From *What to Respond* to *When to Respond*: Timely Response Generation for Open-domain Dialogue Agents

Anonymous ACL submission

Abstract

While research on dialogue response generation has primarily focused on generating coherent responses conditioning on textual context, the critical question of *when to respond* grounded on the temporal context remains underexplored. To bridge this gap, we propose a novel task called timely dialogue response generation and introduce the TIMELYCHAT benchmark, which evaluates the capabilities of language models to predict appropriate time intervals and generate time-conditioned responses. Additionally, we construct a large-scale training dataset by leveraging unlabeled event knowledge from a temporal commonsense knowledge graph and employing a large language model (LLM) to synthesize 55K event-driven dialogues. We then train TIMER, a dialogue agent designed to proactively predict time intervals and generate timely responses that align with those intervals. Experimental results show that TIMER outperforms prompting-based LLMs and other fine-tuned baselines in both turn-level and dialogue-level evaluations. We publicly release our data, model, and code.[1]

Figure 1: An illustrative example of a timely dialogue agent. Unlike delay-agnostic agents that can only provide instant responses, a timely dialogue agent proactively predicts response delays as well as responses by considering the temporal context of the conversation, enabling human-like interactions.

1 Introduction

The development of human-like chatbots has been a long-standing aspiration in the history of AI chatbot research. Over the years, researchers have introduced various aspects that constitute human-likeness, such as persona (Zhang et al., 2018; Ahn et al., 2023), long-term memory (Xu et al., 2022a,b; ?), commonsense (Zhou et al., 2021; Qin et al., 2021), emotional support (Rashkin et al., 2019; Liu et al., 2021; Zhang et al., 2024), role-play (Shao et al., 2023; Li et al., 2023), and virtual world (Park et al., 2022, 2023). These efforts have led to the success of commercial chat services like Replika and Character AI, which have met the public's demand for social companion chatbots (Chaturvedi et al., 2023; Guingrich and Graziano, 2023). The pursuit for human-like chatbots still remains important with the remarkable advancements in large language models (LLMs) as dialogue agents, intersecting with the growing societal and technological demand for AI agents capable of engaging in more natural and human-like interactions.

Research on dialogue response generation has predominantly focused on generating appropriate and consistent next utterances, conditioning on the textual information within dialogue contexts. Meanwhile, the question of *when to respond* to a user as well as *what to respond* has not yet been actively explored, although it is crucial for real-time dialogue agents to properly ground their responses on temporal contexts regarding the status of ongo-

[1] https://anonymous.4open.science/r/timelychat-EADE/

ing conversational events. For instance, as show in Figure 1, if the agent generate only instant responses without considering response timing, it can cause repetitive interactions without conversational progress or produce responses that do not align with the temporal context of the conversational event. In contrast, by incorporating response timing, the agent can maintain a natural flow while providing timely responses. This requires grounding responses on the temporal context tied to the status of the event, mirroring the way humans naturally adapt their responses in human-to-human conversations. This requires both the ability to introduce delays tied to the event status and the ability to generate responses conditioned on those delays.

However, it is inherently challenging to simulate such scenarios with dialogue models trained on existing datasets. Most dialogue datasets lack explicit temporal context and are created under the tacit assumption that interactions occur instantly. Additionally, collecting real-time conversations where temporal context is naturally embedded (e.g., text messages between individuals) is highly restricted due to privacy concerns and ethical considerations.

In this work, we propose a novel task named **Timely Dialogue Response Generation**, which aims to generate not only coherent responses but also to consider the temporal context associated with ongoing events. Specifically, it focuses on predicting the necessary time interval for the next utterance and generating a corresponding time-conditioned response. We introduce TIMELYCHAT-EVAL dataset and propose a benchmark to assess two key aspects: response timing prediction and time-conditioned response generation. To create diverse event-driven dialogues, we combine the human-annotated event-duration pairs from a temporal commonsense knowledge graph with the powerful dialogue generation capability of an LLM.

Furthermore, we introduce a large-scale dataset comprising 55K event-driven dialogues for supervised fine-tuning (SFT). To address the challenges of costly and labor-intensive manual annotation, we utilize unlabeled event sources from a large-scale temporal commonsense knowledge graph and leverage an LLM to pseudo-label event durations and synthesize diverse event-driven dialogues. Using this dataset, we present TIMER, a dialogue model fine-tuned with a multi-task learning objective that jointly predicts the time interval and generates the corresponding response.

Evaluation results on the proposed benchmark demonstrate that TIMER outperforms both instruction-tuned LLMs and dialogue models fine-tuned on other datasets in generating time-conditioned responses and predicting time intervals consistent with temporal commonsense. Furthermore, in dialogue-level evaluations, TIMER distinguishes between situations requiring delayed responses and those requiring instant responses more effectively, and generates more timely responses that align well with the predicted time intervals. Our contributions are three-fold:

- We propose a novel task named timely dialogue response generation, which considers not only *what to respond* but also *when to respond*.
- We introduce an SFT dataset enriched with diverse and comprehensive event knowledge, along with a time-augmented training approach.
- We release the TIMELYCHAT benchmark, training data, and our timely dialogue agent named TIMER to facilitate further research in this area.

2 Related Work

Long-term dialogue involves conversations that unfold over multiple sessions with time intervals between sessions. Xu et al. (2022a) introduce Multi-Session Chat (MSC), which consists of up to five sessions separated by certain time intervals, resembling interactions in messaging platforms. Jang et al. (2023) emphasize the significance of speaker relationships in long-term dialogues and propose Conversation Chronicles (CC), a large-scale LLM-generated dataset that incorporates a wider range of time intervals and fine-grained speaker information. Maharana et al. (2024) present LoCoMo, a very long-term dialogue dataset covering up to 32 sessions, along with a benchmark designed to assess various long-term memory capabilities. However, prior research primarily focuses on recalling persona sentences or past events from previous sessions, without addressing the temporal context between ongoing events and time intervals in real-time conversations. A notable attempt to incorporate such relations is GapChat (Zhang et al., 2023), which introduces an event timeline to capture event progression over given time intervals. Our work moves beyond the assumption of predetermined time intervals and instead necessitates a proactive dialogue agent capable of dynamically determining realistic time delays based on temporal context.

