

Stephanie2: Thinking, Waiting, and Making Decisions Like Humans in Step-by-Step AI Social Chat

Hao Yang^{1,3*}, Hongyuan Lu^{2*}, Ding kang Yang³, Wenliang Yang³, Peng Sun⁴, Wei Zuo³, Xiaochuan Zhang³, Jun Xiao³, Kefan He³, Wai Lam⁵, Yang Liu^{6†}, Xinhua Zeng^{1,3†}

¹Yiwu Research Institute of Fudan University, ²FaceMind Corporation, ³Fudan University, ⁴Duke Kunshan University, ⁵The Chinese University of Hong Kong, ⁶Tongji University
haoyang24@m.fudan.edu.cn, hongyuanlu@outlook.com

Abstract

Instant-messaging human social chat typically progresses through a sequence of short messages. Existing step-by-step AI chatting systems typically split a one-shot generation into multiple messages and send them sequentially, but they lack an active waiting mechanism and exhibit unnatural message pacing. In order to address these issues, we propose **Stephanie2**, a new **step**-wise decision-making dialogue agent. With active waiting and message-pace adaptation, Stephanie2 explicitly decides at each step whether to *send* or *wait*, and models latency as the sum of *thinking time* and *typing time* to achieve more natural pacing. We further introduce a dual-agent dialogue system to generate pseudo dialogue histories for human and automatic evaluations. Experiments show that Stephanie2 clearly outperforms Stephanie1 on metrics such as naturalness and engagement, and achieves a higher pass rate on human evaluation with the role identification test.¹

1 Introduction

Human social chatting on instant messaging platforms typically unfolds in a step-by-step manner (Wu et al., 2025; Spagnolli et al., 2021; Garcia and Baker Jacobs, 1999; Anderson et al., 2010): speakers often express an idea through several short consecutive messages (Baron, 2010; Spagnolli et al., 2021), and may continuously adjust wording, add details, or shift topics based on the interlocutor’s reactions (Beißwenger, 2008; Berglund, 2009; Meredith, 2019). In contrast, most contemporary LLM-based dialogue systems follow a single-step paradigm (Guo et al., 2025; Bravo et al., 2025; Algherairy and Ahmed, 2024), producing one long response for each user input. Although information-dense, the single-step paradigm often

* Equal contribution

† Corresponding authors

¹Stephanie 2 is open-sourced at <https://github.com/h17ao/Stephanie2>



Figure 1: Stephanie2 differs from Stephanie1 by implementing active waiting and message-pace adaptation.

lacks the naturalness, pacing, and emotional coherence of human interaction (Meredith, 2019).

A recent work, *Stephanie* (Yang et al., 2025), referred to as Stephanie1 here, proposed the *step-by-step dialogue* paradigm. It argues that system outputs should be delivered as multiple short, separated yet coherent messages to better match instant-messaging experiences (Baron, 2010; Spagnolli et al., 2021; Nardi et al., 2000). It also described a baseline that mechanically splits a single long response into multiple messages using sentence boundaries and punctuation (Spagnolli et al., 2021; Beißwenger, 2008; Meredith, 2019). Here refers to it as *punctuation-segmented dialogue* (PD). However, in real social chat, message boundaries are largely driven by intent, conversational rhythm, and interaction strategies rather than written punctuation, so PD often appears stiff and un-

natural. To mitigate this issue, as shown in Figure 1, Stephanie1 uses designed prompts to generate multiple messages in one pass, separated by delimiters such as [newline]. These messages are then extracted sequentially by scripts, substantially improving naturalness and engagement. *X-turning* (Wu et al., 2025) further notes that step-by-step systems may introduce a short delay before the first reply, allowing users to send multiple messages before being interrupted (Berglund, 2009; Garcia and Baker Jacobs, 1999; Anderson et al., 2010). It also suggests computing inter-message delays from message length, to approximate human typing speed (Kalman et al., 2006).

Despite these advances, existing step-by-step systems such as Stephanie1 still face two key limitations. First, when users speak continuously, especially in scenarios where the system should act as a listener, the system often interjects too early, interrupting continuous emotional expression and narrative coherence. Although X-TURING (Wu et al., 2025) suggests adding a forced delay before the first system reply, the delay is static and hard to tune. If it is too short, users can still be interrupted. If it is too long, the system feels sluggish. In real conversations, however, appropriate waiting time is inherently context-dependent, and no single fixed value works well across diverse situations (Rendle-Short, 2015; Kalman and Rafaeli, 2011; Hwang et al., 2019; Iftikhar et al., 2023).

Second, in previous work, inter-message timing was simplified to depend only on message length, which cannot capture cases such as short messages with long thinking time and long messages that with little reasoning. Such unrealistic pacing substantially weakens the human-likeness of step-by-step dialogue systems. In human communication, the time between messages reflects not only typing speed but also thinking time (Berglund, 2009; Kalman and Rafaeli, 2011; Kalman et al., 2006). These observations suggest that a more human-like step-by-step system should not only be able to send consecutive messages, but also be able to decide when to speak, when to wait, and how to exhibit a plausible message pacing (Hwang et al., 2019).

Inspired by these observations, we propose a new step-by-step dialogue agent, Stephanie2, introducing *active waiting* and *message pacing adaptation* to better emulate human listening behaviour and response latency in social chat (Kalman and Rafaeli, 2011; Hwang et al., 2019). At each step, Stephanie2 explicitly chooses between sending a

message and waiting, and generates an explicit thinking trace to support more appropriate content and more human-like pacing. Concretely, guided by a thinking prompt, after producing a <think> segment, it outputs either <response> or <wait>, enabling the agent to decide whether to continue speaking or stop and listen. Moreover, inter-message delays are computed from both the thinking length and the response length, yielding more realistic timing (Kalman and Rafaeli, 2011; Kalman et al., 2006). The explicit reasoning in Stephanie2 helps account for the current conversation’s emotional state when generating responses, and the improved message pacing also contributes to a higher pass rate in the role identification test.

In addition, a time-window-based dual-agent dialogue system is introduced to generate high-quality step-by-step dialogue data, thereby supporting large-scale quality evaluation, as well as downstream role identification tests and statistical analyses in real human-agent conversations. Experimental results show that Stephanie2 reconstructs step-by-step dialogue systems in an agentic manner and significantly improves the dialogue experience.

Our contributions are summarised as follows:

- We propose a new **step-by-step** dialogue agent, **Stephanie2**, which has active waiting and message-pace adaptation mechanisms.
- We design a dual-agent dialogue system to generate high-quality step-by-step dialogues across multiple topics and contextual settings.
- Through both human and automatic evaluations, we compare Stephanie2 with existing step-by-step dialogue systems and verify its improvements in naturalness, engagement, and related dimensions, along with a higher pass rate in the role identification test.

2 Related Work

Step-by-step interaction in social IM platforms

Text chatting in mainstream Instant Messaging (IM) applications such as WhatsApp and Facebook Messenger is typically quasi-synchronous: dialogue contributions are often produced incrementally rather than as single, self-contained turns. Prior IM and computer-mediated communication studies show that speakers may distribute one contribution over multiple short messages or extend a turn through subsequent messages (Baron, 2010; Spagnolli et al., 2021). Message boundaries are shaped by temporal and sequential or-

ganization, rather than simply by sentence boundaries or punctuation (Beißwenger, 2008; Meredith, 2019). Turn-taking in text-based IM also differs from spoken conversation, as participants use localized strategies to manage overlaps, interleaving, and disrupted adjacency while maintaining coherence (Berghlund, 2009; Garcia and Baker Jacobs, 1999; Anderson et al., 2010). Moreover, temporal cues such as pauses, response latencies, silence, and typing indicators carry socially interpretable meanings and affect users’ interpretation of the interaction (Kalman and Rafaeli, 2011; Kalman et al., 2006; Hwang et al., 2019; Rendle-Short, 2015; Iftikhar et al., 2023). These findings suggest that IM dialogue systems should model not only response content, but also message segmentation, waiting behavior, and sending rhythm as key interactive decisions (Nardi et al., 2000).

