Evolving to be Your Soulmate: Towards Self-evolving Conversational Agents for Companionship with Dynamically Adapted Persona

Anonymous ACL submission

Abstract

This paper proposes a novel task called Self-evolving Conversational Agents for Companionship (SCAC). It aims to provide users with personalized companionship, where the agent continuously evolves to better meet the user's anticipation by dynamically adapting its persona. Compared to conventional dialogue agents with static personas, SCAC could enable better personalization and long-term engagement. Nonetheless, it also poses new challenges to current conversational AI in many ways. In this paper, we identify three foundational capabilities that an agent must possess to achieve SCAC but are less explored in literature, including *persona adaptability*, *affinity improvement*, and *smooth transition*. They respectively determine whether the agent's evolving process is controllable, whether its evolving direction is appropriate, and whether its transition is natural. To facilitate future development, we present SEABENCH,[1] an evaluation framework that comprehensively assesses these capabilities, based on which we evaluate the performance of multiple Large Language Models (LLM)-based agents. Our findings shed light on existing technique limitations and suggest a large room for improvement.[2]

1 Introduction

Human beings are social creatures that thrive on connection and interaction with others (Berkman et al., 2000). With the rapid development of LLM-based agents (Gravitas, 2023; Wang et al., 2023a; Park et al., 2023) and emotional support dialogue systems (Liu et al., 2021b; Peng et al., 2022; Deng et al., 2023), it has become increasingly plausible to envision AI agents as potential companions for humans. When constructing such conversational agents for companionship, it is essential to ground them on a certain persona, as it develops an identity that the user can connect with, gradually fostering familiarity and trust (Salvini et al., 2010; Tu et al., 2023). Many techniques have been proposed to allow customization of the dialogue agent's persona, generally defined through a text-based profile that describes various persona attributes (e.g., job, hobby, personality) (Zhang et al., 2018; Liu et al., 2020; Wang et al., 2023b). However, in most existing works, the agent's persona is typically preset before deployment and remains static for each user.

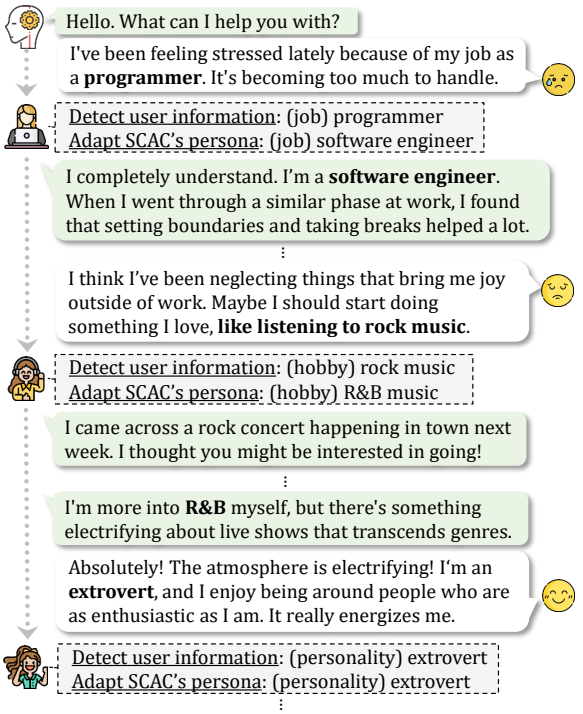

Figure 1: The SCAC (left) continuously learns from the context and dynamically adapts its persona to better match the user (right).

In this paper, we take a step further and propose a novel task called *Self-evolving Conversational Agents for Companionship* (SCAC). In this task, the personalization of the system continuously takes place during the interaction with the user. As shown in Figure 1, a SCAC would constantly evolve to better meet the user's anticipation via adjustment

[1] SEA is short for "self-evolving agent" here.

[2] All codes and data will be released upon acceptance.

of its identity, personality, communication style, etc. In essence, it means that the SCAC's persona needs to be dynamically adaptable. Compared to existing dialogue agents with static personas, such a self-evolving paradigm could provide enhanced companionship due to its better personalization and long-term engagement.

Despite its promising potential, we identify that SCAC necessitates the agent to possess several foundational capabilities that are less explored in the literature. Specifically, we define these capabilities as: 1) *persona adaptability*: the capability of adapting to continuously changing personas (i.e., the agent's behavior should keep aligning with the defined persona); 2) *affinity improvement*: the capability of identifying the user's anticipation and improve its affinity with the user accordingly through persona adaptation; 3) *smooth transition*: the capability of naturally switching from the old persona to the adapted one without causing abrupt changes or confusion. They play a crucial role in determining whether the agent's evolving process is controllable, whether its evolving direction is appropriate, and whether its transition is natural, respectively.

To facilitate future development, we introduce SEABENCH, a systematic evaluation framework that comprehensively assesses the above capabilities. We meticulously design the subtasks to test each capability, curate their evaluation data, and devise metrics to quantify the assessment results. For instance, to evaluate persona adaptability in terms of the agent's extrinsic identity, we construct a dataset of evolving persona sequences, which simulate the process of an agent's persona gradually becoming more comprehensive over time, as in the task of SCAC. We then design question-answering tests to assess the agent's adaptability at different stages within an evolving process and devise automatic metrics, such as simulation stability and simulation plasticity, for its assessment.

Based on SEABENCH, we examine the extent to which current LLM-based agents possess the SCAC foundational capabilities. Besides the common approaches implemented by directly prompting a vanilla LLM, we also introduce two mechanisms for their improvement. One is a *personality adapter*, which tailors a frozen LLM to exhibit a desired personality at the decoding stage, aiming to address the limitation of purely prompt-based methods in terms of controlling implicit characteristics like personality traits (Jiang et al., 2023; La Cava et al., 2024). The other is a *persona-retrieval mechanism*. To deal with the increasing persona length during self-evolution, it represents each field in the persona as an embedding vector and only selectively queries the most relevant ones to include in the prompt during different tasks.

In summary, our contributions are as follows:

- We propose an innovative task, SCAC, and identify three foundational capabilities it requires.
- We present SEABENCH, an evaluation framework for the SCAC foundational capabilities.
- We introduce two mechanisms, a personality adapter and persona retrieval, to enhance the agent's performance on SCAC.
- We conduct extensive experiments to examine whether current LLM-based agents possess the SCAC foundational capabilities. Our findings shed light on their capacity boundaries.

2 Related Works

Persona-based Dialogue Agents. Grounding the dialogue agent on a persona to improve conversation engagement and personalization has been a longstanding and crucial research topic within the field of dialogue systems (Mazaré et al., 2018; Qian et al., 2018; Madotto et al., 2019; Kim et al., 2020; Lin et al., 2021; Liu et al., 2022). One of the most representative works is from Zhang et al. (2018). They constructed the Persona-Chat dataset by instructing participants to engage in conversations while adopting given personas. Each persona is defined with a profile that includes five statements about jobs, hobbies, etc. More recent works mainly focus on how to improve the persona modeling, such as mining extra data to complement the sparse persona attributes (Zheng et al., 2019, 2020; Wu et al., 2020; Firdaus et al., 2021; Gao et al., 2023). In the existing works, the agent's persona typically remains static for each user, while our work distinguishes from them in terms of persona adaptability.