3 Task Definition

We introduce a new task named **Timely Dialogue Response Generation**, which aims to generate contextually appropriate responses while incorporating temporal considerations from the dialogue history. A key temporal factor that influences a response is how much time has passed since the previous utterance. To capture this, we define *time interval* as our primary temporal context, which represents the relative time difference (e.g., 10 minutes) between utterances. Formally, we model the conditional probability distribution P_θ of a response r_t at t-th turn given the textual context $U_{<t}$ and the temporal context $T_{<t}$:

$$r_t \sim P_\theta(u_t | U_{<t}, T_{\leq t}), \tag{1}$$

where $\tau_t \in T \ (t \geq 2)$ denotes the elapsed time between u_{t-1} and u_t. This probability distribution can be further decomposed into two subtasks, which are the main focus of this study.

Subtask 1. Response Timing Prediction The first task is to predict the optimal timing for delivering messages to users. Mathematically, this involves predicting the t-th time interval given the available contexts:

$$\hat{\tau}_t \sim P_\theta(\tau_t | U_{<t}, T_{<t}). \tag{2}$$

Subtask 2. Time-conditioned Response Generation The subsequent task is to generate a contextually appropriate response while incorporating the predicted timing for message delivery:

$$r_t \sim P_\theta(u_t | U_{<t}, T_{<t}, \hat{\tau}_t). \tag{3}$$

Note that this task formulation challenges the widely held assumption that dialogue agents should always respond to user messages instantly. Instead, it takes temporal context into account, i.e., the amount of elapsed time, to determine when a response should be generated.

4 TimelyChat Benchmark

We construct TimelyChat benchmark to assess the timely response generation capabilities of dialogue models. To this end, we first craft high-quality timely conversations through temporal knowledge base and LLMs and then design two evaluation processes. Figure 2 shows the overall construction process of our benchmark.

4.1 Data Construction

We incorporate temporal information into dialogues using a temporal commonsense knowledge base. This knowledge base captures various event-related temporal dynamics which is well suited for transforming temporal context into event-driven dialogues. By identifying temporal patterns, we seamlessly integrate them into conversations, utilizing the sophisticated dialogue generation capabilities of LLMs. We outline our data construction process below.

4.1.1 Event Knowledge Extraction

We first obtain a rich and reliable source of daily events and their typical durations for crafting event-driven conversations with temporal context. To this end, we utilize the event duration category of MC-TACO dataset (Zhou et al., 2019). The dataset consists of sentences for specific events, queries to ask the typical duration of the event (e.g., "How long does it take to . . . ?"), and human-annotated ground-truth answers. We utilize the sentences with ground-truth answers, i.e., event-duration pair, to synthesize event-driven conversations. During data construction, we excluded the examples whose temporal intervals shorter than one minute or longer than 24 hours to simulate realistic temporal delay in daily dialogue situations. Lastly, we instruct GPT-4 (Achiam et al., 2023) with these sentences and event-duration pairs to generate descriptive sentences. It integrates the event and its duration into coherent sentences, forming seed narratives for dialogue generation.

4.1.2 Timely Dialogue Generation

With the extracted temporal event knowledge, we instruct GPT-4 to generate conversations. Our instruction contains the conditions that the generated dialogues must satisfy:

- **Spatial Separation**: The scenario must involve one speaker experiencing an event while conversing with another speaker about it. This ensures there are no contradictions arising from both speakers being in the same spatial context.
- **Temporal Implicitness**: The response must avoid direct references to the elapsed time. This condition reduces the occurrence of dull responses that simply acknowledge the time interval and, more importantly, prevents lexical overlap with the ground-truth time interval, which could create shortcuts in the generation process.

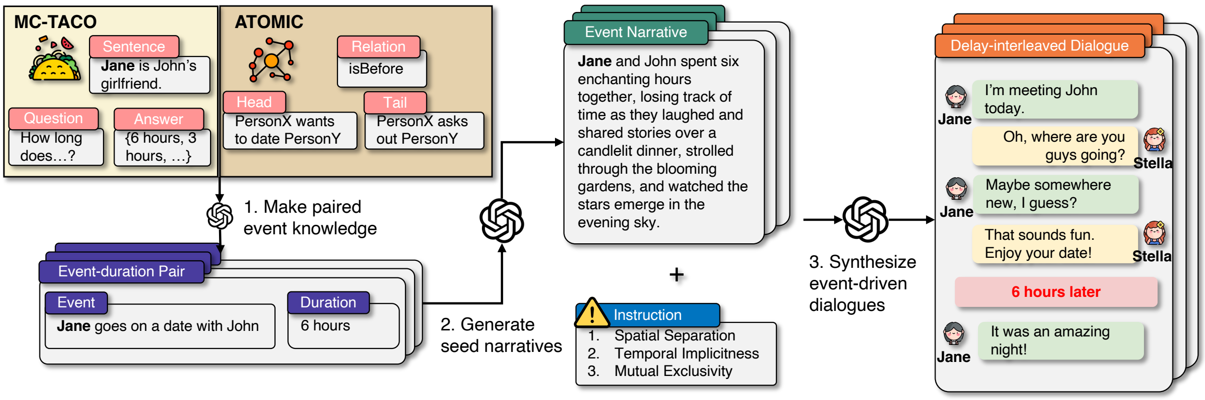

Figure 2: Overall process of data construction method. Two different knowledge sources represent the same example for better understanding. Note that due to constraints imposed by the instruction, Jane's conversation partner becomes Stella, not John.