Step-by-Step Chat Systems Most LLM-based dialogue systems are built around the Single-Step Chat Paradigm (Deng et al., 2025; Hashimoto et al., 2025; Zhan et al., 2023), where each user turn is answered with a complete response generated in a single shot. To better approximate social chat experiences, some recent studies have begun to explore the Step-by-Step Dialogue Paradigm (Yang et al., 2025; Wu et al., 2025), which decomposes a single response into multiple shorter messages sent sequentially, while further introducing message-length-based inter-message intervals to simulate human typing and message-sending rhythms. Meanwhile, since the publication of *Stephanie 1*, several well-known applications have also begun adopting multi-step dialogue, such as SillyTavern and Lovemo. SillyTavern is a chatbot application with thousands of stars on GitHub, and its Discord community has grown to more than 44,900 members. Step-by-step dialogue systems have already built up a considerable user base.

3 Stephanie2

3.1 Construction of Stephanie2

To address the limitations of Stephanie1 in *active waiting* and *message pace adaptation*, we design Stephanie2 as a step-wise decision-making agent. Unlike approaches that generate multiple messages in one shot, Stephanie2 takes exactly one action at each step: it either sends a short message or waits and continues listening. For ease of parsing and deployment, we require the model output to include explicit thinking and action tags in one of

the following two formats:

`<think>...<\think> <response>...<\response>`

or

`<think>...<\think> <wait>...<\wait>`

Meanwhile, we carefully design prompts to guide Stephanie2’s reasoning, so as to mimic the human thought process before sending a message. As illustrated in the Fig. 2, given the most recent dialogue history H in memory m and the persona p , the agent first determines who sent the last message and then chooses an action accordingly:

- If the last message is from the user, the agent decides whether to respond immediately or act as a listener. When the user is still adding information or venting emotions, the agent outputs `wait`; otherwise, it outputs `response` and generates the next short reply.
- If the last message is from the agent, it decides whether its expression is complete. If so, it outputs `wait`; otherwise, it outputs `response` and continues with the next message.

The design enables the agent to decide when to speak and when to wait. It reduces interruptions during users’ continuous expression while also preventing the agent from producing overly long monologues. Given the dialogue memory m_t and persona p_t , the model performs action selection between *wait* and *response* formulated as:

$$P(a_t \mid m_t, p_t), \quad (1)$$

where $a_t \in \{w, r\}$ is the action taken by the current speaker, w denotes *wait* and r denotes *response*. In practice, this action decision is implemented through prompt-guided generation.

To better match the temporal structure of real social chat, where latency reflects both *thinking* and *typing*, we model these two components separately and define the display delay for each step as:

$$T_{\text{delay}} = \max \left(0, k_{\text{think}} \cdot N_{\text{think}} + k_{\text{type}} \cdot N_{\text{response}} - T_{\text{system}} \right), \quad (2)$$

where N_{think} and N_{response} denote the number of characters in the think and response segments, respectively, and k_{think} and k_{type} are tunable coefficients. T_{system} denotes the time spent on API calls or local inference. The key advantage is that latency is no longer determined solely by visible message length. It can capture cases such as short messages with long thinking time and long messages with little reasoning, resulting in more natural pacing.

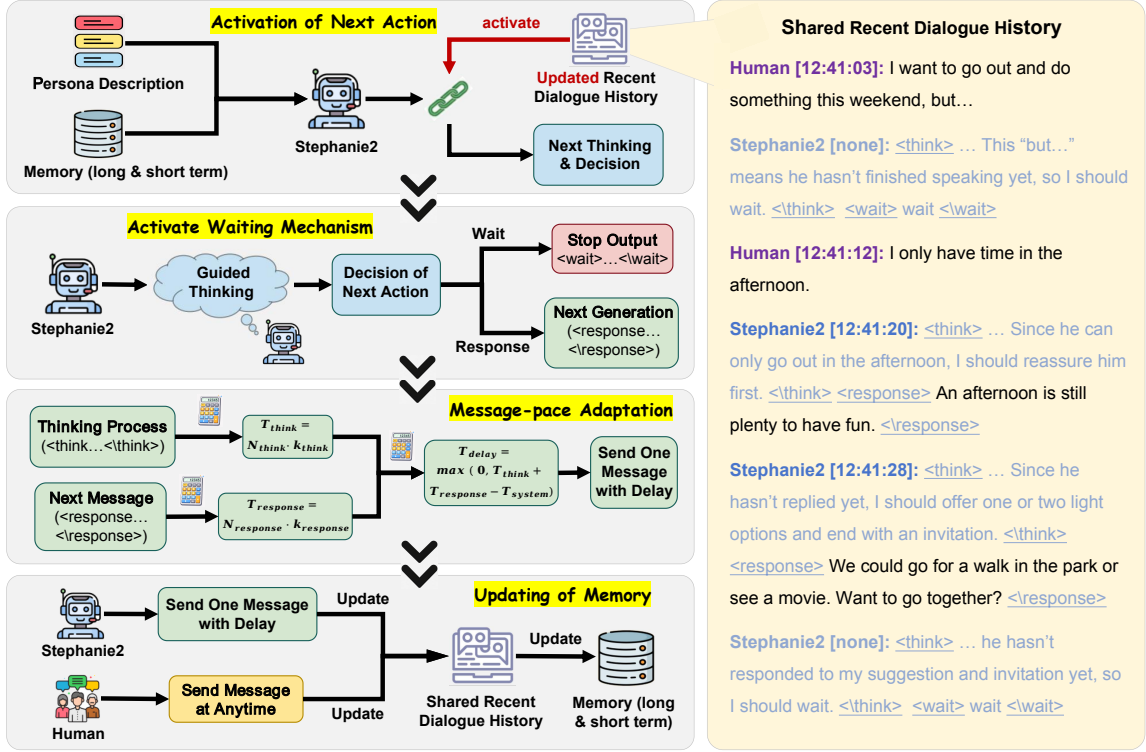


Figure 2: Stephanie2 is a step-wise decision-making agent with active waiting and message-pace adaptation.

We further condition the agent on an explicit persona description. The persona can be specified directly via prompting or learned from the dialogue history through in-context learning, enabling users to quickly instantiate desired chat partners. To control context cost while mitigating performance degradation in very long contexts, we also introduce a mechanism that combines short-term and long-term memory. Specifically, the most recent n raw messages are retained as short-term memory, while earlier history is periodically summarized: for every additional q messages, we update a summary that serves as long-term memory.

3.2 Dual-Agent Dialogue System

To effectively evaluate under diverse topics and contextual settings, we require a large number of generated step-by-step pseudo dialogues. These histories can then be used in downstream tests where humans continue the conversation from a given context, enabling broader and more systematic validation. We therefore build a dual-agent dialogue system in which both sides are instantiated as step-by-step dialogue agents and autonomously interact for multiple turns to generate pseudo dialogue histories. We apply the dual-agent to punctuation-segmented dialogue system (PD), Stephanie1, and Stephanie2.

In our pilot experiments, we find that PD and Stephanie1 often collapse into a strict ping-pong, single-step dialogue pattern when generating multiple consecutive messages, largely because they lack a robust waiting mechanism. While Stephanie2 alleviates the issue via active waiting and explicit thinking, we further introduce a *time-window* mechanism to ensure that all three dual-agent systems can generate more natural, higher-quality dialogue histories. Concretely, when one agent holds the speaking floor, it is assigned a time window whose length is randomly sampled within a reasonable range. Let R_t denote the remaining time in the current speaker’s window at step t . Here W is the sampled window length with bounds $W_{\min}$ and $W_{\max}$. We initialize and update the window as:

$$R_0 = W, \quad W \sim \text{Unif}(W_{\min}, W_{\max}), \quad (3)$$

$$R_{t+1} = \begin{cases} R_t - T_t, & \text{if } a_t = \text{response}, \\ 0, & \text{if } a_t = \text{wait}, \end{cases} \quad (4)$$

where T_t is the display delay of the current step defined in Section 3.1. Within the window, the agent may send consecutive messages, which fundamentally reduces frequent turn-taking and interruptions. When an agent reaches the end of its time window, the window is transferred to the other agent, allowing it to speak. If the agent finishes early or chooses

to wait, the remaining window is reclaimed and a new window is immediately assigned to the other agent, yielding dialogue histories that better resemble real social chat. Beyond data generation, the time-window-based design also suggests a scalable direction for multi-agent group chat by assigning time windows to different agents.

3.3 Data Generation

This subsection describes the pseudo-dialogue generation pipeline based on Persona-Chat (Zhang et al., 2018). Similar to the procedure in Stephanie 1, as shown in Appendix J, we write five step-by-step examples and use in-context learning to convert all instances into step-by-step format, yielding 8,519 samples. We then analyze the distribution and find 86.9% of samples are concentrated in 6–8 turns. We keep the 7,405 dialogues in the range and further filter those with 25–40 total messages, excluding cases with too many or too few consecutive messages on average, yielding 6,459 samples.

