued that LLMs can provide useful data for solving goal-directed conversation tasks by simulating sub-optimal but human-like behaviors, instead of effectively solving such tasks out of the box. Li et al. (2024c) adapted language model agents to plan goal-directed conversations by treat-

ing each utterance as an action, thereby converting conversations into games and leading to easy adaptation of existing method such as reinforcement learning. Zhou et al. (2024a) introduced SOTOPIA to simulate complex social interactions between LLM agents for achieving goals and evaluate their social intelligence. Wang et al. (2024e) further developed SOTOPIA-π, an interactive learning method, to leverage behavior cloning and self-reinforcement training.

Task-Oriented Conversation. Pietquin (2006) developed a user modeling technique for realistic simulation of man-machine goal-directed spoken conversations. Chotimongkol (2008) argued the purpose of a goal-directed conversation is to fill one or more forms and to ensure that the information is consistent. Eshky (2014) modelled the user behaviour in conversations where user goals are unobserved as part of the conversation. Sun et al. (2021) integrated task-oriented and open-domain systems to enhance the functional goals (e.g., booking hotels) in task-oriented conversations. Liu et al. (2022) collected a human-to-human mixed-type task-oriented dialog corpus where an agent provides user-goal-related knowledge to help figure out and achieve specific goals within each session, and then designed a prompt-based continual learning mechanism for model training.

Since the advent of LLMs, Jamakatel et al. (2024) used the goal-directed conversation system to infer the current context in situations requiring emergency landings with LLM data augmentations in safety-critical application. Luo et al. (2024) argued that LLMs were not promising for those responses which effectively guided users towards their goals with intricate constraints and requirements, and employed two LLMs for response generation and verification respectively. He et al. (2024b) proposed the dual-process conversation planning framework and employed two complementary planning systems inspired by dual-process theory in psychology (thinking—intuitive and analytical) to improve the achievement of goals.

Conversational Recommendation. Liu et al. (2020) proposed a conversational recommendation task over multi-type dialogs, with the bots proactively and naturally leading a conversation from a non-recommendation dialog to a recommendation dialog, and developed a human-to-human Chinese dialog dataset DuRecDial. Hayati et al. (2020) created a dataset INSPIRED, which consisted 1,001 human-human dialogs for movie recommendation with measures for successful recommendations. Zhou et al. (2020) built a conversational recommendation dataset which leveraged topic threads to enforce natural semantic transitions towards the recommendation scenario. Bai et al. (2021) developed goal-oriented knowledge copy network, helping discern those knowledge facts which are highly correlated to the dialog goal and context. Zhang et al. (2021a) proposed the knowledge-enhanced multi-subgoal driven recommender system to predict a sequence of subgoals which guided the dialog model to select knowledge. Wang et al. (2022a) aimed to make users accept the goals/targets gradually through conversations, and developed a target-driven conversation planning framework to proactively lead the system to transit between different conversation stages. Deng et al. (2023b) proposed a unified multi-goal conversational recommender system, to unify *goal planning*, *topic prediction*, *item recommendation*, and *response generation* with different formulations into the same sequence-to-sequence paradigm. Zhang et al. (2024d) introduced a goal interaction graph planning framework to model the correlations and order of goals.

Wang et al. (2023c) reviewed the utilization of LLMs for conversational recommendation and showcased the inadequacy of the existing evaluation protocol, and further proposed an interactive evaluation approach via LLM-based user simulators. Li et al. (2024a) aimed to empower LLMs for using external knowledge and goal guidance in conversational recommendation, and decomposed the complex conversational recommendation task into several subtasks through the knowledge retrieval and goal planning. Huang et al. (2024) proposed a evaluation protocol on conversational recommendation based on LLMs simulation, which took both system- and user-centric factors into account by conceptualising three key characteristics and further dividing them into six primary abilities. The authors argued that they regarded conversational recommendation as a social issue to consider its major goal of persuading users to accept the recommendation, instead of just a technical problem.

Takeaway: An increasing number of research is focusing on goal-awareness ability in conversations, which only models the conversational data to optimize towards the goal based on reinforcement learning, or builds the goal related benchmark and evaluation protocol. However, it maybe not promising to lead the model towards a goal solely based on the conversational content without any

Section	Dataset	Parti.	Sce.	Stra.	Go.	Usage	Source	Task	Lang.	Metric
§A.1	DialogNLI (Chen et al., 2023)	✓	✗	✗	✗	Train&Val &Test	Existing Resource	Personal Attribute	EN	
	PersonaChat (Zhang et al., 2018)	✓	✗	✗	✗	Train&Val &Test	Crowdsourcing	Personal Attribute	EN	
	PersonaExt-PeaCoK (DeLucia et al., 2024)	✓	✗	✗	✗	Train&Val &Test	Existing Resource	Personal Attribute	EN	
	FaceAct* (Sakurai and Miyao, 2024)	✗	✓	✗	✗	Test	Existing Resource	Intent	EN	
	Bing Copilot	✗	✓	✗	✗	Train&Val &Test	Existing Resource	General	Multi	
	DialogSum (Chen and Yang, 2021)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Summarization	EN	
	SAMSum (Gliwa et al., 2019)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Summarization	EN	
	CSDS (Lin et al., 2021)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Summarization	ZH	
	MC (Song et al., 2020)	✗	✓	✗	✗	Train&Test	Crowdsourcing	Summarization	ZH	
	Clinical-Conversation (Liu et al., 2023b)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Topic Segmentation	EN	
	SmallDataSet (Canizares et al., 2024)	✗	✓	✗	✗	&Test	Multi-platform	Topic Clustering	EN	
	AnnoMI (Wu et al., 2022)	✗	✗	✓	✗	Test	Crowdsourcing	Strategy	EN	
	ABCD (Chen et al., 2021)	✗	✗	✓	✗	Train&Val &Test	Crowdsourcing	Strategy	EN	
§A.2	RECCON-DD (Poria et al., 2021)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Emotion Cause	EN	
	CEDAR (Zhou et al., 2021)	✗	✗	✗	✗	Train&Val &Test	Crowdsourcing	Commonsense Reasoning	EN	
	Synthetic-Simulated (Chen et al., 2024a)	✗	✓	✗	✗	Train&Val &Test	Rule&LLMs	Conversation Reasoning	EN	
§A.3	LIGHT (Urbanek et al., 2019)	✓	✓	✗	✗	Train&Val &Test	Crowdsourcing	Conversation Consistency	EN	
	TOPDIAL (Wang et al., 2023b)	✓	✓	✗	✗	Train&Val &Test	LLMs	Conversation Consistency	EN	
	ESConv (Liu et al., 2021a)	✗	✓	✗	✗	Train&Val &Test	Crowdsourcing	Emotion Support	EN	
	ESD (Zhang et al., 2024b)	✗	✓	✓	✗	Train&Val &Test	LLMs&Crowdsourcing	Emotion Support	EN	
	LoCoMo (Maharana et al., 2024)	✓	✗	✗	✗	Test	Crowdsourcing	Long-Term	EN	
	MSC (Xu et al., 2022a)	✓	✓	✗	✗	Train&Val &Test	Crowdsourcing	Long-Term	EN	
	DuLeMon (Xu et al., 2022b)	✓	✗	✗	✗	Train&Test	Crowdsourcing	Long-Term	ZH	
	CareCall (Bae et al., 2022)	✗	✓	✗	✗	Train&Val &Test	Existing Resource& Crowdsourcing	Long-Term	Korean	
	CC (Jang et al., 2023)	✗	✓	✗	✗	Train&Val &Test	LLMs& Crowdsourcing	Long-Term	EN	
§A.4	WEAKDAP (Chen et al., 2022)	✗	✗	✗	✗	Train	Existing Resource & LLMs	Data Augmentation	EN	
	PLACES (Chen et al., 2023)	✓	✓	✗	✗	Train	LLMs	Data Augmentation	EN	
	AugESC (Zheng et al., 2023)	✗	✗	✗	✗	Train	Existing Resource & LLMs	Data Augmentation	EN	
	SODA (Kim et al., 2023)	✗	✗	✗	✗	Train&Val &Test	Existing Resource & LLMs	Data Augmentation	EN	
	Synthetic-Persona-Chat (Jandaghi et al., 2024)	✓	✗	✗	✗	Train	LLMs	Data Augmentation	EN	
	CharacterGLM (Zhou et al., 2023a)	✓	✗	✗	✗	Train	Existing Resource & Crowdsourcing & LLMs	Role-Playing	ZH	
	Ditto (Lu et al., 2024)	✓	✗	✗	✗	Train&Test	LLMs	Role-Playing	ZH,EN	
	Sotopia-π (Wang et al., 2024e)	✓	✓	✗	✓	Train	Existing Resource & LLMs	Role-Playing	EN	
	SalesBot 2.0 (Chang and Chen, 2024)	✗	✗	✓	✗	Train	Existing Resource & LLMs	Strategy Injection	EN	
§A.5	PersonaChat* (Tang et al., 2019)	✗	✗	✗	✓	Train&Val &Test	Existing Resource	Open-Domain	EN	
	DuConv (Wu et al., 2019)	✗	✗	✗	✓	Train&Val &Test	Crowdsourcing	Open-Domain	ZH	
	OpenDialKG (Moon et al., 2019)	✗	✗	✗	✓	Train&Val &Test	Crowdsourcing	Open-Domain	EN	
	OTTers (Sevegnani et al., 2021)	✗	✗	✗	✓	Train&Val &Test	Crowdsourcing	Open-Domain	EN	
	TGConv (Yang et al., 2022)	✗	✗	✗	✓	Train&Val &Test	Existing Resource	Open-Domain	EN	
	Sotopia (Zhou et al., 2024a)	✓	✓	✗	✓	Test	Existing Resource& LLM	Open-Domain	EN	
	MultiWOZ (Budzianowski et al., 2018)	✗	✓	✗	✓	Train&Val &Test	Crowdsourcing	Task-Oriented	EN	
	SGD (Rastogi et al., 2020)	✗	✓	✗	✓	Train&Val &Test	Crowdsourcing	Task-Oriented	EN	
	MultiWOZ 2.1 (Eric et al., 2020)	✗	✓	✗	✓	Train&Val &Test	Existing Resource	Task-Oriented	EN	
	ACCENTOR-SGD/ -MultiWOZ (Sun et al., 2021)	✗	✓	✗	✓	Train&Val &Test	Existing Resource	Task-Oriented	EN	
	DuClarifyDial (Liu et al., 2022)	✗	✓	✗	✓	Train&Val &Test	Crowdsourcing	Task-Oriented	ZH	
	DuRecDial (Liu et al., 2020)	✓	✗	✗	✓	Train&Val &Test	Crowdsourcing	Recommendation	ZH	
	DuRecDial2.0 (Liu et al., 2021b)	✓	✗	✗	✓	Train&Val &Test	Crowdsourcing	Recommendation	ZH,EN	
	INSPIRED (Hayati et al., 2020)	✗	✗	✓	✓	Train&Val &Test	Crowdsourcing	Recommendation	EN	
	TG-ReDial (Zhou et al., 2020)	✓	✓	✗	✓	Train&Val &Test	Crowdsourcing	Recommendation	ZH	
	DuConv (Wu et al., 2019)	✗	✗	✗	✓	Train&Val &Test	Crowdsourcing	Recommendation	ZH	

Parti., Sce., Stra., and Go. respectively mean *Participant*, *Scene*, *Strategy*, and *Goal* elements. , , and respectively represent automatic evaluation such as F1, human, and LLM evaluation.

Table 1: Overview of existing benchmarks and metrics related to CA.

external information in a black-box manner. In addition, the goals studied in existing research are relatively shallow and narrow, and lack the universality, most of which aim to achieve a specific goal, such as confirming the harmlessness of the response, or assisting the customer with completing their ticket booking. Differently, CA seeks to analyze out the common, representative and constructive information in conversations for elevating the system's effectiveness in achieving its goals which are generally implicit and grand.

B Benchmark and Evaluation

Table 1 assembles the comprehensive datasets pertinent to *scene reconstruction*, *causality analysis*, *skill enhancement*, *conversation generation*, and *goal-directed optimization*. Despite the availability of numerous relevant datasets, they only contain ample conversational content and lack detailed conversational scene elements, such as participants (user profile), scenarios, strategies, and goals, which hinders the comprehensive and insightful analyses.