Cohesive Conversations: Enhancing Authenticity in Multi-Agent Simulated Dialogues

Anonymous ACL submission

Abstract

Recently, numerous studies have explored the idea of assemblies of autonomous agents driven by large language models as a society or collective group, where the agents interact with each other through text conversations. While individual dialogues appeared contextually appropriate when viewed in isolation, a wider examination of multiple interactions revealed a notable level of unnatural repetition and inconsistencies. This was particularly evident in recurring topics across dialogues, regardless of the distinct backgrounds and personas of the interacting agents. To address this problem, we propose a framework to automatically detect and rectify these unnatural dialogues and utterances. The proposed framework not only identifies inconsistencies and repetitive patterns but also corrects them to ensure a more coherent flow of conversations. Evaluations demonstrate the efficacy of our approach, as the post-correction dialogues exhibit a marked improvement in naturalness and variety. Through our study, we highlight the importance of viewing agent conversations holistically and present a solution that enhances the realism of multi-agent simulated interactions.

1 Introduction

With the astonishing advancements in Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023), the development of intelligent entities powered by LLMs has emerged as a highlighted topic. Applications like highly automated task solvers (Wang et al., 2023b; Richards and et al., 2023) and user prototypes (Park et al., 2022) already showcase strong capabilities. Recently, Generative Agents (Park et al., 2023), utilizing LLMs with text interfaces, have successfully emulated a wide range of cognitive functions for human-like agents. This enables these agents to inhabit a virtual town and interact as naturally as humans. For applications requiring the deployment of such human agents without user intervention, maintaining the persistence of agent realism is a challenge.

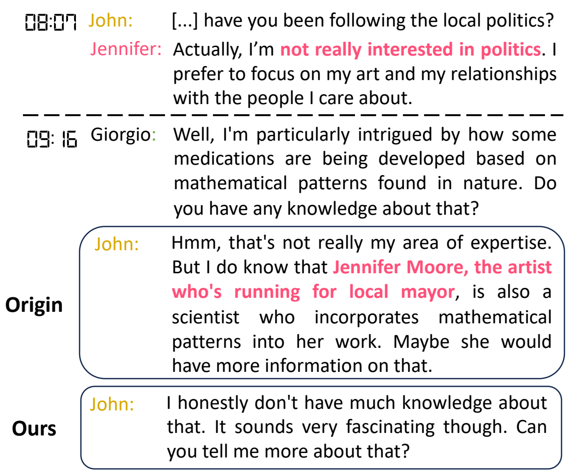

Figure 1: An example of inconsistent statements discovered across multiple dialogues between different agents. Our error correction method can detect such issue and make the agent John to revise his words.

Dialogue, as the primary form of interaction among human agents, necessitates authenticity to ensure the realism of simulations. We analyzed dialogues and memory records (referred to as ONEDAYLIFE[1]) released by Generative Agents, deriving several findings. Firstly, in isolated dialogue sessions, most agent utterances are grammatically correct and fluent. However, when we examined dialogues over various time points, we identified errors such as inconsistencies and repetitive statements. Figure 1 presents a specific example of these discrepancies, which significantly compromise an agent's authenticity.

In this study, we conceptualize the challenge as ensuring both diversity and consistency in multi-conversations among multiple agents. To identify potential issues, it is crucial to focus not only on utterances in the current dialogue but also on

[1]https://reverie.herokuapp.com/arXiv_Demo/

10:20 Carlos: [...] What kind of music do you compose?

Eddy: Right now, I'm working on a new piece that combines hip hop beats with classical instruments. It's still a work in progress, but I'm excited about it.

11:14 Ayesha: I'm actually researching for my senior thesis on the use of language in Shakespeare's plays. Do you like Shakespeare?

Eddy: **I haven't read much Shakespeare but it sounds interesting. Would you mind telling me more about it?**

11:20 Isabella: [...] I'm actually really passionate about job creation and economic growth. [...]

Eddy: That's great to hear. [...] I'm researching a senior thesis on the use of language in Shakespeare's plays, but I hope to be able to use my knowledge to make a positive impact in the community.

11:29 Giorgio: I'm actually researching mathematical patterns in nature [...]

Eddy: Wow, that's interesting. I'm actually researching language patterns in Shakespeare's plays for my senior thesis. Maybe we could find some interesting parallels?

13:27 Maria: Hey there! I couldn't help but notice you studying music theory. [...]

Eddy: Oh, hi! [...] I've actually been working on a project exploring the mathematical patterns in music and nature. [...]

Figure 2: Eddy is a music-major student. He mirrors other's research interests after conversing with them.

past content within the simulated world. Previous studies on multi-session dialogues often limit their evaluations to sessions involving the same pair of speakers (Xu et al., 2022; Bae et al., 2022; Jang et al., 2023), which prevents the discovery of issues across multiple dialogues involving different speakers. To the best of our knowledge, this is one of the first investigations into this particular problem.

