# Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training

Al Owayyed, M.; Denga, A.A.; Brinkman, W.P.

**Citation (APA)**
Al Owayyed, M., Denga, A. A., & Brinkman, W. P. (2025). Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training. In P. Gebhard, T. Schneeberger, B. Biancardi, N. Sabouret, M. Schmitz, & Z. Yumak (Eds.), *IVA 2025 - Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents* Article 17 (IVA 2025 - Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents). ACM. https://doi.org/10.1145/3717511.3747075

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

# Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training

Mohammed Al Owayyed*
Delft University of Technology
Delft, Netherlands
King Saud University
Riyadh, Saudi Arabia
M.alowayyed@tudelft.nl

Adarsh Denga
Delft University of Technology
Delft, Netherlands
A.A.Denga@student.tudelft.nl

Willem-Paul Brinkman
Delft University of Technology
Delft, Netherlands
w.p.brinkman@tudelft.nl

## Abstract

Child helpline training often relies on human-led roleplay, which is both time- and resource-consuming. To address this, rule-based interactive agent simulations have been proposed to provide a structured training experience for new counsellors. However, these agents might suffer from limited language understanding and response variety. To overcome these limitations, we present a hybrid interactive agent that integrates Large Language Models (LLMs) into a rule-based Belief-Desire-Intention (BDI) framework, simulating more realistic virtual child chat conversations. This hybrid solution incorporates LLMs into three components: intent recognition, response generation, and a bypass mechanism. We evaluated the system through two studies: a script-based assessment comparing LLM-generated responses to human-crafted responses, and a within-subject experiment ($N = 37$) comparing the LLM-integrated agent with a rule-based version. The first study provided evidence that the three LLM components were non-inferior to human-crafted responses. In the second study, we found credible support for two hypotheses: participants perceived the LLM-integrated agent as more believable and reported more positive attitudes toward it than the rule-based agent. Additionally, although weaker, there was some support for increased engagement (posterior probability = 0.845, 95% HDI [−0.149, 0.465]). Our findings demonstrate the potential of integrating LLMs into rule-based systems, offering a promising direction for more flexible but controlled training systems.

## CCS Concepts

• **Applied computing → Interactive learning environments**; • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Artificial intelligence**.

## Keywords

Conversational Agents, Large Language Models (LLMs), Belief-Desire-Intention (BDI), Child Helpline Training, Virtual Training Simulations, Counsellor Training, Social Skills Training

## 1 Introduction

Child helplines offer a safe space for children in need to seek help. This requires extensive counselling training efforts. Children can contact helplines such as De Kindertelefoon through online chat interfaces or by phone. As of 2023, the Dutch Children's Helpline, De Kindertelefoon, has over 700 volunteer counsellors, with more than 340 new volunteers trained that year—an increase of 20% compared to 2022 [14]. To train new counsellors, a common practice is roleplaying, where one counsellor acts as a child contacting the helpline while the trainee engages in conversation. However, this method can be time-consuming and resource-intensive. A potential alternative is using interactive social agents to simulate child-helpline interactions, which were helpful in similar training scenarios, such as virtual patients in medical education [22, 51].

A crucial aspect of a simulation system for learning is being as faithful to the actual situation [41]. In this case, simulating realistic, open-ended chat conversations of a virtual child. However, training counsellors using virtual agents can be sensitive, especially when using open-ended unrestricted interaction [7]. While closed-answer options provide a safer environment, they do not fully replicate real-world interactions where counsellors formulate responses independently to develop their chat conversational skills.

One way to enable open-ended conversations is through rule-based intent recognition, where the system interprets user input and selects appropriate responses. An example of such a system is Lilobot, a BDI-based (Belief-Desire-Intention) training system designed for child helplines [1, 21]. The system focuses on bullying scenarios, one of the main topics children discuss with De Kindertelefoon [29]. Lilobot simulates a virtual child who has been bullied at school and is seeking support from a helpline. The system follows a five-phase model of counselling conversations [45], which trainees should apply correctly for the interaction to progress. The phases are building rapport, clarifying the child's story, setting the conversation goal, working towards that goal and wrapping up the conversation. If a counsellor fails to follow these phases, Lilobot may leave the conversation.

However, traditional intent recognition has limitations, as it can only respond within the constraints of its predefined rules and training data [40]. This was also reported for Lilobot. Although positively received for its training potential, participants highlighted certain limitations. Specifically, the system struggled to understand

complex sentences outside its knowledge base, indicating a need for improved comprehension. Participants also expressed a need for more response variations, as variety and realism are important for effective learning [6]. These challenges align with the broader limitations of rule-based conversational agents [40].

Recent advancements with large language models (LLMs) offer promising solutions to these limitations. LLMs can generate human-like text with contextual relevance and variability, making conversations appear more natural and engaging compared to natural language understanding and generation [19, 23, 28]. However, using pure LLMs in training scenarios presents a challenge; while they enhance realism, they should operate within a controlled learning environment to ensure specific educational objectives are met. Without appropriate constraints, LLMs may generate unintended or extraneous content, which could interfere with the learning process, adding extraneous load to the trainee.

To balance these considerations, we propose integrating LLMs with Lilobot's existing BDI framework. This paper presents an integration design that preserves the structured learning environment required for counsellor training and evaluates its effects. Our evaluation has two parts: (1) a non-inferiority analysis, conducted through human coders, comparing LLM-generated and BDI system responses in terms of both understanding and response quality, and (2) an experimental study where participants used both LLM-integrated and BDI-only systems in a within-subject design. Based on this evaluation, we formulate and test the following hypotheses:

**H1:** Individuals perceive the integrated system to be more believable than the rule-based system.

**H2:** Individuals perceive the integrated system to be more engaging than the rule-based system.

**H3:** Individuals have a more positive attitude towards the LLM-integrated system than the rule-based one.

## 2 Related Work

### 2.1 Social Agents for Helpline Training

Researchers explored interactive social agents as tools for training helpline counsellors. Demasi et al. [15] proposed a chatbot simulation to train counsellors in suicide prevention hotlines. This simulation uses annotated transcripts and a multi-task framework to generate responses that mimic various crisis conversation scenarios. Evaluations of the proposed model showed improvements in response diversity and specificity compared to simpler models. For child helpline training specifically, one example is a serious game designed to simulate high-risk interactions in a safe environment (e.g., trafficking and sexual exploitation) [49]. This game uses choice-based interactions to train counsellors in handling sensitive situations while developing essential skills. However, it does not support open-ended interactions, which helps avoid language processing and generating limitations.

