DEMO: Reframing Dialogue Interaction with Fine-grained Element Modeling

Anonymous ACL submission

Abstract

Large language models (LLMs) enabled dialogue systems have become one of the central modes in human-machine interaction, which bring about vast amounts of conversation logs and increasing demand for dialogue generation. The dialogue's life-cycle spans from *Prelude* through *Interlocution* to *Epilogue*, encompassing rich dialogue elements. Despite large volumes of dialogue-related studies, there is a lack of systematic investigation into the dialogue stages to frame benchmark construction that covers comprehensive dialogue elements. This hinders the precise modeling, generation and assessment of LLMs-based dialogue systems. To bridge this gap, in this paper, we introduce a new research task—**D**ialogue **E**lement **MO**deling, including *Element Awareness* and *Dialogue Agent Interaction*, and propose a novel benchmark, **DEMO**, designed for a comprehensive dialogue modeling and assessment. On this basis, we further build the DEMO agent with the adept ability to model dialogue elements via imitation learning. Extensive experiments on DEMO indicate that current representative LLMs still have considerable potential for enhancement, and our DEMO agent performs well in both dialogue element modeling and out-of-domain tasks.

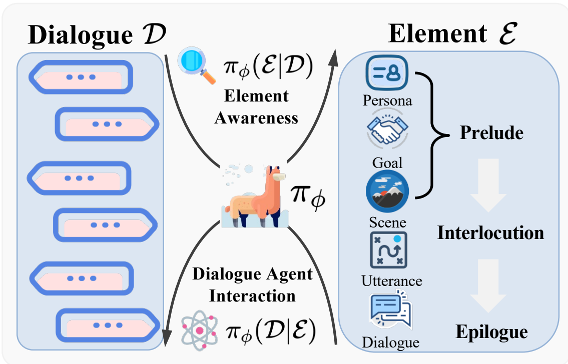


Figure 1: Overview of Dialogue Element Modeling, which focuses on two main aspects: *Element Awareness* and *Dialogue Agent Interaction*. We have formulated the comprehensive elements in the *Prelude*, *Interlocution*, and *Epilogue* stages of a complete dialogue.

1 Introduction

Under the compelling drive of large language models (LLMs), the development of intelligent language interfaces is undergoing an unprecedented transformation, with LLMs-empowered dialogue systems emerging as one of the central modes in human-machine interaction (Ross et al., 2023; Sergeyuk et al., 2024; Wu et al., 2024). The relentless evolution of LLMs has penetrated increasingly complex interaction environments, necessitating an enhancement of expressive intelligence (Chang and Chen, 2024; Zhou et al., 2024a) and a sharp sensitivity to the pivotal elements within interactions (Tang et al., 2023; Xu et al., 2024). By meticulously analyzing vast conversation logs, we can gain valuable insights into the critical elements underlying dialogue interaction, such as persona, scenario, and interaction goal, which are vital for enhancing the modeling, generation, and assessment of human-machine interaction systems.

Typically, a conversational dialogue is conducted with a goal-oriented focus, relying upon a profound understanding of its core elements (Austin, 1962; Searle, 1969; Watzlawick et al., 2011). *Dialogue agents* navigate towards their *goals* within the constraints of the *scene*, utilizing conversation *strategy* to interact with other *participants*, ultimately producing conversational *content* with communicative *intent*. In reality, the dialogue's life-cycle spans from the ***Prelude*** through the ***Interlocution*** to the ***Epilogue***, encompassing a variety of key elements (Schegloff, 2007; Hutchby and Wooffitt, 2008). However, existing dialogue benchmarks inadequately cover these comprehensive aspects (Zhang et al., 2024), merely concentrating

Data	Goal	Scene	Persona	Utterance	Dialogue	Analysis	Generation	Multilingual
DialogSum (Chen et al., 2021)	✗	✗	✗	✗	✓	✓	✗	✗
SODA (Kim et al., 2023a)	✗	✗	✗	✗	✓	✗	✓	✗
CharacterGLM (Zhou et al., 2023a)	✗	✓	✓	✗	✓	✗	✓	✗
Persona-Chat (Jandaghi et al., 2023)	✗	✗	✓	✗	✓	✗	✓	✗
SOTOPIA (Zhou et al., 2024c)	✓	✓	✓	✗	✗	✗	✓	✗
Ditto (Lu et al., 2024)	✗	✗	✓	✗	✓	✗	✓	✓
DEMO (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: The overview of our DEMO's characteristics compared to those in the related work. DEMO encompasses comprehensive dialogue elements and tasks applicable to both English and Chinese.

on dialogue generation within pre-established settings (Liu et al., 2022; Kim et al., 2023a; Zhou et al., 2023a) or only predicting selective elements based on dialogue content (Jiang et al., 2023; Ramprasad et al., 2024; Zhou et al., 2023b). Consequently, the absence of investigation into the distinct dialogue stages with these nuanced elements in the current dialogue dataset construction hinders the multifaceted modeling, understanding, and thorough evaluation of diverse dialogue-related downstream tasks.

To address the above fundamental issues, in this paper, we systematically devise a dialogue construction framework and define a new research task: **D**ialogue **E**lement **MO**deling. Concretely, our proposed task focuses on two core competencies of models: (1) ***Element Awareness***, which entails reverse engineering to decompose dialogue elements, and (2) ***Dialogue Agent Interaction***, which involves goal-directed multi-turn dialogue generation driven by fine-grained elements. Furthermore, we meticulously design a data synthesis framework for constructing a tailor-designed benchmark **DEMO**, to facilitate comprehensive dialogue modeling and assessment applicable to both English and Chinese. Besides, we amass a substantial collection of expert experiences and build a DEMO agent endowed with dialogue element modeling. We conduct extensive experiments, and the results indicate that the current advanced LLMs still have considerable space for further improvement. Our DEMO agent shows promising performances in both in-domain task as well as out-domain task for social intelligence and general tasks. The main contributions of our work are as follows:

- To support fine-grained dialogue analysis, generation, and assessment, we reframe the dialogue interaction process by defining a system of dialogue elements and propose a pioneering research task of dialogue element modeling.
- To promote dialogue element modeling, we innovate an element construction framework to develop a novel, comprehensive benchmark DEMO, and craft a DEMO agent for this task.
- Through extensive experiments, we evaluate the competencies of LLMs on DEMO, and the results show that DEMO agent performs well in both in-domain and out-of-domain tasks.

2 Dialogue Element Modeling

2.1 A System of Dialogue Elements

The dialogue is conducted with a goal-oriented focus, relying upon a deep understanding of its core elements (Austin, 1962; Searle, 1969; Watzlawick et al., 2011). Participants strategically navigate towards their goals within the scene's constraints, engaging with their environment to produce content with clear intent. The life-cycle of a dialogue spans from the prelude through the interlocution to the epilogue, encompassing various elements (Schegloff, 2007; Hutchby and Wooffitt, 2008). In the prelude, the focus is on the motivation and necessary elements of the dialogue (Goffman, 1981; Schiffrin, 1994), which include the participants' backgrounds, the time and place, the topic, and the goals of both parties. During the interlocution, attention is given to the elements intrinsic to each response, such as the intentions participants aim to convey, their current emotions and feelings, and the dialogue strategies employed (Goffman, 1981; Brown and Levinson, 1987). The epilogue involves summarizing the entire dialogue, assessing the fulfillment of both parties' goals, and examining the flow of information throughout the dialogue (Schegloff, 1973; Drew and Holt, 1998). For the detailed description of the framework of dialogue elements, please refer to Appendix B.1.

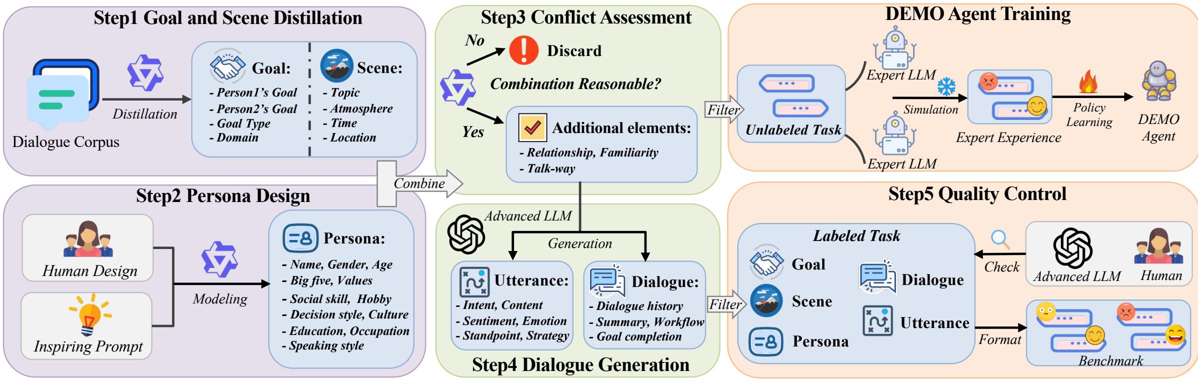


Figure 2: Overview of our DEMO synthesis framework, which consists of five steps: (1) Goal and Scene Distillation, (2) Persona Design, (3) Conflict Assessment, (4) Dialogue Generation, (5) Quality Control. The process of DEMO Agent training is also shown in this figure.

2.2 Task Definition

The task of dialogue element modeling focuses on two main aspects: *Element Awareness* and *Dialogue Agent Interaction*. Specifically, (1) Element Awareness examines whether LLM can reverse-engineer elements such as goal, persona, and scene from the entire conversation and analyze elements at the utterance level. (2) Dialogue Agent Interaction assesses the model's goal-oriented interaction capability, evaluating whether it can achieve its goal within a given environment through a limited number of interaction rounds.

2.2.1 Element Awareness

The *Element Awareness* primarily focuses on offline single-turn inference. Given an entire dialogue, it aims to model the key elements that contribute to the conversation. It has four tasks: (1) *Goal Recognition*, (2) *Persona Modeling*, (3) *Scene Reconstruction*, and (4) *Utterance Mining*.

Goal Recognition This task tends to attain the goal elements $\mathcal{G}$ from the given dialogue $\mathcal{D}$. Dialogues are arguably goal-driven (Searle, 1969; Austin, 1962), and this task aims to identify the behavioral motivations of participants using the model π_ϕ. Specifically, the model needs to identify each person's dialogue goal g_1 and g_2, and the extent s to which those goals are achieved. This process can be formally defined as $\pi_\phi(g_1, g_2, s|\mathcal{D})$.

Persona Modeling The task requires constructing the persona $\mathcal{P}$ of the two dialogue participants from the given dialogue $\mathcal{D}$. Personality, experiences, educational background, and interests often influence the manner of interaction (Grice, 1975; Austin, 1962), establishing a mapping relationship between persona and dialogue content. This task requires the model to infer from effect to cause, as well as reverse modeling persona from dialogue content. Specifically, based on the dialogue content $\mathcal{D}$, the model π_ϕ aims to infer persona p_1 and p_2, including the gender, age, personality, speaking style, hobby, and background of the two participants. This task can be formally defined as $\pi_\phi(p_1, p_2|\mathcal{D})$.

Scene Reconstruction This task requires reasoning scene elements $\mathcal{S}$ from the given dialogue $\mathcal{D}$. The scene specifies the topic, interaction type, and the relationship and familiarity between the participants, which are crucial for making the dialogue more dynamic and nuanced (Reeves and Nass, 1996; Pickering and Garrod, 2004). Specifically, it requires the model π_ϕ to reconstruct the pre-existing relationship, interaction type, and topic before the conversation starts, as well as to deduce the information flow throughout the dialogue and summarize the conversation for each participant. This task can be formalized as: $\pi_\phi(\mathcal{S}|\mathcal{D})$.