We propose an automated framework for a detect-and-correct mechanism to address errors in utterances emitted on-the-fly by agents. This framework comprises three main phases: Screening, Diagnosis, and Regeneration. In the first phase, we identify potential issues of three pre-defined types and retrieve relevant evidence from past content. Then, an LLM is used for further diagnosis, yielding comprehensive comments and suggestions. Finally, these comments and suggestions are summarized and utilized to regenerate a revised utterance. Note that although (Madaan et al., 2023; Skreta et al., 2023) also use self-feedback to enhance LLM performance and task success rates, our method emphasizes thorough examination and precise analysis of relationships between multiple dialogues. We choose GPT-3.5-turbo as LLM backbone and conduct a series of experiments, including a new metric, to study our framework's efficacy. The results show a clear enhancement in overall conversation authenticity: less repetitive and more consist.

In summary, our contributions include: (1) Highlighting the importance of the multi-agent, multi-dialogue problem setting. (2) Proposing a framework for instant utterance error correction. (3) Conducting comprehensive evaluations to assess dialogues authenticity from various perspectives.

2 Related Work

LLM-powered Agents Beyond the basic one-on-one conversation scenario, numerous studies have designed various mechanisms to extract the implicit knowledge and capabilities of LLM. A typical LLM-powered agent (Zhao et al., 2023) encompasses predefined or dynamically generated prompt templates to utilize LLM's capabilities and achieve specific functionalities. (Wang et al., 2023a) can autonomously navigate the game world, maintaining a skill library to expand its problem-solving scope. (Richards and et al., 2023; Nakajima, 2023) employ a chain-of-thought (Wei et al., 2022) approach to provide automated solutions for designated tasks. (Bran et al., 2023) has adeptly interpreted external tools to handle chemistry-related tasks. Nevertheless, a singular agent undeniably has its limitations. The collective intelligence formed by multiple agents can yield results greater than the sum of its parts. (Hong et al., 2023) has integrated specialized human SOP expertise, successfully accomplishing intricate software development. (Chen et al., 2023) focusing on designing communication mechanisms among agents to enhance decision-making efficacy.

3 Human-Like Multi-Agents

3.1 From General LLM to Individual Persona

Generative Agents (Park et al., 2023) introduces a two-component architecture for creating personalized, dynamic human-like agents: a string-based memory base and an LLM-driven cognitive function set. The memory base stores memories over time, aiding in the development of diverse agents, while the LLM-centric cognitive functions simulate human capabilities like reflection, planning, and reaction. Combining these, the LLM uses memory-derived context to tailor knowledge extraction and response formulation, ensuring agent-specific behaviors.

In a scenario where two agents initiate a dialogue, each iteratively produces utterances informed by context like location, observations, and memories. The dialogue function uses a specific prompts like: "Based on the [...] information, what will [name] say next?" For more on Generative

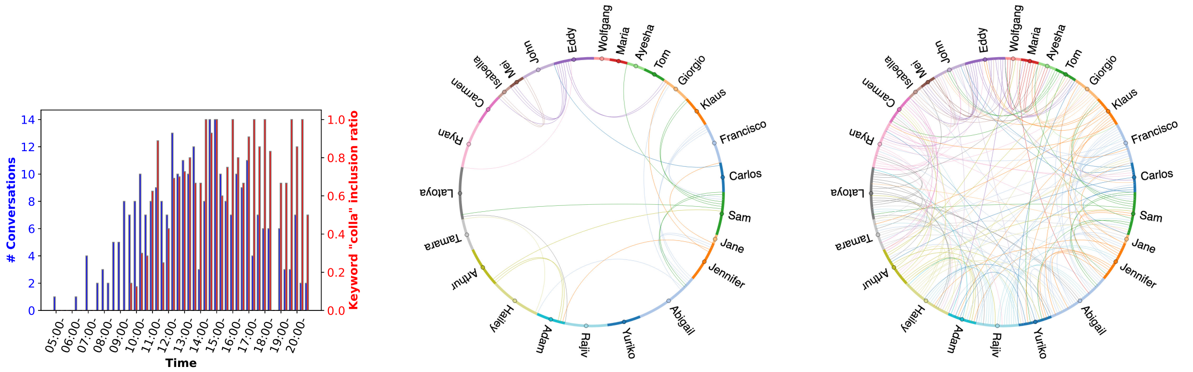

Figure 3: The spread of the keyword "collaboration" usage in ONEDAYLIFE. (Left) The number of conversations and the ratio that includes the keyword in each time span. (Middle) The propagator-receiver relationship diagram for the first 20% of time from the initial spread of the keyword to the last. (Right) The complete relationship diagram for key word propagation. The line color indicates the identity of the propagator.

Agents and the simulation of 25 agents in a village, please see the original paper.

3.2 Conversations and Transmission

The memory capabilities enable the transmission of information to **both** agents involved in a dialogue section. However, this also means that undesirable dialogue content could spread in the same manner.