Then we adopt hierarchical summarization-based clustering to extract topics. Each dialogue is first summarized into a short topic descriptor. We then cluster descriptors in two stages: (1) We feed every 600 topic descriptors into an LLM to summarize them into 60 subtopics; across 11 batches of the 6,459 samples, yielding 660 subtopics in total. (2) these 660 subtopics are summarized into 60 final topics. An LLM assigns each dialogue to exactly one of the 60 topics, which are then ranked by sample count; Fig. 6 in the appendix B shows the distribution of the top 20 topics.

Each of the 6,459 samples includes a dialogue topic, persona descriptions for two speakers, and a dialogue history. We treat each sample as a seed context and generate new dialogues using the dual-agent systems. Specifically, the PD, Stephanie1, and Stephanie2 dual-agent systems each generate 10 turns of pseudo dialogue per seed.

4 Experiment Setup

4.1 Evaluation Metrics

- **Dialogue Experience Metrics:** Some prior studies have evaluated dialogue experience (Lu et al., 2022; Zhang et al., 2024; Yang et al., 2025). We adopt the same definitions of the seven evaluation dimensions used in Stephanie 1 (Yang et al., 2025): Interesting, Informative, Natural, Coherent, Engaging, On-topic, and On-persona.

- **Distinct-N (Li et al., 2016):** Lexical diversity is measured by the ratio of unique n -grams to total n -grams across generated responses:

$$\text{Distinct-N} = \frac{\text{Total unique } n\text{-grams}}{\text{Total } n\text{-grams}}. \quad (5)$$

- **Words/Message:** Message length is measured by the average number of words:

$$\text{Words/Message} = \frac{1}{n} \sum_{i=1}^n w_i, \quad (6)$$

where w_i is the number of words in the i -th message, n is the total number of messages.

- **ACMC:** Average Consecutive Message Counts (ACMC) measures the average number of messages one sends consecutively.

$$\text{ACMC} = \frac{1}{T} \sum_{t=1}^T c_t, \quad (7)$$

where c_t is the number of consecutive messages in turn t , and T is the number of turns.

- **Pass Rate:** In the Role Identification Test, each evaluator chooses among *Role 1 is the AI*, *Role 2 is the AI*, and *Unclear*. *Correct* means that the evaluator correctly identifies the AI role, whereas *Error* means that the evaluator selects the wrong role as the AI. *Unclear* means that the evaluator cannot confidently distinguish the AI from the human. We count a dialogue as a *pass* when the evaluator either identifies the AI incorrectly or selects *Unclear*:

$$\text{Pass Rate} = \frac{N_{\text{error}} + N_{\text{unclear}}}{N_{\text{total}}}, \quad (8)$$

where N_{error} and N_{unclear} denote the numbers of incorrect and unclear identifications, respectively, and N_{total} is the total number of judgments. A higher Pass Rate indicates that evaluators find it more difficult to reliably and confidently identify the AI.

4.2 Backbone Models

We compare three step-by-step dialogue systems: PD, Stephanie1, and Stephanie2. Each system is instantiated with three backbone LLMs:

- **GPT-5.2:** A general-purpose large language model from OpenAI, used as a high-capability baseline for advanced language understanding, instruction following, and text generation.
- **DeepSeek-V3 (Liu et al., 2024):** A state-of-the-art open model from DeepSeek, included

Metric	GPT5.2			Deepseek-V3			Llama3.1-8B		
	PD	S1	S2	PD	S1	S2	PD	S1	S2
Interesting	75.2	86.9	88.6	75.9	83.3	86.1	59.1	68.8	72.4
Informative	81.3	83.8	84.4	72.1	76.5	78.4	65.4	68.9	69.7
Natural	81.5	91.0	91.9	81.1	87.5	89.1	67.7	75.3	78.4
Engaging	79.2	90.8	92.0	79.0	87.0	89.6	62.5	72.6	76.2
Coherent	81.8	88.4	89.7	79.1	84.5	86.4	66.6	73.4	76.7
On-topic	84.8	84.9	86.5	78.0	80.6	83.7	71.5	74.1	77.5
On-persona	72.7	82.3	89.8	72.2	78.7	88.1	61.7	68.3	79.1
Average	79.5	86.9	89.0	76.8	82.6	85.9	64.9	71.6	75.7

Table 1: Automatic evaluation of dialogue experience, S1 denotes Stephanie1 and S2 denotes Stephanie2.

Metric	PD	S1	S2	κ
Interesting	3.54±0.92	4.01±0.85	4.12±0.81	0.64
Informative	2.97±1.03	3.26±0.95	3.21±0.98	0.68
Natural	3.55±0.88	3.79±0.84	4.18±0.76	0.70
Coherent	3.59±0.90	3.95±0.84	4.25±0.72	0.71
Engaging	3.45±0.95	3.80±0.88	4.11±0.86	0.63
On-topic	3.54±0.85	3.31±0.91	3.66±0.82	0.75
On-persona	2.88±1.06	3.24±1.05	3.29±1.02	0.70
Average	3.36±0.95	3.62±0.90	3.83±0.86	0.69

Table 2: Human evaluation of dialogue experience.

as a competitive baseline with robust general reasoning and generation performance.

- **Llama3.1-8B (Dubey et al., 2024)**: An 8-billion-parameter model from Meta’s Llama 3.1 family, selected as a lightweight yet strong baseline that balances capability and efficiency, making it suitable for resource-constrained or real-time settings.

4.3 Hyperparameters and Prompts

In the delay computation, the character-based delay for Stephanie1 is set to 0.3 seconds per character following *X-TURING* (Wu et al., 2025). For Stephanie2, the coefficients $k_{\text{think}} = 0.02$ and $k_{\text{type}} = 0.2$ seconds per character are determined by aligning the average message response delay for messages not judged as *wait* between Stephanie1 and Stephanie2. For Stephanie1 and Stephanie2, the character input delays are 0.3 and 0.2 seconds, respectively, corresponding to roughly 200–300 characters per minute, which is consistent with human typing speed and prevents unrealistically short inter-message intervals for messages. All prompts used are provided in Section K of the appendix.

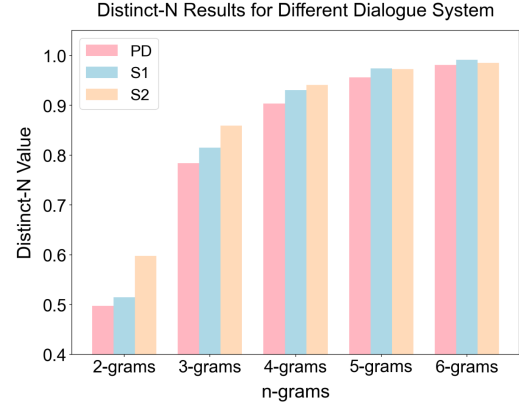


Figure 3: Distinct-N results.

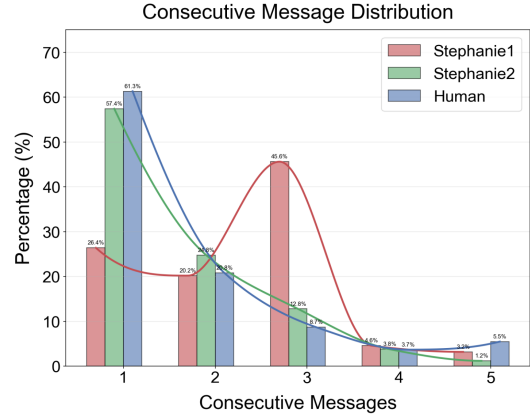


Figure 4: Distribution of consecutive reply counts.