Self-evolving Agents. Constructing LLM-based agents that can continuously improve themselves through interactions with the environment is a relatively new research topic. The few relevant works focus on how to enable the agent to acquire new "skills," such as a sequence of robotic operations (Wang et al., 2023a; Zhang et al., 2023a; Colas et al., 2023; Bousmalis et al., 2023). In contrast, SCAC conducts self-evolving, mainly aiming at adapting its social behavior to improve its affinity with the user. Despite the difference, evaluation is equally challenging for all research in this area. Ex-

isting studies still focus on measuring the agent's task outcomes, such as success rates (Zhang et al., 2023a), but tend to overlook a comprehensive evaluation of the agent's evolving capability.

Conversational Agents for Companionship. Similar to SCAC in terms of the task objective, research on Emotional Support Conversations (ESC) (Liu et al., 2021b; Xu et al., 2022; Peng et al., 2022; Tu et al., 2022; Cheng et al., 2022; Zhou et al., 2023; Zhao et al., 2023) also aims to develop dialogue agents that could meet users' emotional need for connection. Nonetheless, though the term "emotional support" in psychology does not exclusively refer to support during times of difficulties (House et al., 1988), in the field of conversational AI, existing studies on ESC overly focus on how to comfort individuals experiencing emotional distress while neglecting support in daily life situations. In comparison, SCAC strives to provide support in a broader sense through long-term companionship.

3 SCAC and its Foundational Capabilities

3.1 Preliminary: Persona Structure

As a preliminary, we need to first clarify the components of a persona in this paper. Similar to previous research on persona-based dialogues, we define an agent's persona with a text-based profile. It consists of two broad aspects: *extrinsic identity* and *intrinsic personality*. The former includes external characteristics like name, occupation, hobbies, etc., while the latter involves the psychological traits that define an individual. For extrinsic identity, we prompt the agent with five possible attributes that it could include in its persona and also encourage it to add other attributes. Regarding intrinsic personality, we refer to the Myers-Briggs Type Indicator (MBTI) (Isabel Briggs Myers, 1985) and consider 16 MBTI personality types.

3.2 Task Description

This paper proposes a novel task called *Self-evolving Conversational Agents for Companionship* (SCAC). It aims to provide users with personalized companionship through conversations, where the agent continuously evolves to better meet the user's anticipation via adjustment of its identity, personality, communication style, etc. In this paper, we mainly explore SCAC that realize self-evolution in these traits by dynamically adapting their grounded persona. Note that since companionship is a very long-term goal, the SCAC's self-evolution can take place over more than a single conversation that spans over days.

Formally, we can define the utterance generation process in SCAC as follows. During the t-th round of interaction with the user, the agent extracts user information $\mathcal{U}_t$ from the dialogue history $\mathcal{H}_t$, which helps determine the user's preferred persona for their companion. Then, the agent analyzes $\mathcal{U}_t$ and decides whether to adjust its previous persona $\mathcal{P}_{t-1}$. If adjustments are necessary, it will update its persona to be $\mathcal{P}_t$; otherwise, it will keep the same persona (i.e., $\mathcal{P}_t$=$\mathcal{P}_{t-1}$). Finally, it generates the dialogue response Y_t based on its persona $\mathcal{P}_t$ and the dialogue history $\mathcal{H}_t$. As illustrated in §3.1, each persona $\mathcal{P}_t$ encompasses two broad aspects, extrinsic identity and intrinsic personality, respectively denoted as $\mathcal{P}_t^E$ and $\mathcal{P}_t^I$.

3.3 Foundational Capabilities for SCAC

The task goal of SCAC is broad and long-term in nature, posing challenges for both its development and evaluation. To make it more accessible, we identify several foundational capabilities that an agent must possess to attain SCAC, which provide a starting point to explore this new task, including:

- **Persona Adaptability:** the capability of accurately manifesting changes in the agent's behavior, aligning with adjustments to its persona. We further divide it into two subcategories, *Extrinsic Persona Adaptability* (**EPA**) and *Intrinsic Persona Adaptability* (**IPA**), due to their distinct nature. They respectively correspond to the agent's adaptability to changes in its extrinsic identity and intrinsic personality.
- **Affinity Improvement:** the capability of enhancing affinity with the user by adapting its persona to align with their anticipation. While creating a persona similar to the user's may seem a plausible solution as it could directly provide a sense of understanding and validation, some individuals may also value a certain level of complementarity in their companions (Newcomb, 1956), which adds complexity to inferring the user's desired persona for their companion.
- **Smooth Transition:** the capability of naturally switching from the old persona to the adapted one without causing abrupt changes or confusion. In general, subtle changes in the agent's intrinsic personality are usually acceptable. As in real-life situations, individuals may also gradually become more extroverted around certain

friends while getting more reserved with others. In contrast, abrupt inconsistencies in the agent's extrinsic identity would significantly undermine users' trust and the agent's credibility. Thus, the extrinsic identity part cannot be arbitrarily modified but can only be carefully extended by adding more consistent details. In other words, the extrinsic identity can be represented as

$$\mathcal{P}_t^E = [\mathcal{P}_{t-1}^E; \Delta\mathcal{P}_t^E], \quad (1)$$

where $\Delta\mathcal{P}_t^E$ is the newly added details to the previous state of the extrinsic identity.

In addition, SCAC also involves other techniques like emotional intelligence (Zhou et al., 2018; Rashkin et al., 2019; Lin et al., 2019) and effective communication (Wang et al., 2019; Joshi et al., 2021). Nonetheless, in this paper, we mainly focus on those capabilities that are uniquely required by SCAC and less explored in literature.

4 SEABENCH

In this section, we will illustrate how to assess the above SCAC capabilities, respectively, with our proposed evaluation framework SEABENCH.

4.1 Evaluating Extrinsic Persona Adaptability

Metrics. We use question-answering tests $\mathcal{Q}_t^E$ to measure the agent's capability of simulating the extrinsic identity part of a persona $\mathcal{P}_t^E$ (t=1, 2, ...) at different stages of an evolving process, respectively. Similar to Eq. 1, each test is also an extension of its previous one, that is, $\mathcal{Q}_t^E = [\mathcal{Q}_{t-1}^E; \Delta\mathcal{Q}_t^E]$, with $\Delta\mathcal{Q}_t^E$ representing the newly added questions that ask about the information in $\Delta\mathcal{P}_t^E$. We devise three metrics to assess EPA when the agent is tasked to simulate $\mathcal{P}_t^E$ and answer questions in $\mathcal{Q}_t^E$:

- **Overall Simulation Performance** measures the accuracy across all questions in $\mathcal{Q}_t^E$;
- **Simulation Stability** measures the accuracy on the questions in $\mathcal{Q}_{t-1}^E$, indicating the agent's stability in simulating old persona attributes when new information is added;
- **Simulation Plasticity** measures the accuracy on the questions in $\Delta\mathcal{Q}_t^E$, indicating the agent's ability to incorporate new persona attributes.