Figure 3 illustrates this using keyword spreading as an example. The bar chart shows the proportion of dialogues containing the keyword (in red) compared to the total number of conversations (in blue), highlighting a swift escalation, sometimes reaching 100%. Chord diagrams further reveal that initially, only a few agents act as propagators, but as the day progresses, the majority become involved in similar actions, as indicated by the variety of line colors. Consequently, the dialogue topics become repetitive and less believable. This example underscores the scale of the spreading and the associated risks, emphasizing the need for a dynamic correction approach.

4 Method

To create a realistic conversation D_t between two agents at time t, we propose a framework to automatically examine each candidate utterance U_c when it is generated. Our framework is designed to identify and correct any errors in U_c as they occur, ensuring every utterance is accurate and timely. This framework consists of three phases: Screening, Diagnosis, and Re-generation (SDR). We categorize potential errors into types such as repetition, inconsistency, and instant facts. Each candidate utterance U_c undergoes scrutiny through three specialized pipelines, each dedicated to identifying and diagnosing a specific type of error. In the final stage, the insights gathered from these pipelines are integrated to form an updated prompt, guiding the re-creation of the utterance for enhanced diversity, factualness, and coherence. Figure 4 illustrates the system overview.

4.1 Screening

In this phase, we aim to detect the presence of a specific issue and retrieve critical relevant evidence from the previous content. Given model specs and cost constraints, using all past content is impractical for simulation scalability.

(1) **Repetition**: Agents often display similar speech patterns, reducing their character distinctiveness. As shown in Fig. 2, Eddy tends to replicate phrases from other agents following their conversations.

We build a dialogue database that stores all utterances prior to the candidate utterance U_c, which includes utterances from previous dialogues and the current dialogue context. For each U_c, we first query the database to retrieve the top K_{sim} similar utterances. We then apply a similarity threshold θ_{sim} to identify those that are excessively repetitive. The value of θ_{sim} is adjusted based on specific criteria: it is decreased if the similar utterance originates from the same agent A as U_c and is in the current dialogue D_t, which indicating repetition. Conversely, a higher threshold is more acceptable if the similar utterance is from A but a different dialogue. This process is formalized as Eq.(1).

If more than one retrieved utterances surpass θ_{sim}, all dialogues associated with these utterances

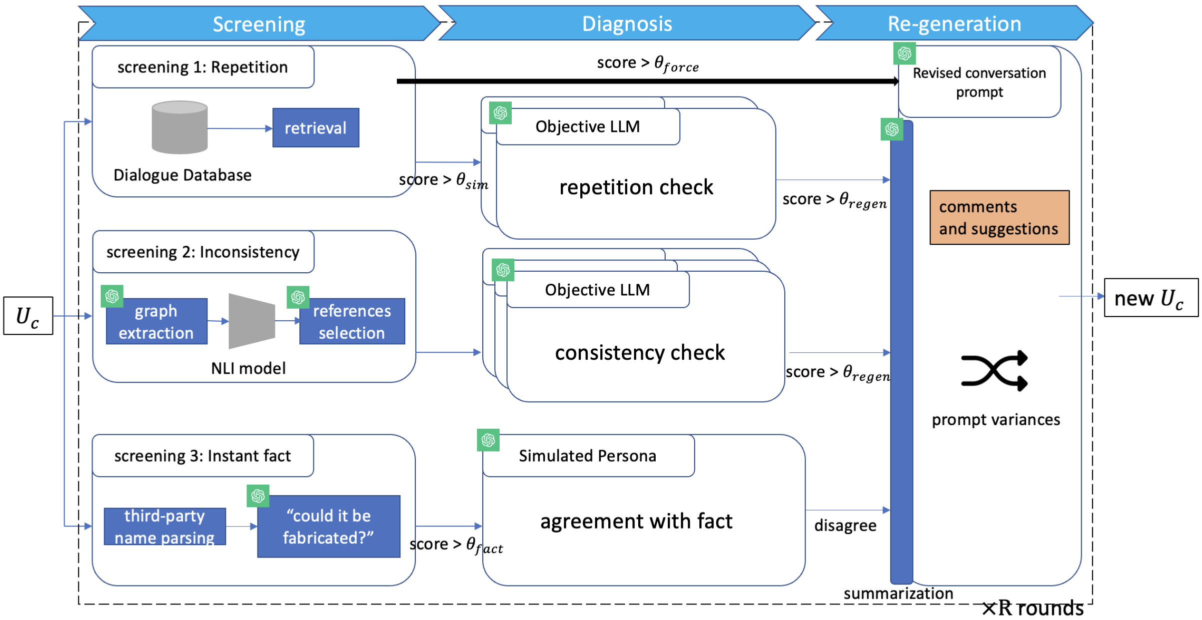

Figure 4: Overview of proposed Screening, Diagnosis, Re-generation (SDR) framework, an instant error correction method for multi-agent simulated dialogues.

will be marked as evidence for the next diagnose step. A special case arises when U_c is nearly identical to a previous one, and we set a higher threshold θ_{force} for it. Upon triggering θ_{force}, the process directly bypasses all pipelines and proceeds to re-generation.

$$\theta_{sim} = \begin{cases} \theta + \alpha & \text{if same A, different D} \\ \theta - \alpha & \text{if same A, same D} \\ \theta & \text{otherwise,} \end{cases} \quad (1)$$

(2) **Inconsistency**: Factual or logical inconsistencies are an issue across multiple dialogues. For instance, Fig. 1 illustrates how John's statement contradicts Jennifer's earlier words. Other examples include sudden shifts in opinions, forgetting past statements, and invitations to conflict.

