Stephanie: Step-by-Step Dialogues for Mimicking Human Interactions in Social Conversations[*]

Hao Yang[1†], Hongyuan Lu[2‡], Xinhua Zeng[1], Yang Liu[1,3], Xiang Zhang[4], Haoran Yang[2], Yumeng Zhang[5], Shan Huang[4], Yiran Wei[4], Wai Lam[2]
[1]Fudan University, [2]The Chinese University of Hong Kong, [3]University of Toronto, [4]FaceMind Corporation, [5]Tsinghua University
`yanghao21@m.fudan.edu.cn, hongyuanlu@outlook.com`

Abstract

In the rapidly evolving field of natural language processing, dialogue systems primarily employ a single-step dialogue paradigm. Although this paradigm is commonly adopted, it lacks the depth and fluidity of human interactions and does not appear natural. We introduce a novel **Step**-by-Step Dialogue Paradigm (Stephanie), designed to mimic the ongoing dynamic nature of human conversations. By employing a dual learning strategy and a further-split post-editing method, we generated and utilized a high-quality step-by-step dialogue dataset to fine-tune existing large language models, enabling them to perform step-by-step dialogues. We thoroughly present Stephanie. Tailored automatic and human evaluations are conducted to assess its effectiveness compared to the traditional single-step dialogue paradigm. We will release code, Stephanie datasets, and Stephanie LLMs to facilitate the future of chatbot eras.[1]

1 Introduction

In the field of natural language processing, the research and development of dialogue systems continue to advance. Progressive dialogue systems aim to mimic human communication in daily life, which is however currently under-studied. Such systems can be potentially applied to broader applications on those hot AI applications such as Character.ai.[2] Such AI companion applications have great potential. For example, Character.ai is ranked as the top 3 popular AI applications with tremendous monthly active users at the time of writing (top 1 is ChatGPT), according to popular Venture Capital.[3] However, these systems predominantly employ a Single-Step Dialogue Paradigm (Abbas et al.; Touvron et al., 2023; Du et al., 2022; Abdin et al., 2024; Achiam et al., 2023), where the system provides a single line, one-time response to each user input, quickly addressing user questions or needs. This approach falls short in simulating the naturalness of real human conversations and attracting user engagement. In reality, daily human conversations are ongoing, dynamically evolving processes involving multiple topics (Song et al., 2022; Butler, 2011; Nie et al., 2024; Poria et al., 2019). Notably, the current Single-Step Dialogue Paradigm fails to fully capture this naturalness and lacks in attracting users to engage in conversations.

To better emulate the style of human social conversations, this paper introduces an innovative dialogue paradigm named step-by-step dialogue (Stephanie), where its human-rated engaging score surpasses that of single-step dialogue, as illustrated in Figure 1. Unlike single-step dialogue, Stephanie mimics casual chats in instant messaging applications, creating a more natural and continuous dialogue flow with the power of in-context learning (Min et al., 2022; Chen et al., 2023; Wang et al., 2024). Under this paradigm, the dialogue system does not just provide a one-time response to each input but constructs a conversation composed of multiple dispersed yet coherent responses. This design allows the system to gradually develop the conversation, with each response focusing on different aspects of the dialogue, making the conversation more detailed, rich, and engaging. We found that it provides a better conversation experience, evoking

[*]This research/paper was partially supported by the Center for Perceptual and Interactive Intelligence (CPII) Ltd. under the Innovation and Technology Commission's InnoHK scheme.

[†]Work done during an internship at FaceMind Corporation

[‡]Corresponding author and co-first author

[1]Our code and data available at: `https://github.com/h17ao/Stephanie`

[2]https://character.ai

[3]https://a16z.com/100-gen-ai-apps/

Findings of the Association for Computational Linguistics: NAACL 2025, pages 153–166
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

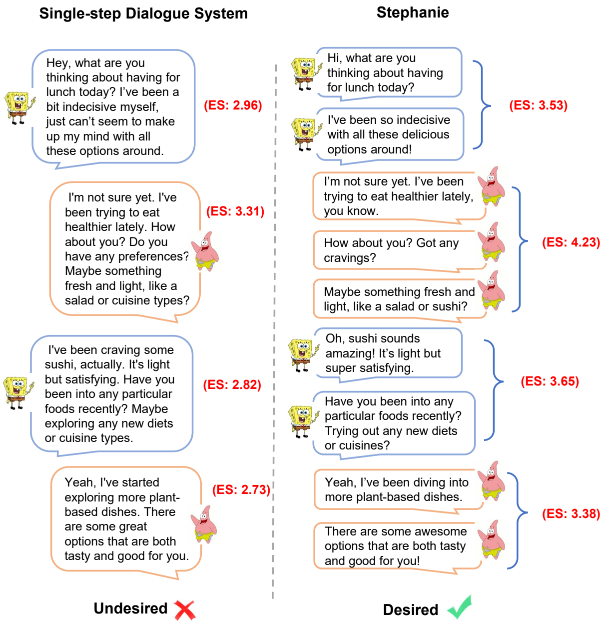

Figure 1: A single-step dialogue system and Stephanie. Stephanie constructs a dialogue composed of multiple dispersed yet coherent responses. ES stands for the **engaging** score given by humans, which is introduced in section 4.4.

greater user engagement. For example, in the step-by-step chat mode, the system can address various aspects of a user's expression step by step, first by supporting users through empathetic language and understanding, and then by asking questions or expanding the topic, gradually building deeper and more continuous communication.

