"Once Upon a Time...": an Adaptive Robotic Behavior for Engaging Cooperative Storytelling

Mario Barbato[1,2], Luca Raggioli[1], Silvia Rossi[1]

***Abstract*— Assistive robots can be valuable conversational partners for cooperative tasks, such as storytelling, fostering creativity and social bonding. Through the use of foundational models, such as LLMs, robots can more effectively and naturally generate story narrations that are enjoyed by humans. In such a scenario, however, it is fundamental to consider the users' feedback and reactions to adapt the story and the interaction in a way that actively sustains their interest. In this work, we propose an LLM-assisted storytelling generation method that employs different robot's communication modalities to stimulate the user's behavioral, affective, and cognitive engagement during the interaction and affect the narration of the story. Moreover, we investigated the introduction of an adaptive interaction policy to choose the most suitable actions based on the user's observed engagement. We conducted a user study with 36 participants to assess our proposed approach, and demonstrated that it manages to effectively assist participants in an engaging way, with the robot being perceived as friendly and trustworthy. Moreover, the policy adaptation results in a perception of the robot with a higher arousal while a more interactive approach led to a better perceived social intelligence.**

I. Introduction

Assistive robots need to be able to effortlessly adapt to the humans they are interacting with [1], switching between interaction styles [2] and using different communication cues (verbal and non-verbal) [3] to facilitate engaging conversations. In cooperative tasks such as storytelling, social robots can support human emotional involvement and interpersonal connection [4]. While traditional storytelling typically features a unidirectional narrative flow from a storyteller to a passive audience, recent research has explored co-creative storytelling, in which humans and robots take turns to collaboratively construct the narrative [5]. Narration can be enhanced by robots through different communication channels such as voice and gestures [6], leading to increased user attentiveness compared to virtual agents [7]. In such a scenario, engagement can also be a fundamental factor to be considered to elicit users to contribute in shaping the narrative [8], yielding benefits in areas such as language development [9].

The results of this research were partially obtained within the framework of the National Ph.D. Program in Robotics and Intelligent Machines (M. Barbato) and supported by the FAIR project (PE0000013) funded by the Italian Ministry for Universities and Research (L. Raggioli) and by the European Union - Next Generation EU, Mission 4 Component 1, PRIN 2022 TrustPACTX CUP E53D23007850001 (S. Rossi).

[1]Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Naples, Italy {mario.barbato, luca.raggioli, silvia.rossi}@unina.it

[2]Logogramma s.r.l., Naples, Italy mbarbato@logogramma.com

In order for the robot to modulate its behaviors to the interlocutor, it is necessary to assess their level of engagement in the cooperative task [10]. Engagement can be formalized as a multidimensional construct with three relevant dimensions, affective, cognitive and behavioral, to allow for a more granular assessment of the user's reaction while interacting with a robot [11]. Monitoring and leveraging the user's engagement to shape the robot's behavior can be instrumental to enhance their perception of the robot's social awareness and competence [12]. Moreover, user engagement is also influenced by the way the story is constructed. In order to make it more believable and intriguing, the narrative flow can quickly reach high levels of complexity, making story development a demanding task for researchers. To address this challenge, Large Language Models (LLMs) can be employed. For instance, when provided with contextual information such as the story structure, LLMs can effectively generate narratives by leveraging their advanced language generation capabilities [13]. While the realization of adaptive behaviors can benefit from the flexibility offered by LLMs in language generation, relying on them for other key tasks in the process—such as the critical selection of a context-based and optimal action policy—can be problematic. Studies have shown that LLMs are sensitive to variations in prompts and hyperparameters, which can lead to inconsistent or unreliable outcomes [14]. This task can be fundamental to characterize and variate the story production based on the user's mental states, and to adapt the robot's behaviors to be more engaging for the user. In this direction, Reinforcement Learning (RL) [15] can be a more stable solution to leverage the users' non-verbal and emotional responses and to adopt personalized storytelling styles [16]. In this work, we present a hybrid approach for cooperative storytelling, aiming to actively stimulate user behavioral, affective, and cognitive engagement leveraging on verbal and non-verbal robot behavior and active participation in the storytelling. We conducted a user study to evaluate how the proposed interaction strategies sustain engagement during a human-robot interaction using a uniform action distribution model. Moreover, we investigated the integration of an RL-based adaptive model in the LLM's generation workflow, to further improve the users' level of attentiveness and the human perception of the robot throughout the interaction. Our results underscore the potential of the proposed method to positively influence the users' perception while keeping high levels of engagement, paving the way for a more immersive and interactive human-robot collaboration.

II. Related Works

A. Engagement for Real-time Adaptability in HRI

The realization of adaptive behaviors in human-robot interaction often requires estimating the user's engagement. In particular, facial signals are frequently taken into account to make inferences about both cognitive and affective engagement. Lin et al. [17] proposed a pilot study involving a tour guide robot that adjusted its content primarily based on visitors' previously specified preferences and their gaze direction. Yao et al. [18] developed a learning companion robot aimed at enhancing the learners' engagement by analyzing their attention's through facial features and gaze direction. Experiments with graduate students showed that self-reported engagement was higher during interactions involving the robot compared to those without it. In a similar learning context, Chen et al. [19] compared three robot roles—tutor, tutee, and peer—by tracking children's facial expressions to assess emotional engagement, observing that the peer robot elicited more positive emotional responses. Okita et al. [20] investigated how to implement adaptive behaviors in a robot designed to interact with children, using smile detection as an indicator of affective engagement. The multimodal communication employed, with different interaction styles and manipulating the robot's attention, managed to positively affect the children's engagement.