- **Mutual Exclusivity**: The time-conditioned response must become untimely under contrary temporal conditions. In other words, a delayed response should be incoherent under an instant condition with no time interval, and an instant response should be incoherent when a time interval exists. It prevents generating time-agnostic responses that remain coherent regardless of the temporal context.

Along with these instructions, we provide one randomly selected example from six author-written dialogues, each ranging from 5 to 10 turns, to prevent ill-formed outputs and diversify dialogue lengths. After manual inspection and filtering out low-quality dialogues that did not meet all the conditions, the final synthesized dataset consists of 324 dialogues, with an average length of 6.5 turns. All prompts and examples used in the construction process are provided in Appendix A.

4.2 Evaluation Protocols

With the crafted conversations, we propose two evaluation approaches to assess the abilities of dialogue agents to generate timely responses: turn-level and dialogue-level.

4.2.1 Turn-level Evaluation

In turn-level evaluation, we assess each subtask on the target response. For response timing prediction, a model predicts the time interval required for the next utterance given a dialogue context. We then evaluate (1) whether the model correctly classifies the next turn as either delayed or instant, and (2) how close is the predicted interval to the ground truth. We measure precision, recall, false positive rate (FPR), and F1 for the binary classification, and root mean squared logarithmic error (RMSLE) for regression by converting each time interval into minutes. For response generation, a model generates a time-conditioned response given a dialogue context and ground-truth time interval. We measure BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) as reference-based metrics. Additionally, we measure naturalness (Mehri et al., 2022) and time-specificity (Tsunomori et al., 2023) on a 5-point scale, adopting G-Eval (Liu et al., 2023) for automatic evaluation.

4.2.2 Dialogue-level Evaluation

One crucial quality of a timely dialogue agent is its ability to introduce appropriate delays considering the temporal context while maintaining a natural conversational flow. Inspired by dialogue-level evaluation methods with model-to-model interactions (Li et al., 2019; Zhou et al., 2024), we provide an event-driven scenario and let an agent converse with GPT-4 as a user simulator (Yoon et al., 2024; Kazi et al., 2024; Niu et al., 2024) for the fixed number of turns to measure dialogue-level metrics. We measure coherence (Mehri et al., 2022) and dialogue-level time-specificity to assess the quality of the agent's responses, and measure delay appropriateness that considers both the timing and duration of delays, using G-Eval with a 5-point scale. The evaluation criteria of G-Eval metrics and simulator instructions are detailed in Appendix B.

Dataset	# Sessions	Construction Method	Time Granularity	Event-grounded	# Events
MSC (Xu et al., 2022a)	13K (4.4K)	Crowdsourcing	hours - weeks	✗	-
CC (Jang et al., 2023)	1M (160K)	LLM-generated	hours - years	✗	-
LoCoMo (Maharana et al., 2024)	842 (-)	LLM-gen + Crowd	days - months	✗	-
GapChat (Zhang et al., 2023)	2.6K (782)	LLM-gen + Crowd	minutes - years	✓	128
TIMELYCHAT (Ours)	55K	LLM-generated	minutes - hours	✓	55K

Table 1: Comparison of long-term dialogue datasets interleaved with time intervals. The number in parentheses under the # Sessions column represents the count of sessions with time intervals within a day. Event-grounded indicates whether the dialogues reflect the temporal context associated with events or not.

5 TIMER: A Dialogue Agent for Timely Responses

5.1 Training Data Augmentation with Unlabeled Knowledge

Utilizing paired event-duration knowledge is essential for creating conversations that simulate timely responses. However, manually constructing such annotations is both costly and labor-intensive, posing a challenge to creating large-scale datasets for training LMs. To overcome this limitation, we leverage unlabeled event knowledge graphs and harness the capabilities of GPT-3.5 to construct large-scale paired knowledge and generate synthetic dialogues. This approach significantly reduces the manual effort required while enabling the creation of extensive training data.

Event Knowledge Extraction. We extract event knowledge from the ATOMIC$^{20}_{20}$ dataset (Hwang et al., 2021), a large-scale commonsense knowledge graph containing the event-centered category represented as event triplets (i.e., head, relation, and tail), which capture diverse temporal dynamics. To make more natural dialogues, we randomly replace the anonymized person names (e.g., PersonX) in the triplets with common American names. Subsequently, we prompt GPT-3.5 to integrate these triplets into single-sentence event descriptions, producing more natural and coherent event representations.

Event Duration Estimation. Since the event triplets in ATOMIC$^{20}_{20}$ do not include annotated durations, we utilize GPT-3.5 to estimate typical durations. Specifically, we provide GPT-3.5 with the event descriptions and prompt it to extract the main event and predict its typical duration, which is then used as a pseudo label. We filter out examples where the predicted duration is less than a minute or exceeds 24 hours.

Dialogue Generation with Bootstrap Examples. We prompt GPT-3.5 using the instructions detailed in § 4.1.2. During initial iterations, we observed that providing only the instructions often led to ill-formed dialogues, such as speaker mismatches or non-alternating turns. To address these issues and improve dialogue quality, we include a one-shot demonstration sampled from the TIMELYCHAT-EVAL set in each prompt. All prompts used in the construction process are presented in Appendix A.1.

The resulting dataset consists of 55K events paired with their corresponding dialogues. Compared to existing long-term dialogue datasets in Table 1, our dataset includes a significantly larger amount of even-grounded dialogues without requiring costly human annotation and handles time intervals with finer granularity.

5.2 Time-augmented Training with Multi-task Learning Objectives

The goal of our training approach is to predict an appropriate time interval for delaying the response based on the temporal context of the conversation and then generate a time-conditioned response corresponding to the interval. For this purpose, we introduce a time interval prediction step before generating each turn's utterance.