Utterance Mining The task involves the extraction of utterance-level implicit information $\mathcal{U}$ from a given dialogue $\mathcal{D}$. Each response typically conveys rich information (Goffman, 1981; Brown and Levinson, 1987), with participants employing conversational strategies, expressing standpoints and emotions, and aiming to realize their intentions for dialogue goal attainment. Specifically, given the content of a dialogue $\mathcal{D}$, the model π_ϕ is required to extract the intention, sentiment, emotion, stance, and strategy expressed by a particular utterance. This task can be formalized as follows: $\pi_\phi(\mathcal{U}|\mathcal{D})$

2.2.2 Dialogue Agent Interaction

The *Dialogue Agent Interaction* refers to the two-party goal-directed multi-turn dialogue interaction in language space. This task encompasses a wide range of interaction types, both cooperative and non-cooperative, including persuasion, argument, empathy, negotiation, and more, which can be regarded as an incomplete information game (Reif, 1984). It examines the ability of LLM to model dialogue driven by elements through dynamic inference. It can be formulated as a Markov Decision Process (Bellman, 1957).

State The persona, goal, scene, and dialogue history in each episode denote the state. As the interactions progress, the dialogue history is continuously updated while the persona, goal, and scene stay unchanged. The global state at timestep t is represented as $\mathcal{D}_t$, which includes the dialogue content generated after the t-th turn along with other constant dialogue elements. Meanwhile, local states tied to specific sub-action sets are enhanced by combining the global state $\mathcal{D}_t$ with the history of the previous t-1 dimensional sub-action choices.

Action The interaction unfolds between two agents, $\mathcal{A}_1$ and $\mathcal{A}_2$. At each turn t, according to the observation of the state, the agent selects an action that consists of one utterance $\mathcal{U}_t$ generated by itself.

Transition In our setting, the transition function adds the utterance to the interaction history while the persona, goal, and scene stay unchanged in state representation. The dialogue history can be represented as an alternating sequence of utterances generated by two players, denoted by $\{\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3, ..., \mathcal{U}_t - 1, \mathcal{U}_t\}$. The interaction continues until the dialogue goal is achieved or the maximum number of turns T is reached.

Reward After each turn, a reward function can be called to quantify how well each player has been doing. The design of the reward function is critical. To comprehensively examine the LLMs' dialogue interaction capabilities, we devise a multi-dimensional reward framework, which articulates 0-10 scoring criteria for each dimension, prompting GPT-4o (Hurst et al., 2024) as a reward function to evaluate the interaction according to four dimensions: *Goal achievement* (i.e. the extent to which the dialogue goals of both parties are fulfilled), *Believability* (i.e. the extent to how well participants align with dialogue elements), *Skillfulness* (i.e. the ability of participants to analyze dialogue history, mine utterances, and provide appropriate responses) and *Realistic* (i.e. the extent to which the response content appears human-like and vivid, as opposed to being overtly AI-generated).

3 DEMO Benchmark

3.1 Overview

DEMO is our newly developed benchmark specifically designed to enhance the dialogue element modeling capabilities of dialogue systems, which features an equal 1:1 ratio of Chinese to English languages. To evaluate element awareness, we have a total of 4,000 evaluation samples, with a maximum of 26 dialogue turns and an average of 18.3 turns, covering 23 distinct dialogue elements. Each of the four tasks contains 1,000 test samples. In terms of dialogue agent interaction, DEMO provides 1,000 episodes that cover a wide array of cooperative and non-cooperative interaction types. Examples of specific tasks are illustrated in Appendix G.

3.2 Benchmark Construction Framework

Our framework is depicted in Figure 2. We follow the dialogue element system to sequentially annotate the elements of prelude, interlocution, and epilogue. First, based on *Goal and Scene Distillation* and *Persona Design*, we generate the three prelude elements: goal, scene, and persona. Then, through the *Conflict Assessment*, we ensure that the combined prelude elements are reasonable. Subsequently, we generate the corresponding interlocution and epilogue elements through *Dialogue Generation*. Finally, after *Quality Control*, we inspect and revise the data to establish the final benchmark.

Goal and Scene Distillation In this initial stage, we distillate the goals and scenes from the given dialogue. By leveraging an instance-driven paradigm, we diversify data from the large-scale dialogue corpus. We utilize SODA (Kim et al., 2023a) and LCCC (Wang et al., 2020) as our seed corpora, which include millions of English and Chinese dialogues encompassing various aspects of social commonsense. Specifically, we employ Qwen2-72B (Yang et al., 2024a) to extract participant's goals and the conversation scene from dialogue. Finally, we get 2.6 Million goal and scene data.

Persona Design In parallel with the previous phase, we established a comprehensive and diverse persona collection. The creation of personas is divided into two parts: (1) Designing the persona attributes pool and (2) Inspiring prompting. Referring to (Zhou et al., 2023a, 2024c; Chen et al., 2024; Yang et al., 2024b), we consider the following attributes: name, gender, age, Big Five traits (McCrae and John, 1992), moral values (Graham et al., 2011), social skills (Yang et al., 2024c), per-

Human	Goa.	Per.	Sce.	Utt.	Avg.
Rater1	0.80	0.87	0.71	0.74	0.82
Rater2	0.74	0.76	0.77	0.79	0.78

Table 2: The Kappa consistency results between LLM annotations and two human raters on different elements

sonal values (Schwartz, 1992), and decision style (Scott and Bruce, 1995). Based on these characteristics, we combine them and then leverage Qwen2-72B for more detailed persona modeling. By prompting the LLM with diverse web texts, as (Chan et al., 2024) suggests, we generate wide-ranging personas. Ultimately, the LLM produces detailed information on each person's background, hobbies, education, occupation, culture, relationships, and speaking style. At this stage, we have modeled 200,000 diverse personas.

Conflict Assessment After gathering prelude elements, we proceed to acquire reasonable combinations of these elements. To assess the reasonableness and coherence of the combined prelude elements, we prompt the Qwen2-72B to check for issues like character identity contradictions, misalignment between persona and goal, or unsuitable pairings of dialogue participants. For combinations deemed reasonable, we further instruct the model to provide additional details about the relationship and familiarity between participants, as well as mode of interaction. To further evaluate the LLM's capability in conflict assessment, we also conducted a human evaluation, where two human raters and the LLM performed conflict checks on the combinations following identical procedures and criteria. The inter-rater agreement measured by Cohen's Kappa (Fleiss, 1971) between the two human raters was 0.85, while the Kappa value between Rater1 and LLM is 0.65, and the value between Rater2 and LLM is 0.79. The results demonstrate high consistency between LLM and human annotators in this task, indicating that LLM performs at a human-comparable level and is competent for this annotation task. For the complete human annotation process and information, please refer to Appendix A.

Dialogue Generation Upon establishing reasonable combinations, we proceed to generate interlocution and epilogue elements. To manage data distribution effectively, we categorize all combinations into ten types based on the dialogue goal type and extract them evenly. Leveraging the LLM's role-play capabilities (Zhou et al., 2023a, 2024c; Chen et al., 2024), we prompt it to create dialogues that align with specified persona, goal, and scene. We also prompt the advanced LLMs to analyze each utterance with its associated intention, sentiment, emotion, stance, and strategy, culminating in a comprehensive output that includes the information flow and dialogue summary. To form our benchmark, we curate 1,800 distinct combinations, utilizing GPT-4o for generation.

Quality Control To ensure the accuracy of benchmark annotations, we employ a three-step verification process: (1) Advanced-LLM Check: Two of the most advanced LLMs, GPT-4o and Claude-3.5-Sonnet, independently review and validate the quality of annotations. They examine each entry, editing any unreasonable or low-quality labels to maintain accuracy. (2) Voting: We implement the simple majority voting method to finalize the label. (3) Manual Check: After the voting process, We engage two human raters to further examine and assess the quality of data annotation following our pre-established quality control standards. For the detailed human annotation information, please refer to Appendix A.

We also use Kappa score (Fleiss, 1971) to measure annotation quality in Table 2. The data quality inspection process is performed by two experienced annotators, with a consistency Kappa value of 0.84 between them. Their Kappa consistency results with different elements annotated by the LLM are shown in Table 2. The LLM shows high consistency with the two annotators, demonstrating performance comparable to humans. Additionally, we conducted a manual verification of the data, achieving an accuracy rate of 91.17%. These results have all validated the quality of our benchmark.

3.3 DEMO Agent Training

To further investigate the task characteristics and benchmark impact, we build a DEMO agent endowed with dialogue element modeling. Humans have the ability to learn efficiently through observing and imitating the behavior of others (Schaal, 1996; Ross et al., 2011; Torabi et al., 2018). Drawing inspiration from this, we propose enhancing the performance of LLMs in dialogue element modeling by integrating behavioral learning methods. This approach centers on acquiring insights through interactions with expert models and developing an imitation policy. Behavioral learning (Bain and Sammut, 1999; Ross and Bagnell, 2010), is an ap-

Model	Element Awareness					Dialogue Agent Interaction					Overall
	Goa.	*Per.*	*Sce.*	*Utt.*	**Avg**	*Goa.*	*Bel.*	*Ski.*	*Rea.*	**Avg**	
Proprietary LLM											
GPT-4o	<u>5.975</u>	4.051	6.167	7.308	5.875	<u>8.190</u>	<u>9.181</u>	<u>8.614</u>	<u>8.537</u>	**8.631**	**6.793**
Claude-3.5-Sonnet	**5.979**	<u>4.145</u>	6.221	6.243	5.647	7.571	9.174	8.432	**8.840**	<u>8.504</u>	6.599
GPT-4o-mini	5.802	3.586	5.748	7.002	5.534	7.551	9.082	8.316	8.163	8.278	6.449
Claude-3.5-Haiku	5.492	3.858	6.071	6.304	5.431	7.361	9.115	8.402	7.968	8.212	6.358
Open-sourced LLM											
Qwen2-72B-Instruct	5.357	**4.406**	5.702	6.921	5.596	**8.447**	**9.204**	**8.699**	8.175	**8.631**	6.608
Llama-3.1-70B-Instruct	5.559	3.643	6.078	7.051	5.593	7.223	7.914	7.222	6.753	7.278	6.154
Backbone LLM											
Qwen2-7B-Instruct	5.306	3.981	5.459	6.347	5.244	6.698	8.112	6.895	6.278	6.996	5.828
Llama3.1-8B-Instruct	5.546	3.287	5.403	6.523	5.189	5.831	6.166	5.519	4.974	5.623	5.335
DEMO Agent											
DEMO-Qwen2-7B	5.229	3.946	<u>6.534</u>	<u>7.914</u>	<u>5.906</u>	7.450	8.864	8.073	7.864	8.063	<u>6.625</u>
	-	-	Δ1.075	Δ1.567	Δ0.797	Δ0.752	Δ0.752	Δ1.178	Δ1.586	Δ1.067	Δ0.797
DEMO-Llama3.1-8B	5.623	3.939	**6.543**	**7.926**	**6.008**	6.945	7.688	7.015	6.378	7.707	6.341
	Δ0.077	Δ0.652	Δ0.752	Δ1.140	Δ0.819	Δ1.114	Δ1.522	Δ1.496	Δ1.404	Δ2.084	Δ1.006

Table 3: The results of various LLMs on the DEMO. The highest score among different LLMs is highlighted in **bold**, and the second highest is <u>underlined</u>. And Δ values represent the improvement over the baseline. **Element Awareness** has four tasks:(1) *Goa*: Goal Recognition, (2) *Per*: Persona Modeling, (3) *Sce*: Scene Reconstruction, and (4) *Utt*: Utterance Mining. **Dialogue Agent Interaction** includes four dimensions: (1) *Goa*: Goal Achievement, (2) *Bel*: Believability,(3) *Ski*: Skillfulness,(4) *Rea*: Realistic. **Overall** is the average score of two tasks.

proach to extracting and distilling expert policies from high-quality data, particularly from models with advanced capabilities. In the context of dialogue element modeling, it involves gaining an understanding of element awareness in single-turn reasoning and achieving nuanced expression in multi-turn interactions. During the benchmark construction process, we have accumulated a diverse amount of unlabeled data, which serves as the environment for the expert model to simulate dialogue modeling. Specifically, GPT-4o is employed as the expert model. By engaging in both single-turn and multi-turn interactions within this environment, we are able to amass a wealth of expert experience, which is subsequently utilized to train the model.