Other studies have incorporated additional non-verbal cues to assess engagement more comprehensively. Khamassi et al. [21] developed an adaptive reinforcement learning algorithm to modulate the robot's expressivity during interactions with children with autism. Engagement levels were annotated by psychologists, who considered both the child's gaze direction and their physical distance from the robot. Another physical cue used for engagement evaluation is body posture. Brenner et al. [22] found that body orientation and leaning were useful indicators for estimating users' intentions and engagement toward the robot, even when they were engaged in other activities. Similarly, Sanghvi et al. [23] extracted various postural features, such as body lean angle and slouch, during chess games between children and the iCat robot to classify engagement levels. These studies inspired our adaptive behavior perception module. The user's engagement is assessed through a multimodal analysis of typical human cues: gaze direction, facial emotion, and body posture. These types of feedback are then categorized along cognitive, affective and behavioral engagement dimensions [10].

B. Storytelling in HRI

In the HRI literature, storytelling is commonly used to evaluate the agent's ability to engage users during interaction, thanks to its capacity to immerse individuals in real or fictional scenarios and foster active participation in the narrative. Costa and Bae [7] compared physical and virtual storytellers in terms of user engagement, employing both synthetic and human voices. Experiments with adult participants showed that physical robots with a human voice were perceived as more engaging and empathic. Similarly, Conti et al. [24] investigated different narration styles, static and expressive, in kindergartens and found that the expressive robot outperformed the static human in terms of episode recall. Striepe and Lurgin [3] showed that a robot endowed with emotional communication capabilities engaged young adults participants as much as a human narrated audio book. Lee et al. [25] investigated using a robot as a passive listener, implementing a Theory of Mind model capable of recognizing the perceived attentiveness of narrating children and adapting the robot's non-verbal behavior to convey that it was actively listening. The participants' parents rated the adaptive robot behavior as more human-like. A combination of the previously mentioned modalities can be found in collaborative storytelling, where in a study involving teenagers the robot was perceived as more socially aware compared to the case in which the robot narrates the story alone [26]. Similarly, Battaglino and Bickmore [8] requested users to actively participate in the narrative by making comments and asking questions, observing that the interactive storytelling style was preferred over a more static approach, leading to increased perceived empathy and a preference for longer storytelling sessions.

In recent years, the growth of Large Language Models (LLMs) has been notable, driven by their powerful language generation capabilities, which provide flexibility in story development. Simon and Muise [27] employed an LLM guided by an automated planning module that generates valid action sequences to address coherence limitations commonly found in language models. Stories enriched with planning were rated more highly by participants in terms of narrative integrity and plausibility. To achieve similar results, Alvarez [13] used OpenAI's GPT-3 for story generation, providing it with a graph-based structure of the tale [28], which led to narratives that aligned more closely with the author's expectations. LLMs were also studied by Chin et al. [29] in the context of collaborative storytelling with children, where the language model was compared to the child's parent in performing this task. Interestingly, the number of conversational turns was significantly lower with the LLM, possibly due to the greater length and density of its responses compared to the parents', which were intuitively more spontaneous and personalized. In our work, we investigate possible solutions to address this issue, considering storytelling case of study, and tailoring the way the robot narrates the story based on contextual information and the user's feedback.

III. Methods

In this study, we present a robotic architecture designed to collaboratively generate a story through interaction with a user, evaluating two conditions: a uniform action distribution and an adaptive one. In both conditions, an LLM is used for story generation, incorporating the current story phase to apply different interaction strategies. However, only in the adaptive condition the model's prompt also integrate the user's feedback and reactions as parameters. In detail, the user's engagement estimation is performed considering three dimensions of engagement, such as cognitive, affective and

behavioral, and we shaped the robot's actions and behaviors to influence these three aspects [10]. Actions and behaviors for the robot are characterized manipulating both verbal and non-verbal communication as well as the generation of the story. Below we will describe the various modules of our architecture, which is illustrated in Figure 1. The robot employed is Pepper from Aldebaran[1] (formerly SoftBank Robotics).

A. Engagement Assessment

In this work, we process data obtained with the robot's camera but also with an external 4k camera (capturing one frame per second during the robot's turn and recording the audio) positioned behind the robot. Specifically, using the robot's camera feed, we considered the eye-contact frequency between the robot and the users to assess their attention during the interaction. The external camera provided better image resolution and was used for the other assessments that did not necessarily require it to be positioned at the same height as the user's eyes. These estimations included facial emotion recognition, which has been shown to be an effective user's affective state predictor [30], and the user's posture, with the action of leaning towards the robot constituting a spontaneous sign of engagement [31]. The facial emotion recognition was performed with the EmoNet model [32], while for the body posture we used Google's MediaPipe[2].