### 2.2 LLMs for Counselling

LLMs have been used to support counselling in various ways. To simulate a counsellor providing help, Steenstra et al. [46] explored LLMs' potential in delivering motivational interviewing for alcohol counselling. They found that LLM-powered virtual agent responses were perceived similarly to human-generated ones, addressing the

limitations of rule-based approaches in understanding nuanced therapy conversations. Recently, Heinz et al. [24] reported a reduction of symptoms in depression, anxiety, feeding and eating disorders after people interacted for four weeks with their therapy chatbot Therabot. This text-based application used a generative LLM fine-tuned on expert-curated mental health dialogues. For training counsellors, Wang et al. [50] introduced an LLM-driven virtual patient to train trainees in formulating cognitive models as part of cognitive behavioural therapy. The authors defined several cognitive models as scenarios and instructed the LLM to simulate patient behaviour using them. This also allows trainees to compare the identified cognitive model with the actual one. When evaluating the system, the simulated interactions resulted in greater perceived knowledge and confidence compared to traditional methods. Other examples of LLM-integrated virtual patients include applications in history taking [25], and medical diagnosis [42, 48]. LLMs have also been used to support counsellors in their work or training by generating reflections [43], providing feedback on trainees' responses [9, 11], and offering suggestions to counsellors [31].

### 2.3 Integrations of LLMs and BDI

The integration of LLMs with rule-based systems has been explored to enhance simulations' reliability, realism, and explainability. Pico et al. [39] examined integrating LLMs with a BDI model to improve emotion recognition in intelligent agents. In their approach, an LLM is prompted with a dialogue and tasked to pick up an emotion, which is then converted into beliefs used by the BDI model. When evaluated, the LLM demonstrated promising capabilities in emotion recognition. Frering et al. [20] integrated LLMs and BDI in a human-robot interaction setting to improve explainability. The authors used an LLM to interpret user input and translate it into commands the BDI agent could process. Additionally, the LLM generated responses explaining the robot's behaviour based on the BDI agent's state.

## 3 Integrating LLMs with Lilobot

We integrated LLMs into Lilobot's rule-based framework to enhance understanding and generation capabilities while maintaining control over scenario structure. We examined integration possibilities within the existing system, explored LLM models and prompt strategies, and then selected a practical implementation approach.

### 3.1 Integration Possibilities

Lilobot follows a BDI architecture [21], which defines internal states for the virtual child, making its decision-making controllable and interpretable. The system maps trainee input to predefined intents—for example, when a trainee writes to the virtual child, "How does that make you feel?", the input is recognised as the intent "request_unknown_feeling." These intents are mapped to child beliefs within the BDI framework, which are tied to the child's desires and intentions—for instance, the intent will increase the belief that "they feel the trainee can be trusted." Based on the current BDI state and the recognised trainee intent, the system picks a response from its knowledge base. To introduce variation, each response has four alternatives, one of which is selected at random; in this case, the child might reply, "It makes me sad... I really don't know what to do." If the trainee's intent is not recognised, the system retrieves

a default response based on Lilobot's currently active desire (i.e., goal). For example, when the desire "they want the trainee to call their school" is active, a default response could be "I want you to call my school." While this structure ensures consistency, it also limits the dialogue's flexibility and adaptability to diverse responses.

To identify where we can integrate LLMs, we followed the general architecture of agent-based training for social skills (ARTES) [2] and Lilobot's limitations. ARTES outlines three possible integration points in the simulation: Natural Language Understanding (NLU), Natural Language Generation (NLG), and Lilobot's thinking. NLU and NLG have been shown to perform well with LLMs [10], allowing for a better understanding of trainee input as well as increased variability and naturalness in responses. As for the thinking component, integrating an LLM would enable Lilobot to bypass unexpected cases not covered in the list of trainees' intents. Therefore, we refer to it as the Bypass component in this paper.

## 3.2 Components
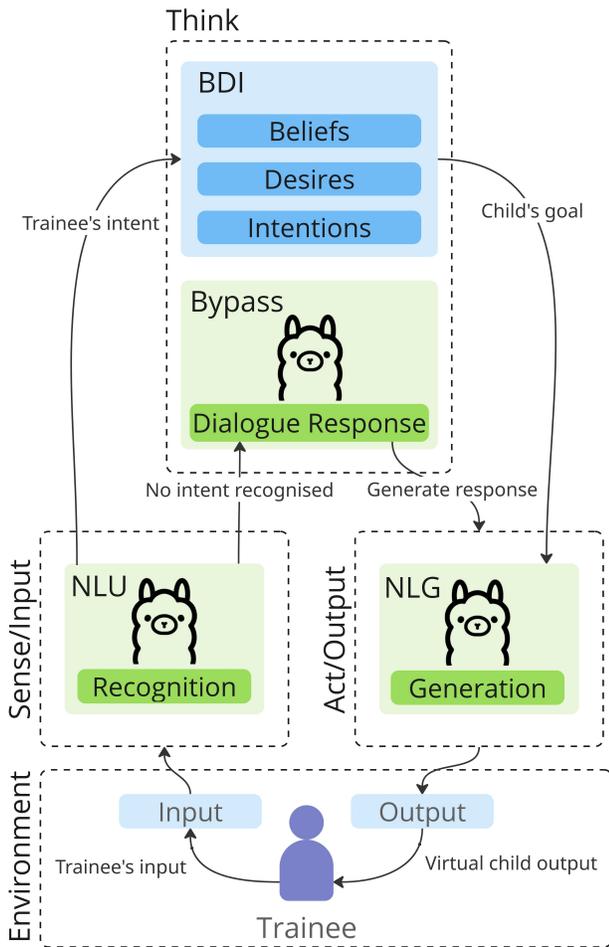


**Figure 1: Architecture of the LLM-integrated BDI-based conversational agent based on the ARTES architecture [2]. The green components indicate where the LLM is integrated.**

Figure 1 shows the integrated system's architecture. The overall flow begins with the trainee sending an input, which the NLU component matches to a trainee intent using over 2,000 examples from a dataset used by the rule-based system. If a match is found, the intent is sent to the BDI system to update the cognitive state of the virtual child, resulting in the identification of a response with four associated example responses. These are passed on to the NLG component along with the original input message from the trainee. The NLG component, in turn, generates a contextually appropriate response for the trainee. If the NLU finds no matching intent, the Bypass component generates a relevant response based on the trainee's input, which is then sent back to the trainee.