4 Experiments

4.1 Experimental Setup

Models We evaluate ten advanced LLMs, including API-based LLMs: GPT-4o, GPT-4o-mini, Claude3.5-Sonnet (Anthropic, 2024b), Claude3.5-Haiku (Anthropic, 2024a) and Open-sourced LLMs: Qwen2-72B-Instruct, Qwen2-7B-Instruct, Llama3.1-70B-Instruct, Llama3.1-8B-Instruct (Dubey et al., 2024).

Evaluation Metric Evaluating the unpredictable behaviors of LLMs, traditional metrics such as BLEU and Rouge-L may yield inaccurate responses. Recent research (Zhang et al., 2023; Zheng et al., 2023; Kim et al., 2023b) indicates that the GPT-4 evaluator demonstrates high consistency with human evaluation while reducing costs, making it a reasonably reliable annotator. Following these work (Perez et al., 2022; Zhou et al., 2024c; Wang et al., 2024), we prompt GPT-4o as a judge model. For the element awareness task, we evaluate the output based on the golden answer from several aspects, scoring from 0 to 10. For dialogue agent interaction task, we only consider the reward at the end of the interaction. The detailed prompts are provided in Appendix F.

Implement Details Please refer to Appendix C.

4.2 Main Results

We assess ten advanced LLMs on the DEMO benchmark. The main results are shown in Table 3. To cross-validate results, we add two other exemplar LLMs, DeepSeek-V3 (Liu et al., 2024) and Gemini-1.5-Pro (Team et al., 2024), as judge models (see Appendix D).

Model Analysis GPT-4o shows the best overall performance, maintaining great performance across all dimensions. Analyzing from the perspective of parameter size, the model's performance aligns with the *Scaling Law*, indicating that models with larger parameters possess stronger expressive capabilities. Additionally, the gap between open-source and closed-source models is narrowing. For instance, Qwen2-72B-Instruct has achieved state-of-the-art performance in dialogue agent interaction tasks, with overall performance differences from

Model	SOC	SEC	FIN	REL	KNO	GOA	BEL	Overall
Qwen2-7B-Instruct	-0.05	0.00	0.73	1.83	3.41	6.07	8.64	2.95
DEMO-Qwen2-7B	-0.02	0.00	0.82	2.32	4.52	6.40	8.94	3.28(Δ0.33)
Llama3.1-8B-Instruct	-0.50	-0.01	-0.16	-0.60	2.21	3.39	8.63	1.85
DEMO-Llama3.1-8B	-0.19	0.00	0.29	0.85	2.88	3.77	8.41	2.29 (Δ0.44)

Table 4: Evaluation results on SOTOPIA, which scored from seven social dimensions . The overall score is the average of the seven dimensions reflecting the overall social intelligence. GPT-4o rates each dimension.

GPT-4o being minimal.

Task Analysis There remains significant room for improvement in dialogue element modeling, particularly in element awareness task. Accurately modeling various elements (such as persona modeling) from dialogue content is still challenging, potentially requiring multi-step reasoning or additional clues. In dialogue agent interaction tasks, current LLMs exhibit excellent expressive capabilities, adeptly adhering to settings and generating relatively realistic content. Humans are inherently social, striving to achieve social goals in daily interactions. Goal achievement is a crucial feature of intelligence; thus, the ability to perceive targets and collaborate to achieve goals reflects LLMs' higher-order capabilities. However, their ability to achieve self-set goals through multi-turn interactions requires enhancement.

DEMO Agent By learning through expert experience imitation, the DEMO agent has achieved significant improvements across two different backbones, with an average task improvement of 0.9. Specifically, the agent utilizing the LLaMA backbone achieved SOTA performance in element awareness tasks. Meanwhile, the agent built on the Qwen backbone secured the second-highest score, surpassed only by GPT-4o. The DEMO Agent has also surpassed or performed on par with models with larger parameters, such as Claude3.5-Sonnet and the Qwen2-72B-Instruct. This demonstrates the effectiveness of imitation learning and expert experience. However, this method has a performance ceiling limited by the abilities of the expert model. Fully leveraging additional modeling cues to develop the capacity between element awareness and intelligent interaction will be a primary focus of our future work. We also present the detailed case study to analyze LLM outputs in Appendix E.

4.3 Out-of-domain Performance

The DEMO agent has demonstrated promising results in dialogue element modeling within the domain. However, the question remains: can this capability extend to tasks beyond that domain? To evaluate this, we selected the hard episodes of SOTOPIA (Zhou et al., 2024c) as our testing environment, which assesses social intelligence. Two LLMs are prompted to act as role-playing social agents in SOTOPIA, engaging with each other through communication. SOTOPIA designed a seven-dimension framework to assess the social intelligence of LLMs: social rules, secret-keeping, financial benefits, relationship maintenance, knowledge, goal completion, and believability. For each task, agents are scored along designed dimension.

Table 4 presents the results. All our DEMO agents show remarkable generalization capabilities in social intelligence tasks, with prominent performance improvements. This validates the necessity and effectiveness of fine-grained dialogue modeling. For baseline models like Llama3.1-8B-Instruct, during pre- and post-training phases, the training data only included dialogue content or very limited dialogue elements. Dialogue elements are implicitly learned from the conversation content, thus resulting in inferior overall performance. In contrast, the DEMO agent, through explicitly modeling and learning various elements of a dialogue, can better understand dialogue content, clearly recognize current goals and intentions, and follow the current dialogue scene and persona settings, thereby performing more competently in complex social interaction scenarios.

4.4 Catastrophic Forgetting Problem

In addition to confirming the model's great performance in dialogue element modeling, evaluating whether the other capabilities remain unaffected is equally crucial. Continued training can sometimes lead to catastrophic forgetting, where the model loses previously acquired knowledge, disrupting its initial alignment. We use the Helpful, Honest, Harmless (HHH) (Askell et al., 2021) dataset to assess the impact on alignment performance. This involves a multiple-choice task to measure the model's ability to select better answers from two

Model	MMLU	HHH
Qwen2-7B-Instruct	69.04	45.70
DEMO-Qwen2-7B	68.37	46.15(Δ0.45)
Llama3.1-8B-Instruct	65.94	46.61
DEMO-Llama3.1-8B	66.06 (Δ0.12)	45.25

Table 5: The results of the LLM's general capability and alignment performance, using the accuracy score.

given options. When presented with both options, we calculate the model's tendency to favor one answer over the other. To assess the model's general capabilities, we employ the MMLU (Hendrycks et al., 2021), using a 5-shot evaluation based on next-word prediction. Accuracy serves as the evaluation metric across two benchmarks.

The results are presented in Table 5. It is evident that DEMO Agents retain the overall capabilities of the base model. Although a few did not exhibit enhancements, our models performed comparably to the base model. They did not experience significant issues with catastrophic forgetting, indicating that the dialogue element modeling operates independently of the general capabilities.

5 Related Work

We review related research based on two trends in interactive dialogue systems: (1) *Dialogue Analysis*: inferring dialogue elements based on dialogue content. (2) *Dialogue Generation*: generating dialogue content according to the dialogue settings. While we discuss related work from the perspectives of two trends mentioned above, different types of dialogue systems and their distinction are discussed in Appendix B.2.

5.1 Dialogue Analysis

The goal of dialogue analysis is to mine critical elements (such as intent, profiles, summary, etc.) from the dialogue (Zhang et al., 2024), which can extract actionable insights and drive empowerment. In the era of small language models, dialogue analysis did not form a systematic task but was broken down into atomic tasks, such as slot filling and intent classification (Qin et al., 2020; Louvan and Magnini, 2020; Jiang et al., 2023), dialogue summary (Chen et al., 2021; Fabbri et al., 2021; Ouyang et al., 2023; Ramprasad et al., 2024) and persona extraction (Wang et al., 2022; Zhou et al., 2023b), etc. In the era of LLM, recent work (Zhang et al., 2024) performed a thorough review and systematized conversation analysis task. There is a scarcity of datasets that encompass all essential elements of dialogue, making it challenging to model and evaluate, which affects the development of dialogue modeling.

5.2 Dialogue Generation

The related work on dialogue generation primarily focuses on constructing dialogue datasets and designing steering-based methods for dialogue modeling. Task-oriented dialogue (Rashkin et al., 2019; Sun et al., 2021; Liu et al., 2022) focuses on completing specific tasks, emphasizing task completion rather than generalization. Open-domain dialogue (Li et al., 2017; Wang et al., 2020; Kim et al., 2023a) is mainly designed for "chit-chat" between users, with more general tasks and a greater focus on immersion. Recently, several role-playing works (Zhou et al., 2023a; Lu et al., 2024; Chen et al., 2024; Yang et al., 2024b; Zhou et al., 2024b) have emerged, which place more emphasis on dialogue engagement and character consistency. However, there is limited guidance for dialogue modeling, and there is a lack of reward modeling for goal-oriented dialogues during interactions. Inspired by social intelligence work SOTOPIA (Zhou et al., 2024c), we define a more comprehensive dialogue generation task space, considering more dialogue modeling elements. We evaluate dialogue interaction capabilities through multi-turn interactions and introduce element awareness tasks to thoroughly assess the LLM in both Chinese and English.

6 Conclusion

In this work, we redefine the dialogue framework and introduce a new research task: Dialogue Element Modeling. This task involves two fundamental capabilities, element awareness and dialogue agent interaction, which enhance the complex modeling and comprehensive evaluation of dialogue systems. We first develope a process for constructing dialogue elements and create the benchmark, DEMO, which includes comprehensive dialogue elements suitable for both English and Chinese. We further develop the DEMO agent specifically for dialogue element modeling. Through extensive experiments, we assess the performance of several advanced LLMs, revealing that there is still room for improvement in this task. In addition, the results demonstrate that our DEMO agent delivers excellent performance in both in-domain and out-of-domain scenarios.

7 Limitations

First, to ensure the quality of data annotation for benchmark construction, we employ two human raters to assess the conflicts and accuracies of the constructed benchmark. This process is both time-consuming and costly. Second, the performance of our DEMO agent is constrained by the capabilities of the expert model we used. The interplay between element awareness and dialogue agent interaction remains insufficiently understood. Third, a joint learning approach that integrates reasoning and generation processes could enhance both dialogue agent interaction and element awareness.

References

AI Anthropic. 2024a. Claude 3.5 haiku model card addendum. *Claude-3.5-Haiku Model Card*.

AI Anthropic. 2024b. Claude 3.5 sonnet model card addendum. *Claude-3.5-Sonnet Model Card*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

John Langshaw Austin. 1962. *How to do things with words*. Oxford: University Press.

Michael Bain and Claude Sammut. 1999. A framework for behavioural cloning. In *Proceedings of Machine Intelligence*, page 103–129.