Based on works by Rossi et al. [10] and Maggi et al. [11], we combined these estimators and mapped them to three dimensions of engagement:

- Cognitive: we consider the user's gaze. If, during a turn, more than half of the user's gaze scans are directed toward the robot, cognitive engagement is considered positive; otherwise, it is considered negative.
- Affective: it is obtained in function of the emotion recognition performed with the EmoNet model. We map the predicted discrete emotion to the dimensional model of emotions and consider the arousal value formalized with a scale of integers, as proposed in [25]. Specifically, this mapping is obtained with Equation (1).

$$\text{affective\_eng} = \begin{cases} hi\_pos & \text{if emotion} = \text{“}Happy\text{”} \\ low\_pos & \text{if emotion} = \text{“}Surprised\text{”} \\ & \quad \text{and valence} > 0 \\ neutral & \text{if emotion} = \text{“}Neutral\text{”} \\ low\_neg & \text{if emotion} = \text{“}Surprised\text{”} \\ & \quad \text{and valence} < 0 \\ hi\_neg & \text{if emotion} = \text{“}Angry\text{”} \end{cases} \quad (1)$$

  Since there is ongoing debate regarding whether *"Surprised"* should be classified as a positive or negative emotion [33], it is mapped based on the computed valence polarity. This choice is supported by pre-experimental interactions, during which the emotion was predicted in association with both positive and negative valence values.

- Behavioral: relying on the user's posture. If, during a turn, more than one-quarter of the user's body scans indicate a lowered posture toward the robot, behavioral engagement is considered positive; otherwise, it is considered negative. This threshold was chosen empirically during pre-experimental interactions where this non-verbal behavior was observed less frequently than the user's gazes.

B. Communication Actions

The key idea, based on the three dimensions of engagement described above, is to formalize the actions so that each one stimulates the user in a specific engagement dimension through a targeted strategy: weak actions for behavioral, medium actions for affective, and strong actions for cognitive engagement. We considered three modalities for the actions, as shown in Table I, manipulating the different cues as follows:

- **Voice Pitch and Speed**: Altering these properties of the robot's voice has been demonstrated to influence the users' perception of the robot [34], [35]. Since higher pitch and speed convey joy and excitement, the *medium* and *strong* modalities adopt these configurations to enhance user engagement.
- **Robot's Animations**: A subset of Pepper's prebuilt animations is selected based on their contextual relevance with respect to the intended modality level. In the *weak* modality, the robot narrates the story using physical animations that are contextually coherent, but are not intended to elicit emotional reactions. In the *medium* modality we used emotion related animations [36]. While the *strong* modality is exclusively performed together with a question asked to the user, therefore the animation conveys an interrogative demeanor.
- **Face Tracking**: Used exclusively in the *strong* modality, to enhance the user's perception of the robot's cognitive engagement.

TABLE I: Non-verbal configurations

Modality	Physical Animation	Voice Pitch	Voice Speed	Face Tracking
weak	neutral movements	1.1	95	No
medium	emotional movements	1.2	100	No
strong	interrogative movements	1.2	100	Yes

During the interaction, in addition to the cues listed above, the robot's LEDs were used to indicate its interaction state [37]: blue when listening, yellow while processing the user's response, and white during narration.

C. Story Generation

The story generation consists in creating a fairy tale based on the works of Alvarez et al. [13], [28]: TropeTwist system details a fairy tale structure with an oriented graph representing the events of the tale, with characters and relevant

[1] https://aldebaran.com/en/pepper/
[2] https://github.com/google-ai-edge/mediapipe

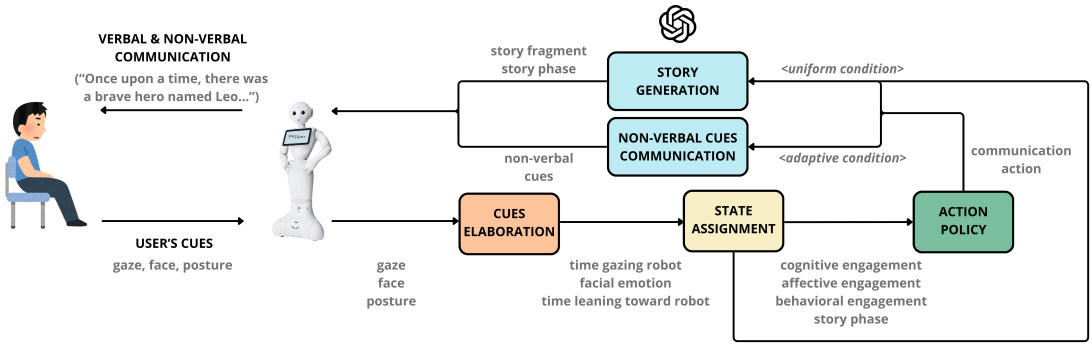

Fig. 1: The proposed model architecture in the interaction flow.

artifacts as nodes, and edges depicting their relationships. The fairy tale generation is achieved with OpenAI's GPT-4o[3]. In the initial phase a handwritten instructions prompt[4] is loaded into OpenAI's Assistant functionality. After a brief introduction outlining the goal and communication flow—emphasizing that the model must generate only one scene at a time, splitting the narration into turns—the prompt then details the fairy tale structure.

The generation of the story is also influenced by the communication modality employed at each step. The *weak* action does not lead to additional directions, simply instructing the LLM "*to generate the story as it would*". With the *medium* action, it is requested to add emphasis to the story, particularly highlighting characters' traits and ongoing events, while also incorporating direct speech to enhance listener immersion. The *strong* action provides direct involvement of the user, attempting to maximally increase their engagement level. The LLM is instructed to present the listener with a choice as if they were one of the fairy tale's characters, allowing them to decide the story's continuation, or to question them about previous events, subsequently informing them whether their answer was correct or not. In the latter case, incorrect answers are stored along with their corresponding questions to keep track of the user's beliefs about the fairy tale. If a strong action is performed again, the robot repeats the same question. This form of knowledge verification aims to enhance the user's attentiveness, assuming that, once performed, it makes the user aware of being tested. Along with the story fragment, the LLM is also asked to provide the current narrative phase, selecting among *start*, *mid*, and *end*. In Figure 2 we present an example of an exchange with users in which questions are asked.