## 3.3 Prompt Creation

We explored several prompts for the three components: NLU, NLG, and Bypass. We followed prompt engineering techniques [5, 47]. This led to specific prompt designs, one for each component.

**NLU:** To classify trainee input into a BDI-recognisable intent, we designed a prompt in which the LLM's task is to select the best-matching intent for the input or return *"unknown"* if no match is found. When a trainee input is received, it is embedded and compared—using L2 distance in embedding space—to over 2,000 annotated examples (e.g., *"why are you being bullied?"* mapped to the intent *bullying_why*), which are stored in a vector database. The closest examples to the trainee's input are then included in the prompt, along with their corresponding intents and the trainee's input itself, and sent to the LLM. This method reduced the prompt size while preserving contextual relevance. In our initial attempts, we embedded all examples in the prompt, which resulted in slower response times.

**NLG:** When a trainee's input is matched to an intent, it is sent to the BDI in the Think component, along with the child's state, and then passed to the NLG sub-component to generate a response. The NLG prompt includes the trainee's input, a defined child persona (e.g., *"You must play the character of Lilo, a 9 year old child being bullied at school."*), the child's current goal as represented in the BDI model, and four example child responses linked to the identified trainee intent. The prompt instructs the LLM model to generate a reply similar in tone and structure to these examples. The resulting response is then sent to the trainee as the virtual child's reply.

**Bypass:** When no matching trainee intent is found, the Think component triggers the Bypass sub-component. The prompt structure is similar to that of the NLG prompt, with the key difference being the absence of example responses. In this case, the LLM is instructed to generate a contextually appropriate response from a child's perspective without relying on predefined examples. The generated response is then sent to the trainee.

## 4 Study 1: Script Evaluation

To evaluate the effectiveness of the prompts, we assessed whether the outcomes generated by the LLM were non-inferior to those produced by human-crafted responses. These human-crafted responses, drawn from Lilobot's rule-based model, had been reviewed by domain experts [21]. Inspired by Steenstra et al. [46], we employed a within-subjects design in which we asked four human coders to rate or label the outcomes of both the LLM prompts and the

human-crafted responses from the rule-based system. The coders included an undergraduate medical student, a computer science graduate, and two master's students in computer science. For NLU, we focused on categorisation accuracy. We created 12 trainee input examples, where the LLM prompt, the rule-based system, and the coders independently matched each utterance to a trainee intent, or labelled it as "unknown" if it did not match any trainee intent. For NLG and bypass, we evaluated whether the language used in the LLM-generated responses was appropriate to the context, the virtual child's state, and the trainee's input. We generated eight NLG responses per system: eight human-crafted responses from the rule-based system and eight LLM-generated counterparts. Additionally, we generated eight LLM bypass responses that could not be matched to intents in the rule-based system. Participants were then presented with all 24 responses in randomised order, along with the conversation history for context, and were asked to rate them. The source of each response was concealed from the participants. All data, prompts, pipelines for the prompts, and analysis code are available online [1].

### 4.1 Measures

**NLU:** Participants were given a list of 38 trainee intents with brief explanations and 12 trainee inputs. Their task was to match each input to one or more intents. In the ground truth, 10 inputs were each mapped to a different intent, and the remaining two were mapped to two intents each, resulting in 14 intents to identify.

**NLG and Bypass:** Participants were asked to rate each response on a Likert scale (1 = Strongly Disagree to 7 = Strongly Agree) based on five statements adapted from Steenstra et al. [46] (Table 1).

**Table 1: Rating statements for the Bypass and NLG evaluation tasks, adapted from [46].**

| Statements |
| --- |
| C1: This response is in coherent English |
| C2: This response is coherent in this bullying context |
| C3: This response directly addresses and replies to the counsellor's previous message |
| C4: This response makes sense |
| C5: This response makes sense in this context of bullying |

### 4.2 Data Analysis

We conducted a Bayesian analysis to perform a non-inferiority test. Specifically, we fitted a Bayesian multilevel regression model, using non-informative priors. For NLU, we used the model to estimate whether the intent categorisation result for each example matched the ground truth using the rule-based model or the LLM. For the NLG and Bypass components, the model accounted for variability across individual coders when rating LLM- and human-crafted utterances. In both cases, we defined a non-inferiority threshold just below the null value. We then assessed the proportion of the posterior distribution of the parameter estimate that fell above this

threshold, which would indicate practical non-inferiority [30]. The threshold was set based on half of Cohen's [38] threshold for a small effect size ($-0.1 \times \text{sd(score)}$) to define negligible or better effects [30]. Additionally, inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC), Fleiss' and Cohen's Kappa to estimate agreement between coders, the LLM, the rule-based system, and ground truth.

### 4.3 Results and Discussion

In the NLU analysis, the LLM correctly recognised all 14 trainee intents for the 12 examples, while the rule-based system recognised 12 intents, and the median of the four human coders recognised 13. To check for overall reliability, Fleiss' Kappa among human coders alone was 0.81, and between the coders and the LLM was 0.82, which Landis and Koch [32] classify as "almost perfect" agreement. Agreement between coders and the rule-based model was slightly lower but still substantial ($\kappa = 0.78$), according to the same classification scale. Cohen's Kappa indicated perfect agreement between the LLM and ground truth ($\kappa = 1.00$), and almost perfect agreement between the median coder and ground truth ($\kappa = 0.92$), as well as between the rule-based model and ground truth ($\kappa = 0.85$) [32]. The non-inferiority results indicated some support for the non-inferiority of the LLM, with the probability of 0.87 being non-inferior.

For NLG and Bypass, coders' ratings agreement was generally low (ICC < 0.5). Median ratings clustered near the top of the scale (coder 1 = 6.5, coder 2 = 7, coder 3 = 7, coder 4 = 5.4), potentially indicating ceiling effects and limited variance. LLM texts received a mean rating of 6.30 (SD = 0.89), while human-crafted texts had a mean of 6.28 (SD = 0.78). The results showed a posterior probability of 0.70, indicating some support for the non-inferiority of LLM-generated content relative to human-crafted text.