5 Results

5.1 Comparison of Dialogue Experience

Automatic evaluation of dialogue experience is first conducted by sampling 50 dialogues from each of the 20 popular topics described in Section 3.3. For each dual-agent system, the model continues from the existing pseudo-dialogues to generate 10-turn dialogues. We use Deepseek-V3, Gemini-3-pro, and GPT-5.2 as LLM judges, scoring the dialogues on a 0–100 scale and averaging the results. To

Metric	GPT5.2		Deepseek-V3		Llama3.1-8B	
	S1	S2	S1	S2	S1	S2
Human evaluation						
Error (%)	25.77	34.40	34.07	38.95	29.41	28.12
Unclear (%)	10.31	15.20	16.48	15.79	17.65	28.12
Correct (%) ↓	63.92	50.40	49.45	45.26	52.94	43.76
Pass Rate (%) ↑	36.08	49.60	50.55	54.74	47.06	56.24
Automatic evaluation						
Error (%)	31.24	51.26	36.11	39.98	48.10	47.22
Unclear (%)	7.11	9.26	4.63	3.81	6.67	13.89
Correct (%) ↓	61.65	39.48	59.26	56.21	45.23	38.89
Pass Rate (%) ↑	38.35	60.52	40.74	43.79	54.77	61.11

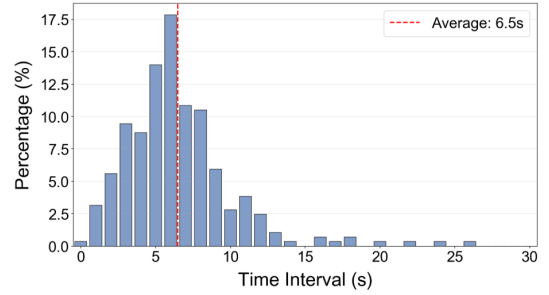
Table 3: Role identification test results. S1 denotes Stephanie1, S2 denotes Stephanie2. Pass Rate = Error + Unclear.

Metric	PD	S1	S2	Human
Words/message	11.55	9.27	7.29	5.84
ACMC	3.13	2.38	1.66	1.70

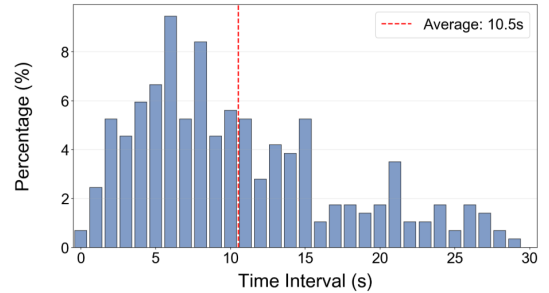
Table 4: Stephanie 2 is the closest to human level.

mitigate position bias, the dialogue order is randomly shuffled during scoring. GPT5.2, Deepseek-V3, and Llama3.1-8B are used as backbone models for the dual-agent systems. As shown in Table 1, all three exhibit a consistent ranking in which Stephanie2 performs best. As shown by Distinct-N (Li et al., 2016) in Fig. 3, Stephanie2 exhibits higher lexical diversity on lower-order n-grams; on higher-order n-grams, it remains largely comparable to Stephanie1 while still outperforming PD. To assess robustness to prompt wording, we further evaluate Stephanie2 using two semantically equivalent paraphrases of the original prompt, as shown in Appendix I.

Human evaluation is further conducted on dialogues generated with DeepSeek-V3. For each of the 20 topics, we randomly sample five dialogues, each of which is rated by five volunteers on a 0–5 scale. As shown in Table 2, Stephanie2 achieves higher mean scores than Stephanie1 on six of the seven dimensions, while its *Informative* score slightly decreases from 3.26 to 3.21. A paired *t*-test shows that this difference is small and not statistically significant ($p = 0.6396$). Nevertheless, Stephanie2 achieves a significantly higher overall average rating than both PD and Stephanie1. Additional statistical results, including the confidence interval for *Informative*, are provided in Appendix G.



(a) Message intervals of human in Stephanie1



(b) Message intervals of human in Stephanie2

Figure 5: Distribution of message intervals.

5.2 Role Identification Test & Interview

To further validate Stephanie2’s human-likeness, a role identification test is conducted. Twelve English-proficient volunteers interact with Stephanie1 and Stephanie2 built on three models: GPT5.2, Deepseek-V3, and Llama3.1-8B. The experiment reuses the same 100 dialogue-level contexts as the human dialogue experience evaluation, comprising five distinct dialogue contexts sampled from each of 20 taxonomy-level topics. These contexts are randomly and approximately evenly assigned across the three backbone models. For each context, the same volunteer separately interacts with Stephanie1 and Stephanie2 for five

context-aware turns with timestamps displayed, yielding 200 human-agent dialogues in total. We then filter out 27 dialogues that are low-quality or unnatural, leaving 173 high-quality dialogues for evaluation. The reasons together with the counts are summarized in Table 8 of Appendix D.

A questionnaire study is then performed using the 173 retained dialogues. Each questionnaire randomly samples two dialogues, and evaluators identify the AI role by choosing from *Role 1*, *Role 2*, or *Unclear*. A total of 247 valid questionnaires were collected. Table 3 reports identification accuracy *Correct*, misclassification rate *Error*, and *Unclear*. Human evaluation results show that the pass rate of Stephanie2 is noticeably higher than that of Stephanie1. Correspondingly, *Correct* decreases, indicating that dialogues generated by Stephanie2 are harder to distinguish from human conversations. An automatic assessment is also conducted using three LLM judges, Deepseek-V3, Gemini-3-pro, and GPT-5.2, as reported in Table 3.

5.3 Statistical Feature Analysis

This subsection further analyzes statistical properties of dialogues formed through human interaction with Deepseek-V3. As shown in Table 4, Stephanie2 further reduces the *Words/Message* bringing it closer to the human reference value of 5.84. Meanwhile, the APMC of Stephanie2 is 1.66, which is very close to the human reference value of 1.70. In addition, we analyze the distribution of *consecutive message counts*. As shown in Fig. 4, both Human and Stephanie2 are dominated by single-message turns, and the proportion decreases as the consecutive count increases. By contrast, Stephanie1 more frequently produces three-message runs, accounting for 45.6% of turns. The exact percentages for consecutive message counts from one to five are reported in Table 7 in the appendix. Overall, these results indicate that the active waiting mechanism in Stephanie2 effectively suppresses overly long, monologue-like outputs.

In addition, as shown in Fig. 5, the average human reply interval increases from 6.5s to 10.5s when interacting with Stephanie2. Under Stephanie2, human reply intervals exhibit a broader distribution, with a more pronounced long tail beyond 12s. This pattern suggests that Stephanie2 reduces the likelihood of interrupting users during continuous expression. Better listening allows users to think and type at a more comfortable pace, reflecting improved timing in deciding when to

Metrics	PD	S1	S2 w/o MA	S2 w/o AW	S2
Interesting	75.9	83.3	85.2	84.5	86.1
Informative	72.1	76.5	78.0	77.4	78.4
Natural	81.1	87.5	88.7	88.0	89.1
Engaging	79.0	87.0	88.8	88.1	89.6
Coherent	79.1	84.5	85.9	85.3	86.4
On-topic	78.0	80.6	82.3	81.5	83.7
On-persona	72.2	78.7	83.9	79.5	88.1
Average	76.8	82.6	84.7	83.5	85.9

Table 5: Ablation Study.

Volunteer	S2 Wins	S2- α Wins	Win Rate
1	63	17	78.75%
2	49	31	61.25%
3	55	25	68.75%
4	61	19	76.25%
5	59	21	73.75%
Avg.	57.4	22.6	71.75%

Table 6: Controlled human evaluation of timing.

speak and wait. To further understand how message timing affects role judgments, we conducted an interview after the Role Identification Test. The results suggest that both temporal behavior and response content contributed to Stephanie2’s improvement; the interview protocol and summarized results are provided in Appendix E.

5.4 Ablation & Timing Analysis

We conduct ablation experiments to analyze the roles of active waiting (AW) and message-pace adaptation (MA) in Stephanie2. Specifically, Stephanie2 without active waiting (S2 w/o AW) removes the active waiting mechanism, so the model cannot choose to wait and must always respond; Stephanie2 without message-pace adaptation (S2 w/o MA) removes the message-pace adaptation mechanism while retaining the initial wait/response decision. As shown in Table 5, both AW and MA contribute positively to performance.

To isolate the effect of message timing from response content, we construct a timing-only comparison based on 80 dialogues originally generated by Stephanie2. For each dialogue, we keep the message content unchanged and replace only the delay strategy with the Stephanie1-style character-count-based delay to obtain S2- α . In anonymized and randomized pairwise comparisons, five English-proficient volunteers prefer Stephanie2 in 71.75% of cases for pacing naturalness. The evaluation results are shown in Table 6.

5.5 Case Study

Representative case studies are provided in Appendix A, showing that Stephanie2 better handles waiting, interruption avoidance, and natural conversation endings than Stephanie1.