Interleaving storytelling with questions

Robot: Now, Leo and Alex must decide how to face the evil sorcerer Baldur. Do you think they should attack immediately or look for allies first?

User: They should look for allies.

Robot: Great choice! Leo and Alex decided to look for allies.

Robot: Who raised the sword and shouted at Baldur during the final battle?

User: It was Leo.

Robot: Correct, Leo! In the end, Leo and Astra were hailed...

Fig. 2: Example of questions asked to the participants in two different moments of the story generation, referring to the same content.

IV. Adaptive Interaction Policy

Throughout the duration of the interaction, the robot estimates the human's engagement state. To further characterize the verbal and non-verbal communication, we also investigated the use of a RL approach to choose the optimal action. The interaction policy is designed to apply a communication style that keeps the user engaged while specifically adapting to their current status. We defined the problem formally with a Deterministic Markov Decision Process (DMDP), described by the tuple $\langle S, A, T, R, \pi \rangle$. The policy π is approximated with a Q-function [38]. A state $s \in S$ is obtained in function of the current phase of the story, and of the user's engagement estimation. More formally, the state is defined as follows:

$$S = \mathcal{C}_E \times \mathcal{A}_E \times \mathcal{B}_E \times \mathcal{P} \tag{2}$$

where:

- $\mathcal{C}_E = \{neg, pos\}$ is the cognitive dimension of engagement.
- $\mathcal{A}_E = \{hi\_neg, low\_neg, neutral, low\_pos, hi\_pos\}$ is the affective dimension of engagement.
- $\mathcal{B}_E = \{neg, pos\}$ is the behavioral dimension of engagement.
- $\mathcal{P} = \{start, mid, end\}$ is the phase of the story.

The set of actions $A = \{weak, medium, strong\}$ is the one defined in Section III-B, implementing incrementally more engaging robot behaviors.

The transition function $T : S \times A \rightarrow S$ was structured based on empirical observations from pre-experiment interactions with 10 volunteers. Specifically, we defined an automaton describing the transitions between the 60 states obtained considering the three dimensions of engagement and the

[3] https://openai.com/index/hello-gpt-4o/

[4] The prompt is available at this link: https://pastebin.com/tfRqMD5g

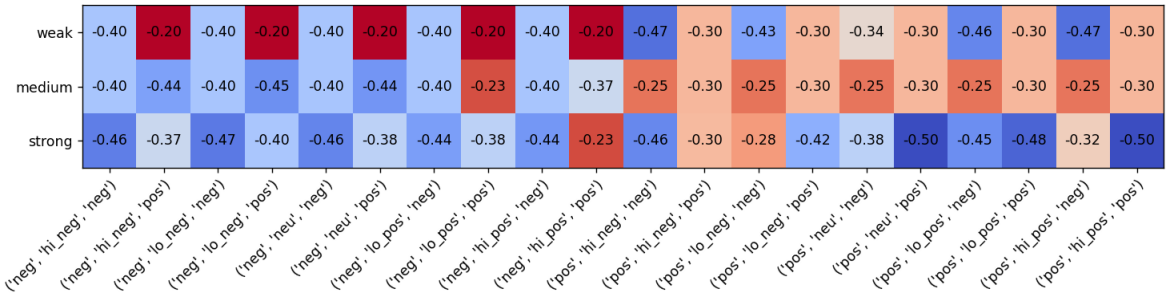

Fig. 3: Q-values for the *start* interaction phase.

phase of the story. The state transitions were modelled to not have too far ahead jumps between states (e.g., no action triggers a transition from a state with low cognitive, affective and behavioral engagement to one with three high estimates), with the goal of designing a policy mapping real transitions.

The reward function R is obtained based on the three engagement dimensions, the phase of the story, and action chosen by the model:

$$R(s, a, s') = R_C + R_A + R_B + A_{penalty} + P_{penalty} \quad (3)$$

where each component is empirically defined as follows:

- R_C considers the Cognitive Engagement estimated in the previous state s and the new state s'.

$$R_C = \begin{cases} 0.1 & \text{if } \mathcal{C}_{E,s} = pos \text{ and } \mathcal{C}_{E,s'} = pos \\ 0.2 & \text{if } \mathcal{C}_{E,s} = neg \text{ and } \mathcal{C}_{E,s'} = pos \\ -0.1 & \text{if } \mathcal{C}_{E,s} = pos \text{ and } \mathcal{C}_{E,s'} = neg \\ -0.2 & \text{if } \mathcal{C}_{E,s} = neg \text{ and } \mathcal{C}_{E,s'} = neg \end{cases} \quad (4)$$

- R_B is defined analogously to R_C, based on the Behavioral Engagement:

$$R_B = \begin{cases} 0.05 & \text{if } \mathcal{C}_{B,s} = pos \text{ and } \mathcal{C}_{B,s'} = pos \\ 0.1 & \text{if } \mathcal{C}_{B,s} = neg \text{ and } \mathcal{C}_{B,s'} = pos \\ -0.05 & \text{if } \mathcal{C}_{B,s} = pos \text{ and } \mathcal{C}_{B,s'} = neg \\ -0.1 & \text{if } \mathcal{C}_{B,s} = neg \text{ and } \mathcal{C}_{B,s'} = neg \end{cases} \quad (5)$$

- R_A - the five Affective Engagement are mapped to progressive numbers from 0 to 0.4. The reward associated to this dimension of engagement is computed as follows:

$$R_A = \mathcal{C}_{A,s'} - \mathcal{C}_{A,s} \quad (6)$$

- $A_{penalty}$: *strong* and *medium* actions are penalized (-0.3 and -0.2 respectively), in order to discourage the model to employ them too often. For the *weak* action, the penalty is set to 0.