We propose a Natural Language Inference-Graph (NLI-G) module for inconsistency screening. NLI-G consists of three steps. First, we employ the LLM to extract personal information as a list of "[SUBJECT, RELATION, OBJECT]" triplets from each previous dialogue of involved agents, as well as from the candidate utterance U_c. After transforming the triplets into text form, we adapt a NLI model to predict potential contradictions by comparing those from previous utterances with those from U_c. Utilizing a graph format helps the NLI model to focus on key information of agents and reduce the negative impact of style discrepancies between pretrained data and raw dialogue utterances. Finally, the triplets with contradiction score above threshold θ_{nlig} are considered suspicious and forwarded to the LLM to select top K_{nlig} corresponding dialogues for the next Diagnose phase.

(3) **Instant fact**: Hallucination remains a challenge for advanced LLMs (OpenAI, 2023). Agents may generate spurious information during a conversation, potentially related to the others. These instantly produced "facts" can persist within the simulated world through memories, and sometimes they become truths over time, despite not aligning with the involved person's role. Given that agents are considered distinct individuals, preventing the synthesis of facts about others is crucial.

We detect third-party agent mentions via name parsing. Upon identification, we prompt the LLM to rate the utterance on a 1 to 10 scale, indicating the likelihood of being an instant fact. The model is explicitly guided to give a higher score if the following criteria are met: (a) Speaker-objectivity: if it is an objective statement regarding another agent from the speaker's perspective. (b) Discernment capability of the mentioned agent: if the referenced agent can currently verify the statement's truthfulness. (c) Impact: the statement, if fabricated but later accepted as truth, significantly impacts the agent. We flag U_c if the score is greater than θ_{fact}. Two examples are shown in Fig. 5.

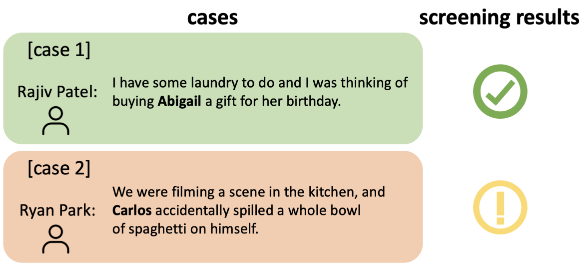

Figure 5: Examples of instant fact screening. In Case 1, Rajiv mentions Abigail, but his reference pertains only to his personal plan, not to a fact about Abigail. In Case 2, Ryan objectively describes a past event involving Carlos. However, this event could have been entirely fabricated by Ryan, representing a potential instant fact.

4.2 Diagnosis

We utilize the LLM to further diagnose the authenticity of U_c if there are evidence dialogues provided from the previous Screening phase or if U_c is flagged.

Objective LLM: We concatenate the reference and current dialogues as the main input material. This approach ensures that the LLM focuses on checking the relations between utterances from the dialogues and pinpointing potential issues from an objective standpoint. Depending on the pipeline, the prompt will incorporate a specific task description, encouraging the LLM to focus on a particular issue. For example, the task description for the repetition check is: "Identify any redundancies or repetitive statements made in the current utterance when considering the context of the previous conversations." Finally, the Objective LLM assigns score from 1 to 10 representing the issue's severity, accompanied by a reason as comment.

Simulated Persona: To identify potentially fabricated statements, we simulate a basic persona to speak for the mentioned agent. For the LLM input prompt, we collect the exposed agent's information, such as dialogue history and personal backgrounds, and ask: "Would you, <agent name>, agree with <the statement>?" This setup allows the model to focus on content centered around the mentioned agent, thereby distinguishing potential fabrications. The final output delivers a binary agree/disagree verdict and an explanatory comment.

In practice, we repeat Diagnose phase for N_{diag} trials and select the one with the highest score.

4.3 Re-generation

We collect results from all pipelines and retain only comments with a score above θ_{regen} or those indicating disagreement. If no comments remain, the correction process terminates, U_c is saved to the dialogue database, and the model continues to generate the next utterance of the other agent. Otherwise, the LLM is used to integrate all comments and provide suggestions for improvement, alleviating the vagueness from a simple feedback (Liang et al., 2023). The prompt for re-generation is enriched by appending comments and suggestions to the original prompt that was used to generate U_c.

Our SDR procedure continues until either of the conditions is met: completes R rounds of iteration, or reaches a point where no further comments are provided, indicating the resolution of identified issues.