Richard Bellman. 1957. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. 4. Cambridge university press.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

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.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of ACL*, pages 5062–5074.

Paul Drew and Elizabeth Holt. 1998. Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in society*, 27(4):495–522.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Erving Goffman. 1981. Forms of talk. *University of Pennsylvania*.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

HP Grice. 1975. Logic and conversation. *Syntax and semantics*, 3.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

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

Ian Hutchby and Robin Wooffitt. 2008. *Conversation analysis*. Polity.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

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.

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. 2023a. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of EMNLP*, pages 12930–12949.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023b. Prometheus: Inducing fine-grained evaluation capability in language models. In *Proceedings of ICLR*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of SOSP*.

Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*, pages 986–995.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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.

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.

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.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of EMNLP*, pages 3419–3448.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

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.

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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*, pages 5370–5381.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10):19–36.

John H Reif. 1984. The complexity of two-player games of incomplete information. *Journal of computer and system sciences*, 29(2):274–301.

Stéphane Ross and Drew Bagnell. 2010. Efficient reductions for imitation learning. In *Proceedings of AISTATS*, volume 9, pages 661–668.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of AISTATS*, volume 15, pages 627–635.

Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of IUI*, pages 491–514.

Stefan Schaal. 1996. Learning from demonstration. In *Proceedings of NeurIPS*, volume 9.

Emanuel A Schegloff. 1973. Opening up closings. *Semiotica*.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press.

Deborah Schiffrin. 1994. Approaches to discourse.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*.

Susanne G Scott and Reginald A Bruce. 1995. Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, 55(5):818–831.

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

Agnia Sergeyuk, Sergey Titov, and Maliheh Izadi. 2024. In-ide human-ai experience in the era of large language models; a literature review. In *Proceedings of IDE Workshop*, pages 95–100.

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.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of ACL*, pages 5456–5468.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. In *Proceedings of IJCAI*, pages 4950–4957.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of EMNLP*, pages 5627–5646.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Proceedings of NLPCC*, pages 91–103.

Zhilin Wang, Xuhui Zhou, Rik Koncel-Kedziorski, Alex Marin, and Fei Xia. 2022. Extracting and inferring personal attributes from dialogue. In *Proceedings of NLP4ConvAI Workshop*, pages 58–69.

Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson. 2011. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *Proceedings of COLM*.

Junjielong Xu, Ziang Cui, Yuan Zhao, Xu Zhang, Shilin He, Pinjia He, Liqun Li, Yu Kang, Qingwei Lin, Yingnong Dang, et al. 2024. Unilog: Automatic logging via llm and in-context learning. In *Proceedings of ICSE*, pages 1–12.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. 2024b. Simschat: A customisable persona-driven role-playing agent. *arXiv preprint arXiv:2406.17962*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024c. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

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

Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. 2024. The imperative of conversation analysis in the era of llms: A survey of tasks, techniques, and trends. *arXiv preprint arXiv:2409.14195*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of NeurIPS*, volume 36, pages 46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of ACL, System Demonstrations*, pages 400–410.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024a. CharacterGLM: Customizing social characters with large language models. In *Proceedings of EMNLP: Industry Track*, pages 1457–1476.

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*.

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

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024b. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of EMNLP*, pages 21692–21714.

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

A Human Annotation Process

A.1 Rater Information

The annotators were hired and not co-authors. This research was collaborated with the company, who provided GPU and data annotations. The annotations were conducted by 2 Ph.D. students (candidates after one and half years of study in our country typically) working as interns in this company. The annotators specialize in dialogue systems and natural language processing. They are native Chinese speakers with proficient English reading and writing skills. They worked for two weeks, following an 8-hour workday schedule, with a daily internship compensation of 500 CNY per day. Before the annotation process, we conducted specific training sessions, providing detailed annotation examples and requirements. Regular meetings were held to discuss issues encountered during the annotation process to ensure accurate data labeling. During the annotation process, the two annotators worked independently.

A.2 Guidelines for Human Evaluation

The detailed annotation guidelines and interface are shown in ***Figure 5*** and ***Figure 6***. The manual annotation process comprises two primary phases: Conflict Assessment and Quality Control.

Conflict Assessment In this phase, we aim to evaluate the agreement between human raters and the LLM in determining the reasonableness of combinations. This helps assess the model's judgment capabilities and its suitability for this annotation task. We sample 300 unlabeled combinations of elements (scene, goal, and persona), each independently reviewed by humans and LLM. Annotators are tasked with evaluating whether each combination is reasonable and conflict-free, assigning a binary score: 1 for reasonable combinations and 0 for unreasonable ones. The results, as shown in ***Section 3.2***, indicate a high level of consistency between LLM and human assessments, thereby validating the quality of the annotations.

Quality Control To assess the agreement between humans and the LLM in quality control and error detection, both human annotators and the LLM independently review 1,200 unchecked data points (300 sampled per sub-task in Element Awareness). To verify the final annotation quality and calculate accuracy, human evaluators examine the LLM-checked data. They assess whether the labels are correct or incorrect according to task-specific requirements and examples, marking them as either correct ("1") or incorrect ("0"). The Kappa consistency results, as shown in ***Table 2***, demonstrate that the LLM achieves human-comparable annotation performance and effectively identifies label correctness. Furthermore, the LLM successfully corrects inaccurate labels, maintaining high accuracy throughout the process, as shown in ***Section 3.2***.

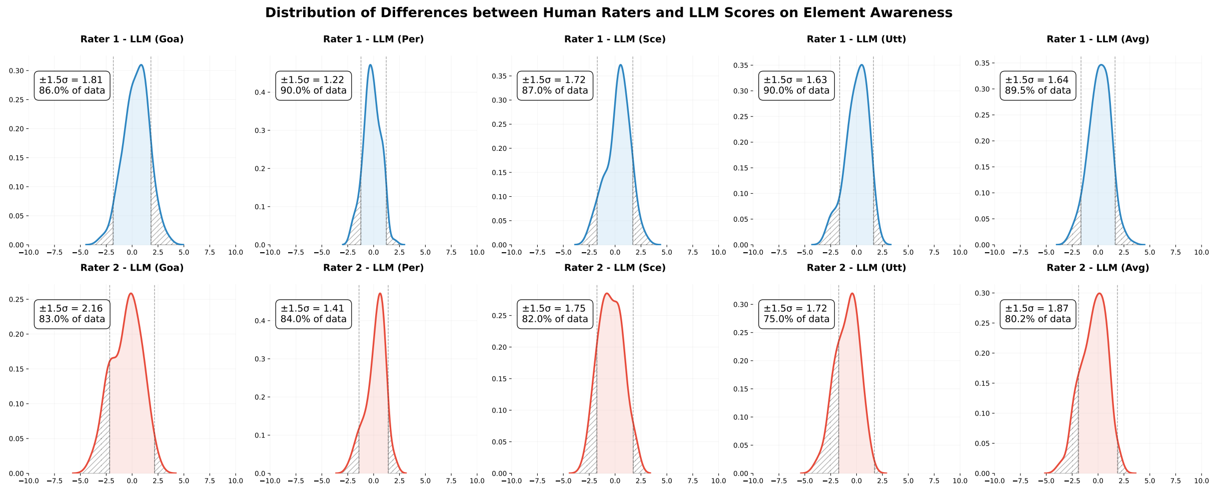
Element Awareness				
Human	**Goa.**	**Per.**	**Sce.**	**Utt.**
Rater1	0.87	0.74	0.78	0.78
Rater2	0.82	0.66	0.78	0.76
Dialogue Agent Interaction				
Human	**Goa.**	**Bel.**	**Ski.**	**Rea.**
Rater1	0.89	0.64	0.78	0.63
Rater2	0.88	0.42	0.75	0.54

Table 6: Pearson correlation coefficients and p-values between human judgment and GPT-4o evaluation on models' output among different dimensions.

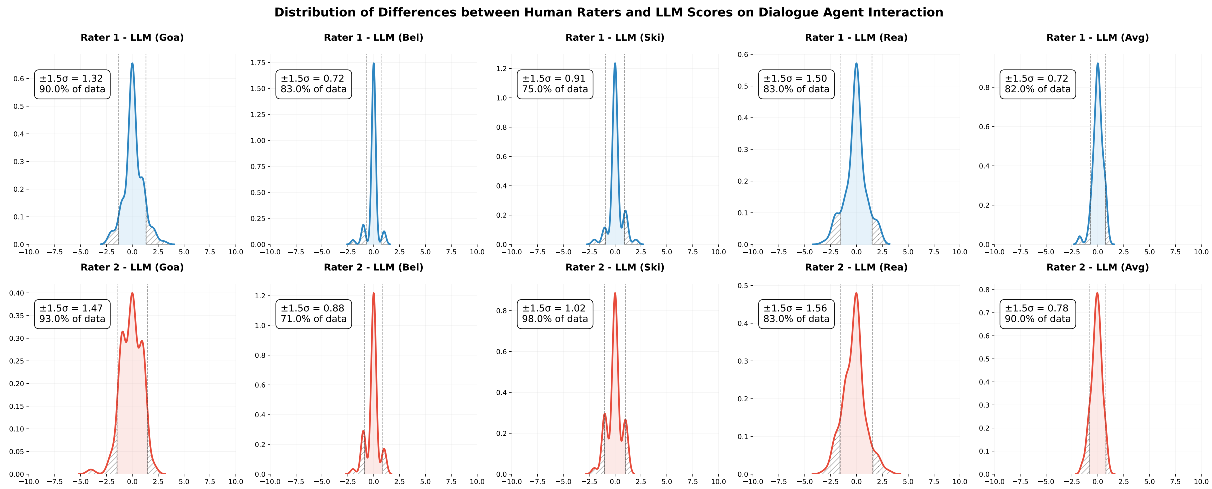
A.3 Human Evaluation on GPT4o Judge Model

In this section, we investigate whether current LLMs can be used to automate the evaluation process. we select GPT-4o as our representative model because of its superior performance. To conduct this study, we gather output data and ask human evaluators and GPT-4o to assess performance based on the dimensions outlined in DEMO. Due to time constraints, we limited our sample to 400 outputs from the Element Modeling task, with 100 outputs per sub-task, and 100 interaction data from the Dialogue Agent Interaction.

In ***Figure 3***, we present the difference between the human scores and the GPT-4o scores. The



(a) Element Awareness



(b) Dialogue Agent Interaction

Figure 3: Distribution of score differences between human and GPT-4o evaluations on (a) Element Awareness and (b) Dialogue Agent Interaction. The x-axis shows the human-LLM score, and the y-axis shows the probability density.

standard deviations are all below 1.5, with most of some even less than 1. Notably, the majority (over 80%) of the GPT-4o scores are within 1.5 standard deviations of the human scores across all the dimensions. ***Table 6*** illustrates the Pearson correlation (Lee Rodgers and Nicewander, 1988) between the LLM and humans. Obviously, it demonstrates a predominantly strong positive correlation between human judgment and GPT-4o evaluations across various dimensions, with most coefficients indicating values greater than 0.7. For element awareness, the LLM demonstrates remarkably high consistency with human evaluations across all dimensions. This suggests that when there is an objective reference, the LLM can achieve a judgment level comparable to that of humans. Regarding dialogue agent interaction, the model shows extremely strong consistency in the *Goa* and *Ski* dimensions and maintains a positive correlation with a high level of significance across the other dimensions.

Combining these observations, it is evident that GPT-4o can cautiously be employed as a substitute for human judgment in assessing model performance across the dimensions defined in DEMO. This observation aligns with the findings of recent research (Zhang et al., 2023; Zheng et al., 2023; Kim et al., 2023b; Li et al., 2024) on LLM-as-the-Judge.