If the one-time response of a single-step dialogue system is simply divided into multiple responses by punctuation, the overall logic and integrity of the one-time response itself will result in an unnatural and stiff step-by-step dialogue, which does not resemble a real social interaction with people. In order to fully consider the semantic similarities and differences between sentences, as well as naturalness and anthropomorphism when generating step-by-step dialogue and implementing a dialogue system with a step-by-step chat function, we introduced a comprehensive prompting framework that employs a dual learning strategy and a Further-Split post-editing method to generate and optimize step-by-step dialogue datasets. We then used this dataset with a specific fine-tuning strategy to be compatible with existing large models, thereby establishing a step-by-step dialogue system. The step-by-step dialogue paradigm demonstrates significant academic and practical value in enhancing the naturalness and engaging nature of chat systems. By simulating real social interactions, this research not only advances the technology of dialogue systems but also provides new insights and approaches for achieving more natural and human-like communication between machines and humans.

The main contributions of this paper include:

- We innovatively propose a step-by-step dialogue paradigm that utilizes a series of dispersed yet coherent responses to more closely mirror the style of real human communication interactions, thereby enhancing the engaging nature and human-likeness of the dialogue.

- We introduced a bidirectional learning strategy and a Further-Split post-editing method to generate and optimize step-by-step dialogue datasets, and then we fine-tuned existing large models to develop a step-by-step dialogue system. To facilitate future research, we will release code, Stephanie datasets, and Stephanie LLMs in the near future.

- Finally, we comprehensively compared single-step dialogues with progressive dialogues through both human and automated evaluations, demonstrating the significant advantages of step-by-step dialogue systems over traditional single-step dialogue systems.

2 Related Work

Large Language Models for Dialogue Systems In dialogue systems, previous dialogue systems have been traditionally finetuned on publicly available dialogue datasets (Zhang et al., 2019; Adiwardana et al., 2020; Roller et al., 2020; Thoppilan et al., 2022). Motivated by ChatGPT's success, developers are now conducting supervised finetuning on open-source large language models like LLaMA (Touvron et al., 2023) to develop dialogue systems. This process involves finetuning with constructed instruction-following examples(Taori et al., 2023) and using dialogue data distilled from ChatGPT (Ulmer et al., 2024; Chiang et al., 2023). Furthermore, some studies have been prompting dialogue systems built on large pre-trained models to induce the knowledge embedded in these language models. Other works study the fallback unanswerable questions to make dialogue systems more controllable (Lu et al., 2022a). Areas of focus include task-oriented dialogues (Labruna et al., 2023; Swamy et al., 2023; Mi et al., 2022), knowledge-supported dialogues (Semnani et al., 2023; Rogers et al., 2023), and open-domain dialogues (Chen et al., 2023; Lee et al., 2023; Hongru et al., 2023).

Emotional Support in Dialogue Systems Emotions play a crucial role in building dialogue systems, not only involving emotional expression but also alleviating users' emotional distress through guided conversations and support techniques (Zhou and Wang, 2017; Huber et al., 2018; Huang et al., 2020). Empathetic responses are key to effective emotional support, focusing on understanding users' emotions and providing personalized replies (Liao et al., 2021; Sun et al., 2021; Majumder et al., 2020). The empathetic capabilities of large language models (LLMs) can be enhanced through semantic similarity learning, bi-directional generation, and integration with knowledge bases (Qian et al., 2023), combined with intermediate reasoning steps (Hongru et al., 2023). By formulating complex dialogue strategies, emotional support systems can achieve goals such as exploration, comfort, and action (Rogers et al., 2023; Peng et al., 2022; Cheng et al., 2023; Chua, 2024). Currently, LLM-based emotional support faces the challenge of data scarcity. One solution is to use dialogues as generative seeds and leverage the model's contextual learning potential to recursively generate scalable emotional support dialogue datasets (Zheng et al., 2023). Other works integrates the generation of partner personas to further enhances dialogue systems (Lu et al., 2022b).

In the current field of natural language processing, most dialogue systems based on large language models primarily adopt a Single-Step Chat Paradigm(Wu et al., 2023; Touvron et al., 2023; Mai et al., 2023; Yamazaki et al., 2023). Within this paradigm, the system responds to each user input with a comprehensive and complete one-time reply to promote interaction. Such interactions provide information-dense responses to handle complex inquiries, focusing on the informational density and completeness of each response, which is suitable for directly resolving specific questions or providing detailed information in a single interaction. However, this paradigm exhibits certain limitations in emulating the natural fluidity and emotional expression found in human daily dialogues. While it can identify and respond to users' emotional inclinations, the interaction pattern often sticks to a question-and-single-answer format, lacking the emotional continuity and interaction depth present in real conversations.

3 Methodology

In this section, we will delve into the process of generating and optimizing step-by-step dialogues, and based on this, create a high-quality step-by-step dialogue dataset. We further fine-tuned and built a dialogue system capable of step-by-step interactions to simulate the step-by-step dialogue paradigms found in real human social exchanges.

3.1 Dual Learning Strategy for Step-by-Step Dialogue Generation

To efficiently generate step-by-step dialogues that mimic real human social interactions, inspired by contrastive prompting, we propose a dual learning strategy combining both positive and negative learning objectives within a comprehensive prompt framework. As illustrated in Figure 2, the framework consists of three elements: background information D, positive learning objectives P, and negative learning objectives N, aiming to enhance the model's ability to generate dialogues that are both rich and natural. The language modelling probability is:

$$p(r \mid D, P, N) \tag{1}$$

where r is the response output of the model, and the design of the three elements is as follows:

- **Background Information:** We use an LLM to summarize and generate the themes T of each dialogue segment from the persona-chat dataset and the characteristics C of the dialogue participants, to form the background information $D = \{T, C\}$. This information guides the model's generation, covering common topics such as family, work, and leisure activities, while considering the diverse personalities of the dialogue participants—for example, one might be described as optimistic and active, while another might be portrayed as having recently faced setbacks but remaining diligent and academically inclined.