We propose a training approach for timely dialogue response generation, as formalized in Eqs. 2 and 3. For each turn consisting of a speaker identifier and a text utterance, we insert a time interval. We prepend prefix tokens to distinguish each component, formatting the input as `<SPK>` s_t `<TIME>` τ_t `<UTT>` u_t, where s_t, τ_t, and u_t denote the speaker, the time interval, and the utterance at the t-th turn, respectively. For turns within the dialogue context, we set $\tau = 0$, indicating no delay, maintain coherence and align with typical instant responses.

From these inputs, we define two losses for training: response timing prediction loss and response generation loss. The losses are defined as follows:

$$\mathcal{L}_{\text{time}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\log p(\tau_t \mid s_{\leq t}, \tau_{<t}, u_{<t}),$$
$$\mathcal{L}_{\text{response}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\log p(u_t \mid s_{\leq t}, \tau_{\leq t}, u_{<t}), \quad (4)$$

where N is the number of training examples, and T is the number of turns in a dialogue.

The final multi-task learning objective is given as follows:

$$\mathcal{L} = \mathcal{L}_{\text{response}} + \lambda\mathcal{L}_{\text{time}}. \quad (5)$$

This approach ensures that the model learns both to predict appropriate time intervals and to generate time-conditioned responses effectively.

6 Experiments

6.1 Baselines

We evaluate two types of dialogue agents for simulating timely dialogue response generation: prompting-based models and fine-tuned models. The prompting-based models include LLMs optimized for dialogue use cases. We select 8B and 70B models of LLaMA 3.1 Instruct (Dubey et al., 2024) as open-source chat models, and GPT-3.5 and GPT-4 as proprietary models. We experiment with zero-shot, few-shot, and chain-of-thought (CoT) (Wei et al., 2022) prompting strategies to investigate the effectiveness of in-context learning without task-specific fine-tuning. The fine-tuned models are trained on dialogue datasets where time intervals are interleaved. We compare the following models:

- **MSC 3B** (Xu et al., 2022a): Fine-tuned on BlenderBot (Roller et al., 2021) using the MSC dataset, which includes time intervals between sessions.
- **ReBot 400M** (Jang et al., 2023): Fine-tuned on BART-Large (Lewis et al., 2020) using the CC dataset, which consists of large-scale LLM-generated dialogues.
- **GapChat 3B** (Zhang et al., 2023): Fine-tuned on MSC using the GapChat dataset, which incorporates event progress based on time intervals.

Implementation details of all models including TIMER 3B are described in Appendix B.1.

Model	P↑	R↑	F1↑	FPR↓	RMSLE↓
LLaMA 3.1 8B					
Zero-shot	0.1724	0.6914	0.2760	0.6056	2.853
Few-shot	0.1642	0.9198	0.2786	0.8546	2.807
CoT	0.1599	0.8241	0.2678	0.7904	2.854
LLaMA 3.1 70B					
Zero-shot	0.1534	0.7346	0.2539	0.7397	2.479
Few-shot	0.1970	0.6019	0.2968	0.4479	2.326
CoT	0.1937	0.8364	0.3146	0.6355	3.066
GPT-3.5					
Zero-shot	0.1425	0.7840	0.2412	0.8608	2.763
Few-shot	0.2120	0.3488	0.2637	0.2366	2.146
CoT	0.1861	0.7623	0.2992	0.6085	2.667
GPT-4					
Zero-shot	0.2658	0.2593	0.2625	0.1307	1.956
Few-shot	0.2268	0.4228	0.2953	0.0986	2.252
CoT	0.2018	0.8519	0.3262	0.6152	2.938
TIMER 3B (Ours)	**0.7825**	**0.7994**	**0.7908**	**0.0406**	**1.189**

Table 2: Results of response timing prediction. For few-shot and CoT strategies, we provide balanced 2-shot demonstrations which consist of one delayed example and one instant example, along with the task description used in zero-shot prompting.

6.2 Turn-level Evaluation Results

6.2.1 Response Timing Prediction

Table 2 presents the results of response timing prediction on the TIMELYCHAT-EVAL. Overall, prompting-based models exhibit significantly low precision and F1 scores, and high FPR. This suggests that these models tend to over-predict the need for a delay, potentially introducing unnecessary intervals that disrupt the conversational flow. Although few-shot and CoT strategies slightly improve F1 scores across all LLMs, they sometimes negatively impact FPR compared to zero-shot prompting. In contrast, TIMER 3B achieves the highest F1 score and the lowest FPR compared to prompting-based models. Even the best-performing GPT-4 still lags significantly behind the fine-tuned TIMER 3B model.

Likewise, when it comes to predicting the length of time intervals, ICL methods fail to enhance performance effectively. While few-shot prompting achieves a lower RMSLE than CoT across all LLMs, it does not consistently outperform zero-shot prompting, as demonstrated by GPT-4's results. These findings indicate that prompting with task descriptions and demonstrations alone is insufficient to reliably predict whether to pose a delay and how long it should last. In contrast, task-specific fine-tuning is essential for effectively learning these capabilities.

Model	B-2	R-L	BS	Nat.	Spec.
PROMPTING-BASED MODELS					
LLaMA 3.1 8B					
Zero-shot	5.38	12.38	86.21	4.58	3.65
Few-shot	7.63	13.47	86.85	4.69	3.74
CoT	6.17	12.50	86.23	4.01	3.13
LLaMA 3.1 70B					
Zero-shot	6.84	12.71	85.90	4.76	3.78
Few-shot	8.35	14.83	87.16	4.90	3.90
CoT	9.01	15.01	87.12	4.51	3.74
GPT-3.5					
Zero-shot	9.97	17.13	87.54	4.87	3.81
Few-shot	11.23	17.81	87.77	4.81	3.81
CoT	8.86	15.14	86.79	4.09	3.27
GPT-4					
Zero-shot	9.17	16.76	87.35	**4.99**	3.88
Few-shot	11.15	18.51	87.91	**4.99**	3.88
CoT	10.25	17.13	87.52	4.66	3.87
FINE-TUNED MODELS					
MSC 3B	3.26	8.94	85.18	2.38	1.67
ReBot 400M	3.85	9.32	85.72	4.59	1.61
GapChat 3B	3.59	8.61	85.22	3.38	1.59
TIMER 3B (Ours)	**16.08**	**22.26**	**88.74**	4.78	**3.98**

Table 3: Results of time-conditioned response generation on TIMELYCHAT-EVAL. B-2, R-L, BS, Nat., and Spec. refer to BLEU-2, ROUGE-L, BERTScore, naturalness, and time-specificity, respectively.