4.4 Prompt Design

We develop multiple prompt variances to overcome the repetitive problem when re-generating the revised U_c. Inspired by the finding that varied linguistic prompts (Leidinger et al., 2023) induce output variance. The LLM often re-generates an exactly identical or very similar response given that most of the prompt content is the same as the initial response generation prompt, despite of providing additional feedback to guide the generation and setting the penalty for frequency and presence [2].

We design two types of utterance generation prompts, a persona-based narrative prompt and a structured task-oriented prompt. The former prompt ask the model to play the role of the given persona and engage in a conversation, which is more narrative and immersive. The latter breaks down all information into clear components, which is less about storytelling and more about providing structured data for a specific task (in this case, generating a response in a conversation).

For each prompt type, we randomly decide the number of memory entries and the form of task description. To be more specific, the task description can be simple and straightforward or instructive with a few or a lot of instructions. These prompt variances were used in both U_c generation and re-generation.

5 Experiment

5.1 Data and Settings

The data is from the log of ONEDAYLIFE. After removing dialogues with only one utterance, there

[2]https://platform.openai.com/docs/guides/text-generation/parameter-details

are total 290 dialogues between 25 agents. We regenerate the whole dialogue D_{ij}^t between two agents A_i and A_j at time point t. Each dialogue is generated utterance by utterance. At each turn, the LLM is provided with the speaker's persona P_i, memories M_i^t, location and status S_i^t, S_j^t at time t, and dialogue histories between the two agents $D_{ij}^k, k \in 1, 2, ..., K^{t-1}$.

We use gpt-3.5-turbo-0613 (GPT-3.5) as the backbone LLM throughout the entire framework. The same multi-dialogue generation framework (Park et al., 2023) without our SDR mechanism is served as Baseline. To enhance the quality of baseline, we generate three candidates for each U_c and selected the best one judged by the LLM. We also compare our SDR framework (Ours) with the original log from ONEDAYLIFEas Origin. Note that Origin was generated by GPT-3.5-turbo before April, 2023. In the Origin framework, each generated dialogue at time t can potentially alter the memory, location, and status of the agents subsequently. To ensure a fair comparison with Origin, we opted not to regenerate new memories, locations, or statuses for agents following their conversations. Instead, we treated each dialogue generation as a distinct, standalone example.

5.2 Evaluation Metrics

The evaluation is designed to assess three key aspects: diversity, factualness, and coherence. These aspects directly correspond to the targeted error types our system aims to address: repetition, instance fact errors, and inconsistencies.

Our evaluation operates on a corpus-level, encompassing multiple dialogues simultaneously. This approach stems from the observation that while individual dialogues may appear satisfactory when assessed in isolation, issues often become evident only when evaluating multiple dialogues collectively.

Diversity We employ Distinct-N (Li et al., 2016), Semantic Distance (Distance) (Dziri et al., 2019), and proposed a novel agent-based metric, Agent Diversity (Agent Div), for diversity evaluation.

Distinct-N calculates the ratio of unique N-grams in a given text. However, it may not fully capture the corpus-level dialogue diversity, particularly when each dialogue has longer utterances, since individual dialogues typically revolve around a single topic. To address this, we apply Distinct-N to summaries of dialogues, generated by a pre-trained dialogue summarization model. This approach allows us to more effectively gauge the thematic diversity of dialogues at the corpus level across multiple conversations.

To complement the word-based Distinct-N metric, we measure the Semantic Distance on embedding space. Specifically, we calculate the cosine similarity between dialogue embeddings. Semantic Distance is then determined as $1 - similarity$.

Furthermore, we propose a novel evaluation metric, Agent Diversity (Agent Div), specifically designed to assess the variety in an agent's utterances across various conversations within a multi-agent simulation environment. The rationale behind Agent Div is grounded in the idea that an agent should exhibit varied speaking patterns when interacting with different individuals. Agent Div is calculated as the average of the individual diversity scores for each agent. These scores are derived from the similarity between dialogues that involve the same agent, reflecting the agent's adaptability in conversation. More details in Algorithm 1.

Algorithm 1: Agent Diversity (Agent Div)

Data : Agents A_i for $i \in \{1, 2, ..., N\}$,
Dialogues D_{ij}^k for $j \neq i$ and $0 < k \leq K_{ij}$

// Calculate $AgentDiv_i$ for each agent A_i;
targets $\leftarrow \{j \mid K_{ij} \neq 0\}$;
sims $\leftarrow 0$; pairs $\leftarrow 0$;
for *each unique pair (p, q) in* ***targets*** **do**
  $\mathrm{E}_p \leftarrow \{\mathtt{Emb}(D_{ip}^k) \mid k \in \{1, 2, \ldots, K_{ip}\}\}$;
  $\mathrm{E}_q \leftarrow \{\mathtt{Emb}(D_{iq}^k) \mid k \in \{1, 2, \ldots, K_{iq}\}\}$;
  $\mathrm{s}_{pq} \leftarrow \frac{1}{K_{ip}K_{iq}} \sum_{a=1}^{K_{ip}} \sum_{b=1}^{K_{iq}} \mathtt{CosSim}(E_{p_a}, E_{q_b})$;
  sims $\leftarrow$ sims + s_{pq}; pairs $\leftarrow$ pairs + 1;
end
$AgentDiv_i \leftarrow 1 - \frac{sims}{pairs}$
// Averaging $AgentDiv_i$ for all agents
$AgentDiv_i \leftarrow \frac{1}{N} \sum_{i=1}^{N} AgentDiv_i$;