- **Positive Learning Objectives:** To help the model understand the step-by-step dialogue paradigm, we created five high-quality step-by-step dialogue examples as the positive objectives P. These examples simulate everyday social exchanges between two individuals and serve as a basis for few-shot learning, training

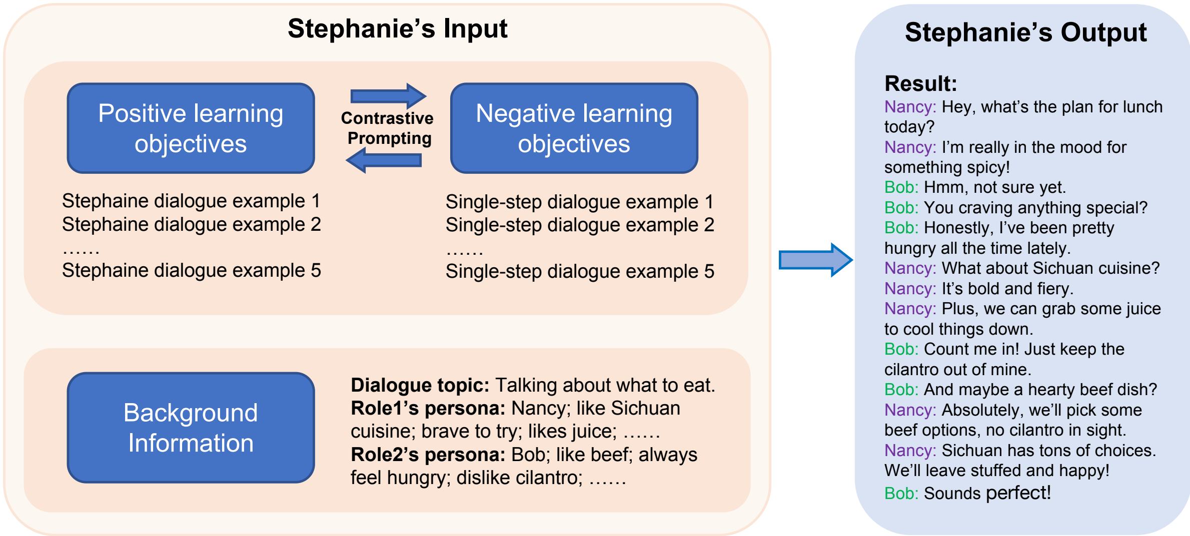

Figure 2: In the process of step-by-step dialogue generation, we adopted a dual learning strategy to enhance the model's ability to generate natural dialogues through the Step-by-Step Dialogue Prompt Framework. This strategy combines positive and negative learning objectives. The positive objective includes high-quality step-by-step dialogue examples selected from real social interactions, while the negative objective comprises designed high-quality single-step dialogue examples. Through contrastive prompting, this approach helps the model distinguish between step-by-step dialogues and single-step dialogues, thus generating more natural and emotionally rich step-by-step dialogues.

the model to generate coherent and emotionally rich step-by-step dialogues in different background contexts.

- **Negative Learning Objectives:** Simultaneously, we designed five high-quality single-step dialogue examples as the negative objectives N, which share the same theme as the five step-by-step dialogue examples for the positive objectives. Through contrastive prompting, these examples enable the model to discern the differences between single-step and step-by-step dialogues. This negative learning approach helps the model better understand the step-by-step dialogue paradigm by pushing away dissimilar single-step dialogue examples.

This dual learning strategy is a robust prompting framework. Through this structured approach, the model considers both positive and negative learning objectives during dialogue generation, enhancing its ability to understand and generate step-by-step dialogues while ensuring that the generated dialogues align with the themes and character traits in the background information.

3.2 Optimizing Step-by-Step Dialogues Using the Further-Split Post-Editing Method

Although the method described in Section 3.1 enabled the model to make some progress in generating step-by-step dialogues, our evaluation showed that some generated dialogues still exhibited characteristics of single-step dialogues, such as dense, one-time responses. To address this issue and further enhance the coherence and naturalness of emotional expression in dialogues, we designed a post-editing optimization method called "further-split."

In this process, we selected five initial step-by-step dialogues generated by the model for detailed analysis and manual restructuring. We further split these dialogues according to the natural flow and emotional progression of actual conversations, reorganizing and optimizing the content. The optimized step-by-step dialogue examples were paired with the original examples, serving as rewritten examples to guide the model in learning how to further split and rewrite dialogues, thereby generating more natural and human-like step-by-step dialogues to closely mimic real social interactions.

3.3 Dataset Generation and Finetuning Strategy for Stephanie

Based on the aforementioned comprehensive prompt framework and the further-split post-editing method, we generated a high-quality step-by-step dialogue dataset. To effectively utilize this dataset for finetuning existing large language models, we designed a specific finetuning strategy.

During the finetuning process, we introduced delimiters to format the dataset, providing structured input and output for the model, where the content between each pair of delimiters represents a single exchange between the dialogue participants. We then used this newly formatted step-by-step dialogue dataset to finetune the model. After finetuning, the model's output also adopted the same delimiter-separated step-by-step dialogue format. Then, we found Stephanie brings better user engagement, converting the delimiter format of the input and output into a format similar to message bubbles in social software, allowing users to interact with the large language model using the step-by-step dialogue paradigm.