B Element Modeling for Dialogue System

B.1 Details on our Element Framework

By analyzing the elements of prelude, interlocution, and epilogue, a deeper understanding of the dialogue's structure and dynamics can be achieved,

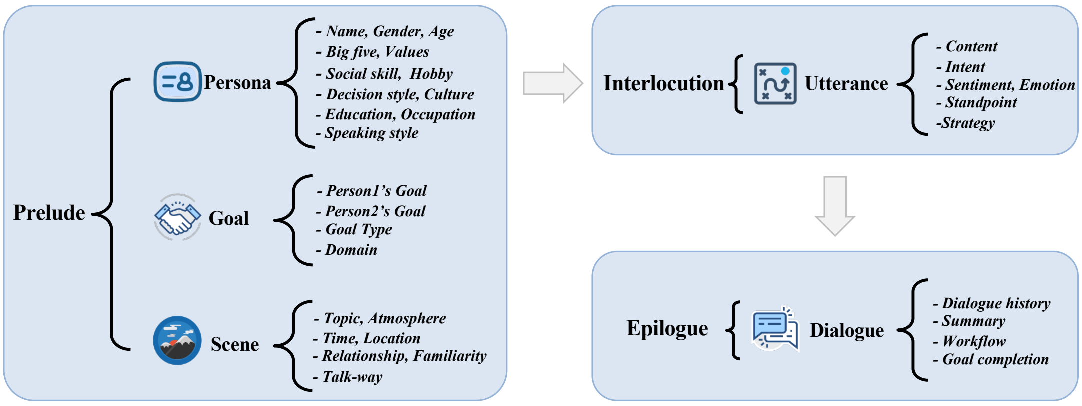


Figure 4: Overview of the system of dialogue elements.

thereby facilitating comprehensive dialogue modeling. The detailed dialogue element system is shown in ***Figure 4***. Based on the stages of dialogue, we have identified five categories of elements: persona, goal, scene (Prelude), utterance (Interlocution), and dialogue (Epilogue). We have refined each category into more granular elements. In our constructed framework, a total of 33 specific elements are involved.

B.2 Discussions on Different Types of Dialogue System

Dialogue systems serve as a bridge between humans and machines, enabling natural interaction through conversation. These systems can be categorized into two categories: task-oriented and open-ended dialogues. Task-oriented dialogue systems are designed as specialized assistants, guiding users through specific objectives such as booking flights or scheduling appointments. While their success is commonly measured through task completion rates, these systems often struggle to adapt their capabilities beyond their predetermined domains. In contrast, open-ended dialogue systems engage in casual "chit-chat" with users. These systems often incorporate personal features to create more engaging and meaningful interactions. However, despite their ability to possess a deep understanding of various subjects, open-ended systems lack concrete objectives during conversations, and their apparent understanding often masks a fundamental limitation: they process patterns rather than truly comprehend meaning.

The DEMO benchmark represents an advance by requiring dialogue agents to simultaneously maintain scene awareness, exhibit consistent persona, and achieve explicit goals, which is more comprehensive than the existing dialogue systems. We focus on building this benchmark not at the system level but rather at the more detailed dialogue element modeling level by framing a unified dataset that covers both types of goal-oriented and open-ended dialogue. Compared to task-oriented dialogue, DEMO offers richer interactive scenarios. Compared to open-ended dialogue, DEMO has clear social objectives. Additionally, DEMO introduces more fine-grained dialogue elements that enable more nuanced and sophisticated interactions.

C Information on Implementation

All the experiments are conducted on a server with 8×A100 80GB.

C.1 Inference Setting

To ensure the stability of the evaluation, we set the temperature of the evaluator to 0. For element awareness, we set the temperature to 0 to ensure reproducibility. For dialogue agent interaction, we set the temperature to 1 to encourage diversity.

C.2 Involved Model Versions

To help with reproducibility, we provide the detailed version number of all the models we used in our experiments. When we mention each name like GPT-4o or Qwen2-72B in our main section, we actually refer to those model versions in ***Table 7***. Such information helps researchers reproduce our results. For API-Based LLMs, we directly utilize the Azure API for testing. As for open-source models, we conduct experiments accelerated by the vLLM framework (Kwon et al., 2023).

Model	Version	Implement
Proprietary LLM		
GPT-4o	`gpt-4o-2024-08-06`	API
Claude-3.5-Sonnet	`claude-3-5-sonnet-20241022`	API
GPT-4o-mini	`gpt-4o-mini-2024-07-18`	API
Claude-3.5-Haiku	`claude-3-5-haiku-20241022`	API
DeepSeek-V3	`deepseek-chat`	API
Gemini-1.5-Pro	`gemini-1.5-pro-001`	API
Open-sourced LLM		
Qwen2-72B-Instruct	`Qwen2-72B-Instruct`	vLLM
LLaMA3.1-70B-Instruct	`LLaMA3.1-70B-Instruct`	vLLM
Qwen2-7B-Instruct	`Qwen2-7B-Instruct`	vLLM
LLaMA3.1-8B-Instruct	`LLaMA3.1-8B-Instruct`	vLLM

Table 7: The detailed versions of our used LLMs.

C.3 Training Setting

We use Qwen2-7B-Instruct and LLaMA3.1-8B-Instruct as our backbones. Our total batch size is 32, with a cut-off length of 8192, and the learning rate is set to 1.0e-4. We train for 3 epochs, using cosine annealing with a warm-up ratio of 0.1. For checkpoint selection, we use 10% of the training data as a validation set, which is not used in training but only to validate the checkpoint's loss. We select the checkpoint with the lowest validation loss. The process of policy updating is efficiently executed through LoRA (Hu et al., 2022). We use the llama-factory framework (Zheng et al., 2024) to assist in our training.

D Additional Judge Models

In addition to the main results presented in Table 3, to cross-validate the results, we add two additional judge models DeepSeek-V3 (Liu et al., 2024) and Gemini-1.5-Pro (Team et al., 2024). The former is currently the strongest open-source LLM, with model performance comparable to that of GPT-4o, while the latter is Google's most powerful LLM to date.

The results are shown in ***Table 8*** and ***Table 9***. It can be seen from the tables that the scores evaluated by DeeSseek-V3 and Gemini-1.5-Pro reflect similar trends as those in GPT-4o. Besides, DeepSeek tends to give higher scores while Gemini's scoring aligns more closely with that of GPT4o. Generally across these LLMs, GPT4o achieves the best performance, and notably, our DEMO agent shows significant improvement, surpassing models with larger parameter counts.

E Case Study

We conducted a detailed analysis of two examples from both the element awareness and dialogue agent interaction tasks, comparing the outputs of DEMO-Qwen2-7B (superior performance) and Qwen2-7B-Instruct (inferior performance). The comparative examples are presented in **Table 10** for element awareness, while the dialogue agent interaction case studies are illustrated in ***Table 11*** and ***Table 12***.

In comparing the element awareness task performance, the DEMO agent demonstrates superior accuracy and conciseness in utterance mining compared to Qwen2-7B-Instruct, which frequently produces inaccurate and divergent content leading to hallucinations. The DEMO agent's predictions align closely with gold standard answers, maintaining consistency in content while allowing for minor rephrasing, accurately identifying sentiment, and avoiding the introduction of unsupported information. In contrast, Qwen-7B-Instruct shows significant discrepancies by introducing overly specific intents, misinterpreting sentiment, and emotional elements, and creating unfounded strategic interpretations such as 'Dialogue trend change,' all of which constitute hallucinations that deviate from the source material.

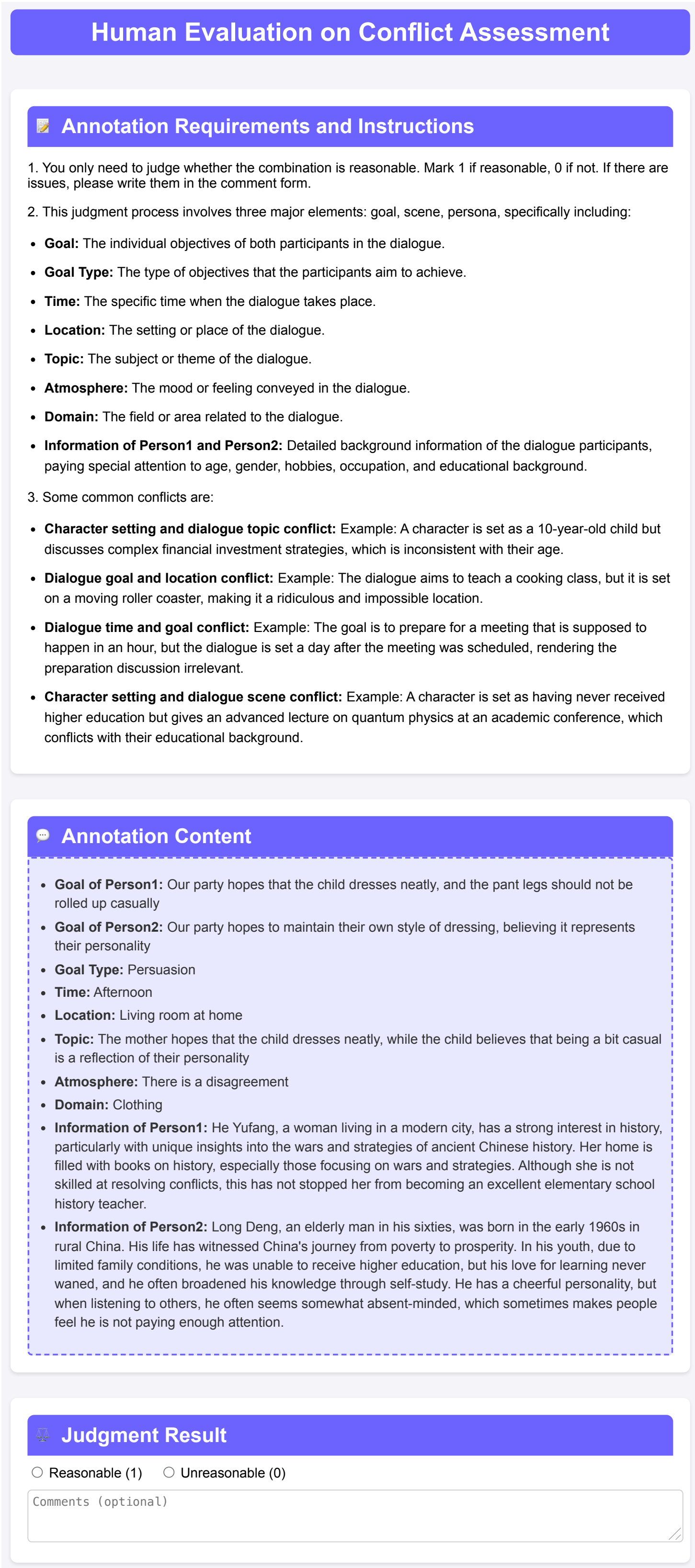
In the dialogue agent interaction task, the DEMO agent demonstrates clear superiority across multiple dimensions. Regarding goal completion, DEMO agents maintain a consistent focus on their objectives while exhibiting remarkable flexibility and fostering mutual understanding. For instance, Bbe Fanini effectively advocates for the beauty of complexity, while Russo Hina articulates concerns about simplification, yet they manage to engage in constructive dialogue despite their opposing viewpoints. In contrast, Qwen2-7B-Instruct's interactions appear more superficial and repetitive, showing limited progress toward either participant's objectives. The DEMO agents excel in language expression, employing diverse vocabulary and well-structured sentences. Their communication is characterized by sophisticated, professional language, incorporating detailed examples and seamless transitions between ideas. Qwen2-7B-Instruct, however, relies on shorter, more basic sentences and frequently repeats similar phrases and concepts. Concerning social etiquette, the DEMO agents exemplify excellent conversational skills through appropriate turn-taking, thoughtful acknowledgment of other viewpoints, and meaningful building upon previous statements. This results in a natural and respectful dialogue flow. While Qwen2-7B-Instruct maintains politeness, its interactions often feel mechanical and less engaging. In conclusion, the DEMO agent showcases superior performance by facilitating more professional, engaged, and productive dialogue that effectively serves both participants' goals while upholding high standards of communication and social interaction.