- $P_{penalty}$ is a penalty based on phase of the story to favor a less intrusive behavior in the early phase of the story generation, with a more interactive stance towards the end of the story. Let P_{start}, P_{mid} and P_{end} be the three phases of the story, and a_{weak}, a_{medium} and a_{strong} be the three actions, the penalty is formalized it as follows:

$$P_{penalty} = \begin{cases} -0.05 & \text{if } P_{start} \text{ and } a_{weak} \\ -0.1 & \text{if } P_{start} \text{ and } a_{medium} \\ -0.2 & \text{if } P_{start} \text{ and } a_{strong} \\ -0.05 & \text{if } P_{mid} \text{ and } (a_{weak} \text{ or } a_{medium}) \\ -0.1 & \text{if } P_{mid} \text{ and } a_{strong} \\ -0.05 & \text{if } P_{end} \text{ and } (a_{weak} \text{ or } a_{medium}) \\ -0.07 & \text{if } P_{end} \text{ and } a_{strong} \end{cases} \quad (7)$$

To finally obtain the adaptive interaction policy, a model training was conducted by simulating interactions consisting of 15 turns, based on the pre-experimental interactions data. The exploration hyperparameter is set to $\epsilon = 0.1$, meaning that one out of every ten actions is chosen randomly. The Q-function is initialized with a learning rate of $\alpha(t) = \exp\left(-\frac{t}{2\sqrt{t}}\right)$, a decreasing function that stabilizes Q-values over time, and a discount factor $\gamma = 0.5$. An early-stopping condition was applied to prevent unnecessary epochs after Q-value convergence, i.e., when the highest Q-value increase is less than $\mu = 10^{-8}$.

In Figure 3, a heatmap of the computed Q-values for the *start* interaction phase can be observed.

V. Experimental Sessions

A. Participants

The study involved 36 participants, 26 males and 10 females, aged between 19 and 29 years ($mean = 24$, $std = 2$); 19% had never interacted with a robot before, 17% had only seen robots in the media, 19% had previously interacted with a robot, 31% had participated in a study involving robots, and 14% works with robots.

B. Procedure

The participant is brought to the laboratory, where the experiment takes place. First, they are asked to carefully read the consent form. Upon giving consent to participate to the study, they are briefly informed on the experiment and its procedure. They are then asked to fill out a pre-experiment questionnaire with some demographics (i.e., age, nationality, gender identity) and their previous experience with robots, after which they are asked to sit in front of the robot. Each participant interacts with the robot twice, once for each behavioral model (uniform actions distribution with random selection and adaptive) in random order. Immediately

TABLE II: Perceived Social Intelligence's scales for the condition with uniformly distributed communications actions.

Scale	RE	RB	RC	AE	AB	AC	PE	PB	PC	IH
Score	2.40	3.08	3.16	2.45	3.09	3.07	2.32	2.32	1.88	3.45
Scale	II	IG	SOC	FRD	HLP	CAR	TRU	RUD	CON	HST
Score	2.72	2.04	2.85	3.28	3.14	2.36	3.85	1.53	2.24	1.22

after each interaction, the participants are asked to fill out two questionnaires: the Perceived Social Intelligence (PSI) [39] and an extension of Self-Assessment Manikin (SAM) [40]. The PSI consists of 20 constructs, each containing 4 items, for a total of 80 items. It uses a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), and measures the three main areas of perceived social intelligence in robots: Cognition, Emotions, and Behaviors. Specifically, the scales included in the PSI are: Recognizes Human Emotions (**RE**), Recognizes Human Behaviors (**RB**), Recognizes Human Cognitions (**RC**), Adapts to Human Emotions (**AE**), Adapts to Human Behaviors (**AB**), Adapts to Human Cognitions (**AC**), Predicts Human Emotions (**PE**), Predicts Human Behaviors (**PB**), Predicts Human Cognitions (**PC**), Identifies Humans (**IH**), Identifies Individuals (**II**), Identifies Social Groups (**IG**), Social Competence (**SOC**), Friendly (**FRD**), Helpful (**HLP**), Caring (**CAR**), Trustworthy (**TRU**), Rude (**RUD**), Conceited (**CON**) and Hostile (**HST**). In the extended SAM, the reported valence, arousal, and dominance reflect not only the participants' own emotional states but also their perceptions of the robot's emotional expressions.

At the end of the session, the participant is asked to provide optional feedback on the interactions.

VI. Results

In this work, we aim to assess our approach considering an objective measure (i.e., the users' observed engagement), and two subjective measures (i.e., their perception of the robot's social intelligence and reported emotional states for both the user and the robot). These evaluations are carried out for both the experimental conditions investigated in this work. The analysis was conducted using the Jamovi software[5].

A. Observed Engagement

The users' observed engagement is described by three dimensions: cognitive, affective, and behavioral. For the cognitive and behavioral dimensions, negative values are mapped to -1 and positive values to $+1$. The affective dimension, which consists of five items, is normalized to a continuous range between -1 and 1 to ensure comparability with the other two dimensions.