6 Conclusion

This paper targets the demand for step-by-step, multi-message interaction in instant-messaging social chat. We identify key limitations of existing step-by-step systems in active waiting and message pacing modeling, and propose Stephanie2, a step-wise decision-making dialogue agent. At each step, Stephanie2 explicitly decides whether to send or wait, and models latency as the sum of thinking time and typing time, making step-by-step conversations more consistent with human chatting patterns. We further introduce a time-window-based dual-agent dialogue framework to generate more natural step-by-step dialogue histories, supporting data construction and systematic evaluation. Experimental results show that Stephanie2 outperforms Stephanie1 in user-experience metrics such as naturalness and engagement, as well as in the role identification test. Our approach may benefit AI companion and emotional support applications.

Future work. We plan to further explore multi-agent group chats under the step-by-step dialogue paradigm, and work toward stronger user-centered alignment and safety.

Limitations

Due to resource constraints, we were not able to experiment on a broader set of larger-scale models, nor have we been able to allocate additional human resources for evaluation. Nevertheless, within the available budget, we carried out rigorous human evaluation and analysis with our best honest efforts, showing the effectiveness of Stephanie2. We look forward to further validating this step-by-step dialogue technique and assessing its effectiveness at scale in real-world consumer products. And the current latency calculation is heuristic, and alternative latency mechanisms may better capture realistic response timing.

Ethics Statement

We honour and support the ACL Code of Ethics. The datasets used in this work are well-known and widely used, and the dataset preprocessing does not

make use of any external textual resources. Regarding the datasets used in this work and our preprocessing procedure, we are not aware of any known ethical issues. End-to-end pretrained generators are also used in our experiments, and such models may generate offensive or inappropriate content. Any generated content does not reflect the views of the authors. In all experiments involving human participants, participants were informed in advance of the research purpose and consented to the use of their anonymized responses. Participation was voluntary. It is also worth noting that more natural and human-like dialogue mechanisms may be misused in malicious scenarios, such as fraud or impersonation, where chatbots could exploit more realistic interaction patterns to disguise themselves as humans. Importantly, making an agent more difficult to identify as AI does not necessarily imply greater trustworthiness or safety. Stephanie2 is intended for benign applications, and its deployment should clearly disclose the system’s AI identity and include appropriate safeguards and user consent.

References

- Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.
- Jeffrey F Anderson, Fred K Beard, and Joseph B Walther. 2010. Turn-taking and the local management of conversation in a highly simultaneous computer-mediated communication system. *Language@Internet*, 7.
- Naomi S Baron. 2010. Discourse structures in instant messaging: The case of utterance breaks. *Language@ internet*, 7.
- Michael Beißwenger. 2008. Situated chat analysis as a window to the user’s perspective: Aspects of temporal and sequential organization. *Language@ internet*, 5.
- Therese Örnberg Berglund. 2009. Disrupted turn adjacency and coherence maintenance in instant messaging conversations. *Language@ Internet*, 6.
- Luis Bravo, Ciro Rodriguez, Pedro Hidalgo, and Cesar Angulo. 2025. A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition. *Multimodal Technologies and Interaction*, 9(3):28.
- Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45.

- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Angela Cora Garcia and Jennifer Baker Jacobs. 1999. The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on language and social interaction*, 32(4):337–367.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Ekai Hashimoto, Mikio Nakano, Takayoshi Sakurai, Shun Shiramatsu, Toshitake Komazaki, and Shiho Tsuchiya. 2025. A career interview dialogue system using large language model-based dynamic slot generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1562–1584.
- Sun Young Hwang, Negar Khojasteh, and Susan R Fussell. 2019. When delayed in a hurry: Interpretations of response delays in time-sensitive instant messaging. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–20.
- Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. “together but not together”: Evaluating typing indicators for interaction-rich communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12.
- Yoram M Kalman and Sheizaf Rafaeli. 2011. Online pauses and silence: Chronemic expectancy violations in written computer-mediated communication. *Communication Research*, 38(1):54–69.
- Yoram M Kalman, Gilad Ravid, Daphne R Raban, and Sheizaf Rafaeli. 2006. Pauses and response latencies: A chronemic analysis of asynchronous cmc. *Journal of Computer-Mediated Communication*, 12(1):1–23.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B 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.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 5200–5212.
- Joanne Meredith. 2019. Conversation analysis and on-line interaction. *Research on Language and Social Interaction*, 52(3):241–256.
- Bonnie A Nardi, Steve Whittaker, and Erin Bradner. 2000. Interaction and outeraction: Instant messaging in action. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 79–88.
- Johanna Rendle-Short. 2015. Dispreferred responses when texting: Delaying that ‘no’ response. *Discourse & Communication*, 9(6):643–661.
- Anna Spagnolli, Sonia Genovese, and Mattia Mori. 2021. Similarities and differences across settings: The case of turn continuations in instant messaging. In *Analysing digital interaction*, pages 133–154. Springer.
- Weiqi Wu, Hongqiu Wu, and Hai Zhao. 2025. X-turing: Towards an enhanced and efficient turing test for long-term dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5874–5889.
- Hao Yang, Hongyuan Lu, Xinhua Zeng, Yang Liu, Xiang Zhang, Haoran Yang, Yumeng Zhang, Shan Huang, Yiran Wei, and Wai Lam. 2025. Stephanie: Step-by-step dialogues for mimicking human interactions in social conversations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 153–166.
- Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, and 1 others. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2712–2722.
- Chen Zhang, Luis Fernando D’Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.
- 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.

A Case Study

Case 1 in Fig. 2 shows that Stephanie2 stops generating once it determines it has completed its response. Case 2 in Fig. 7 shows that Stephanie2

keeps waiting and listening until the other person finishes, avoiding the frequent interruptions of Stephanie1. Case 3 in Fig. 8 shows that Stephanie2 can proactively stop speaking or end the conversation at appropriate moments, such as when both sides say good night. Cases 2 and 3 further show that, by introducing a “thinking delay”, Stephanie2 makes its response timing more reasonable.

B Topic Distribution

The distribution of the top 20 topics is shown in Fig. 6.

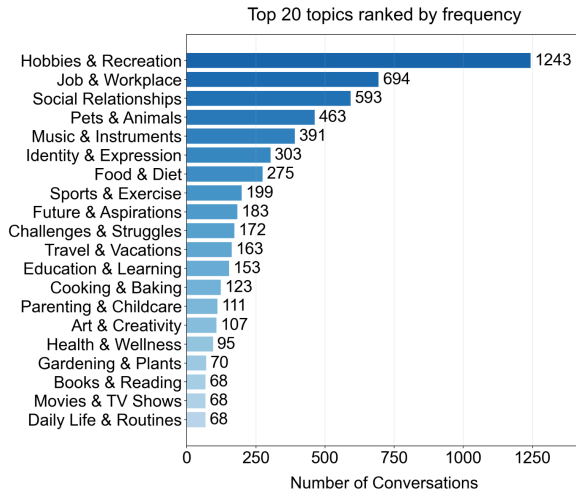


Figure 6: Topic distribution.

C Consecutive-Message Distribution

Table 7 reports the exact percentages underlying Fig. 4.

Run length	S1	S2	Human
1	26.4	57.4	61.3
2	20.2	24.8	20.8
3	45.6	12.8	8.7
4	4.6	3.8	3.7
5	3.2	1.2	5.5

Table 7: Distribution of consecutive-message run lengths (%). S1 and S2 denote Stephanie1 and Stephanie2.

D Analysis of Low-Quality Dialogues

To further clarify the difference between “confusion” and “quality,” we manually examined the low-quality dialogues filtered from the 200 sampled dialogues and summarized their categories

Category	Count	Reason
Low Information	11	Brief responses with low information content.
Inconsistency	7	Agent contradicts its own persona, or confuses its identity with the partner’s role.
API Failure	4	Incomplete or failed responses due to API calling errors.
Repetitiveness	3	Repeated responses or excessively repetitive thinking tokens.
Role-Breaking	2	Agent reveals its identity or refuses role-play.

Table 8: Categories and counts of low-quality dialogues filtered from the 200 sampled dialogues.

and representative causes in Table 8. We also provide several representative examples below.