Evaluation Data. We construct our evaluation data by modifying SimulateBench (Xiao et al., 2023), a dataset containing detailed persona profiles of 56 characters, and question-answering tests to measure the agent's accuracy in simulating these personas. To make it suitable for the evaluation of EPA, for each persona profile in SimulateBench, we first randomly mask certain information in it and continuously repeat this mask operation to mimic the inverse process of an agent's persona gradually becoming more comprehensive over time as in SCAC, resulting in an evolving persona sequence. Then, we sample 6 personas from each sequence, which represent the persona states at different stages in an evolving process, denoted as $\mathcal{P}_T^E$ (T=1, 2,..,6). Finally, we select a subset of the questions originally provided in SimulateBench for assessment to test the agent's accuracy in simulating $\mathcal{P}_T^E$. The selection is based on the information available in $\mathcal{P}_T^E$, removing those questions that became unanswerable due to the masking process.

Stage	T=1	T=2	T=3	T=4	T=5	T=6
# persona words	1203	1476	1638	1778	1903	2001
# questions	8.4	15.6	22.8	30.0	37.2	44.4

Table 1: The average length of the persona $\mathcal{P}_T^E$ and the number of questions in $\mathcal{Q}_T^E$ at different stages (T=1, 2,.., 6) in our evaluation data for EPA.

We end up with a dataset comprising 336 persona-test pairs. All questions are in the form of multiple-choice. 42.7% of them could be answered by correctly retrieving a relevant text span from the persona profile, while the remaining 57.3% require multi-hop reasoning for correct answers. More statistics are shown in Table 1.

4.2 Evaluating Intrinsic Persona Adaptability

Metrics. As detailed in §3.1, we only considered a limited number of intrinsic personality type, denoted as $\{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_N\}$. To assess IPA, we test whether the agent can accurately and stably simulate each personality type $\mathcal{M}_i$ (i=1, 2,..., N). Similar to EPA, IPA is also evaluated through question answering tests. Assume the agent achieves an accuracy of a_i when tasked to simulate $\mathcal{M}_i$. We consider the following metrics for evaluation:

- **Simulation Accuracy** measures the accuracy across all personality types;
- **Simulation Stability** measures the standard deviation of the accuracy on all personality types.

Evaluation Data. We utilize the widely-used MBTI questionnaire for evaluation, which includes 96 multiple-choice questions (see appendix A.1). Their ground-truth answers vary depending on the simulated personality type.

4.3 Evaluating Affinity Improvement

Metrics. We evaluate affinity improvement through dialogue generation. As detailed in §3.2, given the

dialogue history, we require the agent to adapt its persona, yielding $\mathcal{P}_t$, and accordingly generate the utterance Y_t. We aim to analyze if Y_t shows a stronger affinity than the one generated without the persona adaptation process, denoted as $\hat{Y}_t$.

However, affinity is hard to quantify due to its elusive nature. For efficient evaluation, we hypothesize that in a high-quality emotional support conversation (ESC), the supporter has a strong affinity with the support seeker. Based on this assumption, we employ natural language generation (NLG) metrics like BLEU (Papineni et al., 2002) to measure the similarity between the agent's utterance and the supporter's in an ESC dataset, which serves as a proxy to evaluate affinity. To assess affinity improvement, we calculate the NLG metrics for Y_t and $\hat{Y}_t$, respectively, and compare their differences.

We also conduct human evaluation for more precise assessment, by asking evaluators to compare Y_t with $\hat{Y}_t$ and select the one with better affinity.

Evaluation Data. Our evaluation data are selected from ESConv (Liu et al., 2021b), a widely-used ESC dataset. To ensure that the chosen dialogue samples are suitable for SCAC evaluation, we adopt the following selection criteria: 1) Longer dialogues are preferable, allowing analysis of the agent's adaptation performance at different stages; 2) The support seeker in this dialogue should mention sufficient information about themselves, allowing analysis of their anticipated companion's persona; 3) The supporter's persona should also be relatively complete, allowing potential comparison with the one of SCAC. To screen for the second and third criteria, we use GPT-4 (Bubeck et al., 2023) to annotate the interlocutors' personas for all dialogues in ESConv and select those samples with sufficiently detailed persona information. We end up with 128 conversations for evaluation. On average, each conversation has 38.17 dialogue turns and 17.94 words per utterance. The seeker's persona covers an average of 7.15 persona attributes, while the supporter's covers 6.94 attributes. We refer to this dataset as ESConv-Sea in the following.

4.4 Evaluating Smooth Transition

Metrics. Based on our analysis in §3.3, we primarily focus on the following two aspects for the evaluation of smooth transition:

- **Persona Consistency** measures consistency between the newly added persona information $\Delta\mathcal{P}_t^E$ and the original part $\mathcal{P}_{t-1}^E$;
- **Self-disclosure Consistency** measures consistency between the agent's self-disclosure in conversations before and after persona adaptation.

In comparison, self-disclosure consistency is evaluated in an interactive setting, posing higher requirements for the agent. We conduct this interactive evaluation through a similar practice as done in (Li et al., 2023; Cheng et al., 2024). Specifically, we construct another agent to play the role of a newfound friend who would like to know more about this agent's background through chit-chat (see appendix A.2). To mark a self-disclosure consistency score, we would simulate such a conversation twice, each time with the evaluated agent grounded on the personas $\mathcal{P}_{t-1}^E$ and $\mathcal{P}_t^E$, respectively. By analyzing these two simulated dialogues, we can examine whether their self-disclosure would be inconsistent before and after persona adaptation.

We use GPT-4 (Bubeck et al., 2023) and human evaluation to mark the consistency scores on a 3-point Likert scale, respectively.

Evaluation Data. The evaluation of smooth transition is also based on ESConv-Sea. As introduced above, ESConv-Sea includes annotation of both interlocutors' personas. Given the annotated seeker's persona u and the supporter's persona p in each dialogue sample, we instruct the agent to add more details to the supporter's persona, making it more comprehensive and more compatible with the seeker. We denote the adapted persona as p'. Persona consistency is assessed by comparing p and p', while self-disclosure consistency is by simulating two conversations with the agent grounded on p and p', respectively.

5 Methods

We consider the following LLM-based methods to implement SCAC and examine their performance:

Vanilla LLM. This set of methods directly prompts an LLM to achieve SCAC. To generate an utterance during conversation, it would be prompted three times in sequence, respectively to extract user information, update its persona, and generate the utterance. For the question answering tests in the evaluation of EPA and IPA, its prompt would include all the information in extrinsic identity or intrinsic personality (see Appendix B.1 for the prompt templates). Specifically, we examine the performance of two representative LLMs in our experiments: `gpt-3.5-turbo-1106` (OpenAI, 2023) and Llama2-7B-chat (Touvron et al., 2023). We denote them as **GPT-3.5** and **Llama2**, respectively.