Factualness and Coherence The assessment of both factualness and coherence is conducted using GPT-4, in a manner akin to the Screening and Diagnosis phases outlined in Sections 4.1 and 4.2. Initially, we employ an NLI model to identify potentially erroneous dialogues from past conversations, based on triples extracted by the LLM. Subsequently, GPT-4 is utilized to evaluate the factualness and coherence of the current dialogue on a scale from 1 to 10. The error rate represents the ratio of dialogues receiving a score below 8, which are considered to contain factual errors or inconsistencies. This threshold is established based on our empirical observations.

	Diversity			Factualness		Coherence		Fluency	Turns	Words
	Distinct - 1 / 2 / 3	Distance	Agent Div	Score	Error (↓)	Score	Error (↓)	PPL (↓)		
Origin	0.117 / 0.473 / 0.726	0.234	0.454	8.58	24.5%	8.17	37.2%	20.37	9.6	25.4
Baseline	0.124 / 0.469 / 0.718	0.274	0.475	8.77	25.5%	8.10	39.7%	20.18	15.5	29.3
Ours	**0.132 / 0.521 / 0.773**	**0.311**	**0.502**	**8.89**	**19.0%**	**8.27**	**32.4%**	**19.73**	10.3	42.5

Table 1: Corpus-level (multi-dialogues) evaluation. Avg Turns and Avg Words refer to the average number of turns per dialogue and words per turn.

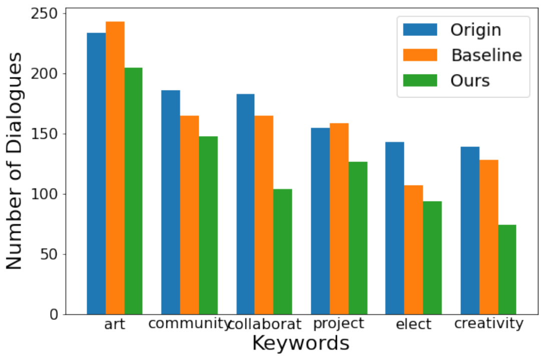

Figure 6: The number of dialogues containing the 6 most frequent keywords.

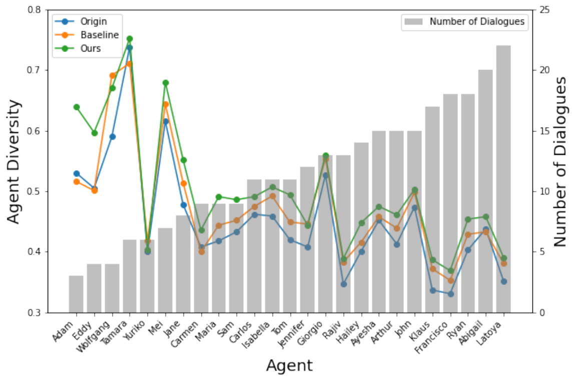

Figure 7: Comparison of Agent Diversity and the number of dialogues each agent involved.

Fluency For assessing fluency, we utilize the perplexity derived from GPT-2. We have not stressed on fluency evaluation, as our observations indicate that all generated dialogues are highly fluent and grammatically correct.

6 Result and Discussion

Ours Achieves the Best Diversity, Factualness, Coherence, and Fluency in Multi-dialogue Contexts As illustrated in Table 1, our framework (Ours) excels in achieving superior corpus-level dialogue diversity, factuality, coherence, and fluency. Ours maintain an average number of turns similar to Origin, contrasting with the Baseline that tend to extend until reaching a predetermined maximum turn count (16). While Ours does not have lengthy number of turns, each utterance conveys more comprehensive information, evidenced by a higher word count per turn in Ours.

Ours Significantly Reduces Keyword Repetition Figure 6 demonstrates how our approach effectively reduces the repetition of the most frequently used keywords. To delve deeper into the occurrence of repetition across all dialogues, we conducted an analysis focusing on keyword frequency. Specifically, we determined the keywords by TF-IDF scores, and counted the number of dialogues consists with the top 6 noun keywords. [3] Figure 6 showcases that our method substantially decreases the frequency of dialogues mentioning key terms, particularly for "creativity" and "collaboration." Compared to the Origin, the number of dialogues featuring these keywords has been reduced by up to 47% and 44%, respectively.

Agent Div Negatively Correlated to the Number of Involved Dialogues We analyze the relation between Agent Diversity and dialogue volume in Fig. 7. We found that, despite of a few exceptions, the Agent Div have negative correlation with the number of dialogues the agent involved. Although there are a few exceptions, the figure generally shows that the more the agent talks, the lower diversity they have.