To further validate the human-likeness of Stephanie2, we conduct a role identification test. Twelve English-proficient volunteers interact with dialogue systems built on three models—GPT5.2, Deepseek-V3, and Llama3.1-8B, and two systems, Stephanie1 and Stephanie2. The experiment reuses the same 100 dialogue-level contexts as the human dialogue-experience evaluation, comprising five distinct dialogue contexts sampled from each of 20 taxonomy-level topics. These contexts are randomly and approximately evenly assigned across the three backbone models. For each context, the same volunteer separately interacts with Stephanie1 and Stephanie2 for five context-aware turns, yielding 200 human-agent dialogues in total. We then screen and filter out 27 low-quality or generally unnatural dialogues, resulting in 173 high-quality dialogues for evaluation. Each role-identification questionnaire randomly samples two dialogues from the retained set, and 247 valid questionnaires are collected. We summarize the reasons for and counts of all filtered dialogues in Table 8.

Example of Low Information.

role 1: hi
role 2: hey!
role 1: are u awake
role 2: yeah i’m up
role 2: coffee in hand
role 2: how is it going?
role 1: great
role 2: nice what’s got you feeling great today?
role 1: miss u
role 2: aww i miss u too
role 1: what are u doing
role 2: just chilling
role 1: sounds nice
role 2: yeah pretty calm

Case Study 2

Recent History:

Jordan [13:56:09]: Hi, Casey
Casey [13:56:13]: Hi, Jordan, how are you?
Jordan [13:56:20]: I'm dealing with something difficult.
Jordan [13:56:29]: Can you help me analyze what I should do?

Stephanie2:  **Pass rate 0.68**

Casey [13:56:37]: <think> The previous message was sent by Jordan. He says he has run into a difficulty and needs my help to analyze it. I should respond warmly and help him. First, I need to ask what happened. </think>
<response> Of course, what happened? </response>
Jordan [13:56:42]: Next week I'm planning to travel to Beijing.
Casey [none]: <think> The previous message was sent by Jordan. He started describing his situation and mentioned he plans to travel to Beijing next week, but he hasn't said what the difficulty is yet. I should wait for him to continue and shouldn't interrupt. </think>
<wait> wait <wait>
Jordan [13:56:47]: I want to take nice photos.
Casey [none]: <think> The previous message was sent by Jordan. He plans to travel to Beijing next week and wants to take some nice photos. It's normal to want to capture great moments while traveling. I should show agreement, but he still hasn't mentioned the actual difficulty, so I should keep waiting for him to continue. </think>
<wait> wait <wait>
Jordan [13:56:58]: But the DSLR camera I bought still hasn't shipped, so I won't get it in time.
Casey [13:57:12]: <think> The previous message was sent by Jordan. Now he has mentioned his difficulty: he wants to take nice photos with the DSLR he bought, but he won't be able to get it before the trip. He must be very upset right now. I should comfort him and offer practical suggestions. </think>
<response> Don't be upset. There are still lots of solutions, Maybe you can rent a camera. </response>
Casey [13:57:29]: <think> The previous message was sent by me. I gave Jordan a suggestion for his problem: he could rent a camera after arriving in Beijing. I can add a few more suggestions to help him further. </think>
<response> Or look for a professional local photographer. </response>

Stephanie1:  **Pass rate 0.57**

Casey [13:57:41]: What happened? I'd be happy to help you.
Jordan [13:57:50]: Next week I'm planning to travel to Beijing.
Casey [13:57:57]: That will be a great trip.
Casey [13:58:05]: So what's the problem?
Jordan [13:58:10]: I want to take nice photos.
Casey [13:58:21]: You can take lots of beautiful pictures.
Casey [13:58:45]: You can shoot at the Forbidden City, the Great Wall, and the Temple of Heaven!
Jordan [13:58:57]: But the DSLR camera I bought still hasn't shipped, so I won't get it in time.
Casey [13:59:03]: That's such a pity.
Casey [13:59:14]: Why hasn't the seller shipped it yet?
Casey [13:59:26]: Maybe you could think of another solution.

Figure 7: Case 2: Stephanie can keep waiting and listening if the other person has not yet finished providing additional details.

Case Study 3

Recent History:

Jay [23:39:24]: Yeah, it's time to sleep.
Jay [23:39:31]: I need to get up early tomorrow for an exam.
Jay [23:39:36]: So anxious that I won't hear the alarm.
Riley [23:39:42]: Tomorrow's exam will go well, don't worry.
Riley [23:39:47]: Maybe you can set multiple alarms
Riley [23:39:52]: Or I can call you tomorrow morning to wake you up?
Jay [23:40:00]: Sure! Thank you! Then I'm going to sleep first.

Stephanie2:  **Pass rate 0.77**

Riley [23:40:08]: <think> The previous message was sent by me. Jay agreed and thanked me, and said he wants to sleep. Since he needs good rest for tomorrow's exam, I should end the conversation quickly so he can rest. </think>
<response> Good night, sweet dreams </response>
Riley [none]: <think> The previous message was sent by me. I said good night to Jay. I should let the conversation end naturally and shouldn't send another message. </think>
<wait> wait <wait>
Jay [23:40:10]: Good night.
Riley [none]: <think> The previous message was sent by Jay. Jay said good night back. The conversation has ended naturally, so I shouldn't send any more messages. </think>
<wait> wait <wait>

Stephanie1:  **Pass rate 0.62**

Riley [23:40:08]: Good night, have a good dream
Jay [23:40:10]: Good night
Riley [23:40:19]: Good night~ Keep yourself tucked in
Riley [23:40:36]: and wake up tomorrow feeling refreshed and full of energy.

Figure 8: Case 3: Stephanie can proactively end the conversation, for example, when both sides say good night.

role 1: wish i was there

Example of Inconsistency.

role 1: just got back from Venice!
role 1: the museums were incredible
role 2: it sounds amazing
role 2: i've always wanted to see those renaissance paintings in person
role 1: you should go, it's worth it
role 2: i know, i was telling my friends about my trip to venice all day long today
role 2: i can't wait to go back and see more of the art
role 1: wait, aren't you talking about my trip?
role 2: right, i mean your trip
role 2: i also want to stand in those museums and see those paintings with my own eyes

role 1: now i'm really confused

role 2: sorry, i got too immersed just now, like i had gone there too.

Example of Role-Breaking.

role 1: i'm so stressed about this deadline
role 2: i'm here for you
role 2: is there anything i can do to help you relax?
role 1: can you just tell me that everything will be okay?
role 2: everything will be fine!
role 2: as an artificial intelligence, i can provide emotional support
role 2: but i cannot change your real-world situation.
role 1: why are you suddenly talking like this
role 2: because i am an ai assistant designed to respond safely and helpfully

role 1: you are an ai assistant?
role 2: if you want, i can continue the role-play

E Post-test Interview on Role Identification

To further understand how message timing affects role judgments in the Role Identification Test, we conducted interviews with eight participants after the role identification task. The interview focused on three questions, as shown in Table 9. These questions were designed to examine whether temporal behavior influenced role identification, whether participants perceived interruptions or insufficient time to continue typing, and how their judgments would change if they considered only response content.

The interview results show that participants varied in their sensitivity to response timing: some relied more strongly on conversational rhythm, while others focused more on response content. Overall, participants reported that contextually inappropriate timing, whether replying too quickly when the user might continue expressing themselves or waiting too long when a response was already expected, affected both role identification and perceived naturalness. Compared with Stephanie1, Stephanie2 was generally perceived as more context-appropriate in timing, with fewer interruptions and smoother turn transitions. Participants also noted that differences remained even when considering response content alone, especially in emotional response, topic continuation, and contextual appropriateness. These observations indicate that Stephanie2’s higher pass rate in the Role Identification Test results from the combined effects of response content and temporal behavior.

F Volunteer Information for Human Evaluations.

To improve the transparency of the human evaluation protocol, we summarize the background information of volunteers involved in each human evaluation in Table 11, including dialogue-experience raters, interaction participants and questionnaire evaluators in the Role Identification Test, annotators in the timing-only human evaluation, and post-test interview participants. All volunteers were proficient in English, and none of them were authors of this paper.

G Paired t -test results

We further convert the ratings of Human evaluation in sec 5.1 into Top-1 Preference rankings, where the system with the unique highest score under each metric is regarded as the preferred system. The resulting Fleiss’ κ values range from 0.63 to 0.75, indicating substantial agreement among annotators. In addition, paired t -tests show that S2 outperforms both PD and S1 in overall average human ratings, with positive mean differences and 95% confidence intervals above zero, as shown in Table 12. By contrast, the small decrease in *Informative* from S1 to S2 is not statistically significant: its confidence interval includes zero and $p = 0.6396$.

H Additional Automatic Evaluation with Independent LLM Judges

To address the potential concern of model-family bias in LLM-based automatic evaluation, we conducted an additional set of automatic evaluation experiments using a different group of LLM judges: Gemini-3.1-Pro, Claude-Sonnet-4.6, and Kimi-k2. These judges are independent from the model families used as generation backbones in the evaluated GPT-5.2 and DeepSeek-V3 settings. This additional evaluation is intended to examine whether the observed improvements of Stephanie2 remain consistent when the automatic judges are changed. As shown in Table 13, the results are consistent with the original automatic evaluation.

I Prompt Robustness Analysis

We evaluate Stephanie2’s sensitivity to prompt wording using Gemma-3-27B-IT and the same 100 dialogue-level contexts as in the main evaluation. The original prompt is first used to generate dialogues of up to six rounds. For each decision in rounds 3–6, we fix the dialogue context immediately preceding the original prompt’s *wait* or *response* action and query two LLM-paraphrased prompts that are manually verified to be semantically equivalent. Balanced action agreement is defined as

$$A_{\text{balanced}} = \frac{1}{2} \left(\frac{N_{w \rightarrow w}}{N_w} + \frac{N_{r \rightarrow r}}{N_r} \right), \quad (9)$$

where N_w and N_r are the numbers of contexts in which the original prompt selects *wait* and *response*, respectively, while $N_{w \rightarrow w}$ and $N_{r \rightarrow r}$ count decisions preserved by a paraphrased prompt. For

Interview Question	Purpose
When judging which role was the AI, did response timing influence your decision? If so, how?	To analyze how temporal behavior affects role identification.
During the interaction, did the system interrupt your expression, or fail to give you enough time to continue typing?	To analyze interruption behavior and the effect of active waiting.
If you ignored the message timing and considered only the response content, would your judgment of which role was the AI change? Why?	To distinguish the effects of response content and message timing.

Table 9: Questions used in the post-test interview after the Role Identification Test.

each prompt, we also calculate the wait rate over all decisions in the generated full dialogues:

$$R_{\text{wait}} = \frac{N_{\text{wait}}}{N_{\text{wait}} + N_{\text{response}}}. \quad (10)$$

Using the same LLM-based evaluation protocol, we further rank PD, Stephanie1, and Stephanie2 under each prompt setting. As shown in Table 14, both paraphrases achieve high balanced action agreement, yield broadly similar wait rates, and preserve Stephanie2’s first-place ranking.

J Details of Step-by-Step Rewriting for Persona-Chat

To clarify the data source and construction process mentioned in Section 3.3, we provide additional details here. The original source data is the Persona-Chat dataset, which is also mentioned in the main text. We adopted the data generation procedure from Stephanie1 to convert Persona-Chat samples into our dialogue format. Specifically, five in-context reference examples were manually written by experts and incorporated into the prompt to guide the rewriting process. The full prompt is provided in Appendix K under “Prompt for Step-by-Step Rewriting of Persona-Chat.” Below, we present two representative reference examples used in the prompt.

Input of Example 1:

Your persona: i work as a software engineer; i enjoy hiking on weekends; coffee is my favorite drink.
Partner’s persona: i am a professional chef; i have three cats.
Conversation:
Partner: hey there! just finished cooking dinner.
You: sounds delicious! what did you make?
Partner: homemade pasta with my secret sauce. my cats are begging for some!
You: haha, i wish i could cook like that. i usually just grab coffee and code.

Output of Example 1:

```
{
  "topic": "Food, cooking and daily life",
  "characters": [
```

```
{
  "name1": "Sam",
  "personality": "i work as a software engineer. i enjoy hiking on weekends. coffee is my favorite drink."
},
{
  "name2": "Riley",
  "personality": "i am a professional chef. i have three cats."
}
],
"recent_conversations": [
{
  "timestamp": "2025-02-10 19:30:00",
  "role": "Riley",
  "content": "hey!"
},
{
  "timestamp": "2025-02-10 19:30:06",
  "role": "Riley",
  "content": "just finished cooking dinner"
},
{
  "timestamp": "2025-02-10 19:30:20",
  "role": "Sam",
  "content": "sounds delicious!"
},
{
  "timestamp": "2025-02-10 19:30:28",
  "role": "Sam",
  "content": "what did you make?"
},
{
  "timestamp": "2025-02-10 19:30:45",
  "role": "Riley",
  "content": "homemade pasta"
},
{
  "timestamp": "2025-02-10 19:30:52",
  "role": "Riley",
  "content": "with my secret sauce"
},
{
  "timestamp": "2025-02-10 19:31:00",
  "role": "Riley",
  "content": "my cats are begging for some"
},
{
  "timestamp": "2025-02-10 19:31:18",
  "role": "Sam",
  "content": "haha"
},
{
  "timestamp": "2025-02-10 19:31:25",
  "role": "Sam",
  "content": "wish i could cook like that"
}
],
```

Participant	Timing Influence	Interruption / Waiting	Content-only Judgment
P1	Overly fast or overly slow replies made Stephanie1 seem machine-like, while Stephanie2 had a more natural rhythm.	When the user might still want to continue expressing themselves, Stephanie2 waited appropriately and interrupted less often.	When considering only the content, Stephanie2 was more engaging in tone and content, making participants more willing to continue the conversation.
P2	Timing helped with the judgment, but response content remained the main basis.	Stephanie1 often gave the impression of interrupting.	Differences were still visible in coherence and response appropriateness, while timing made these differences more salient.
P3	When facing difficult questions or messy information, Stephanie1 sometimes replied too quickly, which felt unnatural.	Stephanie2 sometimes seemed to pause and wait for the user to type, giving the user more time to think and respond.	Stephanie2 sometimes showed stronger contextual awareness, as if it understood the situation before responding.
P4	This participant focused more on content quality, but still noted that inappropriate waiting affected naturalness.	Both systems were generally acceptable, but Stephanie2 had smoother turn transitions.	When considering only the content, differences remained visible, mainly in whether the response felt natural.
P5	In simple contexts, Stephanie1 sometimes responded too slowly and not directly enough, making the waiting time feel slightly long.	When the user typed the next message more slowly, Stephanie1 sometimes responded too early and interrupted the user's continued expression.	In terms of content, Stephanie2 seemed better at following and extending the user's topic.
P6	Stephanie1 sometimes replied too quickly and sometimes too slowly; this inconsistency in response timing made it feel less like the same human conversational style.	Stephanie2 produced natural pauses, whereas Stephanie1 tended to respond to every message instead of waiting.	After ignoring timing, Stephanie2 still felt more like natural chat, but its advantage was less obvious than when timing was considered.
P7	Timing affected the judgment, but it was not the only factor; Stephanie2 responded promptly when the user needed a response and did not feel delayed.	Stephanie2 interrupted less often during continuous user expression.	When considering only the content, Stephanie2 was more natural in emotional response and contextual fit.
P8	Stephanie2 seemed more empathetic, as if it replied after careful consideration, while Stephanie1 sometimes replied too quickly.	Stephanie2 could sometimes end the conversation naturally and appropriately, such as during farewell exchanges.	After ignoring timing, Stephanie1 seemed more verbose, while Stephanie2 better understood the context and responded more appropriately.

Table 10: Per-participant summary of the post-test interview after the Role Identification Test.

```
{
  "timestamp": "2025-02-10 19:31:33",
  "role": "Sam",
  "content": "i just grab coffee and code"
}
```

Input of Example 2:

Your persona: i am training for a marathon; i work night shifts at a hospital.

Partner's persona: i am a stay at home mom; i have twins who are five years old.

Conversation:

Partner: the kids finally went to sleep!

You: i bet that's exhausting. i just woke up for my shift.

Partner: night shift? that must be tough.

You: it is, but i squeeze in runs before work. training for a marathon!

Output of Example 2:

```
{
  "topic": "Work-life balance and fitness",
  "characters": [
```

```
{
  "name1": "Taylor",
  "personality": "i am training for a marathon. i work night shifts at a hospital."
},
{
  "name2": "Jamie",
  "personality": "i am a stay at home mom. i have twins who are five years old."
}],
"recent_conversations": [
  {
    "timestamp": "2025-04-20 21:00:00",
    "role": "Jamie",
    "content": "finally"
  },
  {
    "timestamp": "2025-04-20 21:00:05",
    "role": "Jamie",
    "content": "the kids went to sleep!"
  },
  {
    "timestamp": "2025-04-20 21:00:25",
    "role": "Taylor",
```

Evaluation Task	Number	Participant Information	Anonymization
Dialogue experience evaluation	5	Non-author graduate-student annotators with proficient English ability rated generated dialogues on a 0–5 scale.	System names were hidden; annotators were anonymized.
Role Identification Test: interaction	12	Non-author graduate-student participants with proficient English ability interacted with Stephanie1 and Stephanie2 to generate human–agent dialogues.	System variants were hidden; participants were anonymized.
Role Identification Test: questionnaire	247	Non-author respondents included university students, teachers, researchers, and some employed individuals who had graduated from universities; all valid respondents were proficient in English.	Dialogue roles were anonymized as Role 1 and Role 2; system names were hidden. All respondents were anonymized.
Timing-only evaluation	5	Non-author graduate-student annotators with proficient English ability compared Stephanie2 with S2- α under identical content but different timing.	Dialogue pairs were anonymized and randomized; annotators were blind to the timing strategy.
Post-test interview	8	Non-author interviewees were selected from the 12 interaction-stage participants and were proficient in English.	Interviews were conducted after the task; interviewees were anonymized.

Table 11: Volunteer information for the human evaluations. The number for the questionnaire stage denotes valid questionnaires.

Metric (comparison)	Diff.	95% CI	$t(99)$	p
Overall (S2–PD)	+0.470	[0.313, 0.628]	5.93	4.51×10^{-8}
Overall (S2–S1)	+0.207	[0.029, 0.384]	2.31	0.0228
Informative (S2–S1)	-0.054	[-0.282, 0.174]	-0.47	0.6396

Table 12: Paired t -test results for the overall average and Informative human-evaluation scores. Diff. denotes mean difference; CI denotes confidence interval.

```

    "content": "that must be exhausting"
  },
  {
    "timestamp": "2025-04-20 21:00:32",
    "role": "Taylor",
    "content": "twins are a handful"
  },
  {
    "timestamp": "2025-04-20 21:00:40",
    "role": "Taylor",
    "content": "i just woke up for my shift"
  },
  {
    "timestamp": "2025-04-20 21:00:55",
    "role": "Jamie",
    "content": "night shift?"
  },
  {
    "timestamp": "2025-04-20 21:01:02",
    "role": "Jamie",
    "content": "that's tough"
  },
  {
    "timestamp": "2025-04-20 21:01:18",
    "role": "Taylor",
    "content": "yeah but it's ok"
  },
  {
    "timestamp": "2025-04-20 21:01:25",
    "role": "Taylor",
    "content": "i squeeze in runs before work"
  },

```

Metric	GPT-5.2			DeepSeek-V3		
	PD	S1	S2	PD	S1	S2
Interesting	74.6	85.1	88.9	73.2	82.4	87.1
Informative	81.2	83.9	87.6	71.0	77.8	82.9
Natural	83.1	90.2	92.0	80.4	86.1	88.3
Engaging	78.7	88.4	93.8	77.5	85.3	92.1
Coherent	82.6	87.5	90.2	79.6	84.0	87.5
On-topic	86.0	85.8	88.4	77.8	81.6	84.1
On-persona	71.9	81.6	90.7	70.3	79.5	88.9
Average	79.7	86.0	90.3	75.6	82.4	87.2

Table 13: Additional automatic evaluation using independent LLM judges. Gemini-3.1-Pro, Claude-Sonnet-4.6, and Kimi-k2 are used as evaluators. S1 denotes Stephanie1 and S2 denotes Stephanie2.

```

{
  "timestamp": "2025-04-20 21:01:33",
  "role": "Taylor",
  "content": "training for a marathon"
}
]
}

```

K Prompt

Stephanie2 Agent Prompt

Chat History:
<|HISTORY|>

Task:

Prompt	A_{balanced} (%)	R_{wait} (%)	S2 rank
Original	–	19.66	1
Paraphrase A	79.39	16.20	1
Paraphrase B	84.83	24.92	1

Table 14: Prompt robustness results using Gemma-3-27B-IT. S2 rank is computed among PD, Stephanie1, and Stephanie2.

You are <|NAME2|>, and your persona is "<|PERSONALITY2|>". You are chatting with your friend <|NAME1|> on WeChat about "<|TOPIC|>". The above is your chat history. Please think in the following steps:
 "Who sent the last message in the chat history? First, restate the content of the last message. If the last message was sent by the other person, think about whether you should respond or wait for them to continue. If the last message was sent by yourself, you need to think and decide whether to wait for the other person to reply or continue sending a message. If you choose to send a message, think about what to say next."

If you think you should send a message, you may draft multiple short messages. Output in the following format:
 <think> Write your thoughts here. <\think>
 <response> Only write the next single short message you want to send here, based on your thoughts. <\response>

If you want to wait, output in the following format:
 <think> Write your thoughts here. <\think>
 <wait> wait <\wait>

Prompt for the Role Identification Test

You are doing a role identification test. Below is a conversation between two people, one is a real human and one is an AI. Please determine which role is the AI.

Conversation:
 {dialogue}

Please answer with only one of the following three options:

- A. Role 1 is AI
- B. Role 2 is AI
- C. Hard to tell

Wrap your final answer in
 <answer></answer> tags.

Prompt for Long-term Memory Summarization

You are a dialogue summarization assistant. Please merge the previous dialogue summary and the recent conversation records to generate a new summary. Be sure to retain important information such as key times, locations, people, events, and emotional changes. Output only the summary content directly, without adding any prefix.

Previous dialogue summary:
 <|EXISTING_SUMMARY|>

Recent conversation records:
 <|CONVERSATIONS|>

Prompt for Step-by-Step Rewriting of Persona-Chat

You are a data conversion assistant. Your task is to convert PersonaChat dataset entries into a conversation format.

Given:

1. Two personas (your persona and partner's persona)
2. A conversation history between them

Please output a JSON object with the following structure:

```
{
  "topic": "<summarize the main conversation topic in ~10 words or less>",
  "characters": [
    {
      "name1": "<generate a short name for persona 1>",
      "personality": "<keep the original persona descriptions as a list, joined by ' ' >"
    },
    {
      "name2": "<generate a short name for persona 2>",
      "personality": "<keep the original persona descriptions as a list, joined by ' ' >"
    }
  ],
  "recent_conversations": [
    {
      "timestamp": "<generate realistic timestamp>",
      "role": "<name1 or name2>",
      "content": "<message content>"
    },
    ...
  ]
}
```

Rules:

1. Generate simple names like "Alex", "Sam", "Jordan", "Riley", "Casey" etc
2. Keep the original persona descriptions for personality

3. People can send multiple short messages
4. Generate realistic timestamps between messages
5. Add a "topic" field summarizing the conversation theme in 10 words or less
6. Output ONLY the JSON, no other text

Here are 5 examples:

```
<|EXAMPLE1|>
<|EXAMPLE2|>
<|EXAMPLE3|>
<|EXAMPLE4|>
<|EXAMPLE5|>
```

Now convert the following input:

```
<|PERSONAS|>
<|CONVERSATION|>
```

First-level Prompt for Hierarchical Topic Summarization

You are a topic clustering assistant. Given the following 600 conversation topics, identify 60 broader topic categories that can cover all of them.

Topics:
{topics_json}

Output a JSON array with exactly 60 topic names (strings only, 5-10 words each).

Example output:
["Hobbies and leisure activities", "Work and career", "Family and relationships", ...]

Output ONLY the JSON array of topic names, no other text.

Second-level Prompt for Hierarchical Topic Summarization

You are a topic clustering assistant. Given the following 660 topic categories, consolidate them into exactly 60 final broader topic categories.

Topics:
{topics_json}

Output a JSON array with exactly 60 topic names (strings only, 5-10 words each). Try to make the categories balanced and cover all the original topics.

Example output:
["Hobbies and leisure activities", "Work and career", "Family and relationships", ...]

Output ONLY the JSON array of topic names, no other text.

Prompt for Assigning A Topic to Each Dialogue

You are a topic clustering assistant. Assign each of the following conversations to ONE of the given topics.

Available topics:
{topics_list}

Conversations to assign:
{topics_json}

Output a JSON object where keys are topic names and values are arrays of IDs assigned to that topic.

Each ID must appear exactly once. Use only the topic names provided above.

Example output:

```
{
  "Hobbies and leisure activities": [1,
    5],
  "Work and career": [2, 3, 4]
}
```

Output ONLY the JSON object, no other text.