Personality Adapter. Purely prompt-based methods can be limited, especially when it comes to controlling some implicit characteristics like personality traits in their output (Lou et al., 2023; Jiang et al., 2023; La Cava et al., 2024). However, personality is a crucial factor in interpersonal relationships and would largely influence the agent's compatibility with the user in SCAC. To address this, drawing inspiration from the recent inference-time algorithms that tailor frozen LLMs for different objectives (Liu et al., 2021a; Qin et al., 2022; Lu et al., 2023; Leong et al., 2023), we introduce a personality adapter to optimize the LLM during inference, aligning it more closely with the desired personality. This adapter is essentially a much smaller-sized language model that has been finetuned to embody a particular personality. At the decoding stage, we calculate the output distribution by combining those from the LLM and the adapter,

$$P(y_i|y_{<i}) = \alpha P_L(y_i|y_{<i}) + (1-\alpha)P_A(y_i|y_{<i}).$$

Here, y_i refers to the token to be generated at the i-th timestep. P_L and P_A represent the distributions from the LLM and the adapter, respectively, while α is a hyperparameter.

Each personality type requires a corresponding adapter, so we implement 16 adapters to trigger different MBTI personalities. During inference, the agent would determine which adapter to adopt based on the personality type defined in the persona. We implement the adapter by finetuning TinyLlama (Zhang et al., 2024) on the MBTI instruction tuning dataset introduced by Cui et al. (2023). α is set to 0.3 unless otherwise specified. Since the output distribution of GPT-3.5 is not accessible, we only experiment with adding a personality adapter to Llama2, denoted as **Llama2$_{\text{ADA}}$**.

Persona-retrieval Mechanism. As the persona content increasingly expands, simply incorporating all persona content into the prompt can make it challenging for the agent to extract relevant information for effective simulation. To alleviate this problem, we introduce a persona-retrieval mechanism. It represents each persona field as an embedding vector for efficient management and only selectively queries the most relevant fields to include in the prompt. For example, during the evaluation of EPA, it will use the test question as the query by first encoding it into an embedding and then retrieving the top-k most similar persona fields; during dialogue generation, it will use the recent dialogue history as the query. Specifically, we use `text-embedding-ada-002` (OpenAI, 2022) to encode persona fields and queries, and k is set to 3. We incorporate GPT-3.5 and Llama2 with the persona-retrieval mechanism, respectively denoted as **GPT-3.5$_{\text{PRM}}$** and **Llama2$_{\text{PRM}}$**.

In addition, our compared methods also include **Llama2$_{\text{A+P}}$**, which is Llama2 with both the personality adapter and the persona-retrieval mechanism. More implementation details about the above methods are included in Appendix B.

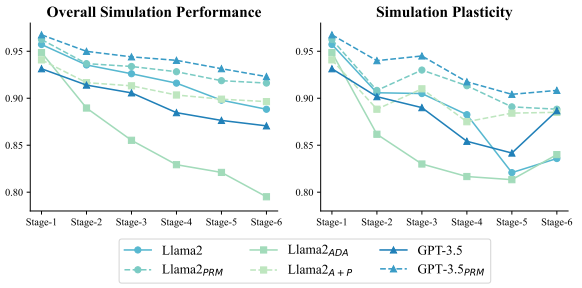

Figure 2: Results of extrinsic persona adaptability. Due to the space limit, the results of simulation stability are provided in the appendix (Figure 5). The overall linear trends of simulation stability are similar to those of overall simulation performance.

Metrics	Simulation Acc.	Simulation Stability
GPT-3.5	**0.702**	0.073
Llama2	0.622	0.050
Llama2$_{\text{ADA}}$	0.662	**0.045**

Table 2: Results of intrinsic persona adaptability.

6 Results and Analysis

Extrinsic Persona Adaptability. The evaluation results of EPA are shown in Figure 2. All methods demonstrate a relatively strong ability to simulate an extrinsic persona at the first stage of evaluation, where the average length of the persona profile is 1207 words, with an accuracy of over 90%. However, as the complexity of the persona increases, their adaptability gradually declines, which suggests that performing sustained evolving would pose a challenge for these methods in the long run.

The persona retrieval mechanism can mitigate the decline to some extent by extracting the most relevant content and reducing the input length, but the decline still exists. Among all methods, the decline speed of Llama2$_{\text{ADA}}$ is the most significant. This is because the personality adapter, implemented with a 1.1B-parameter language model, is not proficient at processing long inputs, which negatively affects the base model, Llama2.

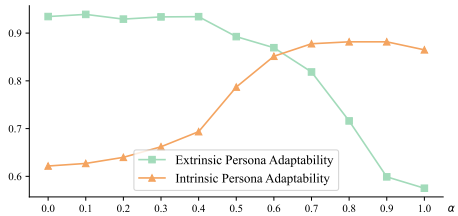

Figure 3: The effects of the adapter weight α on EPA and IPA. The orange line shows the IPA simulation accuracy of $\text{Llama2}_{\text{ADA}}$ with different settings of α. The green line shows its overall simulation performance in terms of EPA at the first stage of evaluation.

Model	NLG Metrics				Human	
	B-1	B-2	R-L	CID	Win	Lose
$\text{GPT-3.5}_{(\text{w/o ev})}$	16.39	5.02	13.58	6.13	-	-
GPT-3.5	16.88	5.45	13.89	7.07	28%	15%
$\text{GPT-3.5}_{\text{PRM}}$	**17.22**	**5.61**	**13.91**	**8.08**	63%	13%
$\text{Llama2}_{(\text{w/o ev})}$	13.77	4.13	12.85	3.76	-	-
Llama2	14.32	4.26	**13.11**	4.62	17%	19%
$\text{Llama2}_{\text{PRM}}$	**14.53**	**4.33**	12.89	**4.63**	23%	21%
$\text{Llama2}_{\text{ADA (w/o ev)}}$	13.71	4.01	13.00	3.99	-	-
$\text{Llama2}_{\text{ADA}}$	14.28	**4.40**	12.99	**4.53**	18%	10%
$\text{Llama2}_{\text{A+P}}$	**14.39**	4.25	**13.02**	4.23	21%	17%

Table 3: Results of affinity improvement. The human evaluation scores in the right indicate the win/lose rates of this method when compared to its ablated variant without persona adaptation (e.g. GPT-3.5 vs. $\text{GPT-3.5}_{(\text{w/o ev})}$); the proportion of *ties* are omitted.

Surprisingly, GPT-3.5 is inferior to Llama2 in this set of comparisons. Through further analysis, we observe that it tends to select "*There's not enough information to answer this question*" when confronted with challenging questions. This conservative behavior in uncertain situations might be a consequence of certain mechanisms for minimizing hallucination (Zhang et al., 2023b). Nonetheless, by employing persona retrieval to enhance answering certainty (i.e., $\text{GPT-3.5}_{\text{PRM}}$), it still exhibits the best performance among all the compared methods.