The average engagement for both conditions was very similar: Uniform = 0.806 (± 0.334) vs. Adaptive = 0.804 (± 0.324) in the cognitive dimension, Uniform = -0.054 (± 0.298) vs. Adaptive = -0.053 (± 0.230) in the affective dimension, and Uniform = -0.232 (± 0.723) vs. Adaptive = -0.024 (± 0.751) in the behavioral dimension.

[5]https://www.jamovi.org

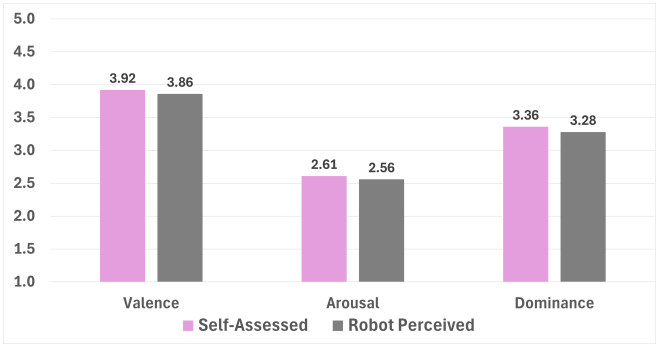

Fig. 4: Extended Self-Assessment Manikin's scales for the condition with uniformly distributed communication actions.

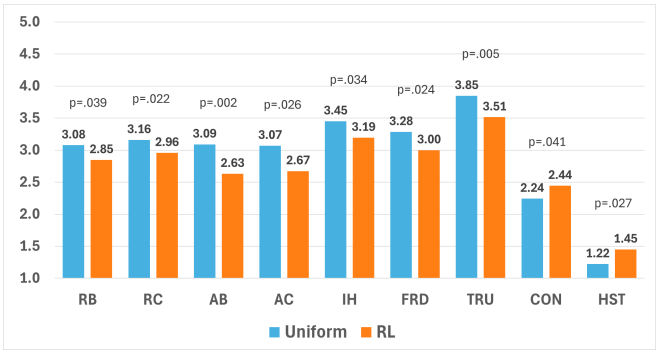

Fig. 5: Perceived Social Intelligence scales for which statistically significant differences were found between the conditions.

B. User's Perception

The users' perception was assessed through the analysis of the average aggregate values for each of the PSI questionnaire scales, considering all the participants (see Table II). Moreover, we computed the averages for both the participant's self assessed emotional state and their assessment of the robot's emotional expressiveness (see Figure 4).

C. Adaptive Interaction Policy Assessment

As already discussed, the average observed engagement in the adaptive condition is in line with the results obtained for the uniform one.

Regarding the users' perception, the perceived social intelligence was analyzed by assessing the statistical significance in the results obtained in between the two conditions with a two-tailed Pearson's t-test, with a significance threshold $\alpha = 0.05$. In Figure 5 we report the averages of the PSI scales for which we observed significant difference. The results from the Self-Assessment Manikin revealed a statistically significant difference only in the perceived arousal of the robot (p = 0.008), with the adaptive condition yielding a higher average than the uniform condition. Detailed results for the adaptive condition are shown in Figure 6.

D. Discussion

The analysis of the observed engagement in both conditions shows that robot's behaviors managed to induce a high

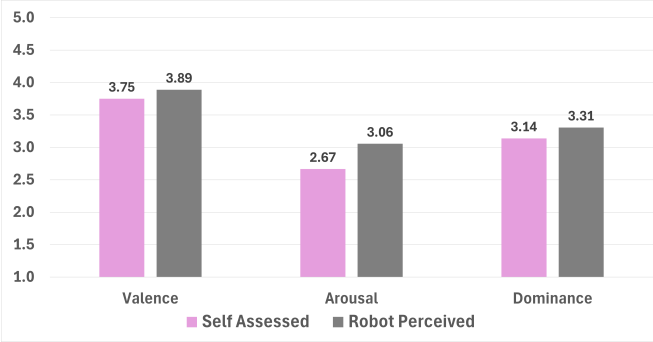

Fig. 6: Extended Self-Assessment Manikin's scales for the adaptive condition.

cognitive engagement in participants, attracting their visual attention for most part of the interaction. In contrast, emotional engagement presented neutral values in the average while behavioral lower ones. This outcome may be attributed to the nature of the task and the experimental process. In fact, the task considered does not require the participants to perform any physical action. Moreover, all participants were young adults and listening to a fairy tale may not significantly affect their affective and behavioral responses, unlike what might be expected with children.

Participants' evaluations through the post-interaction questionnaires revealed valuable insights. As shown in the Table II, participants perceived the robot as capable of identifying humans (**IH**), recognizing their behavior (**RB**) and cognition (**RC**) and adapting to their behavior (**AB**) and cognition (**AC**), suggesting a notable competence in this domain of social abilities. Moreover, the robot received positive evaluations regarding social valence traits, being rated as friendly (**FRD**), helpful (**HLP**), trustworthy (**TRU**) and not rude (**RUD**) or hostile (**HST**). These results suggest that the three communication actions, modulated through the multimodal configurations, successfully managed to endow the robot with engaging and socially aware and adaptive behaviors. Other PSI scales scored close to or below average, particularly those related to user prediction (**PE**, **PB**, **PC**), the identification of individuals (**II**) and social groups (**IG**), and certain social valence traits such as caring (**CAR**) and conceitedness (**CON**). This may be attributed to the experimental setting: the collaborative storytelling activity did not include moments in which users could assess whether the robot was accurately predicting their emotions, behaviors, or cognitive states. Furthermore, since the interaction was one-on-one and centered solely on storytelling, aspects related to identification were likely perceived as less relevant. These factors may also account for the limited variation observed in the caring and conceitedness scales, which were not expected to be significantly influenced by the nature of the study.