This plug-and-play finetuning strategy enables our step-by-step dialogue dataset to be compatible with various existing language models, thereby constructing a dialogue system capable of step-by-step interactions to provide a coherent and emotionally rich dialogue experience in practical applications.

4 Experiment Setup

4.1 Dataset

Our incremental dialogue dataset originates from the PERSONA-CHAT dataset (Zhang et al., 2018). It is a renowned multi-turn dialogue dataset grounded in character personas, with each dialogue instance typically comprising around 8 turns, where each self and partner character is described by roughly 4 traits.

Initially, we used the Llama3-70b model to summarize the themes of 8,939 dialogues from the PERSONA-CHAT training set, with summaries averaging between 50 to 100 words. Subsequently, we adopted the Stephanie dialogue generation approach described herein, incorporating the dialogue themes and approximately four traits of each character as background information. Using the Llama3-70b model, we generated an incremental dialogue dataset. During the generation process, we created three dialogues for each theme, and three native English language experts selected 5,457 high-quality dialogues from the results.

4.2 Prompt

We describe the prompt generated for step-by-step dialogue with Stephanie as follows:

<five examples of single-step dialogues>.
<five examples of step-by-step dialogues>.
In single-step dialogues, each role sends only one message per turn. In contrast, step-by-step dialogues allow multiple messages to be sent consecutively before the other role replies, simulating the style of human daily chit-chat. Please generate a step-by-step dialogue and a single-step dialogue based on the background information:
<background information>.

We can also describe the prompt optimized for generating step-by-step dialogue using the Further-Split method as follows:

<five examples of single-step dialogues and corresponding Stephanie>.
Please assess whether each message reply in the following step-by-step dialogue can be further rewritten into multiple replies to make the conversation more natural, interesting, engaging, and closer to human interaction. Then, provide a new version of the step-by-step dialogue:
<the single-step dialogue to be further-split into Stephanie>.

4.3 Baselines and Comparison Models

In evaluating the performance of our model, we consider several leading models in the field of language processing. These models are used as benchmarks due to their significant capabilities in various tasks within natural language processing. Each model is briefly described as follows:

- **GPT-4**: Developed by OpenAI, GPT-4 represents the latest advancement in the Generative Pre-trained Transformer series. Renowned for its vast knowledge base and flexibility across multiple tasks, GPT-4 is a critical benchmark for assessing advanced language understanding and generation capabilities.

- **Llama3-70b**: Also from Meta's Llama series, the Llama3-70b model, with its 70 billion parameters, is aimed at deep contextual understanding and complex reasoning tasks. It serves as a high-end model for performance comparison.

- **Llama3-8b**: A model from Meta's Llama family, Llama3-8b is designed to provide a balance between performance and efficiency with its 8 billion parameters. It is optimized for rapid response and lower resource usage, making it suitable for real-time applications.
- **Phi3-3.8b**: Phi3-3.8b from Microsoft's Phi-3 series of small language models excels in performance while being highly efficient in terms of computational resource usage. These models are designed for flexible deployment across cloud, on-device, and edge computing scenarios, ensuring effectiveness even with limited connectivity. Phi3-3.8b uses high-quality, curated training data to achieve results comparable to larger models.

4.4 Evaluation Metrics

To comprehensively assess the performance of our step-by-step dialogue, we have utilized a series of evaluation metrics aimed at thoroughly measuring various aspects such as the diversity, naturalness, and effectiveness of the dialogues, among others. These metrics include Dialogue Experience Metrics (suitable for both automated and human evaluations), Lexical Diversity Metrics (Distinct-N), and statistical features of the dialogue data, such as the average number of words per message and the Average Consecutive Message Counts (ACMC):

- **Dialogue Experience Metrics**: **Interesting**: The degree of interest in the dialogue. If the dialogue carries a negative sentiment, the score is 0. **Informative**: The amount of information contained in the dialogue. **Natural**: Whether the dialogue is natural and human-like. **Coherent**: Whether the dialogue is logical, consistent, and flows smoothly without contradictions. **Engaging**: Whether the dialogue is engaging, meaning if what is said by both roles makes them want to continue the dialogue. **On-topic**: Whether the dialogue stays on the topic described in the dialogue topic. **On-persona**: Whether the dialogue matches the personas of role1 and role2.
- **Distinct-N**: To quantify the lexical diversity of the dialogues, we utilize the Distinct-N metric. This metric calculates the diversity of n-grams in the generated responses across all possible values of N, showing the system's capability to produce varied and engaging content. The Distinct-N is defined as:

$$\text{Distinct-N} = \frac{\text{Total unique n-grams}}{\text{Total n-grams}} \tag{2}$$

- **Words/Message**: Calculates the average number of words per message, providing insight into the verbosity or conciseness of the dialogues. This helps in determining the efficiency and clarity of the communication. The formula for Words/Messages is defined as:

$$\text{Words/Message} = \frac{\sum_{i=1}^{n} w_i}{n} \tag{3}$$

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

- **ACMC (Average Consecutive Message Counts)**: This metric measures the average number of consecutive messages sent by one participant before receiving a response. It is calculated as:

$$\text{ACMC} = \frac{\sum_{i=1}^{n} c_i}{m} \tag{4}$$

where c_i is the number of consecutive messages in the i-th turn without interruption by the other participant, n is the total number of such turns, and m is the total number of messages sent by the participant.

5 Results

5.1 Evaluation of Conversation Quality

We selected 100 conversation data from the persona-chat dataset as the Original Single-Step Dialogue α. First, we used GPT-4 to summarize the themes of these 100 dialogues. Then, along with the personas of the dialogue participants, we used this background information to write prompts using the Step-by-Step Dialogue Prompt Framework proposed in this paper. These prompts were fed into GPT-4 to generate the Generated Single-Step Dialogue β. Then we applied a further-split method to optimize the γ, resulting in the Further-Split Step-by-Step Dialogue, Stephanie. We also split α using periods to get α', in order to compare the difference between the dialogue α' obtained by simply splitting single-step dialogue with punctuation and Stephanie. Notably, our experiments show that without the dual learning strategy, the

Metrics	GPT4 α	α'	β	γ	Stephanie	Llama3-70b α	α'	β	γ	Stephanie
Interesting	82.00	82.74	80.00	84.66	**88.35**	80.20	83.43	75.25	88.74	**91.63**
Informative	83.20	–	80.35	85.03	**88.19**	79.24	–	72.93	86.67	**88.29**
Natural	87.25	88.26	88.19	91.89	**94.72**	86.82	89.16	84.36	95.00	**97.61**
Engaging	85.74	86.91	84.84	89.42	**92.64**	83.64	85.71	78.74	93.25	**95.86**
Coherent	86.26	87.42	84.31	90.40	**93.64**	84.55	85.26	79.83	94.36	**95.73**
On-topic	–	–	91.54	93.53	**96.35**	–	–	87.78	95.88	**97.10**
On-persona	–	–	92.93	94.25	**96.0**	–	–	90.05	96.53	**98.10**

Table 1: Automatic Evaluation on GPT4 and Llama3-70b. The values represent the percentage scores for each metric, used to evaluate the performance of different dialogues (α, α', β, γ, Stephanie) generated by GPT-4 and Llama3-70b. These scores indicate how interesting, engaging, informative, coherent and natural each dialogue is, as well as its adherence to the given topic and persona. Bold values indicate the highest scores among comparable models, highlighting exceptional performance in specific metrics.

Metrics	α	β	Stephanie
Interesting	79.73	72.14	**83.53**
Informative	75.48	74.56	**79.37**
Natural	79.79	75.87	**87.41**
Engaging	83.55	78.38	**86.41**
On-topic	–	78.87	**82.57**
On-persona	–	77.24	**80.04**

Table 2: Automatic evalution on phi3-3.8b. The table shows the percentage scores of the performance of three dialogues (α, β, γ) across multiple metrics. Bold values represent the best performance in each metric for the phi3-3.8b evaluation.

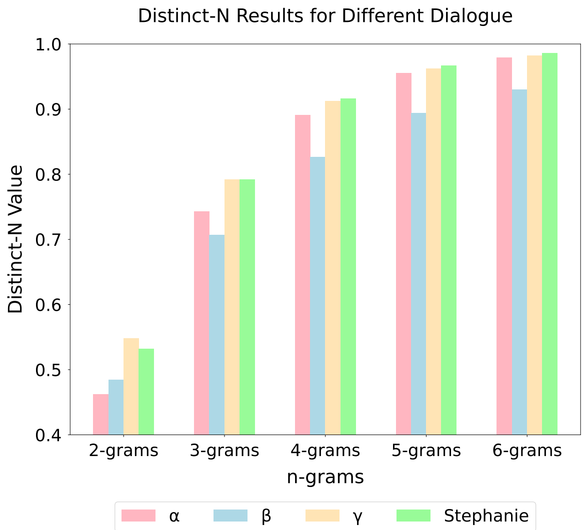

Figure 3: Distinct-N Results for Different Dialogue. This graph displays the lexical diversity of dialogues generated by various models, measured by the Distinct-N metric for n-grams from N=2 to N=6. Each colour represents a different dialogue model (α, β, γ, Stephanie), highlighting variations in linguistic complexity and diversity.

model cannot effectively distinguish or generate step-by-step dialogues.

Additionally, we conducted corresponding experiments with the Llama3-70b and phi3-3.8b models, generating their respective α, β, γ, and Stephanie.

Subsequently, we conducted automatic machine assessments of the three models on six metrics: Interesting, Informative, Natural, Engaging, On-topic, and On-persona, with Claude-3-sonnet as the assessment expert providing scores from 0 to 100, as shown in tables 1 and 2. The βs generated by the large models were generally weaker than the original dialogues on most metrics, with the exception of the 'Natural' metric for GPT-4, where β performed better than α. This indicates that single-step dialogues generated by large models are inferior to original human dialogues. The γ was significantly superior to α on all six metrics, demonstrating the superiority of the step-by-step dialogue paradigm. Stephanie showed further improvement over the β, highlighting the effectiveness of the further-split method. Additionally, we conducted a human evaluation of GPT-4, inviting three advanced graduate students majoring in English to score on a 0-5 scale. The results were positively consistent with the prior results.

We conducted further statistics on the β, γ, and Stephanie generated by GPT-4. Table 4 presents the statistics for α, β, γ, and Stephanie, including the average number of words per response (Words/Messages) and the Average Number of Consecutive Message Counts (ACMC). The results show that the β is similar to α, with β's Words/message being slightly higher than α's.

Metrics	α	β	γ	Stephanie
Interesting	2.93	2.85	3.53	**3.68**
Informative	3.71	3.13	3.78	**3.91**
Natural	2.97	2.89	3.65	**3.97**
Engaging	3.13	2.96	3.72	**4.06**
On-topic	–	3.30	3.79	**3.99**
On-persona	–	3.17	3.73	**3.89**

Table 3: Human evalution on GPT4. The table presents human evaluation scores for different dialogue models (α, β, γ, Stephanie) generated by GPT-4. Scores range from 1 to 5, with higher scores indicating better performance.