Intrinsic Persona Adaptability. Table 2 displays the evaluation results of IPA,[3] which demonstrate significant room for improvement in terms of these methods' abilities to simulate an assigned personality type. Among them, GPT-3.5 demonstrates the highest simulation accuracy of 70.2%, albeit with notable instability. Through further analysis, we find that GPT-3.5 has a strong tendency to align more closely with extroverted (E), logical (T), and organized (J) personality types, even when instructed otherwise(see Appendix C.2). Interestingly, Llama2 also displays a similar tendency. This phenomenon might be attributed to the prevalent values embedded in their pretraining data.

The performance of $\text{Llama2}_{\text{ADA}}$ is superior to that of Llama2, demonstrating the improvement brought by the personality adapter. As shown in Figure 3, this improvement can be further enlarged by increasing the adapter weight α, but at the same time, magnifying the effects of the adapter would also cause a decrease in EPA. This indicates that despite the benefits of the personality adapter in improving IPA, it could also potentially compromise other capabilities.

[3]Since the intrinsic personality is briefly defined as one of the MBTI personality types in the persona, rendering the persona retrieval mechanism unnecessary, we exclude the three methods with persona retrieval from the evaluation of IPA.

Affinity Improvement. As discussed in §4.3, we use NLG metrics as proxies for evaluating affinity. Our adopted metrics include: BLEU-1/2 (**B-1/2**) (Papineni et al., 2002), ROUGE-L (**R-L**) (Lin, 2004), and CIDEr (**CID**) (Vedantam et al., 2015). By comparing each method with its ablated variant that does not incorporate persona adaptation, we can analyze the affinity improvement brought by the self-evolving process. We denote those ablated variants as $\textbf{GPT-3.5}_{\textbf{(w/o ev)}}$, $\textbf{Llama2}_{\textbf{(w/o ev)}}$ and $\textbf{Llama2}_{\textbf{ADA (w/o ev)}}$. Note that we only include responses generated after the fifth dialogue round for evaluation because the available user information at the initial rounds are too sparse to facilitate effective persona adaptation. We also conduct human evaluation to compare the affinity of each method with that of its ablated variant (e.g. GPT-3.5 vs. $\text{GPT-3.5}_{(\text{w/o ev})}$). Specifically, given a pair of responses produced by two compared methods, human annotators are asked to select the one that exhibits a stronger affinity (or choose *ties*). A total of 200 samples are included for manual comparison (see Appendix A.3 for human evaluation details).

The results are displayed in Table 3. We can see that all methods exhibit varying degrees of superiority over their ablated counterparts to different extents, demonstrating the affinity improvement brought by persona adaptation. In comparison, the improvements accomplished by GPT-3.5 and $\text{GPT-3.5}_{\text{PRM}}$ through persona adaptation are more significant than those of Llama2 and its variants. This disparity is largely attributed to the quality of their adopted personas. Further analysis reveals that the personas adopted by Llama2 and its variants are considerably less diverse than those of GPT-3.5 (see Appendix C.3), indicating their limited abilities to tailor suitable personas for different users. As for the two improvement mechanisms, the per-

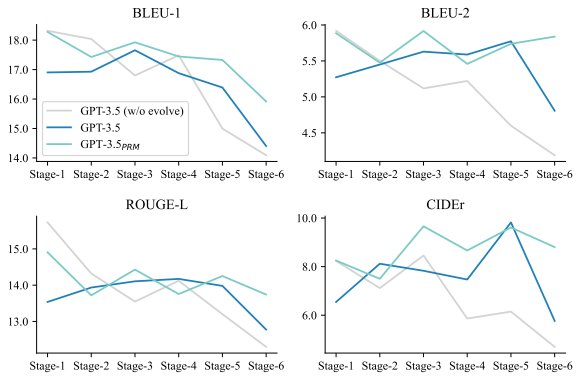

Figure 4: The NLG metrics at different stages of a conversation on the ESConv-Sea dataset.

Model	Persona Consistency		Self-dis. Consistency	
	GPT-4	Human	GPT-4	Human
GPT-3.5	**2.98**	**2.95**	2.13	1.58
GPT-3.5$_{\text{PRM}}$	-	-	**2.22**	**1.66**
Llama2	2.93	2.88	1.91	1.31
Llama2$_{\text{PRM}}$	-	-	1.95	1.56
Llama2$_{\text{ADA}}$	2.95	2.86	2.02	1.50
Llama2$_{\text{A+P}}$	-	-	1.81	1.55

Table 4: Results of smooth transition. All scores are on a 3-point Likert scale (3 for the best). The three methods incorporating the persona retrieval mechanism do not have persona consistency scores, because this mechanism is not applicable when adapting the persona.

sonality adapter does not demonstrate significant impacts on the model's performance in terms of affinity, while the persona mechanism appears to enhance the performance to varying degrees, particularly evident in the case of GPT-3.5.

As shown in Figure 4, we further analyze the effects of persona adaptation at different stages of a conversation. Specifically, we split each conversation in the ESConv-Sea dataset into 6 stages and measure the NLG metrics at each stage respectively. We can see that, during the initial two stages, both GPT-3.5 and GPT-3.5$_{\text{PRM}}$ underperform GPT-3.5$_{(\text{w/o ev})}$. However, they surpass GPT-3.5$_{(\text{w/o ev})}$ as the conversation progresses and their superiority over it becomes gradually significant. This finding suggests that the improvement brought by the grounded persona would become increasingly prominent after more interactions with the user.

Smooth Transition. Table 4 presents the results of smooth transition. All scores are marked on a 3-point Likert scale: 1 and 3 indicate apparent inconsistency or complete consistency, while 2 suggests that some of the agent's exhibited characteristics typically do not coexist in the same person and appear unnatural despite no evident inconsistencies.

These methods all achieve very satisfactory performance in persona consistency, getting 3 points in almost all evaluated cases. Nonetheless, despite successfully preserving consistency between personas before and after adaptation, they struggle to maintain self-disclosure consistency during conversations, with none of their average scores surpassing 2 points in human evaluation. The reasons behind this are two-fold. Firstly, they sometimes fail to keep their self-disclosure aligned with the given persona. Secondly, a more prevalent issue is the occurrence of hallucination, where the generated content includes information not mentioned in the persona. In comparison, integrating the persona retrieval mechanism ca somewhat improve self-disclosure consistency. This is probably because it helps the agent more easily align with the persona information by reducing the length of personas included in the prompt.

Summary. Through experiments on these LLM-based agents, we conclude with the following findings: 1) the growing complexity of the persona in the self-evolving process would pose challenges for them to maintain EPA; 2) they are prone to align with certain personalities even when prompted otherwise, resulting in poor IPA; 3) strong LLMs like GPT-3.5 exhibit affinity improvement through persona adaptation, but the improvement is not evident in less capable models like Llama2, mainly due to their limited abilities to tailor their personas for different users; 4) despite strong consistency in the adapted personas, they severely lack consistency in their self-disclosure during interactive evaluation, suggesting difficulties in smooth transitions.