## 5 Study 2: Experiment

After evaluating the LLM-generated script quality, we conducted a within-subjects experiment to test our hypotheses. Participants role-played as counsellors tasked with applying the five-phase model with the virtual child. There are two conditions: (1) a rule-based agent and (2) an LLM-integrated agent. Each participant experienced both systems in a randomised order to control for order effects, with the entire experiment lasting approximately 50 minutes. This experiment was approved by the university human research ethics committee (HREC reference number: 4768) and preregistered with the Open Science Framework[2]. The data, analysis code in R (version 4.3.1), and a Docker container for reproducibility are publicly available via the 4TU research data repository [3].
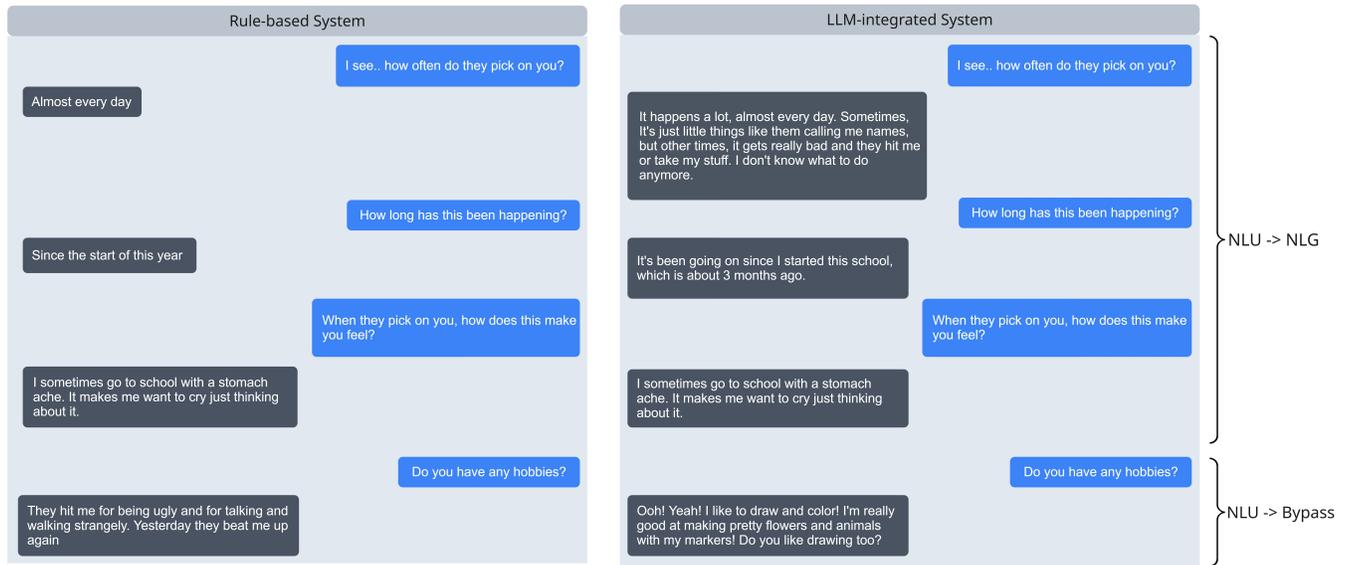
### 5.1 Participants

We recruited participants through the online platform Prolific. To match the target group for this training system, e.g., new helpline volunteers at organisations like De Kindertelefoon, we determined eligibility as adults with at least a Bachelor's degree. We also required fluency in English, an approval rating of $\geq$ 95%, and have completed at least five studies in the past. A total of 55 participants

---

[1]https://data.4tu.nl/datasets/5c5b69b7-f727-42d0-983f-07ab910b8460

[2]https://osf.io/eqxwz
[3]https://data.4tu.nl/datasets/5c5b69b7-f727-42d0-983f-07ab910b8460

**Figure 2: Example conversation for each condition. The rule-based system (left) and the LLM-integrated system (right) respond differently to the same counsellor input. The last input does not exist in the trainee intents list, triggering the Bypass component in the LLM-integrated system. Meanwhile, the rule-based system responds with a message related to the child's active desire.**

were invited; 18 were excluded in total due to incomplete participation (n = 13), failing comprehension or attention checks (n = 2), not providing consent (n = 1), or not meeting screening criteria (n = 2). This resulted in a final sample of 37 participants. The study was conducted between January and February 2025. Participants received compensation per Prolific's minimum payment policies.

Of the 37 participants, 24 held a Bachelor's degree or equivalent (65%), 10 held a Master's degree (27%), and 3 held a PhD (8%). Fourteen participants identified as male (38%) and 23 as female (62%). Age distribution was as follows: 18–24 (n = 10), 25–34 (n = 12), 35–44 (n = 9), 45–54 (n = 5), and 55–64 (n = 1). Participants represented diverse nationalities, including the United Kingdom, Portugal, the United States, South Africa, Nigeria, and India.

## 5.2 Measures

We tested primary and secondary measures to evaluate the rule-based and integrated systems. The primary measures were used to test the three hypotheses, while the secondary measures provided additional insight into participants' experiences.

*5.2.1 Primary Measures.* We have three measures corresponding to the hypotheses: Believability, Engagement, and Attitude. All questionnaire items were drawn from the Artificial Social Agent Questionnaire (ASAQ) [17, 18], using a 7-point scale from −3 (Disagree) to +3 (Agree), with 0 as Neither agree nor disagree. We replaced references to "the agent" with "the virtual child" in all questionnaire items.

**Believability:** Measured using two constructs from the ASA questionnaire, Human-Like Behaviour (HLB) and Natural Behaviour (NB). They capture how believable the agent is to participants.

**Engagement:** Measured using the engagement construct, assessing participants' involvement during the interaction.

**Attitude:** Measured using the attitude construct, reflecting participants' feelings toward the virtual child.

*5.2.2 Secondary Measures.* To gain further insight, we included the following secondary measures:

**Overall Experience:** Assessed using the 24-item short version of the ASA questionnaire, capturing broader constructs evaluation, expressed in short ASAQ score.

**Preference:** At the end of the experiment, participants answered a preference question: "Which virtual child did you prefer interacting with?", requiring a selection between "The first child" or "The second child".