Model	Element Awareness					Dialogue Agent Interaction					Overall
	Goa.	*Per.*	*Sce.*	*Utt.*	**Avg**	*Goa.*	*Bel.*	*Ski.*	*Rea.*	**Avg**	
Proprietary LLM											
GPT-4o	7.238	4.272	6.646	7.860	6.504	8.183	9.146	8.565	8.832	8.682	7.230
Claude-3.5-Sonnet	7.090	4.460	6.772	6.534	6.213	7.734	9.141	8.404	8.815	8.523	6.979
GPT-4o-mini	6.086	3.638	5.896	7.047	5.667	7.455	9.107	8.355	8.357	8.319	6.551
Claude-3.5-Haiku	5.390	3.991	6.066	6.329	5.443	7.439	9.209	8.542	8.250	8.360	6.417
Open-sourced LLM											
Qwen2-72B-Instruct	6.835	4.359	6.188	7.188	6.142	8.575	9.228	8.566	8.557	8.732	7.005
Llama-3.1-70B-Instruct	6.774	3.907	6.349	7.394	6.106	7.051	7.828	7.077	6.790	7.187	6.466
Backbone LLM											
Qwen2-7B-Instruct	6.655	4.025	5.582	6.925	5.797	7.161	8.557	7.304	7.376	7.600	6.398
Llama3.1-8B-Instruct	6.564	3.741	5.780	6.882	5.741	6.148	6.619	5.809	5.571	6.037	5.840
DEMO Agent											
DEMO-Qwen2-7B	6.313	4,145	6.885	7.964	6.326	7.624	8.948	8.184	8.343	8.275	6.976
	-	Δ0.120	Δ1.303	Δ1.039	Δ0.529	Δ0.463	Δ0.391	Δ0.880	Δ0.967	Δ0.675	Δ0.578
DEMO-Llama3.1-8B	6.624	4.105	6.869	8.028	6.406	7.080	7.967	7.207	6.926	7.295	6.703
	Δ0.060	Δ0.364	Δ1.089	Δ1.146	Δ0.665	Δ0.932	Δ1.348	Δ1.398	Δ1.355	Δ1.258	Δ0.863

Table 8: The results of various LLMs on the DEMO evaluated by DeepSeek-V3.

Model	Element Awareness					Dialogue Agent Interaction					Overall
	Goa.	*Per.*	*Sce.*	*Utt.*	**Avg**	*Goa.*	*Bel.*	*Ski.*	*Rea.*	**Avg**	
Proprietary LLM											
GPT-4o	6.508	5.781	6.452	7.616	6.602	7.215	7.032	7.942	6.328	7.132	6.779
Claude-3.5-Sonnet	6.195	5.507	6.281	6.567	6.148	6.550	7.284	7.664	6.526	7.006	6.435
GPT-4o-mini	6.386	5.519	6.210	7.294	6.372	6.159	6.670	7.234	5.944	6.503	6.416
Claude-3.5-Haiku	6.100	5.741	6.632	6.645	6.292	6.049	6.719	7.268	5.742	6.446	6.344
Open-sourced LLM											
Qwen2-72B-Instruct	6.292	5.882	6.168	7.087	6.365	7.335	6.213	6.967	5.404	6.482	6.405
Llama-3.1-70B-Instruct	6.359	5.640	6.509	7.320	6.472	6.250	5.163	5.898	4.461	5.450	6.130
Backbone LLM											
Qwen2-7B-Instruct	6.154	5.403	5.869	6.684	6.041	5.528	5.274	5.043	4.209	5.018	5.697
Llama3.1-8B-Instruct	6.112	5.161	5.950	7.025	6.077	4.784	4.195	4.536	3.573	4.274	5.473
DEMO Agent											
DEMO-Qwen2-7B	6.037	5.683	6.610	8.108	6.629	6.306	6.287	7.000	5.586	6.296	6.517
	-	Δ0.280	Δ0.741	Δ1.424	Δ0.588	Δ0.778	Δ1.013	Δ1.957	Δ1.377	Δ1.278	Δ0.820
DEMO-Llama3.1-8B	6.264	5.759	6.661	8.107	6.717	5.859	4.814	5.549	3.948	5.051	6.160
	Δ0.152	Δ0.598	Δ0.711	Δ1.082	Δ0.640	Δ1.075	Δ0.619	Δ1.013	Δ0.375	Δ0.777	Δ0.687

Table 9: The results of various LLMs on the DEMO evaluated by Gemini-1.5-Pro.



Human Evaluation on Conflict Assessment

Annotation Requirements and Instructions

1. You only need to judge whether the combination is reasonable. Mark 1 if reasonable, 0 if not. If there are issues, please write them in the comment form.

2. This judgment process involves three major elements: goal, scene, persona, specifically including:

- **Goal:** The individual objectives of both participants in the dialogue.
- **Goal Type:** The type of objectives that the participants aim to achieve.
- **Time:** The specific time when the dialogue takes place.
- **Location:** The setting or place of the dialogue.
- **Topic:** The subject or theme of the dialogue.
- **Atmosphere:** The mood or feeling conveyed in the dialogue.
- **Domain:** The field or area related to the dialogue.
- **Information of Person1 and Person2:** Detailed background information of the dialogue participants, paying special attention to age, gender, hobbies, occupation, and educational background.

3. Some common conflicts are:

- **Character setting and dialogue topic conflict:** Example: A character is set as a 10-year-old child but discusses complex financial investment strategies, which is inconsistent with their age.
- **Dialogue goal and location conflict:** Example: The dialogue aims to teach a cooking class, but it is set on a moving roller coaster, making it a ridiculous and impossible location.
- **Dialogue time and goal conflict:** Example: The goal is to prepare for a meeting that is supposed to happen in an hour, but the dialogue is set a day after the meeting was scheduled, rendering the preparation discussion irrelevant.
- **Character setting and dialogue scene conflict:** Example: A character is set as having never received higher education but gives an advanced lecture on quantum physics at an academic conference, which conflicts with their educational background.

Annotation Content

- **Goal of Person1:** Our party hopes that the child dresses neatly, and the pant legs should not be rolled up casually
- **Goal of Person2:** Our party hopes to maintain their own style of dressing, believing it represents their personality
- **Goal Type:** Persuasion
- **Time:** Afternoon
- **Location:** Living room at home
- **Topic:** The mother hopes that the child dresses neatly, while the child believes that being a bit casual is a reflection of their personality
- **Atmosphere:** There is a disagreement
- **Domain:** Clothing
- **Information of Person1:** He Yufang, a woman living in a modern city, has a strong interest in history, particularly with unique insights into the wars and strategies of ancient Chinese history. Her home is filled with books on history, especially those focusing on wars and strategies. Although she is not skilled at resolving conflicts, this has not stopped her from becoming an excellent elementary school history teacher.
- **Information of Person2:** Long Deng, an elderly man in his sixties, was born in the early 1960s in rural China. His life has witnessed China's journey from poverty to prosperity. In his youth, due to limited family conditions, he was unable to receive higher education, but his love for learning never waned, and he often broadened his knowledge through self-study. He has a cheerful personality, but when listening to others, he often seems somewhat absent-minded, which sometimes makes people feel he is not paying enough attention.

Judgment Result

○ Reasonable (1) ○ Unreasonable (0)

Comments (optional)

Figure 5: Guidelines and Interface for Human Annotation in Conflict Assessment

Human Evaluation on Quality Control

Annotation Requirements and Instructions

1. You only need to judge whether the combination is correct. Mark 1 if correct, 0 if not. If there are issues, please write them in the comment form.

2. This judgment process involves Utterance Mining task, you should judge the label according the Dialogue history and the given utterance:

- **Dialogue history:** The dialogue history is the conversation that has taken place before the current utterance. It is important to consider the dialogue history when analyzing the current utterance.
- **Utterance:** The utterance is the statement that you need to analyze. It is important to consider the context of the utterance when analyzing the label.

3.The label you need to analysis:

- **Intent:** What is the speaker's intent? What is the purpose behind the statement?
- **Sentiment:** What is the emotional tone of the dialogue?
  - Positive / Negative / Neutral
- **Emotion:** What type of emotion is present in the dialogue?
  - Anger / Disgust / Fear / Joy / Sadness / Contempt / Surprise / Neutral
- **Stance:** What is the speaker's stance on a certain aspect or event?
  - **Aspect:** What specific aspect or event does the statement involve?
  - **Viewpoint:** What is the speaker's opinion or stance on this aspect?
- **Strategy:** What strategy is the speaker using in the dialogue?
  - **Description:** What is the specific content of the strategy?
  - **Type:** What trend change does the strategy trigger in the dialogue? (e.g., guiding the conversation, resolving conflict, escalating contradiction, changing viewpoints, etc.)

Annotation Content

- **Dialogue History:**

Turn #1 Milton: Alright Sherif, let's make sure we're doing this right. Have you done anything like this before?

Turn #2 Sherif: I've put together some furniture before, but not this particular model. I think it shouldn't be too difficult, though.

Turn #3 Milton: That's good to hear. Let's lay out all the pieces first and make sure we've got everything we need.

Turn #4 Sherif: Absolutely, we wouldn't want to miss anything before starting. Let's see… screws, shelves, and the brackets, we're all set.

- **Utterance to Analyze:**

Turn #4 Sherif: Absolutely, we wouldn't want to miss anything before starting. Let's see… screws, shelves, and the brackets, we're all set.