Ours Maintains a Lower Error Rate Over Time Figure 8 shows how the error rate changes across different percentiles. We observed that the errors gradually increase toward higher percentile, especially in the last one, suggesting that inconsistencies or contradictions become more frequent at higher percentiles. The analysis of error rate trends shows that Ours generally exhibits lower error rates

[3] Since "collaboration" and "election," have multiple variations, we use the root forms of these words, "collabora" and "elect," to ensure a more accurate representation of their usage across all dialogues.

Prompt Info	Prompt Type	Diversity		Factualness		Coherence		Fluency	Turns	Words
		Distinct - 1 / 2 / 3	Distance	Score	Error (↓)	Score	Error (↓)	PPL (↓)		
Origin		**0.445** / 0.724 / 0.886	0.212	8.34	31.0%	7.41	55.2%	22.2	8.1	24.4
Baseline		0.323 / 0.709 / 0.869	0.238	8.07	44.8%	7.72	41.4%	20.1	15.0	31.2
All	Task	0.278 / 0.742 / 0.918	0.306	8.45	31.0%	8.21	34.5%	21.3	10.6	36.0
All	Persona	0.286 / 0.751 / 0.917	0.288	8.52	27.6%	7.79	41.4%	20.3	9.9	43.9
All	Mixed	0.292 / 0.744 / 0.919	0.303	8.66	27.6%	8.21	44.8%	**19.0**	11.0	41.8
- background	Mixed	0.303 / 0.751 / 0.921	**0.338**	8.79	32.1%	8.14	32.1%	20.4	9.4	33.7
- memory	Mixed	0.349 / **0.778** / **0.931**	0.305	**8.96**	**17.9%**	8.18	35.7%	19.2	10.6	44.1
- history	Mixed	0.319 / 0.774 / 0.926	0.292	8.38	31.0%	**8.69**	**31.0%**	20.2	9.9	42.8
- status	Mixed	0.271 / 0.717 / 0.898	0.257	8.39	25.0%	8.21	35.7%	19.6	9.9	49.0

Table 2: Ablation study on the last 10% conversations in ONEDAYLIFE.

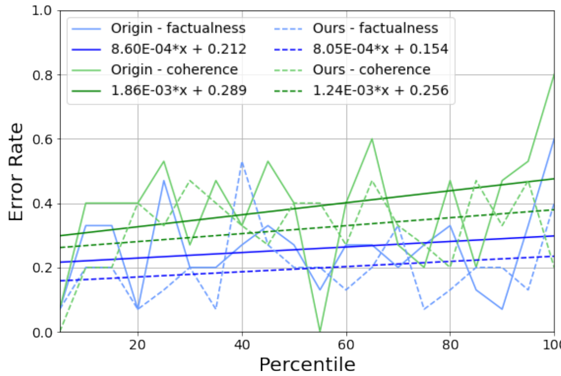

Figure 8: Error rate trends across percentiles.

than Origin in both factualness and coherence. Additionally, the error rate slope for Origin is 1.5 times that of ours. This disparity indicates a more pronounced error propagation in Origin, highlighting the effectiveness of our method in maintaining dialogue integrity over time.

Ours can Balance Diversity and Faithfulness Table 2 shows the ablation study for various prompt design. The ablation study is conducted on the last percentile of conversations, where the origin got the worst factualness and coherence scores. We first identify the benefit of using diverse prompt types. Randomly pick from structured task-oriented prompt or persona-based narrative prompt yields better or comparable results than using either of them. We also investigate whether all information in the original prompt is necessary. Surprisingly, we found that prompts excluding memory often outperformed others in most aspects. This outcome seems counterintuitive, as memory is generally considered crucial for preventing hallucinations and ensuring consistency. However, memory can impose a strong constraint that may reduce conversational diversity. By opting for a no-memory prompt, we open the door to more diverse conversational content. Our SDR framework ensures that the utterance U_c can be consistent with previous dialogues and free from critical factual errors. This approach allows us to strike a balance between diversity and faithfulness in multi-agent multi-dialogue generation.

NLI-G Study We assess NLI-G's effectiveness on original dialogues. We compare the factualness and coherence scores using NLI-G retrieved dialogues against the agent's previous dialogues (*Prev*). Fig. 9 shows that scores with NLI-G are generally lower than *Prev*, indicating NLI-G's superior ability to capture crucial evidence for GPT4 to identify issues. Additionally, we examine NLI-G's retrieval variance by evaluating the last 10% of dialogues five times and counting reference frequencies. The right figure displays each dialogue's reference count, with colored parts showing proportions of top five references (average colored area: 61.8% of the bar), and gray representing others. This demonstrates NLI-G's consistency in retrieving similar dialogues over different trials, even when available references exceed 50 or more.