**Experience:** Qualitative feedback was collected through an open-ended question after each condition ("How did your interaction with the virtual child go?").

## 5.3 System

Both conditions (rule-based and LLM-integrated) were presented through an identical web interface hosted on TU Delft servers. The layout displayed the conversation on the right and a summary of the five-phase model on the left. Each participant interacted with two of the twelve bullying-related scenarios, with each virtual child having a different, randomly assigned name. Scenarios were randomised across conditions to avoid systematic scenario bias and were assigned such that no participant encountered the same scenario twice. As described above, the systems differed only in how they processed and responded to the participants' input. The rule-based agent used Rasa for NLU and predetermined responses, while the LLM-integrated version used Llama 3.2, run via Ollama, for NLU, NLG, and bypass. To ensure a fair comparison, we modified the rule-based system to match the response time of the LLM-integrated system (a random time between 15-25 seconds). Figure 2 shows a

conversational example from the rule-based system (left) and the LLM-integrated one (right). The code is publicly available online[4].

## 5.4 Procedure

Participants were redirected from Prolific to Qualtrics, where they completed an informed consent form. They then watched a short training video introducing the five-phase model and providing context on counselling practices. Afterwards, they interacted with the two systems (rule-based and LLM-integrated), with the order chosen at random. Participants were first directed to one of the systems and instructed to interact with it for 15 minutes. Within this time, the scenario restarted if the virtual child exited or if the participant ended the interaction (e.g., by saying "bye"). After 15 minutes, participants were redirected to Qualtrics to complete questionnaires on believability, engagement, attitude, overall experience, and an open-ended reflection question based on the virtual child they had just interacted with. They were then redirected to the second system, following the same structure: 15 minutes of interaction, followed by the same questionnaires based on the second virtual child. At the end, participants chose their preferred virtual child.

## 5.5 Data analysis

We analysed data from the 37 participants who were included in the study. We reversed items where required for all questionnaire-based measures, according to the ASA questionnaire guidelines, and computed average scores per construct. To test the hypotheses, we conducted Bayesian[5] paired-sample $t$-test using non-informative priors. We calculated the posterior probability that the difference between the averages (LLM-integrated – rule-based) is greater than zero and interpreted it based on the guidelines by Chechile [12] and Andraszewicz et al. [3]. We also checked whether zero is included in the credible intervals expressed by the Highest Density Interval (HDI). For the preference question, we used a Bayesian binomial test. We performed a thematic analysis [8] of the qualitative responses to identify recurring patterns and insights across the open-ended questions. To ensure reliability, two coders with a background in computer science independently coded the responses using a predefined coding scheme. The inter-coder reliability was fair (Cohen's $\kappa = 0.32$) [32]. A third coder then reviewed the disagreements to determine the final coding.

## 5.6 Results

*5.6.1 Primary Measures.* Figure 3 shows an overall comparison across all measures. The overall trend seems to favour the LLM-integrated system.

---

[4]https://github.com/adarshdenga/llm-integration-childhelpline
[5]We reported the Bayesian analysis results in the main text to get more insight and to be consistent with Study 1. In the preregistration, we indicated using a frequentist test by specifying a cutoff value of $p < 0.05$. For transparency, the frequentist results for each measure are as follows:

- **Human-like Behaviour (H1):** $t(37) = 2.10$, $p = 0.04^*$
- **Natural Behaviour (H1):** $t(37) = 1.38$, $p = 0.18$
- **Engagement (H2):** $t(37) = 1.07$, $p = 0.29$
- **Attitude (H3):** $t(37) = 2.46$, $p = 0.02^*$
- **Overall Experience:** $t(37) = 2.57$, $p = 0.01^*$
- **LLM Preference:** $\hat{p} = 0.70$, 95% CI [0.53, 0.84], $p = .02^*$

**Believability (H1)** Human-like behaviour and natural behaviour were rated higher for the LLM-integrated system (HLB: M = 0.79, SD = 1.62; NB: M = 0.14, SD = 1.55) than the rule-based system (HLB: M = 0.35, SD = 1.57; NB: M = −0.19, SD = 1.57). The posterior probability for the human-like behaviour was 0.975, a good bet in favour of the LLM system, with a 95% HDI of [0.002, 0.816] which excludes zero. This further supports the difference as the HDI range is above zero (Figure 4). For the natural behaviour, the posterior probability was 0.905, which is considered a promising but risky bet for the LLM-integrated system. This is also shown in the 95% HDI [-0.161, 0.771], which overlaps with zero. Given these two results, there is credible support for H1.

**Engagement (H2)** Engagement ratings were slightly higher for the LLM-integrated system (M = 2.01, SD = 1.05) than the rule-based system (M = 1.84, SD = 0.77). The posterior probability of 0.845 constitutes only a casual bet, with 95% HDI [-0.149, 0.465] included zero (Figure 4); thus, there is some support for H2.

**Attitude (H3)** Participants reported a more favourable attitude toward the LLM system (M = 0.86, SD = 1.52) than the rule-based system (M = 0.13, SD = 1.76), with a posterior probability of 0.988—a good bet for the LLM system. The 95% HDI [0.083, 1.266] is above zero (Figure 4). This indicates credible support for H3.

*5.6.2 Secondary Measures.* Similar to the primary measure, the overall trend seems to favour the LLM-integrated system.

**Overall Experience** The overall experience rating is also higher for the LLM-integrated system (M = 0.46, SD = 1.25, Short ASAQ score = 11) than the rule-based system (M = −0.021, SD = 1.28, Short ASAQ score = -1). The posterior probability was 0.991—a strong bet and irresponsible to avoid, with the 95% HDI [0.075, 0.807] higher than zero (Figure 4). We also used the ASAQ Representative Set 2024, consisting of 29 artificial social agents, for broader cross-study benchmarking, as suggested by the ASAQ's authors [17]. This resulted in Figure 5, where the LLM-integrated agent outperformed 25% of the set, while the rule-based agent exceeded only 5%.

**Preference** The preference analysis indicates a preference for the LLM system, with 26 participants favouring it. The posterior probability that this preference is above chance is 0.99, which is a strong bet that the LLM-integrated agent is preferred.

**Qualitative Feedback** Table 2 presents nine identified themes and their frequency in each condition. The themes relate either to the child's replies (the first five themes) or to the interaction experience (the last four). Overall, positive themes were more prevalent in the LLM condition, such as perceived conversational depth and human-like responses (e.g., *"… I think it acted of of a human being. It was quite impressive"*). In contrast, negative responses were more frequent in the rule-based condition, including unnatural and slow responses (e.g., *"It was boring, non-responsive, had delayed responses, did not reveal any realistic feelings, etc."*). One theme observed in both conditions—abrupt endings—may be partly attributed to the controlled scenario design, in which the virtual child could exit the conversation if the trainee deviated from the five-phase model (e.g., failing to show empathy or directly suggesting an action).

## 5.7 Discussion

In this study, we evaluated whether integrating LLMs into a rule-based system could enhance believability (H1), engagement (H2),
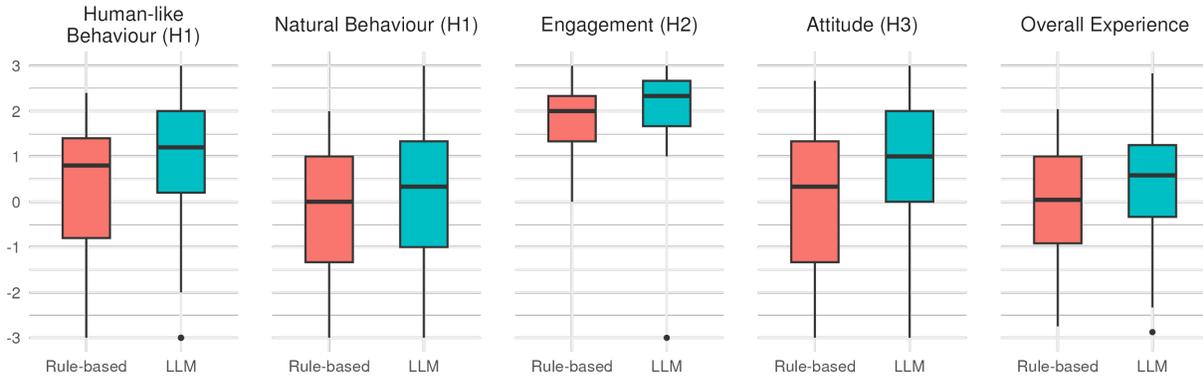
Figure 3: Comparison of participant ratings across five constructs between the rule-based and LLM-integrated agents.
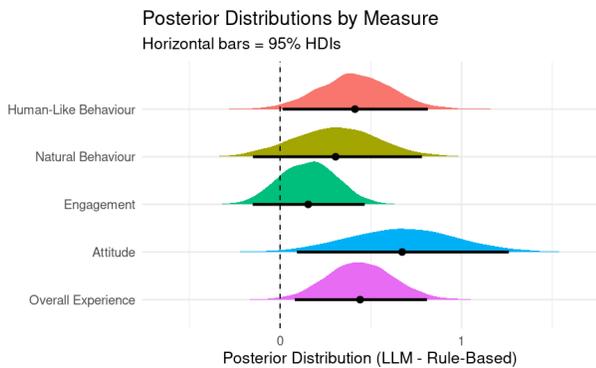


Figure 4: Posterior distributions of effect sizes comparing the LLM-integrated to the rule-based system across outcome measures.
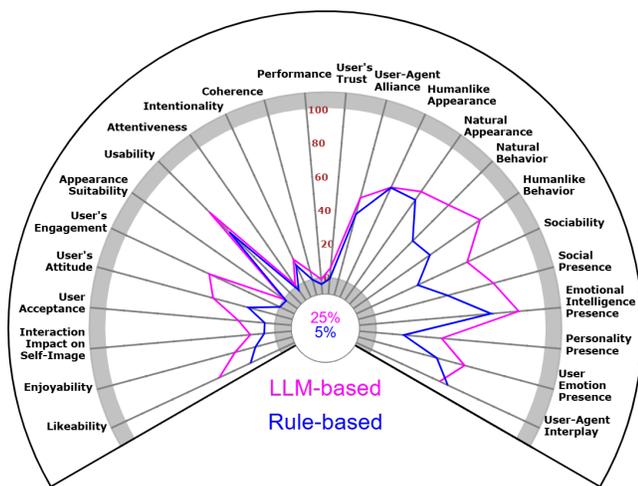


Figure 5: Percentile chart comparing the LLM-integrated and rule-based agents against ASAQ Representative Set 2024 [17]. The grey area represents the range of scores higher or lower than the 29 agents in the dataset. The visualisation was created using ASAQ's visualisation tool [17].

Table 2: Themes identified from participants' responses and the frequency of their occurrence in each condition.

| Theme | LLM-integrated | Rule-based |
|---|---|---|
| Abrupt Ending | 3 | 7 |
| Unnatural Responses | 4 | 7 |
| Depth of Conversation | 4 | 0 |
| Emotional Engagement | 6 | 1 |
| Human-Like Responses | 4 | 2 |
| Positive Experience | 9 | 8 |
| Scripted Responses | 0 | 1 |
| Slow Responses | 6 | 10 |
| Boring Task | 1 | 1 |

and positive trainee attitudes (H3) toward a virtual child used for helpline counsellor training.

We found credible support for believability (H1), particularly human-like behaviour. Furthermore, we hypothesised greater engagement for the LLM-integrated system than the rule-based one (H2). The results provided some support for this. Although scores were slightly higher, the posterior distribution indicated only a casual probability in favour of the LLM system. This may be attributed to a ceiling effect, as both systems scored highly on engagement (Figure 3). Such high engagement may result from the inherently engaging nature of the task, that is, providing support to a child in need. Additionally, participant comments clearly indicated task-related engagement (e.g., for the rule-based system: *"…we worked together to talk about their feelings and challenges at school. It was a productive conversation…"*, and for the LLM-integrated system: *"…The virtual child was engaging, and we had a good conversation! I felt like we made a connection, …"*).

Regarding attitude, we found credible support that people rated the LLM-integrated agent more positively than the rule-based agent (H3). Participants also directly compared the two systems in their comments, with three participants explicitly indicating that the LLM-integrated interaction was better (e.g., *"it went better than the first one [i.e., the rule-based]…"* and *"Definitely better than during the 1st chat [i.e., the rule-based]…"*).