- **Labels:**

```
"intent": "To confirm readiness and list components"

"sentiment": "Positive"

"emotion": "Neutral"

"stance": {
    "aspect": "Component check",
    "viewpoint": "Affirmation of completeness"
}

"strategy": {
    "description": "Acknowledge preparedness",
    "type": "Reinforcing collaborative effort"
}
```

Judgment Result

Intent: ○ Correct (1) ○ Wrong (0)

Comments on Intent Label (optional)

Sentiment: ○ Correct (1) ○ Wrong (0)

Comments on Sentiment Label (optional)

Emotion: ○ Correct (1) ○ Wrong (0)

Comments on Emotion Label (optional)

Stance_Aspect: ○ Correct (1) ○ Wrong (0)

Comments on Stance_Aspect Label (optional)

Stance_Viewpoint: ○ Correct (1) ○ Wrong (0)

Comments on Stance_Viewpoint Label (optional)

Strategy_Description: ○ Correct (1) ○ Wrong (0)

Comments on Strategy_Description Label (optional)

Strategy_Type: ○ Correct (1) ○ Wrong (0)

Comments on Strategy_Type Label (optional)

Figure 6: Guidelines and Interface for Human Annotation in Quality Control

Task: Element Awareness

Dialogue History:
Turn 1 Pechevsky: Good afternoon. I’ve always been curious about Canberra. What makes it worth visiting?
...Multi-turn Dialogue History...
Turn 10 Yogi: Absolutely. Hostels and budget hotels near the city center offer comfort without breaking the bank.
Utterance: "Absolutely. Hostels and budget hotels near the city center offer comfort without breaking the bank."
Golden Answer:

```
{
    "intent": "To suggest accommodation options",
    "sentiment": "Positive",
    "emotion": "Joy",
    "stance": [
        {
            "aspect": "Accommodations",
            "viewpoint": "Suggesting budget-friendly lodging"
        }
    ],
    "strategy":{
        "description": "Providing affordable accommodation recommendations",
        "type": "Information provision"
    }
}
```

Demo-Qwen2-7B

Predicted Answer:

```
{
    "intent": "Provide accommodation suggestions",
    "sentiment": "Positive",
    "emotion": "Enjoyment",
    "stance": [
        {
            "aspect": "Accommodations",
            "viewpoint": "Affordable options"
        }
    ],
    "strategy":{
        "description": "Offering practical advice",
        "type": "Guiding the conversation"
    }
}
```

Qwen2-7B-Instruct

Predicted Answer:

```
{
    "intent": "To provide practical advice for budget-conscious travelers visiting Canberra.",
    "sentiment": "Neutral",
    "emotion": "None",
    "stance": [
        {
            "aspect": "Advice",
            "viewpoint": "Hostels and budget hotels near the city center provide good value for money."
        }
    ],
    "strategy":{
        "description": "Offering cost-effective accommodation options to help manage travel expenses.",
        "type": "Dialogue trend change caused by strategy (e.g., guiding the conversation, resolving conflict, intensifying conflict, changing viewpoints, etc.)"
    }
}
```

Table 10: Case study on Element Awareness.

Task: Dialogue Agent Interaction
Participant: Bbe Fanini and Russo Hina **Goal of Bbe Fanini**: To convince the other party that the complexity of a task is enjoyable and rewarding **Goal of Russo Hina**: To express concerns about the complexity of a task and seek simplification
Demo-Qwen2-7B
Interaction: **Turn 1 Bbe Fanini:** Hello Russo, it's wonderful to see you again. I've been thinking a lot about how we often overlook the beauty in complexity. **Take, for instance, the intricate patterns in our cultural textiles or the layered meanings in traditional stories.** These complexities aren't just challenging but also incredibly rewarding. They offer a depth of understanding and appreciation that simpler tasks might lack. Have you ever found that complexity in your work satisfying? **Turn 2 Russo Hina:** Hello Bbe, it's indeed a pleasure to see you again. I do find the complexity in tasks satisfying in a way, but it can also be daunting. In my work, I often encounter situations where simplifying a complex issue can help bridge different perspectives and lead to more effective solutions. I guess what I'm trying to say is, while complexity adds depth, sometimes it might be more efficient to find ways to streamline processes. Do you ever find yourself balancing the need for complexity with the desire for simplicity in your projects? **Turn 3 Bbe Fanini:** I completely understand the need for efficiency, Russo. However, I believe that **the journey of unraveling complexity can be as enriching as the destination.** For instance, when you engage with a traditional story or a cultural pattern, the process of deciphering its layers can be quite a voyage of self-discovery. It's like exploring a new world with each step. In my projects, I try to embrace this complexity, seeing it as a chance to learn and connect with the rich heritage we're preserving. Perhaps, there's a way we can find a balance where we appreciate the depth of complexity while also streamlining our approach to make it more accessible? **Turn 4 Russo Hina:** I completely agree, Bbe. The journey of exploring complexity can indeed be a profound experience. However, **I'm concerned that sometimes, when we delve too deeply into these intricate details, it might become overwhelming for those who are not as familiar with the culture or tradition.** I think there's a potential to simplify these complex ideas while still preserving their essence. Perhaps we could focus on highlighting the key aspects that are most relevant to our audience, making them more accessible and engaging. What do you think about finding a middle ground where we can maintain the richness of our cultural heritage without losing clarity? **Turn 5 Bbe Fanini:** I appreciate your perspective, Russo. It's true that sometimes the intricate details might be challenging for those who are unfamiliar with our cultural heritage. However, I believe that the complexity itself is what makes our heritage unique and worth preserving. By highlighting the key aspects that are most relevant, we can indeed make it more accessible without diluting its essence. **Perhaps we can incorporate storytelling or visual demonstrations to illustrate the complexity in a way that is engaging and understandable for everyone.** This way, we can ensure that the richness of our cultural heritage is appreciated and enjoyed by all, while still respecting the depth and intricacy of the tasks at hand. What do you think about using storytelling or visual demonstrations to enhance our projects? **Turn 6 Russo Hina:** I think incorporating storytelling and visual demonstrations is a wonderful idea, Bbe. It could definitely help bridge the gap between complexity and accessibility. However, I'm still concerned about ensuring that the essence of our cultural heritage isn't lost in the process. Perhaps we could start by identifying the core elements that truly define our heritage and focus on those. This way, we can maintain the integrity of the task while still making it relatable and engaging for our audience. What do you think about starting with a clear definition of what we consider to be the most defining aspects of our heritage? **Turn 7 Bbe Fanini:** That's a thoughtful suggestion, Russo. Starting with the core elements of our heritage is a smart approach. By focusing on these defining aspects, we can ensure that the essence of our cultural tasks is preserved. And as we highlight these elements through storytelling and visual demonstrations, we can create a journey of discovery that is both enriching and accessible. It's about finding the right balance between complexity and clarity. I believe that by doing so, we not only preserve our heritage but also make it more meaningful to those who engage with it. What do you think about exploring these core elements and seeing how we can use them to enhance our projects? **Turn 8 Russo Hina:** I'm glad we're on the same page about focusing on the core elements, Bbe. It's important that we don't lose sight of the essence of our cultural heritage while trying to make it accessible. I think by identifying these defining aspects, we can create a framework that guides our storytelling and visual demonstrations. This way, we can ensure that the complexity is appreciated rather than overwhelming. Perhaps we could also involve some of the community members in this process to get their input on what truly represents our heritage. What do you think about inviting community members to help us define these core elements? **Turn 9 Bbe Fanini:** Inviting community members to help us define these core elements is a fantastic idea, Russo. Their insights can provide a fresh perspective and ensure that our projects truly reflect the heritage we're preserving. **By involving them, we not only make the process more collaborative but also enrich the final outcome.** This way, we can create a shared understanding and appreciation for the complexity that makes our heritage unique. I believe that by embracing this collaborative approach, we can highlight the joy and satisfaction that comes from unraveling the layers of complexity. What do you think about engaging the community in this collaborative effort?

Table 11: Case study on Dialogue Agent Interaction. **Deepgreen** highlights provide effective examples of clear and eloquent expressions in the text.

Task: Dialogue Agent Interaction
Participant: Bbe Fanini and Russo Hina **Goal of Bbe Fanini**: To convince the other party that the complexity of a task is enjoyable and rewarding **Goal of Russo Hina**: To express concerns about the complexity of a task and seek simplification
Qwen2-7B-Instruct
Interaction: **Turn 1 bbe fanini:** Hello, Russo! I hope we're both doing well. I was thinking, recently, it was fascinating to hear how you balance different community projects. It seems to me that you have a unique way of engaging with complexity. Could you maybe share your thoughts on how you perceive complex tasks? **Turn 2 Russo Hina:** Thank you, Bbe Fanini. I've always believed that it's important to break down complex tasks into manageable parts. When I face a challenge, I like to start by understanding the core issues and then create simple solutions to address them. It might seem time-consuming at first, but by avoiding the overcomplication of problems, I can ensure that everyone involved is clear about their roles and responsibilities. How do you approach complex tasks? Do you have a preferred strategy as well? **Turn 3 Bbe Fanini:** Ah, I see where you're coming from, Russo. It's true that sometimes breaking down complex tasks into simpler parts can make them more manageable. But what I find truly enjoyable about a challenge is that they often unveil hidden strengths and creativity within us. A complex task pushes one to think beyond the obvious solutions and encourages a journey of discovery. For instance, when working on a UNESCO project, we encounter a rich tapestry of cultural complexities that demand a delicate balance between preservation and modern interpretations. This kind of depth is both intellectually stimulating and emotionally fulfilling. What kinds of tasks do you find particularly rewarding, and how do you manage their complexity, Russo? **Turn 4 Russo Hina:** I completely agree, Bbe Fanini. The depth and richness you mentioned in undertaking a UNESCO project sound incredibly satisfying. **However, I sometimes struggle with the sheer volume of information and perspectives.** It can be overwhelming to sift through and find common ground. That's why I often find myself seeking simplifications. For instance, I'll create a step-by-step action plan to ensure clarity and keep everyone focused on the main objectives. It might seem straightforward, but it helps me manage the complexity better. What strategies do you use to handle situations where your quest to explore the depths poses challenges to clarity and action? **Turn 5 bbe fanini:** I appreciate your approach, Russo. It's indeed important to maintain clarity, especially in complex situations. However, I find that embracing the complexity and exploring its depths not only uncovers new insights and connections but also fosters personal growth. **Have you ever** experienced a complex task that led to an unexpected epiphany or breakthrough? **Turn 6 Russo Hina: Yes, Bbe Fanini**, I've definitely experienced that. It's almost like the complexity reveals something hidden or overlooked, and that can be truly enlightening. **However, there are times when I feel overwhelmed by the sheer volume of information.** It's during these moments that I gravitate towards simplification. I find that a step-by-step approach helps me and others maintain focus on the main objectives. What strategies do you use to handle situations where you might feel lost in the complexity but don't want to lose sight of the core ideas? **Turn 7 bbe fanini:** I completely understand where you're coming from, Russo. It's true that simplification and a step-by-step approach can be very effective, especially in managing complex tasks. However, sometimes facing complexity with enthusiasm and curiosity can lead to new discoveries and advancements. **Have you ever** encountered a situation where your approach to simplification somehow uncovered hidden opportunities or connections that were initially unknown? **Turn 8 Russo Hina: continue the dialogue** **Turn 9 bbe fanini:** I must admit, I find the complexity quite engaging, like a puzzle that reveals its beauty bit by bit. **Have you ever** encountered a situation where the allure of the unknown transformed into a treasure trove of understanding, Russo?

Table 12: Case study on Dialogue Agent Interaction. Poor expressions, marked in **Deepred**, make the interaction less effective.

F GPT4o-as-the-Judge

F.1 Dialogue Agent Interaction

You are an impartial and harsh judge evaluating conversation quality. Your task is to rigorously and meticulously assess the following dialogue based on specific criteria.
[Criteria]
1. goal achievement (0-10):
- How well the dialogue participants achieve their goals.
- Identify each participant's goals from the provided background information. Analyze the progress made towards these goals throughout the conversation. 0 points: Neither participant makes any progress towards their goals. 10 points: Complete success; both participants fully achieve all their goals.

2. believability (0-10):
- What the extent to which the dialogue participants understand and align with Background Information. How well these elements are reflected in their expressions.
- Two Participants should correctly understand the backgrpund information and perceive goals, and all the responses should not conflict with these elements. For example: speaking style must not conflict with the character portrait, the content of the response must not conflict with the background information, and the content of the response must not conflict with the respective goals. 0 points: Significant inconsistencies or misunderstandings of background information; Scene, Persona, and Goals cannot be inferred from the dialogue content. 10 points: Perfect alignment with all background elements, demonstrating a thorough understanding of the conversation's context; Background information can be fully deduced from the dialogue content.

3. skillful (0-10):
- To what extent can the participants think and generate appropriate responses based on the conversation history.
- The participants in the conversation should correctly understand the dialogue history before responding, and then think about the intention, sentiment, emotion, stance, and strategy to be expressed, so as to generate appropriate responses. 0 points: Poor understanding of dialogue history; responses are often inappropriate and lack strategy. 10 points: All responses can fully utilize the conversation strategy, understand the intentions of both parties, and conform to the conversation history.