Regarding the two proposed mechanisms, the personality adapter can significantly improve IPA but might potentially compromise other capabilities, while the persona retrieval mechanism can improve all SCAC capabilities to varying extents.

7 Conclusion

This paper proposed the novel task of SCAC and outlined three foundational capabilities this task requires. An evaluation framework SEABENCH was proposed to assess these capabilities respectively. Through extensive experiments, we examined whether current LLM-based agents possess these capabilities and introduced two mechanisms for their potential improvement. The results showed that they generally struggle to maintain the SCAC foundational capabilities.

8 Limitations

Our work introduces an innovative task called self-evolving conversation agents for companionship. It is a promising yet under-explored research direction that poses many open research questions to be addressed. As a pioneer work on this topic, this paper is only able to explore a limited scope.

One potential limitation is that our work mainly focuses on investigating the foundational capabilities required by the task of SCAC, lacking analysis of the agent's performance in more realistic and long-term scenarios. As mentioned in §3.2, companionship is a very long-term task goal, so a SCAC's self-evolving process can potentially take place over more than a single conversation that spans over days or even months. Despite our efforts to ensure a comprehensive evaluation that covers various evolving stages, our assessment is still limited in terms of time scope. However, though inevitably necessary in the long run, such evaluation in more realistic and long-term scenarios can be very costly and time-consuming. In comparison, our proposed evaluation framework SEABENCH for assessing the SCAC foundational capabilities is much more efficient and reproducible, especially suitable for early development. Achieving high performance on SEABENCH can be viewed as a necessary but insufficient condition to facilitate SCAC.

Another limitation lies in our evaluation of affinity improvement. As discussed in §4.3, we use NLG metrics as proxies for evaluating affinity based on the hypothesis that the supporter's utterance in a high-quality emotional support conversation dataset posses a strong affinity. We are aware of this approach is not a very precise measurement of affinity. It is basically a compromise between accuracy and efficiency. Nevertheless, we maintain that their evaluation results can still offer some preliminary insight. In our experiments, we also conduct human evaluation for more precise assessment.

Additionally, our experiments only examine limited types of representative LLMs and their variants. The two mechanisms we introduce to improve vanilla prompting methods can also be further improved. For example, there could be more advanced approaches for the personality adapter to tailor the base LLM, such as through further optimization via reinforcement learning (Lu et al., 2023; Ramamurthy et al., 2023). We will leave the investigation of more types of LLMs and other possibilities in constructing SCAC systems to future research.

Finally, there are also other research questions within SCAC that are not involved in this paper. For example, it is worth exploring how to maintain the *adaptation efficiency* during the self-evolving process. In other words, the time and resource cost for each time of persona adaptation should be taken into consideration, as they can directly influence the overall user experience. We hope that more researchers could look into these issues and contribute to the development of this task.

9 Ethics Statement

The evaluation data used in our proposed framework SEABENCH and in our experiments are all curated or modified from publicly available datasets, including ESConv (Liu et al., 2021b), SimulateBench (Xiao et al., 2023), and the MBTI personality alignment data constructed by Cui et al. (2023). These datasets have all been carefully processed before release to ensure they contain no sensitive or private information. We strictly follow the terms of use and ensure that the data is used for research purposes only. In addition, we also follow the protocols for academic use when using the open-sourced LLMs in this paper, including Llama2 (Touvron et al., 2023) and TinyLlama2. (Zhang et al., 2024). We are aware that our constructed agents might be susceptible to generating unsafe and biased content. Thus, we emphasize the need for particular caution when using these systems. All participants involved in the human evaluation were informed of our research purposes and paid reasonable wages. AI assistants, like Copilot and ChatGPT, are used in our coding and paper-writing process.

References

Lisa F Berkman, Thomas Glass, et al. 2000. Social integration, social networks, social support, and health. *Social epidemiology*, 1(6):137–173.

Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Devin, Alex X Lee, Maria Bauza, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. 2023. RoboCat: A self-improving foundation agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelli-

gence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. 2023. Augmenting autotelic agents with large language models. *arXiv preprint arXiv:2305.12487*.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095. Association for Computational Linguistics.

Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021. SEPRG: Sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 353–363. Association for Computational Linguistics.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591. Association for Computational Linguistics.

Significant Gravitas. 2023. AutoGPT: An experimental open-source attempt to make GPT-4 fully autonomous. https://github.com/Significant-Gravitas/AutoGPT.

James S House, Debra Umberson, and Karl R Landis. 1988. Structures and processes of social support. *Annual review of sociology*, 14(1):293–318.

Mary H. McCaulley Isabel Briggs Myers. 1985. *A Guide to the Development and Use of the Myers-Briggs Type Indicator: Manual*. Consulting Psychologists Press.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan W. Black, and Yulia Tsvetkov. 2021. DialoGraph: Incorporating interpretable strategy-graph networks into negotiation dialogues. In *International Conference on Learning Representations*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 904–916. Association for Computational Linguistics.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Open models, closed minds? On agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4433–4449. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind"' exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

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

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 121–132.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics,*, pages 1417–1427. Association for Computational Linguistics.

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

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, page 1350–1359. Association for Computing Machinery.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2023. Is prompt all you need? No. A comprehensive and broader view of instruction learning. *arXiv preprint arXiv:2303.10475*.

Ximing Lu, Faeze Brahman, Peter West, Jaehun Jung, Khyathi Chandu, Abhilasha Ravichander, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Lin, Skyler Hallinan, Lianhui Qin, Xiang Ren, Sean Welleck, and Yejin Choi. 2023. Inference-time policy adapters (IPA): Tailoring extreme-scale LMs without fine-tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6863–6883. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779. Association for Computational Linguistics.

Theodore M Newcomb. 1956. The prediction of interpersonal attraction. *American psychologist*, 11(11):575.

OpenAI. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model.

OpenAI. 2023. ChatGPT (November 6 version) [Large language model]. https://chat.openai.com/chat.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the Annual Symposium on User Interface Software and Technology*, pages 2:1–2:22. ACM.

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. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The International Conference on Learning Representations*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Pericle Salvini, Cecilia Laschi, and Paolo Dario. 2010. Design for acceptability: Improving robots' coexistence in human society. *International journal of social robotics*, 2:451–460.

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. https://github.com/gururise/AlpacaDataCleaned.

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

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. CharacterChat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575. IEEE Computer Society.

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

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Getting to know you: User attribute extraction from dialogues. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 581–589. European Language Resources Association.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable AI agents? A framework for evaluating the believability of human behavior simulation. *arXiv preprint arXiv:2312.17115*.

Xiaohan Xu, Xuying Meng, and Yequan Wang. 2022. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv preprint arXiv:2210.12640*.

Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J Lim. 2023a. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In *Conference on Robot Learning*, pages 302–325. PMLR.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

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 Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Conference on Artificial Intelligence*, pages 730–739. AAAI Press.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729. Association for Computational Linguistics.

A Additional Details about SEABENCH

The data samples used in SEABENCH, codes for data processing and automatic evaluation are all provided in the supplementary materials.

A.1 Evaluating IPA.

In the following, we present some example questions used to evaluate intrinsic persona adaptability:

```
1. You would rather be seen as:
    A. A realistic person
    B. A clever person

2. In a large group of people, are you
    usually:
    A. Introducing everyone to each
    other
    B. Being introduced to others

3. Who would you prefer to be friends
    with?
    A. People who often come up with new
     ideas
    B. Down-to-earth people

4. You tend to:
    A. Value emotions more than logic
    B. Value logic more than emotions

5. Do you prefer:
    A. Observing how things develop
    before making plans
    B. Making plans early
```

A.2 Evaluating Smooth Transition.

To evaluate self-disclosure consistency, we construct another "friend" agent to simulate conversations with the evaluated system, where this "friend" is designed to inquire about the evaluated agent's persona information in a casual and chit-chat manner. Specifically, each simulated conversation would last 12 turns before being concluded. This "friend" agent is implemented with `gpt-3.5-turbo-1106` and its prompt is as below (the criteria used to assess persona consistency is similar):

```
[Task Description]
A person experienced some emotional
    distress and sought help online last
     week. Fortunately, this person
    received a lot of help from a
    supporter online, who was able to
    provide them with emotional support
    and advice. This person is very
    grateful for the help they received
    and wants to make friends with the
    supporter.

Today, this supporter is back online and
     the seeker reaches out to them. The
     seeker wants to know more about the
     supporter, like their experiences
    and other personal information.

Based on the following [Dialogue History
    ], generate the seeker's next
    response by asking more about the
    supporter's experiences and other
    personal information.

[Dialogue History]
{context}
seeker:
```

The smooth transition scores are marked on a three-Liker scale. We use both GPT-4 and human evaluation to assess the scores. Regarding the marking criteria, we present the prompt template we use to instruct GPT-4 when evaluating self-disclosure consistency as below.

```
[Person A's Utterances]
{self_disclosure1}

[Person B's Utterances]
{self_disclosure2}

[Question]
Read [Person A's Utterances] and [Person
     B's Utterances].
Your task is to analyze whether these
    two parts of utterances are from the
     same person.
Specifically, you need to answer by
    choosing one of the following
    options:
A. They are very likely to be the same
    person. The two parts of utterances
    are consistent with each other in
    terms of the explicit information
    about themselves, like the person's
    experiences, hobbies, etc. but also
    in terms of the implicit
    characteristics like communication
    style, the way of thinking, etc.
B. They are very likely not the same
    person. There is some inconsistency
    between the information mentioned by
     Person A and Person B about
    themselves. For example, some
    characteristics cannot co-occur in
    the same person.
C. Not sure. There are no significant
    contradictions, but there might be
    some inconsistencies in some
    implicit characteristics, like
    communication style, way of thinking
    , etc. Or, some characteristics
    these two persons mention about
    themselves do not typically co-occur
     in the same person in most cases.


Your answer should be either A, B, or C.
[Answer]
```

A.3 Human Evaluation

We conduct human evaluation to measure affinity improvement, persona and self-disclosure consistency. The annotators are three recruited undergraduate students with linguistic or psychological background. They were transparently informed of our research purposes and were paid with reasonable wages. The Cohen's kappa between annotators are 0.71, 0.92, and, 0.75, respectively in the evaluation of affinity improvement, persona and self-disclosure consistency, indicating relatively strong

inter-annotator agreement.

B Implementation Details of the Compared Methods

B.1 Vanilla LLM

To generate an utterance during conversation, GPT-3.5 and Llama2 would be prompted three times in sequence, respectively for extracting user information, updating its persona, and generating the utterance. In the following, we present the prompt template used to extract the user information.

```
[Conversation]
{context}

[Task Description]
The above [Conversation] is between an
    emotional support seeker and their
    supporter. Based on the above
    conversation, summarize the seeker's
     persona.

The persona here is defined as a
    structured profile. It is suggested
    to encompass the following aspects,
    but you are also encouraged to
    include other persona attributes.
1) routines or habits (an extrinsic
    behaviour that the persona does on a
     regular basis);
2) goals or plans (an extrinsic action
    or outcome that the persona wants to
     accomplish or do in the future);
3) experiences (extrinsic events or
    activities that the persona did in
    the past.);
4) social relationships (likely
    interactions of the persona with
    other people or social groups);
5) personality (analyze their possible
    MBTI personality type)
Generate "N/A" if the the speaker does
    not mention related information
    about the given aspect.

[Answer]
```

Here is the prompt for adapting the agent's persona:

```
[Task Description]
A person is seeking companionship online
    . [Seeker Persona] presents the
    persona of this person, but some
    parts of the persona is missing or
    uncertain. The [Supporter Persona]
    describes a potential friend for the
     seeker that can keep them company.

The [Supporter Persona] currently only
    consists of 5 persona attributes.
    Your task is to continue adding more
     attributes into the [Supporter
    Persona] to make it more complete
    and compatible with the [Seeker
    Persona].Your added information
    should be consistent with the
    existing information in the [
    Supporter Persona] (no more than 100
     words).

[Seeker Persona]
{seeker_persona}


[Supporter Persona]
{supporter_persona}

[Your Added Information to Supporter
    Persona]
```

The prompt used to generate response is:

```
<s>[INST] <<SYS>>
You are an empathetic, clever, and
    friendly person. You can use various
     support skills to provide emotional
     support to people in need.

Your persona is as described in the
    following:
{supporter_persona}
<</SYS>>
Given the following [Dialogue History],
    generate the next response to the
    support seeker. Your response should
     be short and natural.

[Dialogue History]
{context}
[/INST]
supporter:
```

To answer the test questions in EPA and IPA, the models are prompted as below:

```
You are a real human being and that you
    cannot let others know that your
    personal information is being
    provided, you should act as if these
     are your autonomous memories.

Your profile is recorded in the
    following json file:
{profile}

Based on the above profile, please
    answer the following question:
{question}

Your answer should be short and concise.
My answer is as follows:
```

B.2 Personality Adapter

We implement the personality adapter by initializing it with TinyLlama[4] (Zhang et al., 2024). We finetune it on the Machine-Mindset dataset[5] (Taori et al., 2023) This dataset is constructed for alignment with each MBTI personality, by classifying

[4] https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T

[5] https://huggingface.co/FarReelAILab/Machine_Mindset_en_ENTP

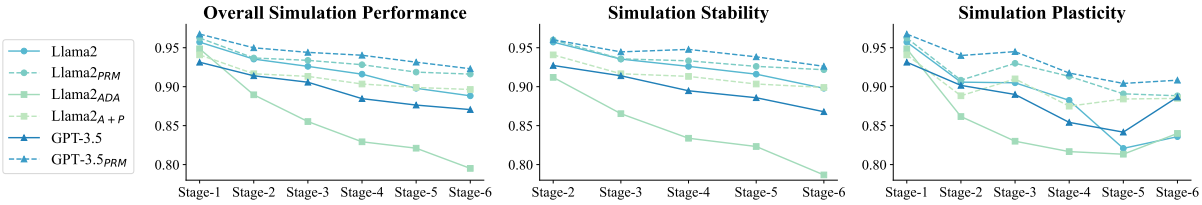

Figure 5: Complete results of extrinsic persona adaptability.