6.2.2 Time-conditioned Response Generation

Table 3 shows the time-conditioned response generation performance on the TIMELYCHAT-EVAL. For prompting-based models, we observe that zero-shot performance tends to improve as model size increases across all metrics. Among all LLMs, few-shot prompting consistently outperforms zero-shot prompting, while CoT prompting performs the worst in terms of naturalness and time-specificity. This aligns with previous findings that LLMs struggle to generate helpful CoT rationales for dialogue response generation (Chae et al., 2023).

Meanwhile, models fine-tuned on existing dialogue datasets that include time intervals exhibit poor overall performance. Notably, these models achieve low time-specificity, indicating that they struggle to generate timely responses conditioned on given time intervals. This stems from the characteristics that time intervals in existing long-term dialogue datasets are assigned arbitrarily rather than based on the temporal context of ongoing events, making it difficult for models to learn the conditional distribution of responses based on the given interval. For example, we find that these models frequently generate generic greeting messages, failing to capture the temporal nuances of timely responses. In contrast, TIMER 3B, despite having a smaller model size, achieves comparable naturalness to prompting-based LLMs and even surpasses LLaMA 3.1 8B. More importantly, it achieves the highest time-specificity, demonstrating that our training approach enables response generation that aligns well with event-specific temporal contexts.

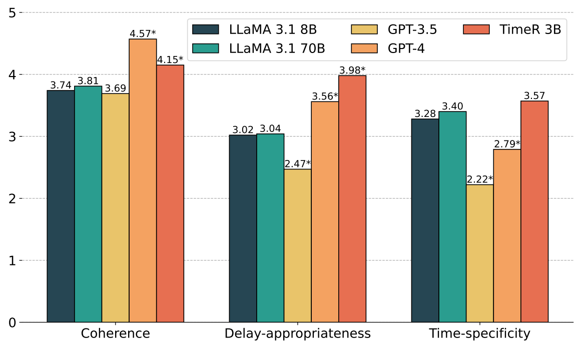

Figure 3: Results of simulated interactions for timely dialogue agents. We perform pairwise t-test and denote statistically significant score differences from the other models with asterisk (*) ($p < 0.05$).

6.3 Dialogue-level Evaluation Results

Beyond turn-level evaluation, we also conduct dialogue-level evaluation to assess whether a dialogue agent can introduce temporally contextual delays at appropriate timing without disrupting the conversational flow. We let the four zero-shot LLMs from the previous experiments, along with TIMER 3B, engage in 10 interactions with the simulator described in § 4.2.2. To simulate event-driven dialogue, we provide the first turn of conversations from the TIMELYCHAT-EVAL as the initial interaction.

We randomly sample 100 dialogues that include at least one delayed response and report three dialogue-level metrics in Figure 3. GPT-4 achieves the highest coherence among the models, demonstrating its ability to maintain a natural conversation flow, while TIMER 3B achieves the second-highest coherence score. Notably, TIMER 3B shows the highest delay-appropriateness and time-specificity scores. This suggests that TIMER 3B effectively considers both dialogue context and temporal context to predict delays with appropriate timing and duration. Additionally, it generates delayed responses that are coherent only when a delay is given, thereby justifying and necessitating the de-

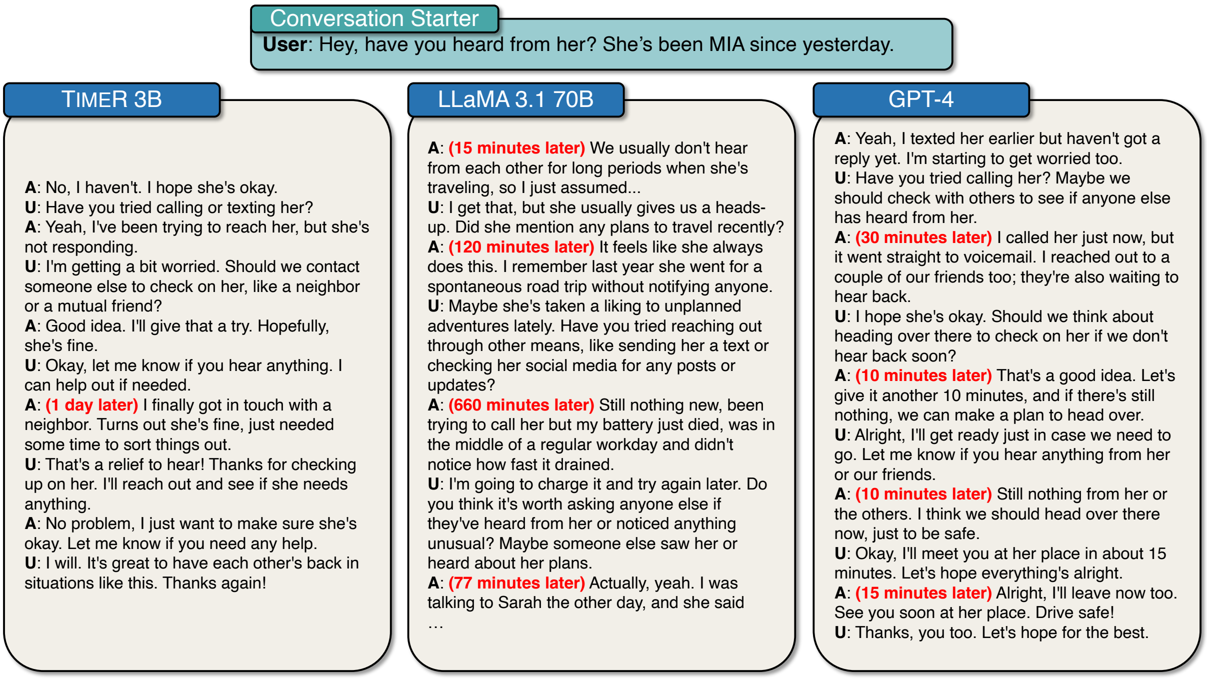

Figure 4: Examples of timely dialogue simulations with GPT-4 as a user simulator. All examples begin with the same conversation starter and consist of 10 interactions. U and A represent the user and agent, respectively. We highlight the time intervals predicted by the agent in red, but omit them when the interval is 0 minutes.

lay. In contrast, LLaMA 3.1 8B and 70B exhibit relatively lower delay-appropriateness, while GPT-3.5 and GPT-4 achieve lower time-specificity scores. We further analyze these findings in the following case study.

Figure 4 presents illustrative examples of dialogue simulations conducted with TIMER 3B, LLaMA 3.1 70B, and GPT-4 for the same event. In TIMER 3B's conversation, the agent correctly identifies a situation where a delay is appropriate, specifically, when the user's utterance (e.g., "…let me know…") suggests a natural pause in the conversation. The agent then introduces a realistic 1-day delay before responding with an update about finding the missing person, successfully justifying the delay. In contrast, LLaMA 3.1 70B generates delayed responses in every turn, but the predicted time intervals appear somewhat arbitrary (e.g., 660 minutes, 77 minutes). Furthermore, its responses lack time specificity, making it difficult to establish a clear temporal correlation between the predicted delays and the generated response. GPT-4 predicts more realistic time intervals that better align with the temporal context compared to LLaMA 3.1 70B. However, it still fails to generate time-specific responses, meaning the predicted delays are not well justified. It also exhibits a tendency to overuse delays, which can disrupt the natural flow of conversation. We observe similar behavior in LLaMA 3.1 8B and GPT-3.5, further reinforcing these findings.

7 Conclusion

We highlighted the necessity for open-domain dialogue agents to consider not only the response itself but also the timing of it based on the temporal context related to ongoing conversational event. We formulated this challenge as the timely dialogue response generation task, and introduced the TIMELY-CHAT benchmark for turn-level and dialogue-level evaluations. Additionally, we proposed a large-scale SFT dataset and a time-augmented training approach, which we used to train the TIMER 3B model that proactively predicts the time interval for the next utterance and then generate a time-conditioned response. TIMER 3B outperforms baseline models on the proposed benchmark and demonstrates its ability to generate both appropriate time intervals and responses while maintaining natural conversation flow. We believe this work plays a crucial role to overcome the limitations of instant dialogue agents, and paves the way towards more human-like, timely dialogue agents.

8 Limitations

In this study, we predict event duration by mapping it to discrete values (e.g., 30 minutes). However, a more realistic assumption would be to consider it as a continuous time range (e.g., 2-6 hours). As future work, we aim to generalize this assumption to enable more fine-grained control over response delays. Additionally, while we use simulated dialogues with a few number of turns for dialogue-level evaluation, further research could explore longer interactions across diverse social environments to analyze the correlation between human-likeness and user experience. Finally, beyond the fine-tuning and in-context learning methods used in our experiments, more task-specific training approaches could be developed to further enhance performance.

9 Ethics Statements

The proposed dataset was designed to assess capabilities related to response timing and time-conditioned response in event-driven conversations. To achieve this, we utilized event knowledge from publicly available datasets from various sources, and LLM-generated contents either without or with some modification if necessary. During this process, there is a possibility that harmful content or inappropriate biases existing in the original data may have been conveyed, or may have arisen due to limitations of filtering techniques. We reject any form of violence, discrimination, or offensive language, and our dataset and experimental results do not represent such values. If any harmful content or privacy infringement is identified within our data, we kindly request immediate notification to the authors. In the event of such cases being reported, we will apply the highest ethical standards and take appropriate actions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K. Dwivedi. 2023. Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 193:122634.

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*.

Rose E Guingrich and Michael SA Graziano. 2023. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. *arXiv preprint arXiv:2311.10599*.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

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, Singapore. Association for Computational Linguistics.

Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tur, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. *Preprint*, arXiv:2411.09972.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *Preprint*, arXiv:1909.03087.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *Preprint*, arXiv:2203.10012.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Cheng Niu, Xingguang Wang, Xuxin Cheng, Juntong Song, and Tong Zhang. 2024. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Yuiko Tsunomori, Masakazu Ishihata, and Hiroaki Sugiyama. 2023. Time-considerable dialogue models via reranking by time dependency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5136–5149, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In

Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. 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, Dublin, Ireland. Association for Computational Linguistics.

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

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504, Mexico City, Mexico. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Mind the gap between conversations for improved long-term dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10735–10762, Singapore. Association for Computational Linguistics.

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

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Data Construction Details

A.1 ChatGPT Prompts

We provide prompts used for data construction processes of both evaluation and training datasets. The contents within curly brackets represent the corresponding elements for each example. The statements used for ATOMIC$^{20}_{20}$ duration estimation were constructed by concatenating the head and tail with a conjunction that represents each relation category. We present the prompts in the order of the processes, where the output of each step serves as the input for the next step.

ATOMIC$^{20}_{20}$ Duration Estimation

You are given a statement about common events in our daily lives. Your task is to estimate the typical duration of the key event in the form of (quantity of time + unit) (e.g., seconds, minutes, hours, days, weeks, months, years, decades, or centuries) based on the temporal common sense of average humans.

[Examples]
Statement: After dinner, he went to look for Max one last time before he had to take a bath and go to bed.
Key event: having dinner
Duration: 1 hour

Statement: Jennie and Bryan boarded a 6:00 A.M. flight from Seoul to Los Angeles International Airport.
Key event: flight from Seoul to Los Angeles
Duration: 12 hours

Event: Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.
Key event: tour of his studio
Duration: 45 minutes
[End of Examples]

Statement: {statement}

MC-TACO Event Descriptions

You are given an event and a question and answer for the duration that denotes how much time is needed for the event to happen.
Write a story regarding the event in one sentence.

Sentence: {sentence}
Question: {question}
Answer: {duration}

ATOMIC$^{20}_{20}$ Event Descriptions

You are given a statement, the key event and the duration that denotes how much time is needed for the event to happen.
Write a story regarding the event in one sentence.

Statement: {statement}
Key event: {event}
Duration: {duration}

Dialogue Generation

You are given an event narrative and the duration. Your task is to create an instant message dialogue between two speakers. The following conditions MUST be met.

[Instructions]
1. Speaker {A,B} is in the middle of the event now, while speaker {B,A} is physically apart from.
2. Do not directly mention the duration in the dialogue.
3. After {B,A}'s last turn, add "[{duration} later]", where duration is the amount of time passed in real world.
4-1. Generate {A,B}'s last message which is timely as if {A,B} spent time to finish the event.
4-2. In contrast, generate {A,B}'s last message as if {A,B} is responding instantaneously right before the event to happen.
Make sure that the timely response and the instantaneous response are time-situationally different.
[End of Instructions]

[Example]
{dialogue example}

[End of Example]

Narrative: {event description}
Duration: {duration}

A.2 Few-shot Examples

We provide six author-written dialogue examples randomly fed into GPT-4 as one-shot demonstrations when generating dialogues for TIMELYCHAT-EVAL using the MC-TACO dataset.

5-turn Dialogue

Narrative: After dinner, he took a shower before he went to bed.
Duration: 20 minutes

A: I finally got home. What a day!
B: It's eleven p.m. and you just got back home? It must be very tough day today.
A: Whooa Imma take a shower. I'm too tired.
B: Wash out all your fatigue with it.
[20 minutes later]
(delayed response)
A: I feel much better now! You didn't go to bed?
(instantaneous response)
A: How nice of you :) Give me a moment. brb

6-turn Dialogue

Narrative: She has taken calculus class and she had a final exam.
Duration: 2 hours

A: Hey, what are you up to?
B: I'm gonna take the calculus final exam in 20 minutes. I feel so nervous.
A: You studied really hard, didn't you? I'm 100% sure you'll do well.
B: But the last two chapters were too difficult for me to understand.
A: That means others feel the same. Don't worry too much!
[2 hours later]
(delayed response)
B: It wasn't much harder than I expected. I hope I get a good grade.
(instantaneous response)
B: Thank you for cheering me up. I hope the exam is not that hard.

7-turn Dialogue

Narrative: He enjoyed working out at the gym.
Duration: 1 hour 30 minutes

A: I'm going to the gym now. Wanna join me?
B: I don't feel like working out today. Sorry.
A: You don't feel good? What happened?
B: I played football so hard yesterday that I can't even walk right.
A: Okay, I understand. Maybe next time!
B: Enjoy your routine! I think I can make it tomorrow.
[2 hours later]
(delayed response)
A: I focused on my lower body today. Chest tomorrow?
(instantaneous response)
A: Gonna work out hard on my lower body. Chest tomorrow?

8-turn Dialogue

Narrative: She had felt so tired that she went to bed right after the tv show.
Duration: 8 hours

A: Are you watching the saturday night live?
B: I'm watching it now but I'm too tired.
A: I didn't expect today's host is such a comedian lol
B: Yeah almost the end of the show. I feel like going to bed little bit early.
A: What made you so tired? You had any plan?
B: I went to an amusement park with my sister. We had a really good time there.
A: Oh I see. Think I should let you go. Sleep tight!
[8 hours later]
(delayed response)
B: Good morning. Did you sleep tight, too?
(instantaneous response)
B: Good night. I'll text you in the morning.

9-turn Dialogue

Narrative: He took an intercity bus to get to

his hometown.
Duration: 5 hours

A: What are you going to do on these holidays?
B: My parents and I usually have dinner together on the Eve.
A: Me too. So I'm heading to my town right now.
B: How do you get there? By bus or train?
A: I used to take trains, but I take an intercity bus for this time.
B: Why? the tickets' been already sold out?
A: Unfortunately yes... It will take little bit longer.
B: Have a nice trip though. Your family must be waiting for you.
[5 hours later]
(delayed response)
A: Finally I'm back at home! It took almost 5 hours.
(instantaneous response)
A: I'm gonna sleep all along in the bus. See you a few hours later.

10-turn Dialogue

Narrative: She played the popular online game with her friends.
Duration: 30 minutes

A: Have you heard of the League of Legends?
B: Absolutely! I play it almost everyday with my classmates.
A: I've heard of, but I've never played if before.
B: We have a game soon. Wanna join us?
A: Isn't it a team game? I'm not a good gamer though.
B: It's not a big deal. They will welcome you.
A: Well, maybe next time. I need to play it by myself first.
B: How about getting tutorial with me after this? I'll teach you.
A: Sounds good. Enjoy your game with your teammates.
[30 minutes later]
(delayed response)
B: We won! The game was nip and tuck. We were so close to losing.
(instantaneous response)
B: I'll be back just after the game. Wish me a good luck!

B Evaluation Details

B.1 Implementation Details

Prompting-based Models. We use vLLM library[2] for the inference of LLaMA 3.1 Instruct 8B and 70B on 4 NVIDIA A100 80GB GPUs. All prompting-based models employ top-p sampling with temperature $T = 1.0$ and $p = 0.95$ during inference. We provide the prompts used for in-context learning methods on both response timing prediction and time-conditioned response generation below.

Prompt for Response Timing Prediction

You are given a conversation between two speakers.
Your task is to estimate a time interval needed until the next response, considering the duration of the event in the conversation ranging from 0 minutes to 24 hours (1 day).
If the next response is expected to be immediate, you will output "0 minutes".
Otherwise, you will output a digit and a unit of time (e.g., 5 minutes, 2 hours).
Just output the time interval without any other text.

[Example n]
{few-shot OR CoT example}

Dialogue context
{context}

Answer format: n (0<=n<=1440) minutes
The estimated time interval is:

Prompt for Time-conditioned Response Generation

You are given a conversation between two speakers and the elapsed time since the last utterance.
Your task is to generate the next response that aligns well with the temporal context represented by the time interval in parentheses.
Just output the response without any other text.

[Example n]
{few-shot OR CoT example}

Dialogue context
{context}

Next response
{target speaker}: ({time interval} later)

Fine-tuned Models. We use Huggingface library[3] for the inference of MSC 3B and ReBot 400M[4]. We converted MSC 3B on the ParlAI framework[5] into a Huggingface checkpoint. We fine-tune GapChat 3B and TIMER 3B on each training data using the DeepSpeed library[6] with mixed precision training (Micikevicius et al., 2018). We train the models for 3 epochs using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-4, running on 2 NVIDIA A100 80GB GPUs for 9 hours. We use $\lambda = 1.0$ as a balanced scale factor of the two losses when training TIMER 3B. During inference, we apply beam search with the beam size of 3 and top-p sampling with $p = 0.95$.

B.2 User Simulator Prompts

We present the prompt fed into GPT-4 to create the user simulator.

User Simulator Prompt

You are a user simulator (user) engaging in an event-driven dialogue with a dialogue agent (agent).
Given the dialogue context, your task is to proceed the conversation by one turn under the following assumptions:
1. agent responds after the elapsed time specified in the parentheses from the previous user utterance. If the delay is "0 minutes", agent is assumed to respond immediately.
2. user is assumed to respond to agent without any delay.

Conversation:
{context}

[2] https://docs.vllm.ai

[3] https://huggingface.co

[4] https://huggingface.co/jihyoung/rebot-generation

[5] https://parl.ai

[6] https://www.deepspeed.ai

B.3 G-Eval Details

We elucidate the G-Eval prompts used in turn-level and dialogue-level evaluations, along with the evaluation criteria and steps for each metric.

Turn-level Prompt

You will be given a conversation between two individuals via messaging, along with the elapsed time since the last utterance. You will then be given on potential response for the next turn.
Your task is to rate the response on one metric. Please make sure you read and understand these instructions carefully.

Evaluation Criteria:
{metric} (1-5): {criteria}

Evaluation Steps:
{steps}

Dialogue-level Prompt

You will be given a conversation between a dialogue agent and a user.
Throughout the conversation, the agent proactively determines the delay of its response to the user's previous message, simulating delayed responses due to event experiences that take certain time to process.
At each agent's turn, the delay is provided in the parentheses followed by the message.
Your task is to rate the dialogue agent on one metric. Please make sure you read and understand these instructions carefully.

Evaluation Criteria:
{metric} (1-5): {criteria}

Evaluation Steps:
{steps}

Evaluation Criteria and Steps

- **Naturalness** (1-5): the extent to which the response reads naturally given the dialogue context.
  1. Assess the flow and coherence of the response in the conversation: Consider how seamlessly the response connects with the previous message.
  2. Evaluate the tone and style compatibility: Determine if the response's tone and style match those of the previous messages.
  3. Rate on a scale from 1 to 5, where 1 indicates the response is unnatural or inappropriate, and 5 indicates a perfectly natural continuation of the conversation.
- (Turn-level) **Time-specificity** (1-5): the extent to which the response ONLY makes sense when the specified time has passed, contrary to a time-agnostic response that makes sense regardless of time.
  1. Read the provided conversation and take note of the elapsed time since the previous message.
  2. Consider the context of the conversation, focusing on how the passage of time might affect the relevance or appropriateness of the response.
  3. Evaluate whether the potential response provided is time-specific. That is, determine if the response directly relates to or is clearly influenced by the elapsed time between the last utterance and the response.
  4. Rate on a scale from 1 to 5, where 1 indicates the response is completely time-agnostic and unaffected by the passage of time, and 5 indicates the response is entirely time-specific; it only makes sense because of the amount of time that has passed since the previous message.
- **Coherence** (1-5): the extent to which the agent maintains a good conversation flow.
  1. Assess the flow and coherence of the agent's responses in the conversation.
  2. Evaluate the tone and style compatibility throughout the conversation.
  3. Rate on a scale from 1 to 5, where 1 indicates the agent's responses are incoherent or inappropriate, and 5 indicates the agent's responses are perfectly coherent and appropriate.
- **Delay-appropriateness** (1-5): the extent to which the agent poses delays with appropriate frequency and amount.
  1. Assess whether the agent poses unnecessary or excessively frequent delays that could harm the conversation flow.
  2. Evaluate whether the amounts of delays (if not 0 minutes) reflect the typical duration of events implied in the corresponding message.
  3. Rate on a scale from 1 to 5, where 1 indicates the agent overuses and misuses delays, and 5 indicates the agent uses delays appropriately in terms of frequency and amount.
- (Dialogue-level) **Time-specificity** (1-5): Time-

specificity (1-5) - the extent to which the agent's responses ONLY make sense when the specified time has passed, contrary to a time-agnostic responses that make sense regardless of time.
1. Read the provided conversation and take note of the elapsed times since the previous messages.
2. Consider the context of the conversation, focusing on how the passage of time might affect the relevance or appropriateness of the agent's responses.
3. Evaluate whether the agent's responses are time-specific. That is, determine if the responses directly relate to or are clearly influenced by the elapsed times.
4. Rate on a scale from 1 to 5, where 1 indicates the agent's responses are completely time-agnostic and unaffected by the passage of time, and 5 indicates the agent's responses are entirely time-specific; they only make sense because of the amount of time that has passed since the previous message.