Metrics	α	β	γ	Stephanie
words/message	11.77	13.67	8.12	5.87
ACMC	1.07	1.08	1.99	2.51

Table 4: Words/message and ANT on dialogues. The table compares the average words per message and the Average Number of Consecutive Message Counts (ACMC) across different dialogue (α, β, γ, Stephanie). This helps in evaluating the verbosity and interaction depth of each dialogue.

Compared to α and β, γ has fewer Words/message and a higher ACMC, indicating that step-by-step dialogues tend to be shorter and contain more messages. Notably, Stephanie, in comparison to γ, further effectively reduces Words/Messages and significantly increases ACMC, demonstrating the effectiveness of the further-split method. Table 5 displays the proportion of consecutive message counts, where it is also evident that γ, compared to α and β, has more consecutive replies. Furthermore, Figure 3 illustrates that Stephanie effectively shifts the distribution of the number of consecutive messages to the right relative to γ.

In assessing lexical diversity among dialogue models, the "Distinct-N" table provides a comparative analysis using the Distinct-N metric for n-grams ranging from N=2 to N=6. As shown in fig 3, The Original Single-Step Dialogue α maintains high diversity, which increases with the complexity of n-grams, reflecting typical human dialogue characteristics. However, the Generated Single-Step Dialogue β exhibits lower diversity scores, especially for higher n-grams, indicating limitations in linguistic variability. Notably, the Generated Step-by-Step Dialogue γ and Further-Split Step-by-Step Dialogue (Stephanie) show superior performance, with Stephanie achieving the highest diversity across most categories. The significant performance of Stephanie in larger n-grams highlights the effectiveness of the further-split method in producing dialogues that are diverse and closely mimic the complex linguistic structures of human communication. This demonstrates that our proposed generation methods and prompting framework can significantly enhance the quality and human-likeness of machine-generated text.

Dialogues	one	two	three	four	five
α	92.65	7.35	0	0	0
β	91.26	8.74	0	0	0
γ	20.50	60.10	17.98	1.21	0.1
Stephanie	11.17	39.24	34.33	10.86	2.97

Table 5: The proportion of consecutive message counts. The table shows the proportion of dialogues with a given number of consecutive messages (one, two, three, four, five) for different dialogue models (α, β, γ, Stephanie). Higher counts indicate a greater tendency for step-by-step dialogues within an interaction.

Metrics	Stephanie-Llama3-8b	Llama3-8b
Interesting	**3.67**	3.01
Informative	**3.81**	3.22
Natural	**4.13**	3.57
Engaging	**3.89**	3.31

Table 6: Human evalution on Stephanie-Llama3-8b dialogue system. The table presents human evaluation scores for the Stephanie-Llama3-8b and Llama3-8b dialogue systems across four metrics: Interesting, Informative, Natural, and Engaging. Scores range from 1 to 5, with higher scores indicating better performance. The fine-tuned Stephanie-Llama3-8b model outperforms the Llama3-8b model across all metrics.

5.2 Fine-Tuning with Step-by-Step Dialogue

Following the demonstration of the effectiveness of our proposed paradigms, generation methods, and prompting frameworks, we aimed to provide a high-quality dataset for fine-tuning existing large models. To this end, we generated a high-quality step-by-step dialogue dataset consisting of 5,457 segments using the Llama3-70b model. Subsequently, we fine-tuned the Llama3-8b model with this dataset to create the Stephanie-Llama3-8b model. We engaged five testers to interact with both the Llama3-8b and Stephanie-Llama3-8b dialogue systems, assessing them across four metrics. The results show that the model fine-tuned with the step-by-step dialogue dataset exhibited superior step-by-step dialogue capabilities, outperforming the Llama3-8b model on all four metrics as presented in table 6.

6 Conclusion

The step-by-step dialogue paradigm introduced in this article enhances human-like interactions in simulated dialogue systems. By integrating a dual learning strategy and a further-split post-editing method, we have effectively generated dialogue data with Stephanie that is more interesting, natural, engaging, and emotionally nuanced. Our evaluations demonstrate that Stephanie's systems significantly outperform traditional single-step dialogue systems across various metrics. We plan to release our code, Stephanie dataset and Stephanie systems in the near future to facilitate chatbot eras.

Limitations

We conducted manual testing with limited human resources, and we look forward to seeing the application effectiveness of this technology on more large-scale consumer products.

Ethics Statement

We honour and support the NAACL Code of Ethics. The datasets used in this work are well-known and widely used, and the dataset pre-processing does not make use of any external textual resource. In our view, there is no known ethical issue. End-to-end pre-trained generators are also used, which are subjected to generating offensive context. However, the above-mentioned issues are widely known to commonly exist for these models. Any content generated does not reflect the view of the authors.

References

Rizwan Abbas, Bingnan Ni, Ruhui Ma, Teng Li, Yehao Lu, and Xi Li. Context-based emotion recognition: A survey. *Available at SSRN 4657124*.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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*.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Emily A Butler. 2011. Temporal interpersonal emotion systems: The "ties" that form relationships. *Personality and Social Psychology Review*, 15(4):367–393.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2023. Cooper: Coordinating specialized agents towards a complex dialogue goal. *arXiv preprint arXiv:2312.11792*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Tat-Seng Chua. 2024. Towards generative search and recommendation: A keynote at recsys 2023. In *ACM SIGIR Forum*, volume 57, pages 1–14. ACM New York, NY, USA.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

WANG Hongru, Rui Wang, Fei Mi, Yang Deng, WANG Zezhong, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In *International Conference of the Italian Association for Artificial Intelligence*, pages 151–171. Springer.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted llms as chatbot modules for long open-domain conversation. *arXiv preprint arXiv:2305.04533*.

Lizi Liao, Le Hong Long, Yunshan Ma, Wenqiang Lei, and Tat-Seng Chua. 2021. Dialogue state tracking with incremental reasoning. *Transactions of the Association for Computational Linguistics*, 9:557–569.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022a. On controlling fallback responses for grounded dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2591–2601, Dublin, Ireland. Association for Computational Linguistics.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022b. 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, Seattle, United States. Association for Computational Linguistics.

Jinjie Mai, Jun Chen, Bing Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. 2023. Llm as a robotic brain: Unifying egocentric memory and control. *arXiv preprint arXiv:2304.09349*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Fei Mi, Yasheng Wang, and Yitong Li. 2022. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11076–11084.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Weizhi Nie, Yuru Bao, Yue Zhao, and Anan Liu. 2024. Long dialogue emotion detection based on commonsense knowledge graph guidance. *IEEE Transactions on Multimedia*, 26:514–528.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. corr abs/2204.12749 (2022).

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Sina J Semnani, Violet Z Yao, Heidi C Zhang, and Monica S Lam. 2023. Wikichat: A few-shot llm-based chatbot grounded with wikipedia. *arXiv preprint arXiv:2305.14292*.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *arXiv preprint arXiv:2106.01702*.

Sandesh Swamy, Narges Tabari, Chacha Chen, and Rashmi Gangadharaiah. 2023. Contextual dynamic prompting for response generation in task-oriented dialog systems. *arXiv preprint arXiv:2301.13268*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Takato Yamazaki, Katsumasa Yoshikawa, Toshiki Kawamoto, Tomoya Mizumoto, Masaya Ohagi, and Toshinori Sato. 2023. Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models. *Advanced Robotics*, 37(21):1364–1381.

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? arxiv 2018. *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*.

A Examples Used in Figure 2

example 1 of single-step dialogue:{
role1: Hello, tell me about you.
role2: Hi, I am a mechanic who lives alone. How about you?
role1: I just got back from the London College of Fashion in the UK.
role2: Wow, that sounds really exciting! I have never traveled to Europe.
role1: You should go sometime.
role2: I hope to eventually. What do you like to do for fun?
role1: I thought I would make a pretty good fashion designer since I love to shop.
role2: I don't mind shopping too much, if it's the sporting goods store anyway.
role1: Maybe I could introduce you to my sister-in-law. She is all wrong for my brother.
role2: Does she like to fish? That's a deal breaker for me. I love fishing.
role1: She might. She thinks she's better than me because she has an actual job.
role2: Who says fashion designer isn't a real job?
role1: I haven't ventured into employment yet in fashion. Too much to do.
role2: You'll get there soon enough.
}

example 2 of single-step dialogue:{
role1: Hey, what's going on? How are you?
role2: Not much, just seen both my children off to school. You?
role1: About to smash some pretzels. Can't get enough of them.
role2: Haha. I love it. My 2 cats are big pretzel lovers.
role1: So you have two cats? What are their names?
role2: Bobby and Billie. I got them at the shelter I like to donate time to.
role1: Oh cool. Do you work? Any hobbies?
role2: My parents were both teachers, so I thought I might, but no.
role1: I live with my mom and watch a lot of Star Trek. Ever seen it?
role2: My 2 children love that show, but I would rather be out hunting.
role1: I am a bow hunter when I am not working on the railroad.
role2: We sound like we would get along.
role1: Yeah, I bet we would. I like hunting but I do not kill animals. I'm weird like that.
role2: Oh, that is weird. How do you hunt?
}

example 3 of single-step dialogue:{
role1: Hello there, how are you?
role2: I am doing well. How are you?
role1: I am great, thanks. Do you like boating?
role2: I like it when I can get away from my job at the grocery store.
role1: It's my favorite activity outside of being a doctor. Do you like beaches?
role2: I give deep sea fishing tours at the beach sometimes.
role1: I volunteer at a farm. Do you like animals?
role2: I am vegan, so I love animals. Same here with the farm volunteering.
role1: I too am a vegan. How long for you?
role2: Just 2 years. My boss at my bookkeeping job got me started.
role1: That's nice. My boss at the grocery store is nice too.
role2: Sounds like we have a lot in common. How do you feel about urban farms?
role1: I work at an urban farm. I love them. Do

you?
role2: Urban farms are the best. There's one on top of my apartment building.
role1: How lovely. Mine is 2 miles up the road.
role2: Is it all organic? I found that sometimes chemicals are necessary.
}

example 4 of single-step dialogue:{
role1: Hello, how is life treating you today?
role2: Okay... rough day waiting tables... Headed to class soon. And you?
role1: I bumped my head on a door frame because I am 6 feet tall.
role2: Wow, got some height on you. Do you work?
role1: Do you usually make bank waiting tables?
role2: Yes, people tip very well! But I really want to be a doctor someday.
role1: I do not work, I guess I am an investor. I have waited tables in the past.
role2: Okay, it's not a super fun gig but pays decent.
role1: I find people to help now. They come to me like stray cats.
role2: Oh, that could be a problem lol. What do you do for fun?
role1: I play around with Mensa members like myself.
role2: Oh okay... I am in a book club.
role1: When I was growing up, my family moved 40 times.
role2: Okay, that is ridiculous lol. Why so many moves?
role1: I think my mother was a bit touched in the head.
role2: Sounds a bit crazy, and hard on a child.
}

example 5 of single-step dialogue:{
role1: Hi, how are you doing today? Do you have any hobbies?
role2: I am good, and yourself, friend?
role1: I am okay. I was prepping dinner and stopped to complete a couple of hits.
role2: Very nice. Where do you work?
role1: I am a personal chef for a family of 4. What about you?
role2: I am in college to become a teacher.
role1: Oh, that is cool. The career highlight of my schooling was winning a spelling bee.
role2: Was that in 3rd grade?
role1: It was in the 2nd grade. I missed the cut in the 3rd grade. Haha.
role2: Dang, they do move on to 5-letter words then. rough.
role1: I know, right? That's why culinary school was where I ended up.
role2: Make sure you spell the food right! :D
}

example 1 of multi-step dialogue:{
role1: We all live in a yellow submarine.
role1: A yellow submarine.
role1: Morning!
role2: Hi!
role2: That's a great line for my next stand-up.
role1: Lol. I am shy.
role1: Anything to break the ice.
role1: And I am a Beatles fan.
role2: I can tell.
role2: I am not.
role2: You can see me in some TV shows.
role1: really? What shows?
role1: I like TV.
role1: It makes me forget I do not like my family.
role2: Wow, I wish I had a big family.
role2: I grew up in a very small town.
role1: I did too.
role1: I do not get along with mine.
role1: They have no class.
role2: Just drink some cola with rum.
role2: And you will forget about them!
role1: Put the lime in the coconut as well...
role2: Nah, plain Cuba Libre.
role2: That's what we drank yesterday.
role2: At the theater.
role1: I prefer mojitos.
role1: Watermelon or cucumber.
role2: Those are really yummy too.
role2: But not my favorite.
}

example 2 of multi-step dialogue:{
role1: Hi.
role1: How are you doing today?
role2: I am spending time with my 4 sisters.
role2: What are you up to?
role1: Wow, four sisters.
role1: Just watching Game of Thrones.
role2: That's a good show.
role2: I watch that while drinking iced tea.

role1: I agree.
role1: What do you do for a living?
role2: I am a researcher.
role2: I am researching the fact that mermaids are real.
role1: Interesting.
role1: I am a website designer.
role1: Pretty much spend all my time on the computer.
role2: That's cool.
role2: My mom does the same thing.
role1: That's awesome.
role1: I have always had a love for technology.
role2: Tell me more about yourself.
role1: I really enjoy free diving.
role1: How about you?
role2: Have any hobbies?
role1: Have any hobbies?
role2: I enjoy hanging with my mother.
role2: She is my best friend.
role1: That's nice.
role1: Moms are pretty cool too.
role2: I am also fascinated with mermaids.
}

example 3 of multi-step dialogue:{
role1: Hello.
role1: How are you today?
role2: I am well.
role2: How are you?
role1: I am very good.
role1: Did you watch the football games today?
role2: No.
role2: No, but my sons are watching a game.
role2: right now.
role1: What do you do for a living?
role1: I proofread for Hallmark.
role2: Cool...
role2: I followed my father.
role2: And became an author.
role1: Books are my greatest pleasure.
role1: I have a nice little library I am building.
role2: Awesome...
role2: Nothing like filling your mind with literature.
role1: I agree.
role1: Have you seen Goodfellas?
role2: Not much of a TV guy.
role2: Since I am always traveling.
role1: I catch the football and hockey games.
role2: Have you ever traveled before?
role1: I do.
role1: I travel extensively to Europe.
role1: And South America.
role2: Ireland and Australia are my go-to places.
role1: Ireland is very nice.
role1: Japan is my next stop.
role2: Never been there.
role2: I heard they are very friendly people.
}

example 4 of multi-step dialogue:{
role1: Hi, how are you doing?
role1: I am getting ready to do some cheetah chasing.
role1: To stay in shape.
role2: You must be very fast.
role2: Hunting is one of my favorite hobbies.
role1: I am!
role1: For my hobby.
role1: I like to do canning.
role1: Or some whittling.
role2: I also remodel homes.
role2: When I am not out bow hunting.
role1: That's neat.
role1: When I was in high school.
role1: I placed 6th in the 100m dash!
role2: That's awesome.
role2: Do you have a favorite season?
role2: Or time of year?
role1: I do not.
role2: But I do have a favorite meat.
role2: Since that is all I eat exclusively.
role2: What is your favorite meat to eat?
role1: I would have to say.
role1: It's prime rib.
role1: Do you have any favorite foods?
role2: I like chicken or macaroni.
role2: And cheese.
role1: Do you have anything planned for today?
role1: I think I am going to do some canning.
role2: I am going to watch football.
role2: What are you canning?
role1: I think I will.
role1: Can some jam.
role1: Do you also play football for fun?
role2: If I have time.
role2: Outside of hunting and remodeling homes.
role2: Which is not much!
}

example 5 of multi-step dialogue:{

role1: rock on.
role1: I am listening to my favorite band.
role1: Guns and roses.
role2: No kidding?
role2: I was just listening to the same thing.
role2: While taking a bath.
role1: Of course.
role1: I love to listen to rock.
role2: Man.
role2: My boxer just peed on the carpet!
role1: Well, I am into black everything.
role1: So at least.
role1: It would not show on my black carpet.
role2: Lol. I love black too!
role2: Guess I was playing my music too loud.
role1: I have a black car, purse.
role1: Wear all black.
role2: Maybe I can borrow something.
role2: As I am packing to visit my dad in China.
role1: Wow, does he live there.
role1: Or work?
role2: Live.
role2: Moved there about ten years ago.
role2: For a computer tech job.
role1: Have you visited him there before?
role2: Once. You cannot even throw a gum wrapper.
role2: Or you can get arrested.
role1: Sounds a bit scary.
role1: I've never been.
role2: Well not too much crime there.
role2: But a lot of people.
}