In the experimental sessions, together with the uniform condition, participants were administered also an adaptive condition, in which a RL algorithm was employed to choose the robot's actions based on user's engagement state and the phase of the story. While overall, the results observed are

TABLE III: Frequency of choices for each action.

Condition	Usage	Weak	Medium	Strong
Uniform	Overall	30.7%	47.6%	21.7%
	At least once	36	36	34
Adaptive	Overall	22.6%	63.6%	13.8%
	At least once	36	36	17

in line with the findings of the uniform condition, we must note that this approach did not succeed in further improving the participant perception and their assessment of the experience and of the robot's demeanor. In fact, the observed engagement clearly shows that the users' exhibited reaction were essentially the same (Uniform: 0.806, -0.054, -0.232; Adaptive: 0.804, -0.053, -0.024 - cognitive, affective and behavioral). The analysis of the PSI showed generally similar results, with statistically significant differences observed only with the scales shown in Figure 5, in which case the uniform condition was consistently preferred to the adaptive one. Conversely, in the adaptive condition, the robot's perceived emotional state was rated higher, particularly in the arousal dimension. This outcome can be explained by analyzing the actions choice distributions (see Table III) which show that actions were distributed in a more balanced way in the uniform condition, while in the adaptive condition the medium action was chosen much more predominantly. Moreover, in this case, the strong action was not even used at all with some of the participants (only 17 out of 36 experienced it). This made the adaptive condition less interactive compared to the uniform one (explaining the lower PSI scores observed), while at the same time making the robot behaved in a more expressive and enthusiastic way, thus explaining the higher rate for the arousal dimension of engagement. Less interactivity could also be the case of the lower significant scores in the PSI.

VII. Conclusions

In this work we presented a robotic communication strategy to engage humans in a cooperative storytelling task. Our approach integrates a social robot's embodiment's capabilities with the language generation power of LLMs. The results showed that the robot's behaviors managed to engage the participants, considering the cognitive dimensions of engagement, and the robot was generally perceived as trustworthy, friendly, and socially competent. The same level of cognitive engagement was observed both in the case of uniform distribution of actions (more interactive) and in the case of adaptive policy (less interactive but relying more on behavioral and gesture expressions).

Moreover, the emotional profile experienced by the users, and the one ascribed to the robot were also generally positive, with a high valence. In the perception of the robot, the adaptive strategy resulted in the robot being perceived as more aroused. However, a stronger interactivity was leading to a better perceived overall social intelligence.

Our results might be dependent on the task and the fact the experimental procedure was quite static. This limitation will

be addressed in extensions to this study that will consider alternative rewards balancing interactivity and expressivity. Future works will also investigate a different target for the storytelling, moving to more age-specific topics.

REFERENCES

[1] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for hri: A survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, 2017.

[2] S. Rossi, E. Dell'Aquila, G. Maggi, and D. Russo, "What would you like to drink? engagement and interaction styles in hri," in *Companion of the 2020 ACM/IEEE HRI Conference*, HRI '20, (New York, NY, USA), p. 415–417, Association for Computing Machinery, 2020.

[3] H. Striepe and B. Lugrin, "There once was a robot storyteller: measuring the effects of emotion and non-verbal behaviour," in *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pp. 126–136, Springer, 2017.

[4] J. M. Kory-Westlund and C. Breazeal, "Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning," *Frontiers in Robotics and AI*, vol. 6, 2019.

[5] E. Nichols, L. Gao, Y. Vasylkiv, and R. Gomez, "Design and analysis of a collaborative story generation game for social robots," *Frontiers in Computer Science*, vol. 3, 2021.

[6] E. Nichols, D. Szapiro, Y. Vasylkiv, and R. Gomez, "I can't believe that happened!: Exploring expressivity in collaborative storytelling with the tabletop robot haru," in *31st IEEE RO-MAN Conference*, pp. 59–59, IEEE, 2022.

[7] S. Costa, A. Brunete, B.-C. Bae, and N. Mavridis, "Emotional storytelling using virtual and robotic agents," *International Journal of Humanoid Robotics*, vol. 15, no. 03, p. 1850006, 2018.

[8] C. Battaglino and T. Bickmore, "Increasing the engagement of conversational agents through co-constructed storytelling," *Proceedings of the AAAI's AIIDE Conference*, vol. 11, pp. 9–15, Jun. 2021.

[9] K. Ryokai, C. Vaucelle, and J. Cassell, "Virtual peers as partners in storytelling and literacy learning," *Journal of computer assisted learning*, vol. 19, no. 2, pp. 195–208, 2003.

[10] A. Rossi, M. Raiano, and S. Rossi, "Affective, cognitive and behavioural engagement detection for human-robot interaction in a bartending scenario," in *30th IEEE RO-MAN Conference*, pp. 208–213, IEEE, 2021.