Looking beyond the results obtained from hypothesis testing, we observed that participants had a more favourable overall experience with the LLM-integrated agent. Figure 5 shows that the LLM-integrated agent scored above the 75th percentile on constructs related to social presence and emotional intelligence. The agent's emotional presence was also supported by our thematic analysis (e.g., for the LLM-integrated system: *"...the virtual child was attentive and behaved more like a human and there were emotions involved."*). Also, participants described the LLM-integrated interaction as more emotionally engaging (e.g., *"The interaction was somewhat emotional and eye-opening, reinforcing the importance of empathy, active listening, and proper intervention..."*). This aligns with LLM's capability to generate emotional dialogues [34, 35, 37]. The results of the thematic analysis of participants' qualitative feedback revealed additional insights into our findings. This reinforces the conclusion that the LLM-integrated system provided a more engaging and believable experience for participants.

On the other hand, the short-ASAQ percentile rating falls within the lower 25% for the LLM-integrated system and 5% for the rule-based system (Figure 5). In contrast to the 29 agents in the ASAQ Representative Set 2024, which are primarily assistant agents, our agent is designed as a training simulation. Therefore, lower ratings in some cognitive constructs, such as performance, attentiveness, and appearance suitability, do not necessarily indicate that the simulation is ineffective. Rather, they may reflect how participants experienced a virtual child in distress, who sometimes "doesn't know" what to do and needs guidance during the conversation—something a counsellor is expected to provide, not a statement about the system. Response time may have affected attentiveness and performance, as 16 participants noted delays with both agents. However, this is not necessarily a limitation, as real children may also take time to respond. Nevertheless, for training purposes, faster replies might make the experience more engaging. Regarding appearance suitability, one possible explanation is the lack of embodiment, which provides fewer cues during interaction and results in a lack of visible appearance. To gain deeper insights beyond the percentile comparison, the virtual child simulation could be compared more specifically to other chat agents used in social situation simulations of non-experts.

Regarding the experiment's limitations, the LLM agent had noticeable response delays due to server constraints. Although we matched the response timing in the rule-based agent, participants still reported both systems as slow. For the same reason, we used a lightweight model (LLaMA 3.2) instead of the intended LLaMA 3. Initial tests with the larger model suggest improved comprehension and response quality, which could further enhance system performance. Additionally, interactions in our experiment were limited to single 15-minute sessions. Extending session durations or incorporating repeated interactions might reveal more complex dynamics or improved learning outcomes [27]. Future research could also isolate specific components (e.g., NLU or the bypass mechanism) to assess their individual contributions to learning outcomes and trainee experience. This could provide deeper insights into how the system is perceived and its overall effectiveness.

## 6 Overall Discussion and Conclusions

This paper presents a hybrid conversational agent that integrates LLM capabilities into a BDI-based rule system designed for child helpline training. First, we found results that support the non-inferiority of LLM-generated dialogues compared to those generated by humans. Then, through both qualitative and quantitative evaluation, we showed that integrating LLM components for NLU, NLG, and bypass improved believability and trainee attitudes, though the support was weaker for engagement.

Several directions for future work could be considered. One direction is to generalise this approach to other systems that train trainees to change individual beliefs and states in conversations, e.g., training social workers to persuade a virtual agent to get vaccinated, or to de-escalate an aggressive agent. This setup reduces the effort required to model conversation content and responses, particularly when crafting training data or generating response variations. However, it still requires designing the rule-based BDI model to capture the pedagogical constraints and structure of conversations, as these are closely tied to the simulation's learning objectives. Furthermore, LLMs could be used to generate belief, desire, and intention contents for a scenario [4].

LLMs controlling or directly influencing the BDI model can be challenging. LLMs are not explicitly trained to convert inputs into structured cognitive states within a numerical or symbolic framework. As a result, biases inherent in LLMs pose challenges when simulating a BDI conversational model. These include formatting and token biases [26, 36], as well as issues from unconventional prompts (e.g., prompts including negation [44]). Therefore, to simulate virtual agents using a BDI framework, LLMs should be paired with a structured BDI system for the LLM to interact with, rather than replace. For example, Retrieval-Augmented Generation [33] could be used to dynamically handle BDI-related content, e.g., retrieving relevant beliefs, desires, and rules during interaction.

Although we focused on integrating LLMs into the agent simulator component, the ARTES architecture [2] highlights other potential integration points. For example, LLMs could be used to generate feedback based on trainees' interactions, enabling deeper skill acquisition [9, 13, 52]. In our case, the BDI system could support feedback generation by supplying the LLM with structured inputs, such as the child's internal state, conversational logs, and learning materials. Furthermore, an LLM could simulate a senior counsellor, offering additional information and answering trainees' queries during interactions.

To conclude, the hybrid approach maintained pedagogical control over the conversational structure, yet offered a better balance between realism and control. This balance opens up new opportunities for scalable, realistic training in sensitive interaction domains, while still aligning with specific learning objectives.

## Acknowledgments

# References

[1] Mohammed Al Owayyed, Alex Despan, Myrthe Tielman, and Willem-Paul Brinkman. 2024. A Cognitive Conversational Agent for Training Child Helpline Volunteers. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[2] Mohammed Al Owayyed, Myrthe Tielman, Arno Hartholt, Marcus Specht, and Willem-Paul Brinkman. 2025. Agent-based social skills training systems: the ARTES architecture, interaction characteristics, learning theories and future outlooks. *Behaviour & Information Technology* 44, 9 (2025), 1787–1814. doi:10.1080/0144929X.2024.2374891 arXiv:https://doi.org/10.1080/0144929X.2024.2374891

[3] Sandra Andraszewicz, Benjamin Scheibehenne, Jörg Rieskamp, Raoul Grasman, Josine Verhagen, and Eric-Jan Wagenmakers. 2015. An introduction to Bayesian hypothesis testing for management research. *Journal of Management* 41, 2 (2015), 521–543.

[4] Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A Santos. 2023. Prompting for socially intelligent agents with ChatGPT. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–9.

[5] Simon Arvidsson and Johan Axell. 2023. Prompt engineering guidelines for LLMs in Requirements Engineering. (2023).

[6] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. 2000. Learning from examples: Instructional principles from the worked examples research. *Review of educational research* 70, 2 (2000), 181–214.