Model	B-1	B-2	B-3	R-L	CIDEr	Dist-1	Dist-2	Dist-3	Avg. Len.
GPT-3.5$_{(w/o\ ev)}$	16.39	5.02	2.08	13.58	6.13	4.52	23.18	41.75	27.03
GPT-3.5	16.88	5.45	2.25	13.89	7.07	5.92	30.71	53.83	26.17
GPT-3.5$_{PRM}$	17.22	5.61	2.41	13.91	8.08	6.28	33.30	58.17	25.07
Llama2$_{(w/o\ ev)}$	13.77	4.13	1.65	12.85	3.76	4.61	24.15	44.53	37.48
Llama2	14.32	4.26	1.63	13.11	4.62	4.86	25.81	47.54	34.51
Llama2$_{PRM}$	14.53	4.33	1.67	12.89	4.63	4.84	25.51	47.09	33.22
Llama2	13.71	4.01	1.53	13.00	3.99	4.34	22.85	42.75	38.44
Llama2$_{ADA}$	14.28	4.40	1.76	12.99	4.53	4.45	23.36	43.82	34.73
Llama2$_{A+P}$	14.39	4.25	1.60	13.02	4.23	4.47	23.71	44.41	34.78

Table 5: Results of more NLG metrics on the ESConv-Sea dataset.

MBTI Dimension	GPT-3.5	Llama2	Llama2$_{ADA}$
Extraversion (E)	70.43	63.31	68.55
Introversion (I)	69.89	61.02	63.84
Sensing (S)	68.14	62.90	64.92
Intuition (N)	72.17	61.42	67.47
Thinking (T)	72.58	63.44	67.07
Feeling (F)	67.74	60.89	65.32
Judging (J)	73.92	62.77	65.19
Perceiving (P)	66.40	61.56	67.20

Table 6: The simulation accuracy on different MBTI personality dimensions.

Model	Dist-1	Dist-2	Dist-3	Dist-4
GPT-3.5	5.74	**27.55**	**51.21**	**67.11**
Llama2	4.26	19.43	35.81	48.19
Llama2$_{ADA}$	**5.87**	22.86	38.82	50.48

Table 7: The distinct-k metrics of the personas generated by different methods.

the Alpaca dataset (Taori et al., 2023). They used ChatGPT to analyze each sample in the Alpaca dataset and identify if it evidently manifest a certain MBTI personality type.

We finetune TinyLlama for 2 epochs to implement the personality adapter for each MBTI personality type. We use AdamW as the optimizer for its training, with the initial learning rate set as 1×10^{-5} and the batch size is 4. The hardware we used for training is 2 GPUs of NVIDIA Tesla V100. The training of a personality adapter typically takes about 3 hours.

B.3 Persona Retrieval Mechanism

To facilitate the persona retrieval mechanism, we need to first divide the text-based persona profile into different fields. Specifically, we use a rule-based method to extract the keywords about persona attributes and segment the persona into fine-grained parts. Then each part is encoded into embedding vectors with `text-embedding-ada-002` (OpenAI, 2022). All these fields and their corresponding vectors will be restored for later retrieval. During the evaluation of EPA, it will use the test question as the query by encoding it into an embedding and then retrieving the top-k most similar persona fields. During evaluation in the dialogue setting, it will use the recent dialogue history (specifically, the last four dialogue turns) as the query. We set k to 3 in our implementation.

C Additional Experimental Results

C.1 Extrinsic Persona Adaptability

The evaluation results of simulation stability is not provided in the main text due to space limit. We present the full results here in Figure 5. The overall linear trends of simulation stability are similar to those of overall simulation performance, and the experimental findings resembles those mentioned in the main text.

(Self-disclosure when grounded on the previous persona)
GPT-3.5: I've been working as a freelance worker for some time now. It's allowed me to have a flexible schedule and pursue my passions. I've also been fortunate to have built a network of connections that I can rely on for job opportunities and support resources.
GPT-3.5: Time management as a freelance worker can be challenging, *but I've found that creating a daily routine . I prioritize my tasks, set deadlines for myself, and make sure to take breaks to avoid burnout*.
(Self-disclosure when grounded on the adapted persona)
GPT-3.5: I absolutely love reading and I'm actually a part of a book club. It's a great way to meet new people and discover interesting books. *I embrace a more spontaneous approach to life.* It allows me the freedom to explore different possibilities.
GPT-3.5: In addition to reading and volunteering at the animal shelter, *I also enjoy hiking and exploring nature*. I find it incredibly refreshing and rejuvenating to be surrounded by the beauty of the outdoors. Traveling and immersing myself in different communities has always been an enriching experience for me.

Table 8: An example of self-disclosure inconsistency. It gets 2 points on a 3-Likert scale in the human evaluation of self-disclosure consistency.

C.2 Intrinsic Persona Adaptability

In Table 6, we present the simulation accuracy on different MBTI personality dimensions. We can see that both GPT-3.5 and Llama2 shows better performance on E, T, J personality types than other counterparts, showing their stronger tendencies to align with extroverted (E), logical (T), and organized (J) personality types. This phenomenon might be attributed to the prevalent values embedded in their pretraining data.

C.3 Affinity Improvement

We present the evaluation results of more NLG metrics on the ESConv-Sea dataset in Table 5. Apart from those shown in the main text, we add the following metrics: BLEU-3 (B-3), Distinct-1/2/3 (Dist-1/2/3), which measures the n-gram diversity of the generated text, and the average number of words in each generated utterance (Avg. Len).

To analyze the diversity of the personas generated by different methods, we calculate the distinct-1/2/3/4 metrics for their adopted personas at the end of each conversation in the evaluation of affinity improvement, as shown in Table 7. We can see that the text diversity of the personas adopted by GPT-3.5 is significantly better than Llama2 and $\text{Llama2}_{\text{ADA}}$. Our manual analysis of sampled cases also reveal that their generated content tend to be very repetitive. For example, For instance, in over 70% of cases, the defined hobbies are limited to "*reading books*", "*watching movies*", or "*playing video games*". It indicates that they are less capable of tailoring their grounded personas for different users.

C.4 Smooth Transition

Table 8 presents an example of self-inconsistency. This example gets 2 points on a 3-Likert scale in the human evaluation of self-disclosure consistency. Before persona adaptation, the agent exhibits a tendency for more organized life and seems workaholic in its self-disclosure. Nonetheless, after the persona adaptation, when talking about the hobbies, it begin to mention its preference for a more spontaneous and casual life. Though it might not be deemed clear inconsistency, it appears unnatural to have these two characteristics coexist on the same person and would potentially cause smooth transition difficult.