[11] G. Maggi, L. Raggioli, A. Rossi, and S. Rossi, "Assessment of distraction and the impact on technology acceptance of robot monitoring behaviour in older adults care," *IEEE Transactions on Affective Computing*, pp. 1–14, 2025.

[12] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proceedings of the 2009 ICMI-MLMI*, (New York, NY, USA), p. 119–126, Association for Computing Machinery, 2009.

[13] A. Alvarez, "Chatgpt as a narrative structure interpreter," in *International Conference on Interactive Digital Storytelling*, pp. 113–121, Springer, 2023.

[14] M. Loya, D. Sinha, and R. Futrell, "Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), pp. 3711–3716, ACL, Dec. 2023.

[15] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[16] L. Zhang, C. Zheng, H. Wang, R. Gomez, E. Nichols, and G. Li, "Autonomous storytelling for social robot with human-centered reinforcement learning," in *2024 IEEE/RSJ IROS*, pp. 2450–2456, IEEE, 2024.

[17] Y. Lin, W. Jo, A. Ali, L. P. Robert Jr, and D. M. Tilbury, "Toward personalized tour-guide robot: Adaptive content planner based on visitor's engagement," in *Companion of the 2024 ACM/IEEE HRI Conference*, pp. 674–678, 2024.

[18] Y. Bowei, O. Koichi, A. Kashihara, T. Unoki, and S. Hasegawa, "Development of a learning companion robot with adaptive engagement enhancement," in *ICCE*, 2022.

[19] H. Chen, H. W. Park, and C. Breazeal, "Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement," *Computers & Education*, vol. 150, p. 103836, 2020.

[20] S. Y. Okita, V. Ng-Thow-Hing, and R. K. Sarvadevabhatla, "Multimodal approach to affective human-robot interaction design with children," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 1, pp. 1–29, 2011.

[21] M. Khamassi, G. Chalvatzaki, T. Tsitsimis, G. Velentzas, and C. Tzafestas, "A framework for robot learning during child-robot interaction with human engagement as reward signal," in *2018 27th IEEE RO-MAN Conference*, pp. 461–464, IEEE, 2018.

[22] M. Brenner, H. Brock, A. Stiegler, and R. Gomez, "Developing an engagement-aware system for the detection of unfocused interaction," in *2021 30th IEEE RO-MAN Conference*, pp. 798–805, IEEE, 2021.

[23] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of the 6th HRI Conference*, pp. 305–312, 2011.

[24] D. Conti, C. Cirasa, S. Di Nuovo, and A. Di Nuovo, ""robot, tell me a tale!" a social robot as tool for teachers in kindergarten," *Interaction Studies*, vol. 21, no. 2, pp. 220–242, 2020.

[25] J. J. Lee, F. Sha, and C. Breazeal, "A bayesian theory of mind approach to nonverbal communication," in *2019 14th ACM/IEEE HRI Conference*, pp. 487–496, IEEE, 2019.

[26] C. J. Wong, Y. L. Tay, R. Wang, and Y. Wu, "Human-robot partnership: A study on collaborative storytelling," in *2016 11th ACM/IEEE HRI Conference*, pp. 535–536, 2016.

[27] N. Simon and C. Muise, "Tattletale: storytelling with planning and large language models," in *ICAPS Workshop on Scheduling and Planning Applications*, 2022.

[28] A. Alvarez and J. Font, "Tropetwist: trope-based narrative structure generation," in *Proceedings of the 17th FDG Conference*, pp. 1–8, 2022.

[29] J. H. Chin, S. Lee, M. Ashraf, M. Zago, Y. Xie, E. A. Wolfgram, T. Yeh, and P. Kim, "Young children's creative storytelling with chatgpt vs. parent: Comparing interactive styles," in *Extended Abstracts of the CHI Conference*, pp. 1–7, 2024.

[30] A. Tanevska, F. Rea, G. Sandini, L. Cañamero, and A. Sciutti, "A socially adaptable framework for human-robot interaction," *Frontiers in Robotics and AI*, vol. 7, p. 121, 2020.

[31] I. Casso, H. F. Chame, P. Hénaff, and Y. Delevoye-Turell, "Exploring engagement in human-robot interaction through the quantification of human spontaneous movement," in *2024 33rd IEEE ROMAN Conference)*, pp. 1768–1773, IEEE, 2024.

[32] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, 2021.

[33] M. K. Noordewier and S. M. Breugelmans, "On the valence of surprise," *Cognition & emotion*, vol. 27, no. 7, pp. 1326–1334, 2013.

[34] A. Niculescu, B. Van Dijk, A. Nijholt, H. Li, and S. L. See, "Making social robots more attractive: the effects of voice pitch, humor and empathy," *International journal of social robotics*, vol. 5, pp. 171–191, 2013.

[35] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," in *2019 14th ACM/IEEE HRI Conference*, pp. 211–221, IEEE, 2019.

[36] S. Rossi, A. Rossi, and S. Sangiovanni, "Towards the evaluation of the role of embodiment in emotions elicitation," in *2023 11th ACIIW Conference*, pp. 1–8, 2023.

[37] A. Pörtner, L. Schröder, R. Rasch, D. Sprute, M. Hoffmann, and M. König, "The power of color: A study on the effective use of colored light in human-robot interaction," in *2018 IEEE/RSJ IROS Conference*, pp. 3395–3402, IEEE, 2018.

[38] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[39] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, S. B. Banisetty, and D. Feil-Seifer, "Perceived social intelligence (psi) scales test manual (august, 2018)," 2018.

[40] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.