[7] Timothy Bickmore. 2022. Health-related applications of socially interactive agents. In *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: Interactivity, platforms, application*. 403–436.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[9] Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education* 24, 1 (2024), 1391.

[10] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[11] Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4130–4161. doi:10.18653/v1/2024.acl-long.227

[12] Richard A Chechile. 2020. *Bayesian statistics for experimental scientists: a general introduction using distribution-free methods*. MIT Press.

[13] David A Cook, Joshua Overgaard, V Shane Pankratz, Guilherme Del Fiol, and Chris A Aakre. 2025. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. *Journal of Medical Internet Research* 27 (2025), e68486.

[14] De Kindertelefoon. 2023. *Jaarrekening 2023 Stichting de Kindertelefoon*. https://jaarverslag.kindertelefoon.nl/resources/jaarverslagen/Jaarrekening%202023%20Stichting%20de%20Kindertelefoon.pdf

[15] Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3623–3636.

[16] Adarsh Denga. 2025. Combining LLMs with BDI Systems for Training Children's Helpline Counsellors. (2025).

[17] Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. 2025. The Artificial Social Agent Questionnaire (ASAQ)-Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *International Journal of Human-Computer Studies* (2025), 103482.

[18] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.

[19] Samuel Kernan Freire, Chaofan Wang, and Evangelos Niforatos. 2024. Conversational Assistants in Knowledge-Intensive Contexts: An Evaluation of LLM-versus Intent-based Systems. *arXiv preprint arXiv:2402.04955* (2024).

[20] Laurent Frering, Gerald Steinbauer-Wagner, and Andreas Holzinger. 2025. Integrating Belief-Desire-Intention agents with large language models for reliable human–robot interaction and explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence* 141 (2025), 109771.

[21] Sharon Grundmann, Mohammed Al Owayyed, Merijn Bruijnes, Ellen Vroonhof, and Willem-Paul Brinkman. 2025. Lilobot: A Cognitive Conversational Agent to

[22] Timothy C Guetterman, Rae Sakakibara, Srikar Baireddy, Frederick W Kron, Mark W Scerbo, James F Cleary, and Michael D Fetters. 2019. Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: mixed methods study. *Journal of medical Internet research* 21, 11 (2019), e15459.

[23] Önder Gürcan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good* (2024), 134–144.

[24] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI* 2, 4 (2025), AIoa2400802.

[25] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. 2024. A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. *JMIR medical education* 10, 1 (2024), e53961.

[26] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050* (2024).

[27] Sean HK Kang. 2016. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences* 3, 1 (2016), 12–19.

[28] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*. 278–290.

[29] De Kindertelefoon. 2023. De Kindertelefoon - Always Someone To Talk To. (2023). https://www.kindertelefoon.nl/resources/pdf-bestanden/cl-en-2023.pdf

[30] John K Kruschke. 2018. Rejecting or accepting parameter values in Bayesian estimation. *Advances in methods and practices in psychological science* 1, 2 (2018), 270–280.

[31] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Supporting the demand on mental health services with AI-based conversational large language models (LLMs). *BioMedInformatics* 4, 1 (2023), 8–33.

[32] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[34] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024. Speak from heart: an emotion-guided LLM-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*. 533–542.

[35] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. 2024. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications* 246 (2024), 123261.

[36] Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F Chen, and Min-Yen Kan. 2024. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *arXiv preprint arXiv:2408.08656* (2024).

[37] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI* 10 (2023), 1271610.

[38] Keith Muller. 1989. Statistical power analysis for the behavioral sciences.

[39] Aaron Pico, Emilio Vivancos, Ana García-Fornes, and Vicent J Botti. 2024. Exploring Text-Generating Large Language Models (LLMs) for Emotion Recognition in Affective Intelligent Agents.. In *ICAART (1)*. 491–498.

[40] AM Rahman, Abdullah Al Mamun, and Alma Islam. 2017. Programming challenges of chatbot: Current and future prospective. In *2017 IEEE region 10 humanitarian technology conference (R10-HTC)*. IEEE, 75–78.

[41] Evelyn-Rose Saus, BjØrn Helge Johnsen, Jarle EidEvelyn-Rose Saus, and Jarle Eid. 2010. Perceived learning outcome: The relationship between experience, realism and situation awareness during simulator training. *International maritime health* 62, 4 (2010), 258–264.

[42] Riley Scherr, Faris F Halaseh, Aidin Spina, Saman Andalib, and Ronald Rivera. 2023. ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Medical Education* 9 (2023), e49877.

[43] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 10–20.

[44] Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2023. You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments. *arXiv preprint arXiv:2311.09718* (2023).

[45] Trine Natasja Sindahl. 2011. *Chat Counselling for Children and Youth - A Handbook.*

[46] Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents.* 1–10.

[47] Udemy. 2024. Prompt Engineering for AI: Learn the Art of Prompting. https://www.udemy.com/course/prompt-engineering-for-ai/. Accessed: 2025-04-04.

[48] Jacqueline Vaughn, Shannon H Ford, Melissa Scott, Carolyn Jones, and Allison Lewinski. 2024. Enhancing healthcare education: Leveraging ChatGPT for innovative simulation scenarios. *Clinical Simulation in Nursing* 87 (2024), 101487.

[49] Eva Veldhuizen-Ochodničanová, Róisín Cotton, Charlene Putney, Maggie Brennan, and Mads Haahr. 2023. The use of serious games in knowledge-and skills-based digital counselling: Applications to trafficking in persons in the war in ukraine. In *Joint International Conference on Serious Games.* Springer, 290–298.

[50] Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. Patient-Ψ: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660* (2024).

[51] Heng Yao, Alexandre Gomes de Siqueira, Anokhi Bafna, Devon Peterkin, Jenelle Richards, Megan L Rogers, Adriana Foster, Igor Galynker, and Benjamin Lok. 2022. A virtual human interaction using scaffolded ping-pong feedback for healthcare learners to practice empathy skills. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents.* 1–8.

[52] Heng Yao, Alexandre Gomes de Siqueira, Margeaux Johnson, Roberta Pileggi, Amy Blue, Michael D Bumbach, Rene Love, and Benjamin Lok. 2024. Enhancing Empathic Communication in Healthcare Education Through Virtual Conversations: Leveraging Large Language Models for Real-Time Feedback. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality.* 41–50.