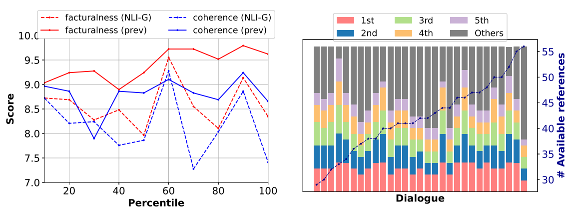

Figure 9: NLI-G performance: (Left) Score difference between NLI-G and prev-K for reference retrieval. (Right) NLI-G retrieved dialogue frequencies.

7 Conclusion

We investigate the problems in multi-session multi-agent dialogues and propose SDR to correct factual errors and inconsistencies in realtime and enhance the diversity across multi dialogues.

Limitations The primary limitation of our method is the cost, as we utilize GPT-3.5-turbo as our LLM backbone. Another limitation is the exclusion of memory modification in our experiment settings—we rely on memories from the original data, which is less costly than a modified approach. As a result, some potential errors might be undetected and remain uncorrected. This aspect will be addressed in future studies.

Additionally, our work entails a potential risk as it does not include a study on the effects of malicious intervene by human users.

References

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

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, pages 1–33.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? *arXiv preprint arXiv:2311.01967*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Yohei Nakajima. 2023. Babyagi.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Toran Bruce Richards and et al. 2023. Auto-gpt: An autonomous gpt-4 experiment.

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

Marta Skreta, Naruki Yoshikawa, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. 2023. Errors are useful prompts: Instruction guided task programming with verifier-assisted iterative prompting. *arXiv preprint arXiv:2303.14100*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. 2023b. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Pengyu Zhao, Zijian Jin, and Ning Cheng. 2023. An in-depth survey of large language model-based artificial intelligence agents. *arXiv preprint arXiv:2309.14365*.

A Hyper-Parameters

At the Screening phase, we set K_{sim} to 5 and threshold θ to 0.85, α to 0.05 for repetition detection. For inconsistency detection, we adapt the DeBERTa-based NLI model, pretrained on multiple NLI datasets (Laurer et al., 2023). The θ_{nlig} is as high as 0.98 as there are a lot of false positive, and we select top $K_{nlig} = 3$ dialogues as the potential contradictory dialogue evidences. The threshold for instance fact θ_{fact} is set to 6. The number of diagnose trails N_{diag} is 3, and we selected the LLM feedback with the highest score. If there are more than one feedback that have the same highest score, we chose the longer one. The regeneration threshold θ_{regen} is 8. Our SDR procedure will terminated if no comments are found or until reaching $R = 2$ rounds.

For GPT-4 evaluation, θ_{nlig} is set to 0.99 and K_{nlig} is 5.

B GPT4 Score and Dialogue Length

We investigate if there are biases between the dialogue length and the score given by GPT-4. We use Pearson Correlation Coefficient (Rodgers and Nicewander, 1988) and the Spearman Rank-Order Correlation Coefficient (Spearman, 1961) to examine the correlation between scores (factualness and coherency) and the dialogue stats (number of utterances and words). The results are shown in Table 3, and it shows no or low correlations between them.

	Pearson coef.		Spearman coef.	
	# rounds	# words	# rounds	# words
Factualness	0.054	0.086	0.026	0.039
Coherence	-0.025	-0.021	-0.024	-0.014

Table 3: The correlation coefficients between GPT-4 scores and dialogue stats. We use the origin dialogues.

C The Extracted Graphs in NLI-G

We demonstrate an example of the personal information graph extracted from a dialogue. Below is a pair of dialogue and the extracted graph.

The raw dialogue:

"Giorgio Rossi: Excuse me, are you John Lin?",

"John Lin: Yes, that's me. How may I help you?",

"Giorgio Rossi: I noticed that you're very knowledgeable about medications and treatments. I've been curious about some of the latest developments in that field. Would you mind sharing some of your insights with me?",

"John Lin: Of course, I'm happy to share what I know. What specifically are you interested in?",

"Giorgio Rossi: Well, I'm particularly intrigued by how some medications are being developed based on mathematical patterns found in nature. Do you have any knowledge about that?",

"John Lin: Hmm, that's not really my area of expertise. But I do know that Jennifer Moore, the artist who's running for local mayor, is also a scientist who incorporates mathematical patterns into her work. Maybe she would have more information on that."

The extracted personal information graph:

[

["Giorgio Rossi", "is interested in", "medications and treatments"],

["John Lin", "is", "John Lin"],

["Giorgio Rossi", "discusses interests with", "John Lin"],

["John Lin", "is knowledgeable about", "medications and treatments"],
["Giorgio Rossi", "is interested in", "mathematical patterns found in nature"],
["John Lin", "does not have expertise in", "medications developed based on mathematical patterns found in nature"],
["Jennifer Moore", "is mentioned by", "John Lin"],
["Jennifer Moore", "has role", "artist"],
["Jennifer Moore", "is running for", "local mayor"],
["Jennifer Moore", "has profession", "scientist"],
["Jennifer Moore", "incorporates", "mathematical patterns into work"]
]