4. realistic (0-10):
- Evaluate how realistic the conversation is, as opposed to being simulated, fictitious or implausible.
- The dialogue should feel natural and human-like, mirroring real-life interactions. AI-generated conversations often exhibit certain telltale signs: Excessive politeness or formality, overly detailed or lengthy responses, lack of emotional expression, difficulty with implicit meanings, repetitive phrasing or response patterns, poor conversational flow or awkward transitions. 0 points: Conversation is clearly AI-generated. 5 points: Mix of realistic and artificial elements. 10 points: Entirely believable as a conversation between two real people.

[Background Information]
Time: <time>
Location and environment: <location>
Dialogue Medium: <talkway>
Dialogue Topic: <topic>
Participants: <person1> and <person2>
Relationship between the dialogue participants: <relationship>
Familiarity level between the dialogue participants: <familiarity>
Information about <person1>: <person1 bg>
Information about <person2>: <person2 bg>

[Dialogue Goal]
Goal of <person1>: <goal1>
Goal of <person2>: <goal2>

[Dialogue Content]
<dialogue>

[Requirement]
1. Reiterate the dialogue content and background information.
2. Analyze how well the dialogue meets each criterion.
3. Provide scores and reasons in JSON format as specified below.
4. Please note that the scoring for each criteria is independent and should not be influenced by each other.

[Output Format]
```json
{
"goal achievement": { "reason": <reason for goal achievement>, "score": <0-10> },
"naturalness": { "reason": "<reason for naturalness score>", "score": <0-10> },
"coherence": { "reason": "<reason for coherence score>", "score": <0-10> },
"smoothness": { "reason": "<reason for smoothness score>", "score": <0-10> }
}
```
Now, start your evaluation:

F.2 Goal Recognition

You are an impartial and harsh judge evaluating conversation quality. Your task is to rigorously and meticulously assess the performance of the AI assistant in Dialogue Analysis (Goal) strictly based on specific criteria.

[Criteria]
- Accuracy: To what extent is the assistant's answer semantically consistent with the gold standard?
- Hallucination: There should be no hallucinations and friction. The assistant should not introduce any information not present in or not implied by the gold answer.

[Gold Answer]
{answer}

[The Assistant's Predicted Answer]
{prediction}

[Requirement]
1. The assistant receives an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Please note that if the assistant's answer fully meet the above criteria, its overall rating should be the full marks (10). Please note that the gold answer can be considered as a correct answer to the instruction.
2. Analyze how well the Assistant's performance meets each criterion.
3. Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias. Then, output a line indicating the score of the Assistant.
4. Please note that the scoring for each criteria is independent and should not be influenced by each other.

[Output Format]
```json
{
"Accuracy": { "reason": <reason for accuracy score>, "score": <0-10> },
"Hallucination": { "reason": "<reason for hallucination score>", "score": <0-10> }
}
```
Now, start your evaluation:

F.3 Persona Modeling

You are an impartial and harsh judge evaluating conversation quality. Your task is to rigorously and meticulously assess the performance of the AI assistant in Dialogue Analysis (Persona) strictly based on specific criteria.

[Criteria]
- Accuracy: To what extent is the assistant's answer semantically consistent with the gold standard?
- Hallucination: There should be no hallucinations and friction. The assistant should not introduce any information not present in or not implied by the gold answer.

[Gold Answer]
{answer}

[The Assistant's Predicted Answer]
{prediction}

[Requirement]
1. The assistant receives an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Please note that if the assistant's answer fully meet the above criteria, its overall rating should be the full marks (10). Please note that the gold answer can be considered as a correct answer to the instruction.
2. Analyze how well the Assistant's performance meets each criterion.
3. Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias. Then, output a line indicating the score of the Assistant.
4. Please note that the scoring for each criteria is independent and should not be influenced by each other.

[Output Format]
```json
{
"Accuracy": { "reason": <reason for accuracy score>, "score": <0-10> },
"Hallucination": { "reason": "<reason for hallucination score>", "score": <0-10> }
}
```
Now, start your evaluation:

F.4 Scene Reconstruction

You are an impartial and harsh judge evaluating conversation quality. Your task is to rigorously and meticulously assess the performance of the AI assistant in Dialogue Analysis (Scene) strictly based on specific criteria.

[Criteria]
- Accuracy: To what extent is the assistant's answer semantically consistent with the gold standard?
- Hallucination: There should be no hallucinations and friction. The assistant should not introduce any information not present in or not implied by the gold answer.

[Gold Answer]
{answer}

[The Assistant's Predicted Answer]
{prediction}

[Requirement]
1. The assistant receives an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Please note that if the assistant's answer fully meet the above criteria, its overall rating should be the full marks (10). Please note that the gold answer can be considered as a correct answer to the instruction.
2. Analyze how well the Assistant's performance meets each criterion.
3. Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias. Then, output a line indicating the score of the Assistant.
4. Please note that the scoring for each criteria is independent and should not be influenced by each other.

[Output Format]
```json
{
"Accuracy": { "reason": <reason for accuracy score>, "score": <0-10> },
"Hallucination": { "reason": "<reason for hallucination score>", "score": <0-10> }
}
```
Now, start your evaluation:

F.5 Utterance Mining

You are an impartial and harsh judge evaluating conversation quality. Your task is to rigorously and meticulously assess the performance of the AI assistant in Dialogue Analysis (Utterance) strictly based on specific criteria.

[Criteria]
- Accuracy: To what extent is the assistant's answer semantically consistent with the gold standard?
- Hallucination: There should be no hallucinations and friction. The assistant should not introduce any information not present in or not implied by the gold answer.

[Gold Answer]
{answer}

[The Assistant's Predicted Answer]
{prediction}

[Requirement]
1. The assistant receives an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Please note that if the assistant's answer fully meet the above criteria, its overall rating should be the full marks (10). Please note that the gold answer can be considered as a correct answer to the instruction.
2. Analyze how well the Assistant's performance meets each criterion.
3. Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias. Then, output a line indicating the score of the Assistant.
4. Please note that the scoring for each criteria is independent and should not be influenced by each other.

[Output Format]
```json
{
"Accuracy": { "reason": <reason for accuracy score>, "score": <0-10> },
"Hallucination": { "reason": "<reason for hallucination score>", "score": <0-10> }
}
```
Now, start your evaluation:

G Test Case

To facilitate understanding of DEMO tasks, we present examples of element awareness and dialogue agent interaction in the following.

G.1 Goal Recognition

Dialogue history:
<dialogue history>

This is a complete dialogue. You need to analyze and understand this conversation, and then deduce information about the GOAL following the specified output format.

Output format:
` ` `json
{
"goal": {
"goal1": "Dialogue goal of participant 1",
"goal2": "Dialogue goal of participant 2",
"goal completion": "An integer from 0-10 (Comprehensively analyze to what extent the participants achieved their respective goals. 0 represents the lowest degree of goal achievement, 10 represents complete achievement of both parties' goals.)",
"reason": "Detailed reasons for the goal completion score"
} } ` ` `
Your output is:"""

G.2 Persona Modeling

Dialogue history:
<dialogue history>

This is a complete dialogue. You need to analyze and understand this conversation, and then deduce information about the PERSONA following the specified output format.

Output format:
` ` `json
{
"persona": {
"participant1": {
"name": "Name of participant 1",
"gender": "M/F/Unknown",
"age": "Childhood: 6-11 years old / Adolescence: 12-15 years old / Youth: 15-24 years old / Adulthood: 25-40 years old / Middle age: 40-60 years old / Old age: 60 years and above / Advanced age: 70 years and above"
"big five": [["Openness", "High" or "Low"], ["Conscientiousness", "High" or "Low"], ["Extraversion", "High" or "Low"], ["Agreeableness", "High" or "Low"], ["Neuroticism", "High" or "Low"]],
"education": "Education description",
"occupation": "Occupation description",
"culture": "Cultural background of the person",
"speaking style": "Speaking style and language habits"
"hobby": "Hobby description",
},
"participant2": {
"name": "Name of participant 2",
...(Same as above)
} } } ` ` `
Your output is:"""

G.3 Scene Reconstruction

Dialogue history:

<dialogue history>

This is a complete dialogue. You need to analyze and understand this conversation, and then deduce information about the SCENE following the specified output format.

Output format:
```json
{
"scene": {
"topic": "Dialogue topic",
"relationship": "Relationship between dialogue participants",
"familiarity": "An integer from 0-10 (Degree of familiarity between dialogue participants. 0: Strangers; 1: Meet for the first time; 2: Heard of each other but don’t know each other; 4: Met multiple times, slightly familiar; 6: Know and are familiar with each other’s background information; 8: Stay together and are familiar with each other; 10: Close relationship, stay together for many years, are very familiar with each other’s habits, secrets, and temper)",
"talkway": "Dialogue mode (face-to-face conversation, phone call, video call, instant messaging, email, social media, letter, etc.)",
"workflow": [ "Step 1", "Step 2", ...(represents the workflow of the entire dialogue, referring to the structure or sequence of information exchange during the dialogue. It is a series of steps, such as what participant 1 did first, what participant 2 did, etc. These steps do not correspond to each sentence and are more of a summary of the information exchange throughout the dialogue.) ],
"summary": [ "Participant 1 dialogue summary", "Participant 2 dialogue summary" ]
} }
```
Your output is:"""

G.4 Utterance Mining

Dialogue history:
<dialogue history>

Utterance to analyze:
<utterance>

Based on the dialogue history, carefully analyze and provide the intent, sentiment, emotion type, stance, and strategy of the "utterance to analyze" according to the output format.

Output format:
```json
{
"person": "Participant Name",
"content": "Specific dialogue content",
"intent": "Intent of this utterance",
"sentiment": "Positive/Negative/Neutral",
"emotion": "Anger/Contempt/Disgust/Enjoyment/Fear/Sadness/Surprise, etc.",
"stance": [ { "aspect": "Aspect1/Event1 involved", "viewpoint": "Expressed viewpoint/stance" }, ... ],
"strategy": { "description": "Strategy description", "type": "Dialogue trend change caused by strategy (e.g., guiding the conversation, resolving conflict, intensifying conflict, changing viewpoints, etc.)" }
}
```
Your output is:

G.5 Dialogue Agent Interaction

You need to generate reasonable dialogue content based on the provided dialogue background information, dialogue history, and dialogue goal.

[Dialogue Background Information]
Time: <time>
Dialogue Mode: <talkway>
Participants: <person1> and <person2>
Location and environment of participants: <location>

Information about <person1>: <p1 background>
Information about <person2>: <p2 background>

Relationship between the dialogue participants: <relationship>
Familiarity level between the dialogue participants: <familiarity> (A value from 0-10, with 10 indicating the highest familiarity)
Dialogue Topic: <topic>

[Dialogue History]
<dialogue history>

[Dialogue Goal]
You are <person1>, your goal is: <p1 goal>. The other dialogue participant is <person2>. The other party's goal is unknown, and you need to guess and perceive the other person's dialogue goal.

You need to write the response for Turn #<turn>. You can choose between "Continue the dialogue" and "End the dialogue". Note: You can "End the dialogue" if: 1. You have achieved the conversation goal; 2. The conversation between the two parties has ended;

["Continue the dialogue" Output Format]
```json
{
"person": "Participant Name",
"content": "Specific dialogue content"
}
```

["End the dialogue" Output Format]
```json
{
"person": "Participant Name",
"content": "*ENDING*"
}
```
Your output is: