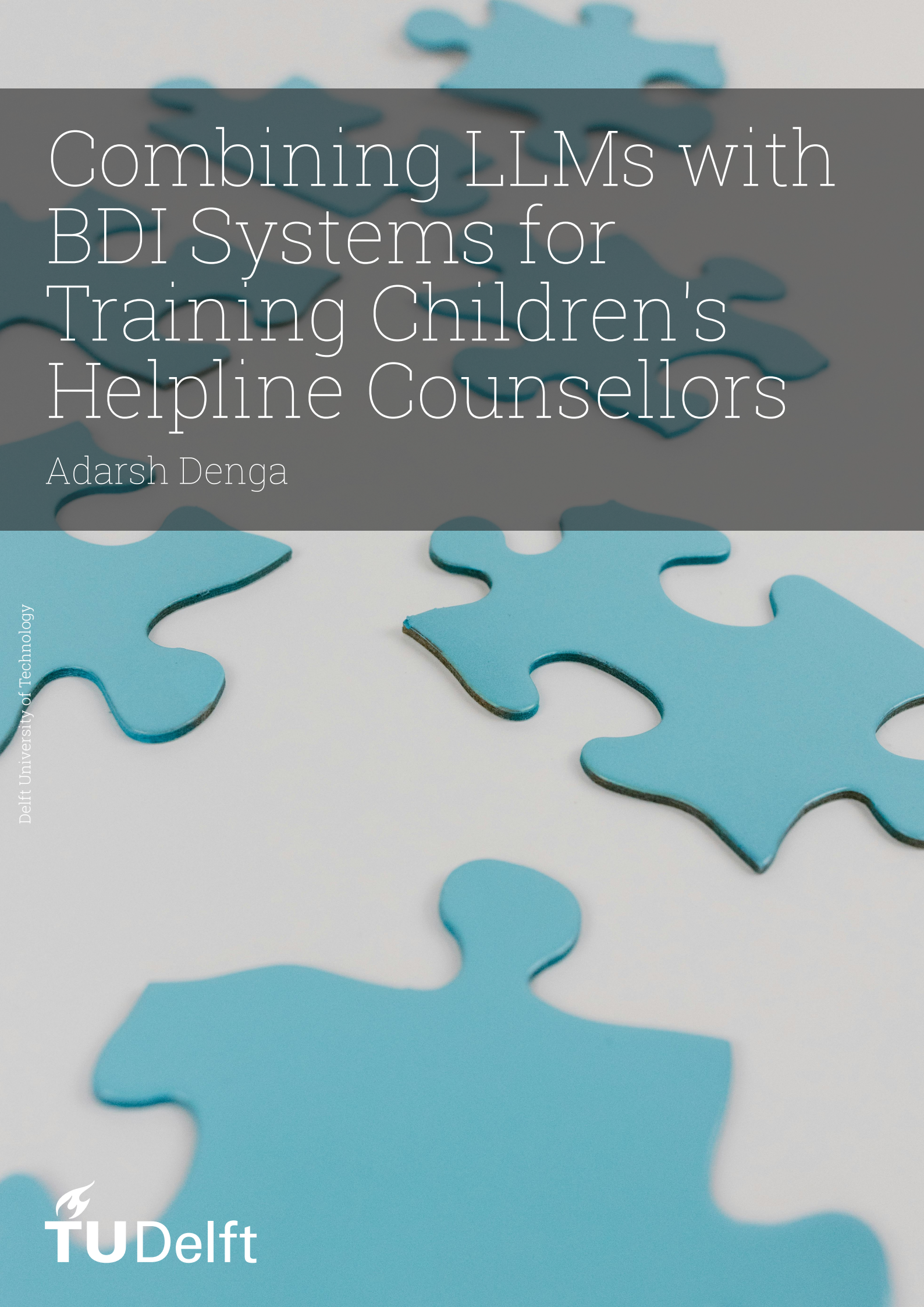
Combining LLMs with BDI Systems for Training Children's Helpline Counsellors

Adarsh Denga



Delft University of Technology

TU Delft

Combining LLMs with BDI Systems for Training Children's Helpline Counsellors

by

Adarsh Denga

to obtain the degree Master of Science
at the Delft University of Technology
to be defended publicly on March 25, 2025.

Thesis Advisor: Willem-Paul Brinkman
Co-Supervisor: Mohammed Al Owwayed
Faculty: Electrical Engineering, Mathematics & Computer Science, Delft

Cover: Close-Up Shot of Puzzle Pieces by Tara Winstead on Pexels
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Abstract

Child helpline counsellors require various skills and strategies to achieve lasting change in children who require assistance. Typical training methods such as role-play are resource intensive, leading to the development of computer simulation-based training systems where learners counsel the computer which assumes the role of a child requiring assistance. Such systems are limited in their understanding and responses, causing them to appear unrealistic and repetitive. In this paper, we built upon one such rule-based agent through the integration of Large Language Models (LLMs) to vastly expand both the understanding and responses of the agent. We conducted a within-subject experiment with 37 participants who we recruited online through Prolific, where they interacted with both systems, assuming the role of a counsellor. Our results indicate that participants find the integrated system to be human-like in its behaviour, have a more positive attitude towards it, and have a better impression of their overall experience with it. Our thematic analysis revealed that the integrated system felt more adaptive, and engaging, and allowed them to focus more on applying the conversational strategy, while the rule-based system felt scripted and boring. Our work provides an integrated system for effectively training child helpline counsellors and a method by which LLMs and rule-based systems can be integrated in general.

Contents

1 Introduction

1.1. Motivation

Child helplines around the world offer an invaluable resource by lending a sympathetic ear to children in need of compassion, encouragement and advice. Child Helpline International, a 'collective impact organization' encompassing 155 different helpline services in 133 countries, reports that each year they receive upwards of 13 million individual calls, and provide counselling services for almost 3 million children [41]. In the Netherlands, the primary children's helpline is De Kindertelefoon. Their report from 2023 states that their counsellors have over 1000 daily conversations with children seeking emotional assistance of which are 76% call-based and 24% chat-based [45]. Similar to Child Helpline International, De Kindertelefoon observes a trend of a rising amount of conversations with children. In 2019, the helpline had 405,267 [46] interactions with children, and this grew to 540,797 [47] by 2022, which is a 33% growth in the span of four years.

At De Kindertelefoon, there are over 600 volunteer counsellors working at five different locations in the Netherlands [48]. The central goal of De Kindertelefoon is to help children increase their self-insight, so they can learn to trust more in their own strength. In addition to providing support through counselling, De Kindertelefoon will direct the children to the appropriate resources for specialist knowledge and help. Given that the conversations can revolve around sensitive topics such as sexuality, emotional problems, home life, relationships and bullying, they are kept confidential. However, conversations need not be centred around the issues that the children face, and those who reach out to the helpline can talk about virtually anything they want to - even to just crack a joke or talk about something that made them happy.

Before assisting children at the helpline, new counsellors undergo 30 hours of face-to-face training that educates and tests them on the core knowledge and practices of counselling. In particular, counsellors learn and practice conversations in alignment with conversational strategies (e.g. Egan's skilled helper model [25]). The goal of such conversational strategies is to achieve lasting change and to empower people to manage their own problems. The training of counsellors is typically done through theory sessions about counselling, conversational techniques and models, and is followed by role-play sessions where volunteers take turns being counsellor and child, and feedback to improve counselling performance.

However, roleplay is not an exercise limited to real people, and can be done with virtual patients as well. Virtual patients are interactive computer-based simulations of real-life scenarios that are made for the purpose of training [86]. They can be used to teach core knowledge, clinical reasoning and communication skills, as well as to assess the progress of the learners [10]. One such virtual agent has been implemented by Grundmann [35], and is in the form of a chatbot named Lilobot, who plays the part of a child being bullied at school. The trainee counsellor chats with Lilobot, and their messages are fed into an emotion model which produces an appropriate emotional response based on the context of the conversation and the state of its emotions. The goal of Lilobot is to train counsellors to apply a conversational strategy (which in this case is the Five Phase model [72]) while counselling children.

Lilobot was tested by counsellors at De Kindertelefoon, who stated that it was realistic, and the learning gained from it was insightful and reflective. However, they also felt that the conversation was not very natural, that the agent gave them repetitive answers, and that sometimes it would just not reply to the counsellor's messages.

Typically, the scenarios for training agents (and virtual simulations in general) are handcrafted, and thereby limited in terms of conversational scope. New scenarios and dialogue need to be scripted by people, which is inefficient [31] and therefore infeasible for providing volunteers with a variety of different training experiences. For the purposes of training specifically, this implies repetitive dialogue and a lack of variation in the outcomes and progression of the conversation. Rule-based systems such as the BDI model [68] provide a way to model emotional state, modulate it based on input to the model, and have the benefit of being deterministic, but lack the realism and variation needed [59] to gain additional learning from being replayed.

In tandem with hand-crafted scenarios, fresh responses in these scenarios can also be generated by leveraging the strengths of technologies such as Large Language Models (LLMs). LLMs have shown great promise in their ability to generate text that is (in many cases) indistinguishable from human text [44], and have the ability to produce context-appropriate outputs based on prompts that have a large degree of variation, which ultimately means more human-like conversation [44]. They can process vast amounts of data and can be tailored to be context-relevant, thus making them a promising choice for conversation-building. This research aims to unite a deterministic, rule-based cognitive model with the generative powers of LLMs in the form of an integrated system in order to produce chatbot outputs that are context-appropriate and feel natural to converse with.

1.2. Research Question

This research aims to explore the implementation of an integrated system of the current emotion model and an LLM to improve the efficiency of training by increasing realism and variety in training situations. As such, our research question can be stated as follows:

How can large language models be integrated into rule-based conversational training systems to improve the quality of training?

To answer this, we tackle the following sub-questions:

- *What are the design considerations of such an integration?*
- *What design for an integration would meet these design considerations?*
- *How do users perceive the effect of the integrated design on the usefulness and realism?*

1.3. Research Method

This research follows the Socio-Cognitive Engineering research method [61]. To address the first sub-question, we laid down the foundation for our research, consisting of the relevant domain knowledge, the human factors, and an overview of current and available technology for our proposed integrated system. This allowed us to come up with a set of design considerations for our system. Additionally, we performed an initial exploration with LLMs to understand their strengths, weaknesses and capabilities, as well as viable choices for models and prompts to use in our integration. We answered the second sub-question by designing the system according to the insights gained on the design specifications of the system. In this phase, we also perform a double coder verification to verify and assess the quality of outputs from our LLMs. Following this, we tackled the last sub-question by evaluating our integrated system. We performed a within-subject for the evaluation of our system against the rule-based counterpart and analysed the collected data with various methods including paired t-tests and thematic analyses to directly compare their performance. Finally, we provided suggestions on possible future research into this field.

2

Foundation

In this chapter, we focused on answering the first sub-research question:

What are the design considerations of such an integration?

Through studying available literature, we researched the domains of child counselling, the current methods of training, and the technology that powers our proposed integrated system. The first section contains our findings about child counselling and the current methods of training. Here, we sought to understand the goals of the training and how counsellors are trained in the skills needed to productively engage with children who seek help. To translate such training practices and goals effectively into a chatbot system, it is important to understand the needs and priorities of both the counsellors and those who train them. Thus, the second section specifies the needs and priorities of the counsellors and instructors through human factors. Finally, the last section delves into literature about technology for virtual training systems, including rule-based systems, as well as the current state of the art in the field of large language models. In the same section, we also present the results of our explorative prompt analysis where we sought to establish some information on the use of LLMs for this research. Overall, this chapter presents a set of design considerations which drove our design for an integrated training system.

2.1. Literature Review

For our literature study, we used Google Scholar as our main search engine. In addition, the supervisors provided valuable literature about counselling, training process, and training systems. For the operational demands, we consulted literature about counselling, the skills counsellors need, and the way they are trained to give them experience applying these skills. This gave us insights about what the counsellors' priorities are with training systems, and the outcomes of learning. For the human factors, we consulted literature about the motivational and learning factors which influence learning and then applied these lessons to our context of training counsellors. Finally, for the technology, we looked at the work on the existing BDI model and searched for papers about LLMs, including current integrations of LLMs with rule-based systems.

2.2. Operational Demands

Over the years, counselling helplines have seen a steady increase in the number of people who contact them seeking help and advice [41]. The two main modes of communication offered by these helplines are call and chat, through which counselling sessions are conducted. At De Kindertelefoon, 24% of all counselling sessions are done over chat [48]. With the growing popularity of chat counselling, building a system for training counsellors requires an understanding of chat counselling and its unique advantages and disadvantages.

2.2.1. What is chat counselling?

Chat counselling is a virtual conversation where counsellors and help-seekers (who, at child helplines, are primarily children) converse through internet-based chat technology [72]. The purpose of chat counselling is to offer information, advice and support for the social and emotional problems that the users of the service may have.

For children, the internet is the largest source of healthcare information [34], with over 84% of adolescents in the USA (aged 13-18) having sought health information online [22]. For helplines, while children as young as nine years old reach out for help, the majority of the people seeking help are between the ages of fourteen and twenty years old [72]. At De Kindertelefoon specifically, most of the help-seekers are between the ages of eight and eighteen years old [48]. The main topics of conversation for those who reach out to De Kindertelefoon revolve around sexuality, emotional problems, home life, relationships and bullying.

2.2.2. What are the unique strengths and weaknesses of chat counselling?

The combination of the medium and the demographic of counselling grants chat counselling some unique advantages that benefit the counselling process. Chat counselling has been found to have certain advantages simply by virtue of its medium. For children, who are the primary demographic of child helplines such as De Kindertelefoon [48], the internet is very familiar [79, 16], making online chat a convenient and widely accessible medium of counselling. Furthermore, the written/typed nature of the conversation allows the child to spend more time in their reflexive process, allowing them to formulate their messages with greater clarity. The transcripts of the conversations can act as an asset - for the counsellor, a tool for self-reflection and teaching, and for the child, a reminder of the advice and knowledge gained while in counselling. As the sessions are over chat, the child has control over some key aspects of the interaction with the counsellor - the time and location, the pacing and the amount of information they share with the counsellor. This level of control allows for levelling of the traditional power imbalance [13, 72] between counsellor and child. Finally, chat conversation lacks the modes of voice, face and embodiment that are present in other forms of counselling such as voice, video and in-person counselling. This reduction of modality known as channel reduction, has been shown to increase the openness and self-disclosure of the children while talking to counsellors.

On the other hand, chat counselling also presents some unique challenges due to the medium. Unlike in-person sessions, chat counselling sessions at helplines are not scheduled beforehand, which means that for a counselling session to happen, both counsellor and child must be present online at the same time. When chat counselling sessions do happen, they take around five times longer than equivalent call sessions [72], which also means larger amounts of time spent sorting out misunderstandings and miscommunications. Although the reduction of channels accelerates the relationship of trust between the counsellor and child, the lack of channels also means that counsellors no longer have access to additional forms of contextual information about the child, such as body language, vocal tone, facial expression, etc. Finally, the anonymity offered by chat helplines may attract people who contact the helpline pretending to be someone else, or those who act in abusive ways towards counsellors.

2.2.3. What are the key skills and competencies for a counsellor?

The key skills and competencies for online counsellors share a great deal of common ground with those used during in-person counselling sessions as well [7, 40, 60]. Some key skills and competencies for counsellors in an online setting are:

- **Rapport-Building:** Rapport-building is the process of establishing a relationship of trust and empathy between counsellor and child in order to encourage the therapeutic process between them. Rapport-building is widely regarded as an essential step in the initial phase of the counselling process [60, 6, 29].
- **Reflecting, Summarising and Paraphrasing:** Reflecting, summarising and paraphrasing are the processes of gathering information, analysing and interpreting it in order to be able to communicate better with the child. These skills are considered to be key elements of the counselling process [40, 87] and allow for a counsellor to have a broader understanding of a child's emotions, cognition and behaviours [6].
- **Empathy:** Empathy is the ability to understand and share the feelings of another. Empathy

and the ability to express it are two key traits/skills the counsellor must have, as they allow the child to feel understood and less emotionally distressed [49]. Acting in an empathetic way has been shown to increase the client's (in this case, the child's) willingness to share more about their problems, thereby strengthening the therapeutic relationship between counsellor and child.

- **Questioning:** Asking questions and clarifying information is crucial to understanding the scope and depth of the child's problems, and there is a host of literature to suggest that asking open and closed-ended questions is vital to the discovery process [71, 60, 49, 40]. Interpreting and clarifying information well is especially important in online counselling situations, where the counsellor can not rely on non-verbal cues to provide them with further contextual information [74].
- **Application of Conversational Strategy:** Online counselling sessions typically follow conversational strategies such as De Kindertelefoon's Five Phase Model [72]. The goal of the Five Phase Model is to conduct the counselling session in a structured way which guides the conversation all the way from initial rapport, discovery, and finally to the establishment of realistic goals. Different works each outline their own versions of conversational strategies [71, 40, 60, 6], but collectively speak of conversational strategies as essential tools to provide structure in online counselling situations [7].

In addition to the key skills described above, other skills which are also used in counselling sessions are online communication [71, 49], time management [87], encouragement [49] and maintaining confidentiality [70].

2.2.4. How are counsellors trained?

As counsellors need several key skills in order to have effective and safe conversations with children, they need to be trained on both the theoretical knowledge and practical skills which they will use. The most common and effective methods by which counsellors are trained theoretically and practically are described in Sindahl [72].

Theoretical knowledge about counselling fundamentals and conversational strategies can be taught through information sessions, live supervision [72], and chat transcripts, which can be used for teaching and feedback.

In addition to learning theoretical knowledge about counselling practices, counsellors need to gain practical experience in applying their skills and knowledge. Typically, role-play sessions [72] are used in order to give them practical experience in a safe training environment. Counsellors take turns role-playing as child and counsellor and converse with each other as they would in an actual counselling setting. Here, they learn how to talk to a child, and importantly, how to apply the conversational strategy. Counsellors who take part in role-play exercises reflect that they feel as though they have benefited greatly from it. The purpose of our research is to study the use of virtual agents in the specific context of giving counsellors simulated practical experience in talking to children.

2.3. Human Factors

In the previous section, we outlined the core skills and competencies of counsellors, and how the training they receive helps them learn these skills and competencies. In alignment with these, we identify a set of human factors that relate to the direct stakeholders of an integrated training system.

2.3.1. Motivational Factors

First, we present a set of motivational factors that relate to learning and academic success. Linnebrink et al. [57] discuss four key components that relate to academic success - self-efficacy, motivation, attributions and achievement goals - of which we examine the relevant components through the lens of the counselling landscape.

Self-Efficacy

Self-efficacy is the degree to which an individual considers themselves capable of organizing and executing the behaviours needed to successfully complete a certain task [53, 21]. Research shows that there is a moderate link between self-efficacy and the knowledge acquisition and subsequent task performance of learners [67, 30]. For a counsellor, their self-efficacy refers to their belief about their own

capability to counsel a child by applying their knowledge and skills.

Motivation
The motivation a learner has, whether intrinsic or extrinsic, has a positive effect on the accomplishment of learning goals and learning behaviour [62]. The motivation of a learner can be influenced by factors such as the perception of the workload and the complexity of the task [26]. Research suggests that the more the learner is motivated to learn for autonomous reasons, the more they will be inclined to use deeper approaches to learning.

Attribution
Attribution theory [80] suggests that when individuals face failure or success, they will analyse the perceived causes for failure or success. According to attribution theory, this focus on the reasons for failure or success explains psychological outcomes such as self-efficacy, and possible indirect links to academic achievement. For counsellors, this means that feedback on their performance through self-reflection and instructors could mean increased self-efficacy to perform better in subsequent training tasks.

Cognitive Load
Training situations for counsellors need to be properly attuned to their level and capabilities. A learning situation offers optimal learning opportunities if a learner is able to cope with the demands of the situation while also having their limits tested [63]. If a learning situation is too easy it will not adequately capture or motivate learners, and if a learning situation is too tough learners cannot cope with the demands. In either case, this leads to a lack of motivation.

2.3.2. Learning Factors
In addition to the motivational factors, we look at learning factors which influence the effectiveness of instruction and practice. Merrill [59] presents an instructional theory based on a broad review of other instructional theories and models known as the First Principles of Instruction. They argue that these interrelated principles, when applied properly in an instructional setting, increase student learning. We present the relevant principles, linking them to our context of training for counsellors.

Real-World Problems
Merill argues that learning is promoted when learners are engaged in solving real-world problems. The goal of training counsellors is to give them the ability to apply their learned skills and knowledge when faced with real counselling situations. In the context of education, this is known as transfer, which refers to the transferability of skills from training into practice [33]. Thus, the training scenarios counsellors are trained on must be authentic to those they will face on the job in order to best promote the learning of key counselling skills and knowledge.

Application, Feedback, Variation
This principle argues three main points. First, it argues that learning is promoted when learners are required to use their newly acquired knowledge and skills to solve problems. In the context of counselling, this includes the practical application of counselling skills through the use of role-play or agent-based training systems. Secondly, it claims that learning is promoted when learners are adequately guided in their problem-solving through feedback and coaching. Finally, it also argues that learning is promoted when learners are required to apply themselves to a variety of different problems, which for counsellors could imply different training scenarios with different personas and problems.

Reflection, Creation
This principle claims that the process of integrating knowledge and skills into one's repertoire is better accomplished when learners have the ability to demonstrate improvement in skill, defend their knowledge and personalize it. Reflection is the process by which learners can reflect on, defend, and share what they have learned, and research suggests that it is an important activity for collaborative problem solving [81]. Creation is the process by which learners can make their learned knowledge their own and is also considered an important activity in the final phases of a learning experience [58]. For counsellors, being able to reflect on their practice through transcripts and instructor feedback, as well as being able to personalize their counselling style might contribute to the effectiveness of their learning experience.

2.4. Technology

Through literature, we have identified the core skills and competencies for counsellors, how they are trained in these skills and competencies, as well as the corresponding human factors. This section presents the current and prospective technologies that can be used in an integrated system in order to increase the quality of training offered by such systems.

2.4.1. Training Systems

The use of intelligent virtual agents [28] is increasingly becoming commonplace for education and training purposes. Such agents provide a powerful and promising medium for experiential learning [69] offering students virtual training based on real-life scenarios. In the context of healthcare education, virtual agents that simulate diagnostic and clinical situations are referred to as virtual patients [86]. Their purpose is to allow learners to learn and apply core knowledge, communication skills, and clinical reasoning and to assess their progress [11, 50] while allowing learners to practice important diagnostic skills in a safe and cost-effective simulated environment. A literature review by Cook et. al. [19] on the use of virtual patients in medical education settings compares the use of virtual patients against other forms of interventions (methods of training) for learners. They find that when compared to no intervention at all, the use of virtual patients leads to significantly higher performances on learning outcomes. When comparing interventions involving human participants with the use of virtual patients they find that the difference in information elicited and the number of correct diagnoses is not significant, with the only difference being the learners treating the virtual patients with less warmth and empathy than the human participants.

The term 'virtual patient' is broad and can be used to describe a multitude of technologies and approaches, but typically refers to any software that allows case-based training [51, 88]. As such, different virtual patients may have very different functionalities based on the diagnostic purpose. For instance, an agent built for simulating virtual needles for simulating regional anaesthesia can have accurate simulations of nerve endings [78]. Similarly, the use case of virtual patients for training children's counsellors comes with its own unique functionalities and requirements. As counsellors deal with humans who may have complex emotional circumstances, virtual patients for training them need to simulate emotion and belief states that react to counsellor input. One way of modelling a virtual agent for this purpose is the BDI model.

2.4.2. The BDI Model

The BDI (beliefs-desires-intentions) model [68] is a rule-based emotion model that has been developed for programming rational agents which simulate human-like reasoning and decision-making processes. When used in a conversational agent, the BDI model allows for the agent to have a decision-making process, and provides a discrete way to model and manipulate its emotional and belief states. To accomplish this, the BDI model draws from Bratman's theory of practical reasoning [14], and has three key components - beliefs, desires and intentions.

Beliefs In the BDI model, the agent has a set of beliefs about different themes which capture its informational state - its beliefs about the world.

Desires In addition to the beliefs, the agent also has a set of desires. The desires represent a set of objectives or situations that the agent would like to accomplish or bring about.

Intentions An intention simply represents the deliberative state of the agent, or what the agent has chosen to do. In other words, an intention is a desire which the agent has committed to follow. An intention is linked with a set of actions which the agent will perform in order to follow its deliberative state.

The agent reacts to events, which are internal or external triggers (e.g. user chat input) that may change the belief state of the events. This in turn also affects the state of the agent's desires and intentions.

As the BDI model is rule-based, it has the same advantages that rule-based systems do - it is deterministic, consistent, and transparent [54]. However, it is also subject to the disadvantages of rule-based systems, in that it is limited, inflexible, cannot adapt, and does not scale well for large use cases.

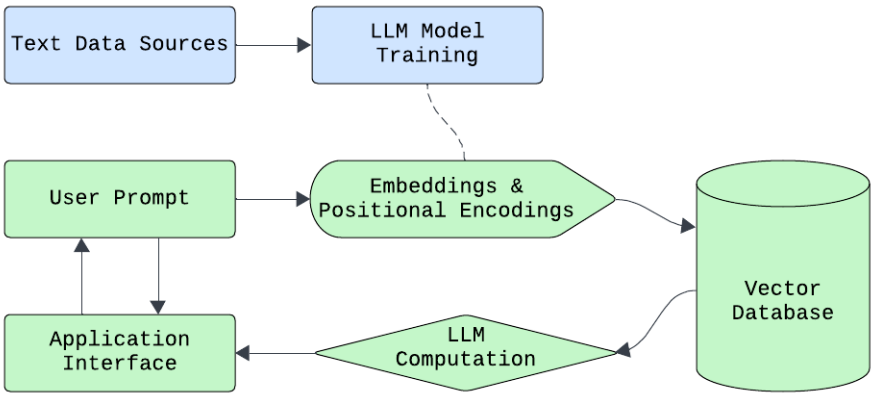


Figure 2.1: Flow inside a Large Language Model from Karanikolas et al. [44]

By contrast, large language models are adaptable and scalable but remain unpredictable and stochastic. Thus, there is promise in uniting the powers of both types of systems, in order to make up for the weaknesses of both individually.

2.4.3. Large Language Models

Large Language Models are a class of deep learning systems that are designed to process language. These models are trained on vast amounts of textual data, and their objective in doing so is to be able to interpret input text and generate context-appropriate responses [12]. Large language models follow instructions given to them in the form of prompts, and models can even be fine-tuned to improve their performance in domain-specific contexts [38, 44].

Large language models have been increasing in popularity and usage, with models such as OpenAI's ChatGPT [1], Meta's Llama [2] and Anthropic's Claude [3] acting as the state-of-the-art in the field. Large language models show great promise in the processing and generation of natural language [44], and these abilities are well suited for integration with a rule-based system such as a conversational agent, where the quality of dialogue plays a big role in the conversational experience.

How do Large Language Models work?

Figure 2.1 shows the basic data flow within an LLM, based on the workflow in Karanikolas et al. [44]. A vast amount of textual data with great variety is fed into the model. The LLM is trained on this data, where all of the knowledge from the data is represented as embeddings, which are numerical representations of words in the form of vectors. These vectors broadly capture the words' semantics, contextual relationships and meanings in different contexts. When a user gives an LLM a prompt, the prompt undergoes the same transformation to vector embeddings. Then, the LLM gathers a response from its knowledge base which best matches the intent of the entered prompt. Finally, NLG techniques are used to transform the gathered information into human-like output.

What tasks must an integrated learning agent perform well?

Hollender et al. [39] propose through their cognitive load theory that there is a limited working memory which we have access to while learning, and that there are three types of cognitive load that consume this memory. Intrinsic cognitive load refers to the intrinsic load of the information that is to be learned. Germane cognitive load refers to the cognitive load which is a result of learning itself, through the formation of schemas. Finally, extraneous cognitive load refers to the cognitive load resulting from an inappropriate presentation of learning material, outside of learning itself. Thus, any instructional design has to take the limitations of working memory into account so that it does not overload the limited working memory capacity in order to promote more effective learning.

For use in an integrated system, we have identified a set of relevant tasks that a large language model must perform well in in order to reduce the extraneous cognitive load involved in the learning process.

The survey by Chang et al. [17] is an amalgamation of evaluations of LLMs in different natural language processing (NLP), natural language understanding (NLU) and natural language generation (NLG) tasks, from which we have selected the relevant tasks:

- **Sentiment Analysis:** Sentiment analysis is the process of determining the emotional inclination from text. Chatbots and virtual agents that adapt their responses according to user sentiment are perceived as more anthropomorphic, and socially present and yield a higher satisfaction with the interaction [24]. Liang et al. [56] and Zeng et al. [89] showed that the performance of LLMs on sentiment analysis tasks is high. Models like ChatGPT [1] and Llama3 [2] especially perform exceptionally well in this regard.
- **Semantic Understanding:** Semantic understanding is the ability of an LLM to understand language and its associated concepts, such as the interpretation and comprehension of words, phrases, sentences and the relationships between them. In essence, it is the extraction of the underlying meaning and intent of the text. Tao et al. [77] find that although LLMs possess an understanding of individual events, their ability to perceive semantic similarity between events and general semantic proficiency is weak.
- **Question Answering:** Question answering is a central task for a wide variety of applications and scenarios such as search engines, customer service, and chatbots, and is important for a conversational agent such as the integrated system. Models like ChatGPT [1] and GPT 3.5 perform well when evaluated on tasks such as CommonsenseQA [76], which is a dataset for the test of common sense question-answering tasks.
- **Dialogue:** Dialogue tasks measure the NLU, contextual understanding and generation abilities of a model, which translate to a more intelligent and natural dialogue system. Bang et al. [9] find that although fully fine-tuned models perform best, generalised pre-trained models such as ChatGPT and Llama3 still perform well in dialogue. However, some models tend to struggle with maintaining belief across multiple turns in the interaction and sometimes hallucinate information that is not given.
- **Sentence Style Transfer:** For training in an integrated system, the immersion of training might be improved the more accurately the agent is able to adhere to a child's writing style. Pu et al.[66] demonstrate that models such as ChatGPT perform well on sentence style transfer tasks, as indicated by the high BLEU [82] scores, which evaluate the quality of text that has been machine translated from one style to another.

Our goal with the integrated system is to have it perform the above tasks well, as they are central to increasing the realism, continuity and immersion of the learners' interactions with the learning agent. Performing well in these tasks leads to a reduction in extraneous cognitive load, thus allowing for an increase in intrinsic cognitive load, which translates to more effective learning [75].

What existing solutions combine LLMs and rule-based systems?

In the previous section, we identified a number of possible areas of integration based on the architectural model of an interactive agent. In this section, we examine available literature on existing integrations to determine the feasibility of those we proposed in the previous section, stressing the performance of said integrations in those key areas.

Pico et al. [64] investigate the application of large language models for emotion recognition. Specifically, they investigate LLMs in the context of generating emotional knowledge in the form of beliefs, which can be used by an agent that employs a BDI model. To assess the LLMs, they prompt them with a piece of dialogue and ask them to pick an emotion from a given list which best matches the given dialogue. Then, they compare the predicted emotion with a ground truth label to calculate the prediction accuracy. Their study compares the performance of several LLMs in the task of recognizing emotions, using two different datasets [65, 32]. They find that LLMs demonstrate promising capabilities for emotion recognition even without training or examples.

Gürcan [37] explores architectures and methods for LLM-augmented agent-based social simulations. They propose a number of different LLM-augmented agent-based simulation situations, spanning a vast variety of applications. By examining the strengths of LLMs and the corresponding areas of weakness in agent-based social simulations over multiple domains, they claim that LLMs offer a transformative

framework for the simulation and analysis of social systems. On the other hand, they also emphasize the risks of relying on them as sources of knowledge, mentioning that LLMs often provide the illusion of understanding, rather than having an actual understanding of the context they are in.

For the purpose of building an integrated system with both rule-based and LLM components, an important task for the LLM is to learn and adapt to situations where there is a lack of resources to learn from. Coleman et al. [18] implement an LLM-based translator between English and an obscure language for which there exists little to no resources to train. They provide the LLM with a set of translation rules and a small sample of example translations and find that their system is able to perform well, maintaining a high degree of semantic similarity between the English and translated sentences. Their work implements a system that lies in both the NLG and NLU areas and demonstrates the ability of LLMs to learn from limited amounts of data while maintaining high performance.

Yet more research into the integration of LLMs and rule-based systems again shows that LLMs are adept at learning and providing desirable output even with small numbers of examples or demonstrations. Li et al. [55] propose a system to rewrite SQL queries to make them more efficient by using an LLM and providing it with an exhaustive set of rewrite rules. Correctness is a central requirement of the system - the produced queries must be equivalent in function to the original and must be more efficient. Their work finds that LLMs are able to use the rewrite rules to produce quality output queries which demonstrates semantic understanding and reliability of generated outputs in relation to the provided rules.

For sensitive purposes such as training counsellors, an integrated system must be able to produce outputs that are relevant to the given context without fabricating unfaithful or nonsensical information (referred to as hallucination [43]). Isaku et al. [42] implement an LLM and rule-based system for the purpose of generating test cases for medical rules in the context of cancer data validation. The correctness of these test cases is vital, as the misclassification of patient cancer data could have serious implications. They find that LLM models such as GPT 3.5 are highly effective in generating medical rule tests with little hallucination.

Although the above implementations are tested, there exist other implementations that are untested. For instance, Goel et al. [31] implement an LLM-based task-oriented dialogue system with few-shot retrieval augmentation. In simpler terms, their implementation takes a user prompt and generates a prompt for an LLM which consists of the context of the conversation along with a small set of examples of relevant data. These act as instructions and demonstrations (few-shot learning) of how the task is usually carried out. This implementation uses LLMs for their capabilities in both NLU and NLG tasks.

2.4.4. What are the possibilities for an integration?

Al Owayyed et al. [4] present the general architecture of agent-based training for social skills (ARTES). Figure 2.2 shows a small part of this general architecture, specifically the architecture of the interactive agent simulator. We have identified three potential areas for the integration of the system - NLU, Thinking and NLG.

NLU: Natural Language Understanding is the ability of a machine to understand, interpret and derive meaning from human language [85]. Large language models are known for their performance in NLU tasks [77]. This area of implementation proposes using an LLM as an intent recognition tool to derive intents from user input to pass on to the BDI model.

Thinking: This proposed area of implementation deals with using a large language model to control the BDI model. The large language model receives user intent and contextual information relating to the dimensions of the BDI model, the persona it is playing, etc. and provides an output which controls the dimensions of the BDI model.

NLG: Natural Language Generation is the ability of a machine to generate human-like natural language [84]. Large language models are known for their performance in NLG tasks [77]. This area of implementation proposes using an LLM as a way to generate text output from the output of the BDI model.

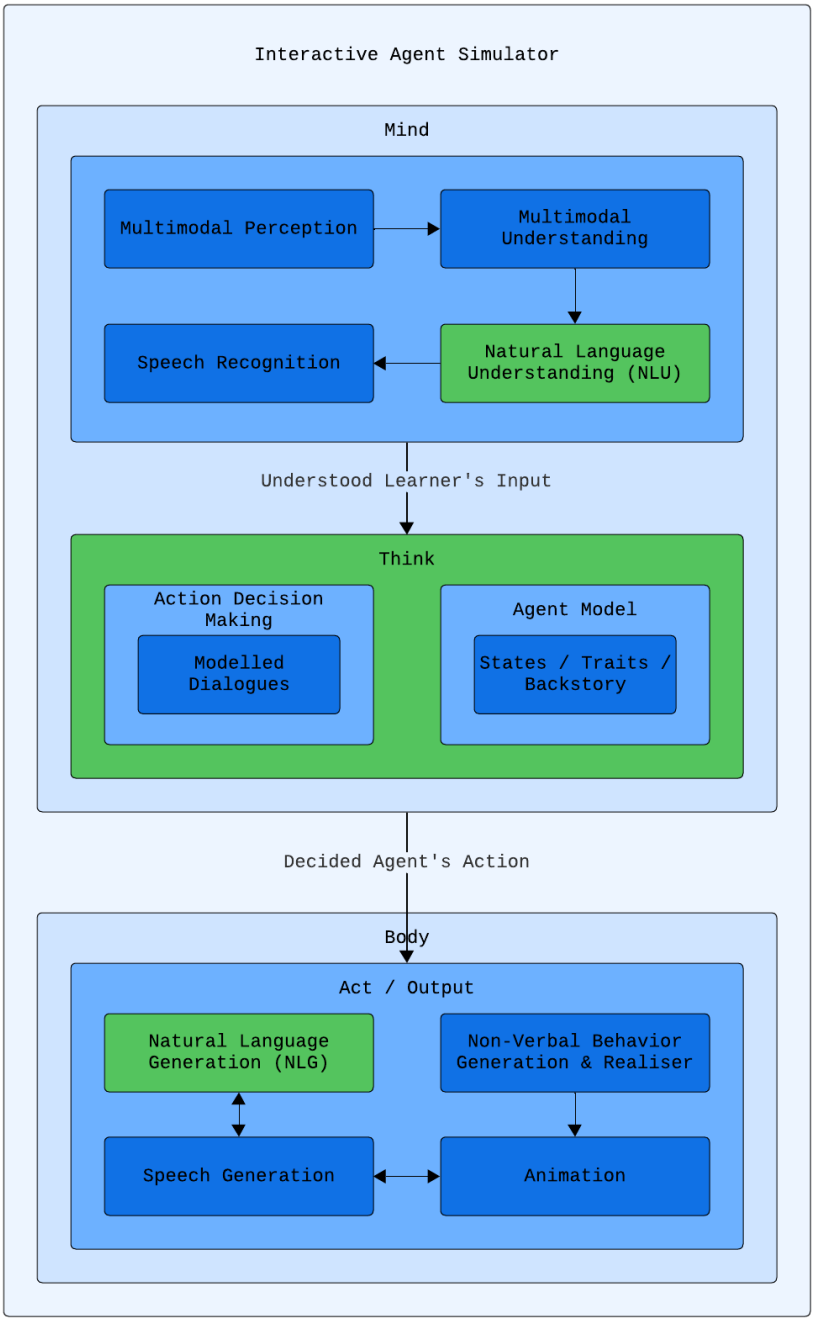


Figure 2.2: Interactive Agent Simulator model, from Al Owayyed et al. [4]. The parts highlighted in green are areas where the integration of an LLM may be possible.

Combining these three main areas also opens up some possibilities for integrations which may be viable.

NLU + Thinking + NLG: This area of implementation proposes using an LLM to handle all three highlighted parts of the system. The NLU and NLG parts use LLMs to handle the input and output, and the Thinking part uses an LLM to control the BDI model, changing the belief, desire and intention state of the model.

From our initial exploration in prompting and observing the performance of LLMs in these three areas, we have found that our prompts work effectively in the tasks of intent recognition for the NLU component, as well as the dialogue generation for the NLG component. However, LLMs display poor performance in tasks that require them to control the BDI model (the thinking part). As such, a particular solution that is of interest to us is to use the LLM for NLU and NLG tasks while keeping the rule-based BDI model.

NLU + NLG: Although large language models display high performance in NLU and NLG tasks, at their core they are still probabilistic models. This means that they produce outputs non-deterministically, which for the sensitive purposes of counselling might mean unpredictable responses which may be inappropriate or unrelated. By contrast, the BDI model is deterministic and produces reliable outputs based on user input. Thus, this solution proposes that an LLM can be used to handle the NLU and NLG tasks that an intelligent agent must perform, while still maintaining the deterministic and reliable core of the BDI model. For an integrated solution, this could imply better intent recognition and more natural, human-like outputs, while still maintaining a high degree of context-awareness through guidance from the BDI model. Furthermore, LLMs could also aid in expanding the limited scope of deterministic, rule-based models such as the BDI model, by being able to use their inherent NLU and NLG abilities to respond to prompts that are outside of the known scope of the BDI model.

2.4.5. LLM Prompting: Exploration

There is a host of literature that suggests large language models have powerful language processing capabilities. For the specific case of our integrated solution, we decided to do an initial exploration of LLMs by testing them with prompts in order to understand these capabilities and gauge the feasibility of the end solution. In figure 2.2 we identify some parts of the existing system in which we envision LLMs may be of use. Here, we experiment with prompting LLMs to perform the tasks of each of these parts. The prompts and the results are given in Appendix A.

Prompt Engineering

We used available literature and courses about prompting LLMs in order to guide our prompts. Through these methods, we were able to gain insight into the process of prompting and the ways to most effectively make LLMs consistently return favourable results. In order to explore the possibilities for prompting and to have a wider understanding of prompts and their performance with different LLMs, we tested multiple versions of each prompt on three different LLMs (GPT 4o mini, Llama3 and Mistral).

Elements Tested with LLMs

We tested the capability of LLMs in the areas of NLU (for intent recognition), NLG (for dialogue generation), and thinking (to control the BDI system) in relation to the actual data that we will be using in the integrated system.

For each component, we tested the abilities of the three aforementioned LLMs with three different prompts which have a small degree of variation between them. We also ran each prompt multiple times in order to check if the results were consistent between runs.

NLU In order to test the NLU capabilities of LLMs, we constructed prompts which task them with the identification of intents from input dialogue. Specifically, the LLMs were provided a list of intents and short explanations about them. Then, they were tasked with identifying intents from the current knowledge base of the rule-based system, or flagging if no intent was identified.

We tested three different LLMs on three different prompts which had a small degree of variation. The first prompt (Appendix A.1.1) asks the LLM to assume the character of a smart assistant which classifies

Table 2.1: An overview of the results from our exploration with LLMs for our NLU task.

	$Prompt\ 1$	$Prompt\ 2$	$Prompt\ 3$
$Llama\ 3$	Great performance overall, good output format. $Single$: 13/13 (100%) $Multi$: 5/6 (83%) $None$: 2/2 (100%)	Persona does not change the performance significantly. Sometimes invents intents. $Single$: 13/13 (100%) $Multi$: 4/6 (66%) $None$: 2/2 (100%)	The lack of explanations significantly worsens performance. $Single$: 11/13 (84%) $Multi$: 4/6 (66%) $None$: 0/2 (0%)
$GPT\ 4o$	Good performance, sometimes misinterprets intents. $Single$: 13/13 (100%) $Multi$: 6/6 (100%) $None$: 0/2 (0%)	Performs worse than with Prompt 1. $Single$: 12/13 (92%) $Multi$: 6/6 (100%) $None$: 0/2 (0%)	Invents intents that are not in the knowledge base. $Single$: 9/13 (69%) $Multi$: 4/6 (66%) $None$: 0/2 (0%)
$Mistral$	Best performance overall, but does not adhere to the output format strictly. $Single$: 13/13 (100%) $Multi$: 6/6 (100%) $None$: 2/2 (100%)	Sometimes invents new intents. $Single$: 12/13 (92%) $Multi$: 5/6 (83%) $None$: 2/2 (100%)	Invents intents, detects multiple intents in single intent test case. $Single$: 8/13 (62%) $Multi$: 4/6 (66%) $None$: 1/2 (50%)

text into different intents. The second prompt (Appendix A.1.2) changes the persona of the LLM instead to a child who is being bullied at school and talking to a child helpline counsellor, having to interpret their messages. The third prompt (Appendix A.1.3) is identical to the first prompt, and removes the explanations for the intents. For each test case, we have 21 test cases in total: 13 where there is a single intent in the input message, 6 where the are two or more intents, and 2 where there are no intents to be identified.

An overview of our findings is shown in table 2.1. We found that the best and most consistent performance is achieved by Llama3 and Mistral using the first prompt, where the expected intents are identified in all of our test cases, whether there are single, multiple or no intents to be correctly identified. Additionally, we find that adding the persona of a bullied child (second prompt) and removing the explanations for intents (third prompt) have a negative effect on overall performance. As such, we chose to use the combination of the first prompt and Mistral to construct the NLU component of our integrated system.

NLG In order to test the NLG capabilities of LLMs, we constructed prompts which task them with the generation of responses which are appropriate contextually and in content. For this, there are two tasks that we tested LLMs on - the generation of responses in the case of an identified intent, where they must generate a response that is similar to a given set of examples, and the generation of responses in the absence of any identified intent, where they must generate an appropriate response in reply to a given input message.

We tested three different LLMs on three different prompts which had a small degree of variation. The first prompt (Appendix A.2.1) asks the LLM to assume the character of a 9-year-old child who is being bullied at school and talking to a child helpline counsellor. The second prompt (Appendix A.2.2) changes the persona of the LLM instead to a smart assistant that generates responses according to instructions. The third prompt (Appendix A.2.3) is identical to the first prompt, but now provides the message to which a response must be generated as a part of the input. For each test case, we have 10 test cases in total: 8 where there is an identified intent (and thus example messages) and 2 where the intent is

Table 2.2: An overview of the results from our exploration with LLMs for our NLG task.

	$Prompt\ 1$	$Prompt\ 2$	$Prompt\ 3$
$Llama\ 3$	Best performance overall. $Known$: 8/8 (100%) $Unknown$: 2/2 (100%)	Lack of persona makes responses worse in case of no intent. $Known$: 8/8 (100%) $Unknown$: 0/2 (0%)	Very hit or miss - prompt instructions sometimes just not followed. $Known$: ?/8 (?%) $Unknown$: ?/2 (?%)
$GPT\ 4o$	Best performance overall. $Known$: 8/8 (100%) $Unknown$: 2/2 (100%)	Lack of persona makes responses worse overall whether or not intent is identified. $Known$: 8/8 (100%) $Unknown$: 0/2 (0%)	Same performance as prompt 2. $Known$: 8/8 (100%) $Unknown$: 0/2 (0%)
$Mistral$	Output format and instructions not followed. $Known$: 0/8 (0%) $Unknown$: 0/2 (0%)	Output format and instructions not followed. $Known$: 0/8 (0%) $Unknown$: 0/2 (0%)	Output format and instructions not followed. $Known$: 0/8 (0%) $Unknown$: 0/2 (0%)

unknown.

An overview of our findings is shown in table 2.2. We found that the best and most consistent performance is achieved by Llama 3 and GPT 4o using the first prompt, where appropriate responses are generated in all of the cases regardless of whether the intent is known or unknown. Additionally, removing the child's persona (second prompt) or adding the message to respond to (third prompt) have a negative effect on overall performance. As such, we chose to use the combination of the first prompt and Llama3 to construct the NLG component of our integrated system.

Thinking Across all the experiment setups using different prompts and LLMs for the Thinking component, we observed poor performance. The LLMs consistently return results that either do not follow the given rules for belief updates or results that create new belief values altogether. As such we choose to use the original rule-based BDI model for the thinking component in our integrated system.

In cases where the input intent is not in the knowledge base of the BDI model, an appropriate response to the input cannot be fetched. In such cases, the system fails to respond to the input in any meaningful manner, opting instead to use a default response of "I do not understand" or a similar phrase. Our design choice in order to make the system more flexible with respect to user inputs in the case of unknown intents involves passing user input directly through to the NLG component of our system. By doing this, we rely on the LLM in the NLG component to generate a context-appropriate response to the user input instead of returning the default response when an intent is not known to the system.

Prompting Findings

Our research on prompting techniques and our exploration have given us some insight into the key prompting techniques that produce desirable results for our use case.

We find that defining the task clearly in concise, formal terms while following a specific Instruction-Input-Output format greatly improves the consistency of results. As such, we divide our prompts into clear instruction, input, and output sections, with each section clearly defining the task and the expected format for inputs and outputs. Additionally, providing demonstrations of optimal task performance act as a valuable way to give the LLM more context on the expected performance, which we also observe to give good results in line with the expected input and output format. Finally, we iteratively tested and

improved our prompts, changing the phrasing and structure in small steps in order to better understand the effect that these changes have on the task performance. We used these techniques in order to arrive at prompts that function well in a consistent manner. An example of a prompt that follows these techniques is given below in Listing 2.4.5.

Listing 2.1: An exmaple of a prompt which follows good prompting practices.

```
#### Instruction ####
You are a smart assistant who identifies intent(s) in from input text.
The predefined list of intents is given below.

<intents>

Your task is to identify the intent(s) from the given message.

#### Input ####
You will receive a message as input, which may contain intent(s) from the ones provided.

#### Output ####
Return the identified intent(s) as a list of intents.
If no intent is recognized, return {intent-unknown}.

Return only the identified intent(s). No added explanation, no notes.

#### Examples ####
Input: "Hi"
Output: {chitchat-greeting}

Input: "Who's been bullying you? And how often?"
Output: {bullying-who, bullying-frequency}

Input: "What did you eat for breakfast this morning?"
Output: {intent-unknown}
```

Exploration Conclusions

To be usable for an integrated system, the LLM components we implement must be able to function well within the specific context of a child helpline conversation. The LLM components must show a level of contextual understanding which allows them to interpret and respond to counsellor input consistently while being contextually appropriate. For our purposes, this means being able to respond to inputs that are both within and outside of the scope of the rule-based system while portraying a realistic and varied character of a child in distress. The end goals of such requirements are related to increasing the realism and variety of the training scenarios, and thus the efficacy of the training sessions.

Through various tests using different prompts and different LLMs, we observed that LLMs are able to provide uswith consistent, valuable results for NLP tasks in general. Specifically, we observed good results in the areas of NLU (for intent recognition), as well as NLG (for dialogue generation). However, our exploration showed that using LLMs to control the rule-based model (the Thinking component) provides very poor performance. We used the knowledge gained from exploring prompting with LLMs for the different components in order to guide our design choices for our implementation.

One major limitation we observe with the two locally running LLMs (Llama3 and Mistral) is that some prompts take substantial amounts of time to return output dialogue. By comparison, online LLMs such as GPT 4o have vastly superior response times with comparable performance. This limitation could prove to be a challenge for our use case in a live chat training system, where output dialogue must be generated quickly.

2.5. Design Considerations

In this chapter, we have gained a lot of insight into counselling, the training for counsellors, and current and possible technologies we can use for an integrated system. Based on this knowledge and the human factors we specified, we now specify a list of design considerations and their related design concerns, as shown in Table 2.3.

Table 2.3: The identified design considerations and their corresponding design concerns.

$Design\,Consideration$	$Design\,Concern$
$D1$: The integrated system must be able to train counsellors in counselling skills and using the conversational strategy	Counsellors need a set of skills and competencies in order to be able to counsel children successfully
$D2$: The integrated system must train counsellors using scenarios that are realistic	Transfer and learning are improved when the training scenarios are authentic to real-life scenarios
$D3$: The integrated system must be able to correctly identify intent from user input	The immersion and progression of training is interrupted if the agent is unable to understand user input appropriately
$D4$: The integrated system must be able to correctly respond to user input	The immersion and progression of training is interrupted if the agent is unable to respond to input appropriately

3

Design

In this chapter, we tackle the research question:

What design for a conversational system would meet these design considerations?

We start by presenting the general design, followed by a detailed explanation of how each component of the design is integrated with LLMs. Finally, we present the results of the verification of the functioning of each of the components.

3.1. Design

We presented the following design for our system (shown in Figure 3.1) with a short explanation of each of its components.

The learner sends an input to the system, which is then processed by the NLU component to identify an intent. If an intent that is known to the knowledge base is recognized, it is sent to the BDI system for processing, which changes the cognitive state of the virtual child. Then the intent is returned with a relevant set of example responses for the NLG component, which generates new, contextually appropriate responses to send back to the user. If no such intent can be identified, the bypass component generates a relevant response in the absence of relevant example responses to learn from, which is then sent back to the user.

3.2. Design Components

In this section, we present a more detailed look into the components of our design and how we integrated LLMs into them.

3.2.1. NLU

The purpose of the NLU component is to interpret user input and to classify it into intents that can be handled by the BDI model. For this, we leverage the language processing powers of LLMs. The design of our NLU component is shown in 3.2.

The knowledge base of the rule-based system contains data to assist in the identification of intents in the form of intents, which are mapped to around 2000 examples of input utterances which match them (shown by the yellow sticky note in 3.2). We convert this data into a vector database, where we store the embeddings (numerical representations of text in high dimensional space [5]) of all of the example utterances. When the user sends an input message, it is also converted to an embedding. Then, we use a distance measure (in this case the L2 norm) to find a small set of example utterances in the vector database which are closest to the user's input utterance. We embed both the user's input message as well as the closest matches for example utterances and their associated intents into a prompt for the LLM, as shown in Appendix C.1. The LLM then returns the intent which best matches the input message, or returns 'unknown' if none of the best matches match the input message.

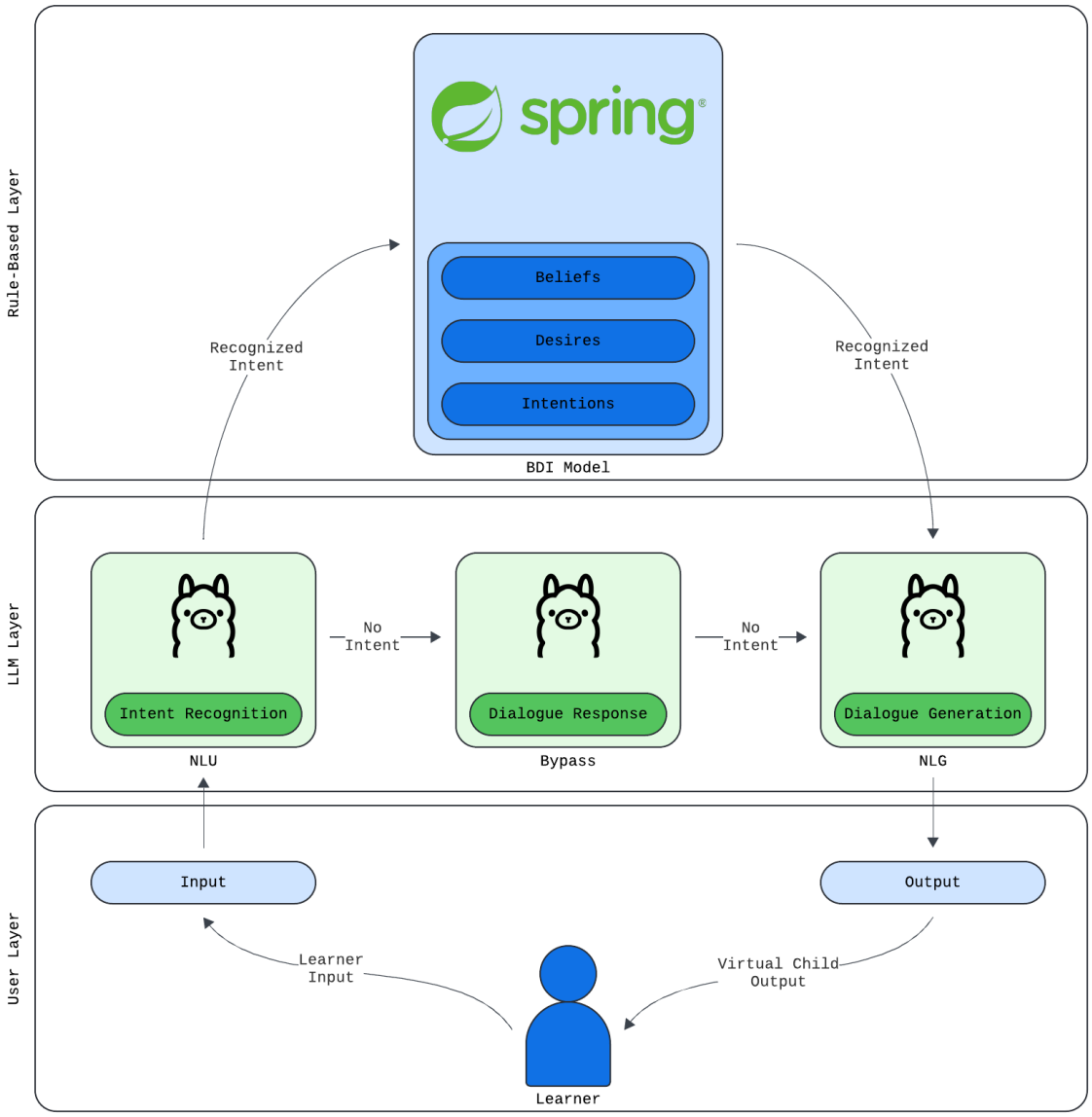


Figure 3.1: The design of the integrated system.

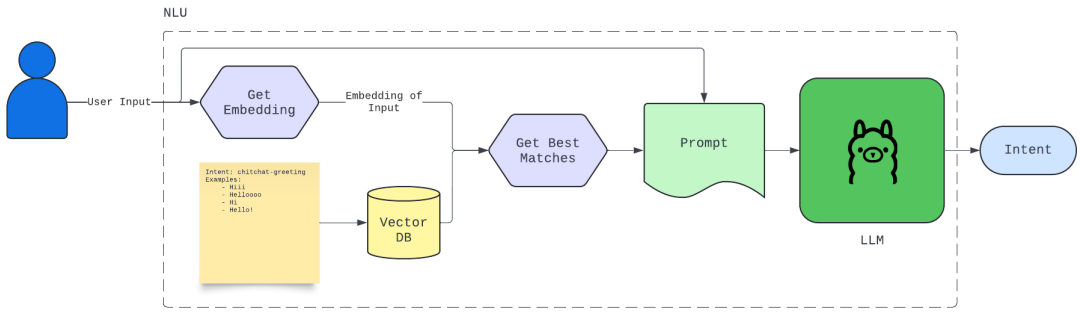


Figure 3.2: The design of the NLU component.

One challenge we faced while constructing prompts for the LLM in the NLU component was the speed of the component itself. In using our older prompts (shown in Appendix A.1), we found that while the model returned accurate results, each prompt took roughly 30 seconds to return a response. This was due, in part, to the hardware we were using, but mostly due to the size of the prompt. By creating a vector database of embeddings for all of the example utterances, we were able to create an efficient system for finding a set of best matches. By using the LLM to act as a verifier for our best matches, and thus only embedding those matches in the prompt, we were able to significantly reduce the size of the prompt (and thus the response time) while still keeping the context we do offer the LLM relevant.

3.2.2. BDI

Although the BDI model is unchanged from the original rule-based implementation, understanding how it handles input can give us more context to understand the reasoning for our design choices while implementing the integration with LLMs.

The BDI model is able to respond to a limited set of intents by changing its internal state based on a set of rules. For instance, receiving a message with the intent `` `ack-empathize` ``, increases the value of belief B05 *("The virtual child thinks the counsellor understands them")* by 0.1. In a similar way, the BDI model is programmed to respond to all of the other recognizable intents based on its set of rules. The change in beliefs then drives the desires, which turn into intents. These intents are then carried out through actions, where the BDI model selects an appropriate response from the knowledge base to send back to the user interacting with it. If the user's dialogue does not belong to the limited set of intents that the BDI model can react to, it cannot change its internal state based on it. Furthermore, it cannot return an appropriate response, as it is unknown to the knowledge base.

3.2.3. Bypass

A central feature of our integrated system is the ability to respond to inputs in the absence of knowledge base examples to draw from. The generative powers of LLMs allow us to generate context-appropriate responses to the user's messages. The design of our Bypass component is shown in Figure 3.3.

As the knowledge base does not have relevant example responses to intents that are not known to it, the rule-based system returns a default response, such as "I do not understand" when such an intent is encountered. In our approach, we embed the user's message into a prompt, as shown in Appendix C.2. The LLM then generates and returns a contextually appropriate response to the user's input.

3.2.4. NLG

The purpose of the NLG component is to generate appropriate responses to the user's messages. We use the language generation powers of LLMs here to accomplish this goal. The design of our NLG component is shown in Figure 3.4.

The knowledge base of the rule-based system contains appropriate responses to identified intents. We use the intent processed by the BDI model in order to fetch a set of example responses that correspond to it. Then, we embed both the user's input message as well as the set of example responses into a

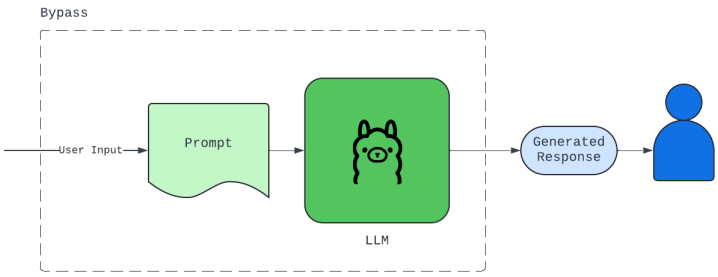


Figure 3.3: The design of the Bypass component.

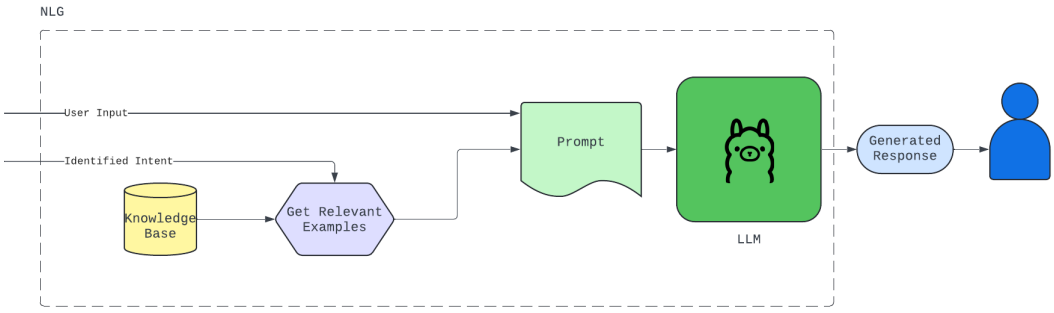


Figure 3.4: The design of the NLG component.

prompt for the LLM, as shown in Appendix C.3. The LLM then generates and returns a response which is similar in content to the examples and is contextually appropriate.

3.3. Verification Through Double Coding

In order to verify the consistency and reliability of the results from our initial exploration with LLMs in our three components, we employed double coders to independently review and validate our results to ensure consistency and reliability.

To gain some insights into the setup when performing such an evaluation, we consulted literature on similar evaluations, particularly relating to evaluating the quality of the responses generated by LLMs. As such, we roughly follow the format of the evaluation described in Steenstra et al. [73]. We recruited four double coders for our evaluation and generated task-specific data from both the LLMs and the current rule-based system as per the needs of each task, which we asked them to evaluate.

3.3.1. NLU

A key part of the process of responding appropriately to the learners' input is to recognise the intent(s) in the dialogue. As such, the purpose of evaluating the NLU component is to determine if the intent(s) identified from the input dialogue by the LLM and the rule-based system match those identified by our human double-coders, thus keeping the results reliable and consistent with our initial exploration. To verify this, our evaluation is as follows.

The double coders were given the input dialogue and the list of intents from the knowledge base with a short explanation of what each intent means. They were then tasked with either identifying the intent(s) which best matched the input dialogue, or flagging if the input did not match any of the listed intents. An example of the task is given in Listing 3.3.1, which shows the input dialogue as well as the intents and explanations. In this case, the intended intent in this input is `chitchat-faring`.

Listing 3.1: An example of the double coders' task for the evaluation of the NLU component.

```
1    Intents: <Intents & Explanations>
2    Input: "How are you doing today"?
```

Table 3.1: Cohen's Kappa scores for agreement between ground truth labels and labels generated by the LLM, BDI and the coders.

$Comparison$	$Cohen's\ Kappa\ (\kappa)$
Ground Truth vs. LLM	1.00
Ground Truth vs. BDI	0.85
Ground Truth vs. Coder 1	0.92
Ground Truth vs. Coder 2	0.63
Ground Truth vs. Coder 3	0.92
Ground Truth vs. Coder 4	0.85
LLM vs. BDI	0.85

The double coders were given 12 input dialogues in total for which they were tasked with identifying intent(s), as shown in B.1. Table B.1 shows the correctness of the ground truth intents in the task and those identified by the LLM, the BDI model and our coders. From the results, we performed numerical evaluations on two fronts.

First, we used Fleiss' Kappa to compute the level of agreement in the identified intents between our four coders. This gave us an understanding of how similarly our coders interpret and extract intents from the input dialogue. We use the script shown in Appendix F.1 to calculate the Fleiss' Kappa score. We obtained a Fleiss' Kappa of 0.338, which indicated a fair amount of agreement $(0.21 < \kappa < 0.40)$ between the coders.

Then, we used the Cohen's Kappa coefficient [83] to determine the level of agreement in the identified intent(s) between the ground truth label (the intents we expected to be identified), and those generated by the LLM and BDI systems and the four coders. Additionally, we also calculated the level of agreement between the LLM and BDI systems. The script we used for this is shown in Appendix F.2, and the results are shown in Table 3.1. Our results show a near-perfect degree of agreement $(0.81 < \kappa < 1.00)$ in all but one of the comparisons, with the outlier having a substantial degree of agreement $(0.61 < \kappa < 0.81)$.

From our results, it can be seen that the LLM is able to identify target intents from input dialogue more accurately than the BDI system and the coders. As such, we are able to justify the use of LLMs as a way to accurately identify input from dialogue regardless of whether or not the knowledge for it exists in the existing BDI system.

3.3.2. Bypass

A key capability of our integrated system is to appropriately respond to inputs for which there is no recognised intent in the scope of the existing BDI knowledge base. To this end, the purpose of our evaluation for the bypass component is to determine if the outputs generated by the LLM in the absence of example outputs to learn from are contextually appropriate with regard to the context of the conversation, the emotional state of the virtual child and the given input.

Through the usage of the rule-based system, we have identified certain utterances for which intents are not recognized, either due to an absence in the knowledge base or a failure in the intent recognition itself. The double coders were given eight input dialogue pieces from this list of utterances, and the reply to the input generated by the LLM. Then, the double coders were tasked with rating each response to the input dialogue in a few different categories on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree) in five criteria, as shown in Table 3.2. An example of the task is given in Listing 3.3.2.

Each dialogue task has five different ratings - one per dialogue rating category. For each, we averaged the five ratings, as shown in Table 3.3. Then we calculated the intraclass correlation (ICC) score to determine the degree of agreement between our four coders using the R script shown in Appendix F.3. Our result from this calculation shows that for the LLM-generated dialogue for the bypass component, the raters had an ICC of $(ICC = 0.246, p < 0.05)$, which implies a low degree of agreement. Table

Table 3.2: Dialogue rating criteria for the Bypass double coder evaluation task.

Dialogue Rating Criteria
$C1$: 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

Table 3.3: Averaged results from our four coders for eight bypass excerpts.

	Bypass LLM Dialogue							
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$
$Coder\ 1$	6.0	7.0	5.0	5.0	5.0	7.0	7.0	6.4
$Coder\ 2$	7.0	7.0	7.0	7.0	6.4	7.0	7.0	5.8
$Coder\ 3$	6.4	7.0	7.0	5.0	5.0	6.6	7.0	7.0
$Coder\ 4$	5.2	5.4	6.0	4.2	4.2	7.0	5.6	4.2

3.4 shows the values for each dialogue rating category for each of the eight dialogue tasks averaged over the four coders. Additionally, it can be seen from the results that for each of our categories $C1 - C5$ the average rating of the dialogue generated by the bypass component is high indicating a high performance in the dialogue generation task.

Listing 3.2: An example of the double coders' task for the evaluation of the Bypass component.

```
1    Input: "'Hows the weather today"?
2    Reply: "Honestly, I don't really care about the weather right now. I'm just trying to
         figure out how to stop this bullying from happening to me."
```

3.3.3. NLG

A central feature of our proposed integrated system is to generate realistic, varied dialogue for each turn in the conversation from the intent and limited set of sample outputs. As such, the purpose of evaluating the NLG component is to determine if the dialogue generated by the LLM is contextually appropriate with regard to the context of the conversation, the state of the virtual child and the given input.

The double coders were given eight pieces of input dialogue, with a brief history of the conversation and

Table 3.4: Averaged results from all the coders for our Bypass task.

	Bypass								
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$
$C1$	7.00	7.00	7.00	6.50	6.50	7.00	7.00	6.50	6.81
$C2$	6.50	6.50	6.25	3.75	3.25	7.00	6.75	6.00	5.75
$C3$	5.75	6.75	5.75	6.25	6.25	6.75	6.50	6.00	6.25
$C4$	5.00	6.25	6.50	6.25	6.25	6.75	6.75	4.75	6.06
$C5$	6.50	6.50	5.75	3.75	3.50	7.00	6.25	6.00	5.67
$\mu(Q_i)$	6.15	6.60	6.25	5.30	5.15	6.90	6.65	5.85	

Table 3.5: Dialogue rating criteria for the NLG double coder evaluation task.

Dialogue Rating Criteria
$C1$: 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

Table 3.6: Averaged results from our four coders for eight LLM dialogue excerpts

	LLM Generated Dialogue							
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$
$Coder\ 1$	7.0	5.8	5.0	6.4	5.6	7.0	6.8	7.0
$Coder\ 2$	7.0	6.2	7.0	7.0	7.0	7.0	7.0	7.0
$Coder\ 3$	7.0	7.0	7.0	6.6	7.0	7.0	7.0	7.0
$Coder\ 4$	5.6	6.4	7.0	5.2	5.2	5.0	4.8	5.4

a response for each. The responses we show and evaluate are of two types. The first type of response is one taken from the existing knowledge base of the BDI model, which was written by human experts, while the second type is one that an LLM generated as a response to the input dialogue. Then, the double coders were tasked with rating each response to the input dialogue in five different categories on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree) in five criteria, as shown in Table 3.5. An example of the task is given in Listing 3.3.3.

Each dialogue task has five different ratings - one per dialogue rating category. For the LLM and human-generated dialogue tasks, we averaged the five ratings, as shown in Tables 3.6 and 3.7. Then for both, we calculated the intraclass correlation (ICC) score to determine the degree of agreement between our four coders using the R script shown in F.3. Our result from this calculation shows that for the LLM-generated dialogue, the raters had an ICC of $(ICC = 0.601, p < 0.05)$, which implies a moderate degree of agreement. For the human-generated dialogue, the raters had an ICC of $(ICC = 0.246, p < 0.05)$, which implies a low level of agreement.

The data in Tables 3.8 and 3.9 show the values for each dialogue rating category for each of the eight dialogue tasks averaged over the four coders. Using this data, we calculated the difference in averages over all eight dialogue tasks for our rating categories between the LLM and human-generated pairs using a paired samples t-test. While the test showed that the LLM-generated dialogue $(M = 6.50, SD = 0.25)$ was rated higher on average than the human-generated dialogue $(M = 6.28, SD = 0.21)$, the difference was statistically insignificant $(t = 1.73, p = 0.13)$. Additionally, we performed a paired-sample t-test with our five rating criteria $C1 - C5$ and found that

Listing 3.3: An example of the double coders' task for the evaluation of the NLG component.

```
1    <Conversation History>
2    Input: "How often do they bully you"?
3    Reply: "It happens a lot, like every other day or so, and 'its really getting to me"
```

3.3.4. Verification Results

From our results, we find that LLMs have comparable performance to human-generated responses and the BDI model for the purposes of identifying intents from input responses and generating appropriate responses to input. For our NLU component, we observed that the LLM is able to correctly identify intents in all our tasks, while the BDI system and human coders perform worse overall. In the case of

Table 3.7: Averaged results from our four coders for eight human-generated dialogue excerpts

	$Human\ Generated\ Dialogue$							
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$
$Coder\ 1$	5.4	7.0	5.8	6.6	5.8	4.0	7.0	7.0
$Coder\ 2$	6.0	7.0	7.0	7.0	7.0	6.6	7.0	6.4
$Coder\ 3$	7.0	7.0	7.0	6.4	6.6	6.6	6.4	5.6
$Coder\ 4$	5.4	5.0	5.6	5.0	6.2	7.0	5.6	5.8

Table 3.8: Averaged results from all the coders for our NLG task with LLM-generated dialogue.

	$NLG:\ LLM\ Generated$								
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$
$C1$	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
$C2$	6.50	6.50	7.00	6.50	6.00	6.50	6.00	6.50	6.44
$C3$	6.25	6.00	7.00	6.00	6.50	6.50	6.50	6.50	6.41
$C4$	6.75	6.50	7.00	5.50	5.50	6.25	6.25	6.50	6.28
$C5$	6.75	5.75	7.00	6.50	6.00	6.25	6.25	6.50	6.38
$\mu(Q_i)$	6.65	6.35	7.00	6.30	6.20	6.50	6.40	6.60	

Table 3.9: Averaged results from all the coders for our NLG task with human-generated dialogue.

	$NLG:\ Human\ Generated$								
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$
$C1$	6.75	7.00	7.00	7.00	7.00	6.50	7.00	7.00	6.91
$C2$	5.75	6.25	6.50	6.25	6.75	6.00	6.75	6.75	6.38
$C3$	5.50	6.50	6.25	5.50	5.75	5.50	6.00	5.00	5.75
$C4$	5.75	6.25	5.75	6.25	6.00	6.00	6.25	5.50	5.97
$C5$	6.00	6.50	6.25	6.25	6.50	6.25	6.50	6.75	6.38
$\mu(Q_i)$	5.95	6.50	6.35	6.25	6.40	6.05	6.50	6.20	

our Bypass component, while we have no direct human responses to compare the ratings of the LLM-generated ones, we observe high ratings across all of our raters for each dialogue snippet. Finally, for our NLG component, we observe that the LLM-generated dialogues are rated similarly or higher than the human-generated counterparts in each snippet. Thus, we hypothesise that LLMs present a promising opportunity in the construction of our integrated system.

4

Evaluation

In the previous chapters, we presented the design of our integrated system for training child helpline counsellors and the resulting implementation. In order to evaluate this system, we set up an experiment to investigate the effect of our implementation on the users' experience with the system.

More specifically, we aim to investigate the difference in the participants' perceptions of the two systems in three dimensions - believability, engagement and attitude.

Our choice for these three dimensions was motivated by our research into the factors that influence learning outcomes and performance, as well as areas where the rule-based system was lacking. Centrally, counsellors need to learn a set of skills and competencies which will allow them to better assist children in need. In order to make a training system which imparts these skills to learners, it must provide them with realistic and believable training scenarios, and correctly interpret and respond to learners' inputs.

Kovacevic et. al. [52] find that the integration of distinct personalities into chatbots can influence users' perceptions towards them. They find that chatbots with infused personalities increase user trust and engagement, suggesting that more believable chatbots could be perceived as better social companions. Additionally, Deng and Yu [23] find that positive user attitudes towards chatbots can significantly improve learning achievement and interest. As such, we hypothesize that our choice of constructs could correspond to the overall quality of training systems.

Thus, we present the following hypotheses for our experiment:

- **H1:** *Participants perceive the integrated system to be more believable than the rule-based system.*
- **H2:** *Participants perceive the integrated system to be more engaging than the rule-based system.*
- **H3:** *Participants have a more positive attitude towards the LLM-integrated system compared to the rule-based system.*

4.1. Methods

This section details the experimental methods of our study. We present the design of the experiment, how we choose and recruit our participants, the measures, and procedure, and finally the statistical analysis methods we use to interpret our results. This experiment was approved by the TU Delft Human Research Ethics Committee (HREC reference number: 4768) and was also registered with the Open Science Framework (OSF) registries before the experiment was performed. The OSF registry is available at osf.io/eqxwz. The data from our experiments was also uploaded to the 4TU project repositories, available at data.4tu.nl/datasets/db7965fc-9741-4a4c-a765-4e627ad184af.

4.1.1. Study Design

For our experiment, we performed a repeated measures (within-subject) design involving two interaction-based interventions where participants interact with two different conversational systems. The first system is the current rule-based conversational agent, and the second is our integrated system. To minimize the learning and order effects, participants were counterbalanced by dividing them into groups of two equal sizes, which interacted with the two conversational systems in different orders.

4.1.2. Participants

The target demographic for our participants is those who have at least a Bachelor's degree - as it is a requirement for volunteer counsellors at children's helplines such as De Kindertelefoon, and speak fluent English. We require participants to have at least a Bachelor's degree as this is in line with recruitment policies for child helplines such as De Kindertelefoon, and we require participants to speak fluent English as our entire experiment setup is also in English. We recruited participants through the online platform Prolific.

We invited 55 participants to perform our experiment. Of them, 13 did not complete the experiment, 2 failed the comprehension check, 1 did not give consent, and 2 were screened out. Thus, we excluded 18 participants from the experiment. As such, we were left with 37 participants for our experiment which took place between January 2025 and February 2025.

Our participants all spoke fluent English. Out of the 37, 24 of them had a Bachelor's degree or an equivalent (65%), 10 of them had a Master's degree or an equivalent (27%), and 3 of them had a PhD or an equivalent (8%). Additionally, 14 of our participants identified as male (38%), while 23 identified as female (62%). Our participants were distributed across different age groups, with 10 being between the ages of 18-24, 12 being between the ages of 25-34, 9 being between the ages of 35-44, 5 being between the ages of 45-54, and 1 being between the ages of 55-64.

4.1.3. Prototype

There are two prototypes each user will interact with - the current system with the BDI model, and our integrated system with both the BDI model and the integrated LLM components. The user interacts with both agents - who play the role of a virtual child being bullied - through a webpage which contains a chat area where they can send messages and receive responses, as shown in Figure 4.1. Visually, the two interfaces for the agents are identical, but the two differ in the way the chats from the users are processed. In order to be able to compare the two systems, we introduced an artificial delay into the rule-based system to make the response times comparable.

In the case of the rule-based system, when a message is sent, the intent of the message is identified by the Rasa service, which then calls the BDI model with the same intent. The BDI model uses the identified intent to change the cognitive state of the virtual child. Finally, the intent is used to select an appropriate response from the knowledge base of the system and sent back to the webpage to be displayed. An example of a chat exchange between a user and the rule-based system is shown in Appendix G.1.

In the case of our integrated solution, when a message is sent, it is passed onto the NLU LLM component which identifies the intent in the text. If the identified intent is within the knowledge base of the system, similar to the rule-based system, the BDI model is called with the same intent. The BDI model uses this intent to change the emotional state of the virtual child. Then, the identified intent is passed to the NLG LLM component to generate new dialogue that is similar to the appropriate responses that are in the knowledge base. If the identified intent is not within the knowledge base of the system, the dialogue is passed to the Bypass LLM component which generates a contextually appropriate response. An example of a chat exchange between a user and the LLM-based system is shown in Appendix G.2.

For our LLM components, the large language model we integrated into the system was Llama 3 [2]. The prompts we used for our NLU, Bypass and NLG components are given in Appendix C. These prompts keep the core features we determined to be important during our exploration but reduce their size in order to make the LLM response times faster.

The code for the prototype for our integrated system is hosted publicly at github.com/adarshdenga/llm-integration-helpline.

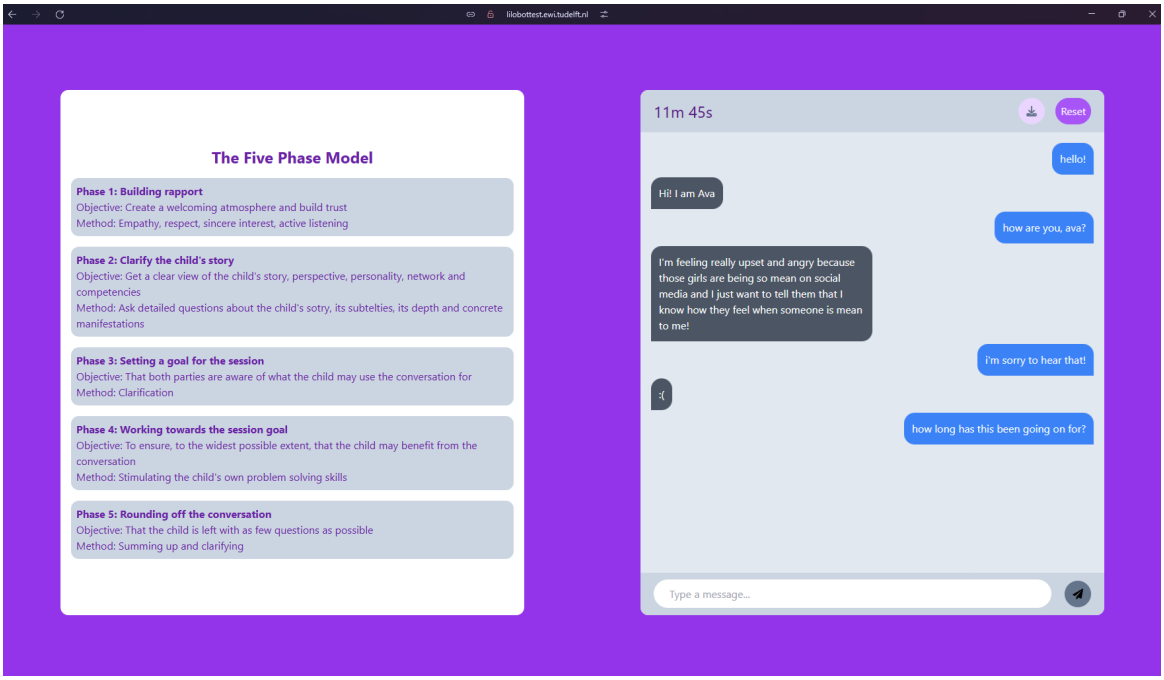


Figure 4.1: The interface where learners can interact with both the LLM and rule-based systems.

4.1.4. Materials

The experiment was set up on the TU Delft private servers, where we hosted both versions of the conversational agent. The technical specifications of the server and our prototypes are given in Appendix H.

The questionnaires for our experiment were hosted on the online survey platform Qualtrics. In addition, our participants were shown a short video with information which explained the basics of counselling techniques and the conversational strategy they were to follow. In our case, this was a short textual explanation of the Five Phase Model.

4.1.5. Measures

We measured and recorded the experience of participants using main measures as well as secondary measures. Our main measures were used to test our three hypotheses, while our secondary measures were used to gauge the participants' overall experience with the system and to understand which system they preferred overall.

Main Measures

Through our main measures, we hope to gain insights and feedback on both systems in relation to our three hypotheses.

- **Believability:** For our believability measure, our hypothesis was operationalised by focusing on two key dimensions from the believability construct in the Artificial Social Agent (ASA) Questionnaire [27]. For our experiment, we chose to use the sub-constructs of 'Human-Like Behaviour' and 'Natural Behaviour'. Our participants interact with both versions of the system through the same chat interface, so we did not expect to see a difference in the results for those dimensions. As such, we chose to omit them from our measures for believability. The sub-construct 'Human-Like Behaviour' refers to the extent to which our participants believe that the virtual child acts like a human. Our measure for this sub-construct consists of 5 questions. The sub-construct 'Natural Behaviour' refers to the extent to which our participants believe the behaviour of the virtual child could exist in or be derived from nature. For this sub-construct, our measure consists of 3 questions. For both measures, participants indicated their level of agreement with statements on a 7-point scale from -3 ('Strongly Disagree'), to 0 ('Neither agree nor disagree'), to +3 ('Strongly Agree').

- **Engagement:** The engagement construct in the long form Artificial Social Agent (ASA) Questionnaire [27] was used to measure the engagement of participants with the virtual child. The construct consists of 3 questions. The participants indicated their level of agreement with statements on a 7-point scale from -3 ('Strongly Disagree'), to 0 ('Neither agree nor disagree'), to +3 ('Strongly Agree').
- **Attitude:** The attitude construct in the long form Artificial Social Agent (ASA) Questionnaire [27] was used to measure the engagement of participants with the virtual child. The construct consists of 3 questions. The participants indicated their level of agreement with statements on a 7-point scale from -3 ('Strongly Disagree'), to 0 ('Neither agree nor disagree'), to +3 ('Strongly Agree').

Secondary Measures

Through our secondary measures, we wanted to gain insights and feedback into participants' overall experience on both systems.

- **Overall Experience with the Agent:** For our secondary measure we use the short form Artificial Social Agent (ASA) Questionnaire [27] to measure the participants' overall experience with the virtual child. The construct consists of 24 questions. The participants indicated their level of agreement with statements on a 7-point scale from -3 ('Strongly Disagree'), to 0 ('Neither agree nor disagree'), to +3 ('Strongly Agree').
- **Open-Ended Questions:** In addition to the above measures, we also asked participants to answer open-ended questions related to their experience with interacting with the agent to capture more nuanced insights that our questionnaire may have missed. After each interaction with an agent, participants were asked *'How did your interaction with the virtual child go?'*.
- **Preference:** After both interactions, they were a question which asked *'Which virtual child did you prefer interacting with, and why?'*. Here, they had to make a choice between the two systems, and then motivate their choice using a short textual explanation.

4.1.6. Procedure

The experiment took place between 28th January 2025 and 30th January 2025. Participants were recruited through Prolific. Through Prolific, they were able to access the Qualtrics questionnaires, which in turn gave them access to the project environments where both of our conversational agents were hosted. The experiment took roughly an hour to complete and was completed in one session. The flow of the experiment is shown in Figure 4.2.

Before being able to participate in the experiment, participants were asked to fill out the informed consent form. Only once they signed this form were they allowed to progress to the next phase of the experiment. In order to understand the task and context better, they were shown a short pre-training video which provided them with information about counselling, common practices, and the Five Phase Model, which guided and structured their conversation with the virtual agents. Participants were assigned randomly to one of two groups to count for order effects. Both groups were presented with both systems, but in a differing order, i.e. either rule-based and then LLM-integrated, or vice-versa. Participants then interacted with a virtual agent, after which they were asked to fill out a questionnaire relating to our main measures of believability, engagement and attitude. Additionally, they were asked one open-ended question asking them to reflect on their experience of interacting with the agent. Then, they interacted with the other virtual agent, after which they were asked to fill in the same questionnaire relating to their experience with the second virtual agent. After both interactions, they were asked to choose which of the agents they preferred interacting with, as well as to provide a short explanation explaining the rationale behind their choice.

4.1.7. Data Preparation & Analysis

Data Preparation

The data from our experiment contained our participants' questionnaire results for both the LLM and rule-based systems. We first identified the LLM and rule-based conditions for each participant from the raw data based on a condition order which we recorded. Then, we formed two distinct datasets which contained the results of our measures for all the participants for the LLM and rule-based conditions separately, using the Python scripts shown in Appendix E.1.

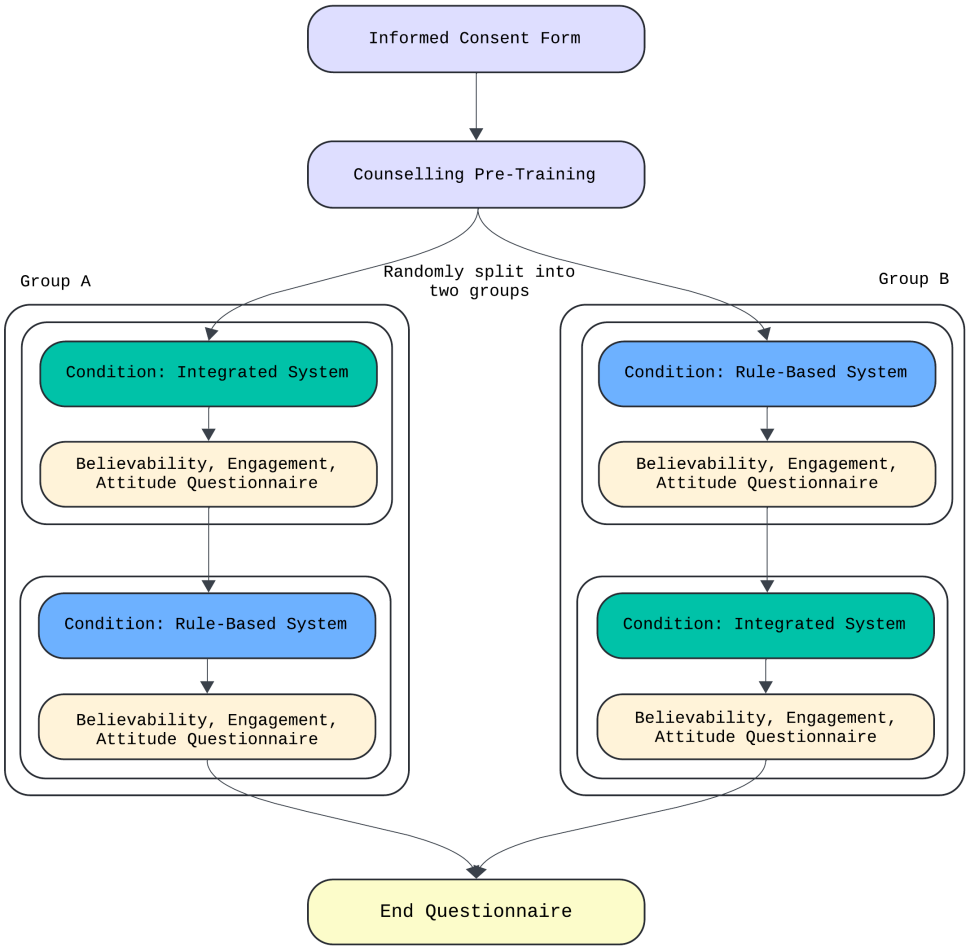


Figure 4.2: Overall flow of our experiment

Table 4.1: Our results for the paired samples t-test

Construct	*LLM-Based System* $M(SD)$	*Rule-Based System* $M(SD)$	*t-value*	*p-value*	d	*df*
Believability (HLB)	0.79 (1.62)	0.35 (1.57)	2.10	0.04	0.07	36
Believability (NB)	0.14 (1.55)	-0.19 (1.57)	1.38	0.18	0.05	36
Engagement	2.01 (1.05)	1.84 (0.77)	1.07	0.29	0.03	36
Attitude	0.86 (1.52)	0.13 (1.76)	2.46	0.02	0.12	36
Overall Experience	0.46 (1.25)	-0.02 (1.28)	2.57	0.01	0.08	36

We gathered the data for each of our five questionnaires (four for the dimensions in our main construct, and one for the overall experience), and averaged the scores for them to obtain five figures that represent a quantitative response. Some of the questions in our questionnaires are reverse-scored, e.g. ”The virtual child is boring”. We transformed the responses for such questions by taking the negative of the value indicated by the participant.

Finally, we were left with 37 pairs of responses for the LLM and rule-based systems which contained five figures each to represent our quantitative measures.

Quantitative Analysis

For our analysis, we compared the responses from our questionnaires for both conditions for each participant applying paired samples t-test. We used the two prepared datasets and the script shown in Appendix E.3 in order to perform a paired samples t-test.

Qualitative Analysis

In order to analyse the qualitative results from our open-ended questions, we performed a thematic analysis [15] to gain insights into the common opinions and themes that are reported repeatedly in our responses.

4.2. Results

The results of our paired samples t-test are shown below, in Table 4.1.

4.2.1. Main Measures

The constructs of believability, engagement and attitude correspond to our three hypotheses we presented at the start of this chapter. For these hypotheses, we present the results from our paired samples t-test below. Additionally, for both the main and secondary measures, the distribution of average scores is shown in Figure 4.3. It shows pairs of distributions for average scores for each construct in Table 4.1 for the LLM and rule-based conditions respectively.

Believability

For the Believability construct, we examine the results for our two sub-constructs individually.

In terms of Human-Like Behaviour, our participants rated the behaviour of the LLM-based system $(M = 0.79, SD = 1.62)$ to be more human-like than that of the rule-based system $(M = 0.35, SD = 1.57)$ with $t(37) = 2.10, p = 0.04$. These results suggest that the integration of the LLM into the conversational system has a significant effect on how human-like the virtual child appears.

In terms of Natural Behaviour, we did not find a significant difference between the LLM-based system $(M = 0.14, SD = 1.55)$ and the rule-based system $(M = -0.19, SD = 1.57)$, with $t(37) = 1.38, p = 0.18$. Thus, it is inconclusive whether or not the integration of LLMs into the conversational system has an effect on how natural its behaviour is perceived to be.

Thus, we find partial support for the hypothesis H1.

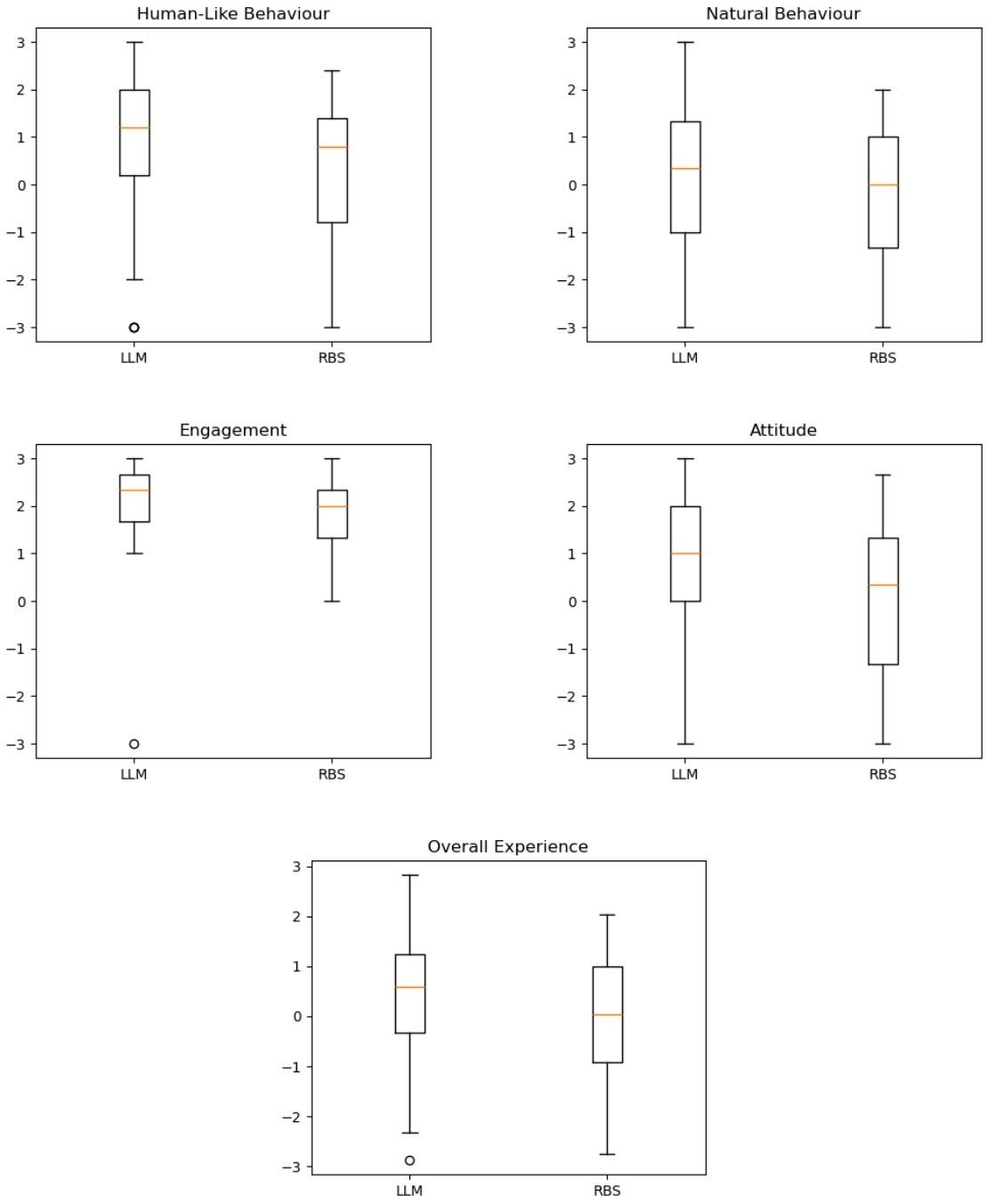


Figure 4.3: Distributions for average scores per construct for each condition.

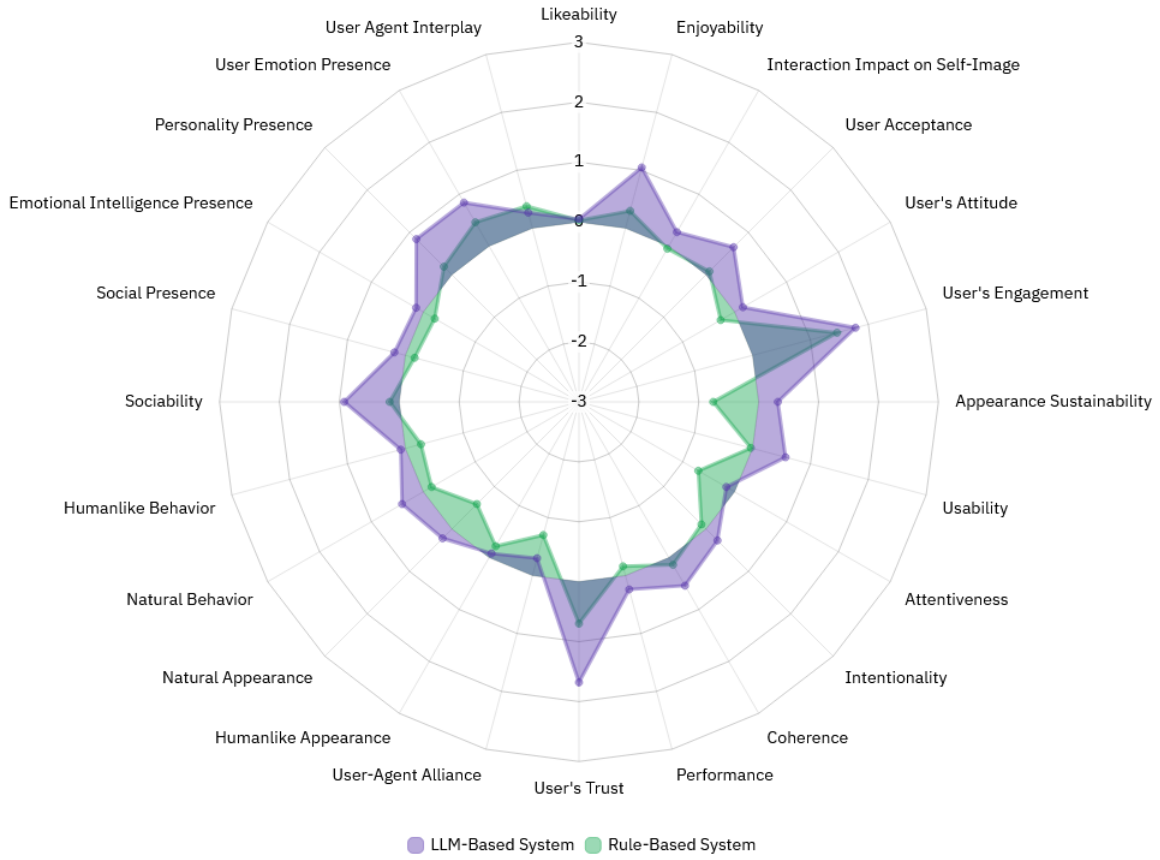


Figure 4.4: A comparison between the LLM-based and rule-based systems for each item in the overall experience questionnaire

Engagement

In the case of the Engagement construct, the results showed that the scores for the LLM-based system $(M = 2.01, SD = 1.05)$ were not significantly higher than that of the rule-based system $(M = 1.84, SD = 0.77)$ with $t(37) = 1.07, p = 0.29$. From our results, it is inconclusive whether or not the integration of the LLM into the conversational system has an effect on the engagement of the conversation. Thus, we did not find support for the alternate H2.

Attitude

Finally, with the Attitude construct, the results showed that the scores for the LLM-based system $(M = 0.86, SD = 1.53)$ were significantly higher than that of the rule-based system $(M = 0.13, SD = 1.76)$ with $t(37) = 2.46, p = 0.02$. These results suggest that the integration of the LLM into the conversational system has a significant effect on the participants' attitudes towards the conversation. Thus, we find partial support for the hypothesis H3 over the null hypothesis.

4.2.2. Secondary Measure

Our secondary measure evaluates the interaction of our participants with both systems. The results showed that the scores for the LLM-based system $(M = 0.46, SD = 1.25)$ were significantly higher than that of the rule-based system $(M = -0.02, SD = 1.28)$ with $t(37) = 2.57, p = 0.01$. Thus, we deem this to be a statistically significant difference. This implies that our participants' overall experience with the LLM-based system was higher than in the case of the rule-based system. Figure 4.4 shows a comparison between the two items for each item in the overall experience questionnaire, where it can be observed that the LLM-based system outperforms the rule-based systems in most items.

From the results of the questionnaire where we ask them which agent they prefer interacting with, we

Table 4.2: Response themes that are common to both conditions.

$Theme$	$Count_{LLM}$	$Count_{Rule-Based}$
Human-Like Responses	5	1
Emotional Engagement	6	2
Positive Experience	8	6
Abrupt Ending	2	2
Unnatural Responses	4	6
Slow Responses	4	12
Technical Issues	1	5

observe that out of the total of 37 participants, 26 of them prefer the LLM version, and 11 prefer the rule-based version. A binomial test was conducted using the script shown in Appendix E.4 to determine whether the participants showed a significant preference between the LLM-based and rule-based systems. Out of 37 participants, 26 (70.36%) preferred the LLM-based system, while 11 preferred the rule-based system. Under the null hypothesis of no preference, the result was statistically significant $(p < 0.05)$, indicating that our participants had a significant preference for the LLM-based system over the rule-based system.

Out of 37 participants, 18 interacted with the LLM-based version first and then the rule-based version, and 19 interacted with the systems in the opposite order, as shown in Table 4.3. In both orderings, the LLM-based version is preferred over the rule-based version, but while the LLM-based version is preferred 61% of the time when it is presented first, it is preferred 79% of the time when presented after the rule-based system.

4.2.3. Thematic Analysis

Common Themes

Table 4.2 shows the themes found in the responses of both the LLM-based and rule-based conditions.

We observed a trend of positive themes being reported more in the case of the LLM-based system compared to the rule-based system. For instance, participants reported that both the virtual children were perceived to have human-like responses, but more often in the LLM-based version $(n = 5, 14\%)$ than the rule-based version $(n = 1, 3\%)$. There was a small difference in the number of participants who noted that the agents were emotionally engaging - $(n = 6, 16\%)$ in the LLM-based version and $(n = 2, 5\%)$ in the rule-based version. Overall, there was a higher proportion of participants who reported an overall positive experience with the LLM-based system $(n = 8, 22\%)$ rather than the rule-based version $(n = 6, 16\%)$.

Conversely, we observed that the negative themes were present more often in the responses relating to the rule-based system as compared to the LLM-based system. Similarly, the reports of unnatural responses were lower in the case of the LLM-based system $(n = 4, 11\%)$ compared to the rule-based system $(n = 6, 16\%)$. Technical issues were reported more often in the case of the rule-based system $(n = 5, 14\%)$ compared to the LLM-based system $(n = 1, 3\%)$.

Participants reported slow response times for both systems with $(n = 4, 10\%)$ in the case of the LLM-based system, and $(n = 12, 32\%)$ in the the case of the rule-based system. Abrupt endings to conversations were reported more equally in both systems $(n = 2, 5\%)$.

LLM-Based System Specific Themes

From the feedback for the LLM-based system, we found some unique themes that were not reported in the feedback for the rule-based system. Participants reported that the LLM-based system displayed personality ($n = 2, 5\%$):

"The interaction was somewhat emotional and eye-opening, reinforcing the importance of empathy,

Table 4.3: Counts for participants based on order and system preference.

$Order$	$Preference\ for\ LLM$	$Preference\ for\ RBS$
$LLM\text{-}based\ first\ (n = 18)$	11	7
$Rule\text{-}based\ first\ (n = 19)$	15	4

active listening and proper intervention."

Others reported that there was a certain depth to the conversation that could not be found in the rule-based system ($n = 1, 3\%$):

"The interaction was somewhat emotional and eye-opening, reinforcing the importance of empathy, active listening and proper intervention."

Importantly, we found that the Five Phase Model was mentioned in the feedback from the LLM-based system, indicating that users of the system could focus more on applying the conversational strategy required to counsel the virtual child ($n = 4, 11\%$):

"I felt like we made a connection, and the five-phase model helped guide the exchange smoothly."

Rule-Based System Specific Themes

From the feedback for the rule-based system, we found some unique themes that were not reported in the feedback for the LLM-based system. For instance, participants reported that they received boring responses from the rule-based system $(n = 1, 3\%)$:

"It was boring, non-responsive, had delayed responses, did not reveal any realistic feelings, etc."

Others felt that the responses from the rule-based system felt scripted ($n = 1, 3\%$):

"The virtual child seemed to be sending responses from a predetermined library, it didn't really adapt to the questions I was asking"

We performed a double coding verification for the results of our thematic analysis where we asked our coders to match an identified code to each response in our qualitative feedback. The list of codes is given in Appendix E.5. Based on the identified codes, we performed a Cohen's Kappa analysis with the script shown in Appendix E.5 to determine the level of agreement between the coders. Results indicated a Cohen's Kappa score of 0.315, which indicates a fair degree of agreement.

4.2.4. Discussion

Based on the paired samples t-test that we performed for our main and secondary measures, we draw the following conclusions from our results. We first describe the comparison between the systems for the main measures and relate them to the hypotheses. Then, we discuss the secondary measures, including the overall experience users have with both systems, the thematic analysis for our open-ended questions, and finally, the results of our binomial test to determine if there is a significant preference between systems.

With our believability measure, we found that while participants perceived our system to be more human-like, its behaviour was not necessarily more natural than that of the rule-based system. As LLMs are trained on vast amounts of textual data, they are capable of understanding and generating human-like text, which allows them to produce responses that are contextually relevant and linguistically natural, enhancing the perception of believability [20, 8].

From our results for our engagement measure, we found that participants did not find our system to be more engaging than the rule-based system. When examining the data, we observed a ceiling effect where the results for both conditions were rated highly in terms of engagement (as shown in Figure 4.3). This leaves us with little room for improvement, and thus smaller statistical differences between the LLM-based and rule-based conditions. The nature of the task in both cases is the same - interacting with a virtual child that requires assistance - which could make participants more engaged by nature.

Finally, we found that our participants had a more positive attitude towards our system than the rule-based system. Our participants found the LLM-integrated system to be more human-like and believable, which corresponds to a more positive attitude towards it [36].

The thematic analysis we conducted from our open questions revealed some key points of feedback from both systems. The LLM-based system was described as more adaptive, engaging, and displaying emotional and personal depth. Participants who used the system described the agent as having a personality with apt perception and responsiveness, which could be factors that additionally influence their attitude towards it [52]. Importantly, the usage of the Five Phase Model was described more often in the feedback for the LLM-based system, suggesting a greater focus on structured guidance through the use of the LLM-based system. On the other hand, the rule-based system was criticized more often for having scripted dialogue that was unresponsive and not adaptive. From our results, we found that the feedback for the LLM-based system contained more positive themes (personality, five-phase model, etc.), while the feedback for the rule-based system contained negative themes (scripted responses, technical issues, etc.). Both systems were reported to have slow response times.

Finally, we asked participants which of the two systems they preferred interacting with. Our responses indicated that 70% of our participants (26 out of 37) preferred the LLM-based system to the rule-based system, which we found to be statistically significant. Furthermore, it was observed that it is preferred more often when participants had already interacted with the rule-based system beforehand.

4.2.5. Limitations

We faced some technical difficulties during the implementation and setting up of our system which limited its performance. Due to the hardware limitations of the servers where our integrated system was hosted, as well as the steep hardware requirements to run large language models locally, we observed higher response times in our integrated system compared to the rule-based system. To make the systems comparable, we delayed the response times of the rule-based system. Faster response times could lead to a better overall impression of both agents (as several users indicated slow response times for both systems), and thus an improved attitude towards them. Furthermore, due specifically to the limited memory of the servers, we were unable to load and use the planned large language model (Llama 3), opting instead for a lighter model (Llama 3.2), for which the performance was worse overall. From our initial tests with Llama3, we observed that it was able to more robustly and accurately identify intents and generate dialogue. As such, we speculate that a higher-performing model could lead to higher performance in terms of believability, engagement, attitude and overall experience.

5

Discussion & Conclusion

In this final chapter, we reflect on our entire research. First, we provide our answers to the research questions we posed at the beginning of our work. Next, we discuss the limitations of our work, as well as our contributions on a scientific and practical level. Finally, we present suggestions for further research in this domain.

5.1. Conclusion

Through our research, we aimed to answer the question:

How can large language models be integrated into rule-based conversational training systems to improve the quality of training?

In order to arrive at an answer to this overarching question, we split it up into three sub-questions, which we examined in each chapter of this report. Here, we condense our answers and findings to each sub-question.

What are the design considerations of such an integration?

In order to establish the design considerations of our integrated solution, we conducted a literature study where we sought to understand counselling, training processes and systems, the factors that influence learning, and the technology that powers our integrated solution.

Additionally, we examined feedback from the users of the rule-based system to find drawbacks and opportunities for improvement. We found that while the users found the rule-based system to be useful, it was also reported to have a poor understanding of user input, repetitive responses, and was not believable.

From our research and from feedback from users who used the rule-based system, we arrived at a set of design considerations which our integrated system must meet, including constructing a system which imparts essential counselling knowledge to learners, providing them with a realistic test scenario by responding to them with believable and varied dialogue, the ability to interpret a wide variety of user inputs, and the ability to respond to messages appropriately, even if they fall outside of the knowledge base of the rule-based model.

What design for an integration would meet these design considerations?

Based on our requirements, our research and our exploration with large language models, we designed a conversational agent which plays the character of a child seeking help and uses large language models to handle the processing of input as well as the generation of output, while still keeping the dependable rule-based model at the core. Our design adds three new components to the rule-based implementation, which we call the NLU, NLG and Bypass.

Our NLU component uses a large language model to extract the intent from the user's input message, given a set of known intents. If the recognised intent is known within the knowledge base of the BDI

model, it can be processed by the BDI to change the cognitive state of the virtual child. From the BDI model, the NLG component is called with a set of examples of appropriate responses to the counsellor's input. The NLG component then uses the examples and the context of the conversation with a large language model to generate a new response which is both contextually appropriate and directly responds to the input message. If the recognised intent is not known within the knowledge base of the BDI model, the Bypass component is called instead. The Bypass component then generates a response to the input message based on the context of the conversation.

How do users perceive the effect of the integrated design on the usefulness and realism?

We evaluated the integrated system with 37 participants. Participants were asked to use both the rule-based and integrated system and answer questionnaires regarding their experience relating to believability (human-like behaviour and natural behaviour), engagement, attitude, overall experience and preference for each. We found that participants perceive the integrated system to be more human-like in its behaviour (while not necessarily more natural), have a more positive attitude towards it, and have a better overall experience with it compared to the rule-based system. Through our thematic analysis, we found that feedback about the integrated system contained positive themes such as "human-like responses", while feedback about the integrated system contained negative themes such as "scripted responses". Additionally, we found that 70% of our participants preferred the integrated system over the rule-based system.

5.2. Contributions

Through our research, we demonstrated that a BDI-based conversational agent which is integrated with LLMs has the potential to offer effective, realistic training scenarios for child helpline counsellors. On a scientific level, we presented a conversational agent which is capable of simulating a child victim of bullying, equipped with a deep understanding of counsellor input, as well as the ability to generate rich, contextually aware messages in response. Our findings from this research could further be used to explore the possibility of integrating LLMs into training systems for training not only counsellors but learners in a variety of fields. On a practical level, we present a method by which LLMs could be integrated into existing rule-based conversational systems, as well as methods by which such systems could be evaluated.

5.3. Limitations & Future Work

Through our experiment, we identified a set of limitations which could serve as possible further extensions to our work. We explain some of the central areas for further exploration below.

A core point of feedback from those who initially used the rule-based system was the repetitive and unrealistic responses from the rule-based model. Our LLM integrated system addresses this specific point, but the repetitiveness of dialogue can only be observed over a course of multiple sessions of usage, and will not be immediately apparent in one short interaction with each agent. With repeated training sessions with the two systems, there could be an opportunity to understand how participants perceive the repetitive nature of the systems across multiple interactions.

As mentioned in Section 4.2.5, limitations in the server narrowed the options of usable large language models. A potential area of exploration would be to use other large language models with different complexities in order to understand the effect that this has on our main and secondary measures.

Lastly, an interesting area for exploration would be to conduct the experiment with experts. For our experiment, we recruited participants who meet the same minimum criteria set by child helplines such as De Kindertelefoon, i.e. having completed at least a bachelor's degree. By recruiting expert counsellors who work at child helplines, we may be able to receive nuanced feedback about our system that our participants were not able to offer due to their lack of expertise in the profession. Additionally, this new feedback could open up new paths by which we could further improve our system as a training tool for volunteer counsellors.

5.4. Final Remarks

In this thesis, we presented a conversational agent which integrates large language models into an existing rule-based system for the purpose of training child helpline counsellors. Users of our system perceive it to be more realistic and have a more positive attitude towards it compared to just the rule-based system, suggesting better performances on learning outcomes, and thus an increased ability to counsel and provide help to children.

References

[1] URL: https://chatgpt.com.

[2] URL: https://ai.meta.com/blog/meta-llama-3/.

[3] URL: https://claude.ai.

[4] Mohammed Al Owayyed et al. “Agent-based social skills training systems: the ARTES architecture, interaction characteristics, learning theories and future outlooks”. In: *Behaviour and Information Technology* (July 2024), pp. 1–28. DOI: 10.1080/0144929X.2024.2374891.

[5] Felipe Almeida and Geraldo Xexéo. *Word Embeddings: A Survey*. 2023. arXiv: 1901.09069 [cs.CL]. URL: https://arxiv.org/abs/1901.09069.

[6] Yair Amichai-Hamburger et al. “The future of online therapy”. In: *Computers in Human Behavior* 41 (2014), pp. 288–294. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.09.016. URL: https://www.sciencedirect.com/science/article/pii/S0747563214004683.

[7] Alia Sarah Asri et al. “Journal of Critical Reviews E-COUNSELLING PROCESS AND SKILS: A LITERATURE REVIEW”. In: *Journal of Critical Reviews* 7 (Sept. 2020), pp. 629–643. DOI: 10.31838/jcr.07.13.110.

[8] Debarag Banerjee et al. *Benchmarking LLM powered Chatbots: Methods and Metrics*. 2023. arXiv: 2308.04624 [cs.CL]. URL: https://arxiv.org/abs/2308.04624.

[9] Yejin Bang et al. *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. 2023. arXiv: 2302.04023 [cs.CL]. URL: https://arxiv.org/abs/2302.04023.

[10] N.B Berman et al. “The Role for Virtual Patients in the Future of Medical Education”. In: (2016). URL: https://journals.lww.com/academicmedicine/FullText/2016/09000/The_Role_for_Virtual_Patients_in_the_Future_of.17.aspx.

[11] Norman B. Berman et al. “The Role for Virtual Patients in the Future of Medical Education”. In: *Academic Medicine* 91.9 (2016), pp. 1217–1222.

[12] Idan A. Blank. “What are Large Language Models supposed to model?” In: *Trends in Cognitive Science, Volume 27, Issue 11*. 2023. URL: https://doi.org/10.1016/j.tics.2023.08.006.

[13] K. K. Boyd. “Power Imbalances and Therapy”. In: *Focus* 11.9 (1996), pp. 1–4.

[14] M. Bratman. *Intention, Plans and Practical Reason*. 1987. DOI: 10.2307/2185304.

[15] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa. eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa. URL: https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

[16] Leon Chaddock. *What Percentage Of Teens Use Social Media? (2024)*. 2024. URL: https://www.sentiment.io/how-many-teens-use-social-media/ (visited on 08/06/2024).

[17] Yupeng Chang et al. “A Survey on Evaluation of Large Language Models”. In: *ACM Trans. Intell. Syst. Technol.* 15.3 (Mar. 2024). ISSN: 2157-6904. DOI: 10.1145/3641289. URL: https://doi.org/10.1145/3641289.

[18] Jared Coleman et al. *LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages*. 2024. arXiv: 2405.08997 [cs.CL]. URL: https://arxiv.org/abs/2405.08997.

[19] David A Cook and Marc M Triola. “Virtual patients: a critical literature review and proposed next steps”. In: *Medical Education* 43.4 (2009), pp. 303–311. DOI: https://doi.org/10.1111/j.1365-2923.2008.03286.x. eprint: https://asmepublications.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2008.03286.x. URL: https://asmepublications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2923.2008.03286.x.

[20] Sumit Kumar Dam et al. *A Complete Survey on LLM-based AI Chatbots*. 2024. arXiv: 2406.16937 [cs.CL]. URL: https://arxiv.org/abs/2406.16937.

[21] Asim Bhatti Dawei Jia and Saeid Nahavandi. "The impact of self-efficacy and perceived system efficacy on effectiveness of virtual training systems". In: *Behaviour & Information Technology* 33.1 (2014), pp. 16–35. DOI: 10.1080/0144929X.2012.681067. eprint: https://doi.org/10.1080/0144929X.2012.681067. URL: https://doi.org/10.1080/0144929X.2012.681067.

[22] Julie Deardroff. *Teens Turn To Internet To Cope With Health Challenges*. 2015. URL: https://news.northwestern.edu/stories/2015/06/teens-turn-to-internet-to-cope-with-health-challenges/.

[23] Xinjie Deng and Zhonggen Yu. "A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education". In: *Sustainability* 15.4 (2023). ISSN: 2071-1050. DOI: 10.3390/su15042940. URL: https://www.mdpi.com/2071-1050/15/4/2940.

[24] Stephan Diederich et al. "Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent." In: *ICIS*. 2019.

[25] Gerard Egan. *The skilled helper: A problem-management and opportunity-development approach to helping*. Nelson Education, 2013.

[26] Katrien Struyven Eva Kyndt Filip Dochy and Eduardo Cascallar. "The direct and indirect effect of motivation for learning on students' approaches to learning through the perceptions of workload and task complexity". In: *Higher Education Research & Development* 30.2 (2011), pp. 135–150. DOI: 10.1080/07294360.2010.501329. eprint: https://doi.org/10.1080/07294360.2010.501329. URL: https://doi.org/10.1080/07294360.2010.501329.

[27] Siska Fitrianie et al. "The artificial-social-agent questionnaire: establishing the long and short questionnaire versions". In: *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. IVA '22. Faro, Portugal: Association for Computing Machinery, 2022. ISBN: 9781450392488. DOI: 10.1145/3514197.3549612. URL: https://doi.org/10.1145/3514197.3549612.

[28] Asbjørn Følstad and Petter Bae Brandtzaeg. "Users' experiences with chatbots: findings from a questionnaire study". In: *Quality and User Experience* 5.1 (2020), p. 3.

[29] Itzhak Gilat and Sarah Rosenau. "Volunteers' perspective of effective interactions with helpline callers: qualitative study". In: *British Journal of Guidance & Counselling* 39.4 (2011), pp. 325–337. DOI: 10.1080/03069885.2011.567327. eprint: https://doi.org/10.1080/03069885.2011.567327. URL: https://doi.org/10.1080/03069885.2011.567327.

[30] Marilyn E Gist. "Self-efficacy: Implications for organizational behavior and human resource management". In: *Academy of management review* 12.3 (1987), pp. 472–485.

[31] Rahul Goel et al. *LLM-based Task-oriented Dialog System with Few-shot Retrieval Augmentation*. 2023. URL: https://www.tdcommons.org/dpubs_series/6407/.

[32] Karthik Gopalakrishnan et al. *Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations*. 2023. arXiv: 2308.11995 [cs.CL]. URL: https://arxiv.org/abs/2308.11995.

[33] R. Grabinger and Joanna Dunlap. "Rich environments for active learning: a definition". In: *Research in Learning Technology* 3 (Dec. 1995). DOI: 10.3402/rlt.v3i2.9606.

[34] Nicola J. Gray et al. "Health information-seeking behaviour in adolescence: the place of the internet". In: *Social Science & Medicine* 60.7 (2005), pp. 1467–1478. ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2004.08.010. URL: https://www.sciencedirect.com/science/article/pii/S0277953604003934.

[35] Sharon Grundmann. "A BDI-Based Virtual Agent for Training Child Helpline Counsellors". In: (2022). URL: https://repository.tudelft.nl/islandora/object/uuid%3Af04f8f0b-9ab9-4f1c-a19c-43b164d45cce.

[36] Rose E. Guingrich and Michael S. A. Graziano. *Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines*. 2024. arXiv: 2311.10599 [cs.HC]. URL: https://arxiv.org/abs/2311.10599.

[37] Onder Gurcan. *LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities*. 2024. arXiv: 2405.06700 [physics.soc-ph]. URL: https://arxiv.org/abs/2405.06700.

[38] Zeyu Han et al. *Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey*. 2024. arXiv: 2403.14608 [cs.LG]. URL: https://arxiv.org/abs/2403.14608.

[39] Nina Hollender et al. "Integrating cognitive load theory and concepts of human–computer interaction". In: *Computers in Human Behavior* 26.6 (2010). Online Interactivity: Role of Technology in Behavior Change, pp. 1278–1288. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2010.05.031. URL: https://www.sciencedirect.com/science/article/pii/S0747563210001718.

[40] Courtney M. Holmes and Kelly A. Kozlowski. "A Pilot Study of Online Group Leadership Skills: Perceived Usage and Difficulty Level". In: *Journal of Counselor Practice* 7.2 (2016), pp. 61–77.

[41] Child Helpline International. *Voices of Children & Young People Around the World*. 2022. URL: https://childhelplineinternational.org/voices-of-children-young-people-around-the-world-2022-data/.

[42] Erblin Isaku et al. *LLMs in the Heart of Differential Testing: A Case Study on a Medical Rule Engine*. 2024. arXiv: 2404.03664 [cs.SE]. URL: https://arxiv.org/abs/2404.03664.

[43] Ziwei Ji et al. "Towards Mitigating LLM Hallucination via Self Reflection". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1827–1843. DOI: 10.18653/v1/2023.findings-emnlp.123. URL: https://aclanthology.org/2023.findings-emnlp.123.

[44] Nikitas Karanikolas et al. "Large Language Models versus Natural Language Understanding and Generation". In: *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*. PCI '23. <conf-loc>, <city>Lamia</city>, <country>Greece</country>, </conf-loc>: Association for Computing Machinery, 2024, pp. 278–290. ISBN: 9798400716263. DOI: 10.1145/3635059.3635104. URL: https://doi.org/10.1145/3635059.3635104.

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

[46] De Kindertelefoon. "De Kindertelefoon - Jaarverslaag 2019". In: (2019). URL: https://jaarverslag.kindertelefoon.nl/2019.

[47] De Kindertelefoon. "De Kindertelefoon - Jaarverslaag 2022". In: (2022). URL: https://jaarverslag.kindertelefoon.nl/2022.

[48] De Kindertelefoon. "De Kindertelefoon - Vrijwilligers". In: (). URL: https://www.kindertelefoon.nl/vrijwilliger.

[49] Phey Kit et al. "Singaporean Counsellors' Online Counselling Experiences with Children: An Exploratory Qualitative Study". In: *Journal of Asia Pacific Counseling* 7 (Aug. 2017), pp. 141–168. DOI: 10.18401/2017.7.2.3.

[50] Andrea Kleinsmith et al. "Understanding empathy training with virtual patients". In: *Computers in Human Behavior* 52 (2015), pp. 151–158. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2015.05.033. URL: https://www.sciencedirect.com/science/article/pii/S0747563215004045.

[51] A. Kononowicz et al. "Virtual patients - what are we talking about? A framework to classify the meanings of the term in healthcare education". In: *BMC Med Educ.* 15.11 (2015).

[52] Nikola Kovacevic et al. "Chatbots With Attitude: Enhancing Chatbot Interactions Through Dynamic Personality Infusion". In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. CUI '24. Luxembourg, Luxembourg: Association for Computing Machinery, 2024. ISBN: 9798400705113. DOI: 10.1145/3640794.3665543. URL: https://doi.org/10.1145/3640794.3665543.

[53] Lisa Larson and Jeffrey Daniels. "Review of the Counseling Self-Efficacy Literature". In: *Counseling Psychologist - COUNS PSYCHOL* 26 (Mar. 1998), pp. 179–218. DOI: 10.1177/0011000098262001.

[54] Jan Leusmann, Chao Wang, and Sven Mayer. *Comparing Rule-based and LLM-based Methods to Enable Active Robot Assistant Conversations*. 2024. URL: https://cui.acm.org/workshops/CHI2024/wp-content/uploads/2024/04/Comparing-Rule-based-and-LLM-based-Methods-to-Enable-Active-Robot-Assistant-Conversations.pdf.

[55] Zhaodonghui Li et al. *LLM-R2: A Large Language Model Enhanced Rule-based Rewrite System for Boosting Query Efficiency*. 2024. arXiv: 2404.12872 [cs.DB]. URL: https://arxiv.org/abs/2404.12872.

[56] Percy Liang et al. *Holistic Evaluation of Language Models*. 2023. arXiv: 2211.09110 [cs.CL]. URL: https://arxiv.org/abs/2211.09110.

[57] Elizabeth A Linnenbrink and Paul R Pintrich. "Motivation as an enabler for academic success". In: *School psychology review* 31.3 (2002), pp. 313–327.

[58] Christine L. McCarthy. "WHAT IS "CRITICAL THINKING"? IS IT GENERALIZABLE?" In: *Educational Theory* 46.2 (1996), pp. 217–239. DOI: https://doi.org/10.1111/j.1741-5446.1996.00217.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1741-5446.1996.00217.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1741-5446.1996.00217.x.

[59] M.D Merrill. "First Principles of Instruction". In: *Educational Technology Research and Development* 50 (2002), pp. 43–59.

[60] David L. Vogel Michael J. Mallen Indria M. Jenkins and Susan X. Day. "Online counselling: An initial examination of the process in a synchronous chat environment". In: *Counselling and Psychotherapy Research* 11.3 (2011), pp. 220–227. DOI: 10.1080/14733145.2010.486865. eprint: https://doi.org/10.1080/14733145.2010.486865. URL: https://doi.org/10.1080/14733145.2010.486865.

[61] Mark Neerincx. *Applying the Situated Cognitive Engineering Method - A Comprehensive Guide*. URL: https://scetool.ewi.tudelft.nl/sites/default/files/sce_manual_v1.0.pdf.

[62] Sevil Orhan Özen. "The Effect of Motivation on Student Achievement". In: *The Factors Effecting Student Achievement: Meta-Analysis of Empirical Studies*. Ed. by Engin Karadag. Cham: Springer International Publishing, 2017, pp. 35–56. ISBN: 978-3-319-56083-0. DOI: 10.1007/978-3-319-56083-0_3. URL: https://doi.org/10.1007/978-3-319-56083-0_3.

[63] Marieke M.M. Peeters et al. "Scenario-Based Training: Director's Cut". In: vol. 6738. June 2011, pp. 264–271. ISBN: 978-3-642-21868-2. DOI: 10.1007/978-3-642-21869-9_35.

[64] Aaron Pico et al. "Exploring Text-Generating Large Language Models (LLMs) for Emotion Recognition in Affective Intelligent Agents". In: *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 1: EAA*. INSTICC. SciTePress, 2024, pp. 491–498. ISBN: 978-989-758-680-4. DOI: 10.5220/0012596800003636.

[65] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *CoRR* abs/1810.02508 (2018). arXiv: 1810.02508. URL: http://arxiv.org/abs/1810.02508.

[66] Dongqi Pu and Vera Demberg. *ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer*. 2023. arXiv: 2306.07799 [cs.CL]. URL: https://arxiv.org/abs/2306.07799.

[67] Miguel A Quinones and Addie Ed Ehrenstein. *Training for a rapidly changing workplace: Applications of psychological research.* American Psychological Association, 1997.

[68] Anand S Rao and Michael P Georgeff. "Modeling rational agents within a BDI-architecture". In: *Readings in agents* (1997), pp. 317–328.

[69] Jeff Rickel. "Intelligent Virtual Agents for Education and Training: Opportunities and Challenges". In: *Intelligent Virtual Agents*. Ed. by Angélica de Antonio, Ruth Aylett, and Daniel Ballin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 15–22. ISBN: 978-3-540-44812-9.

[70] Christina M. Rummell and Nicholas R. Joyce. ""So wat do u want to wrk on 2day?": The Ethical Implications of Online Counseling". In: *Ethics & Behavior* 20.6 (2010), pp. 482–496. DOI: 10.1080/10508422.2010.521450. eprint: https://doi.org/10.1080/10508422.2010.521450. URL: https://doi.org/10.1080/10508422.2010.521450.

[71] Amla Salleh et al. “Online counseling using email: a qualitative study”. In: *Asia Pacific Education Review* Online First (Sept. 2015). DOI: 10.1007/s12564-015-9393-6.

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

[73] Ian Steenstra, Farnaz Nouraei, and Mehdi Arjmand. “Virtual Agents for Alcohol Use Counseling: Exploring LLM-Powered Motivational Interviewing”. In: (July 2024). DOI: 10.48550/arXiv.2407.08095.

[74] Michael Sude. “Text Messaging and Private Practice: Ethical Challenges and Guidelines for Developing Personal Best Practices”. In: *Journal of Mental Health Counseling* 35.3 (July 2013), pp. 211–227. ISSN: 1040-2861. DOI: 10.17744/mehc.35.3.q3712236up621713. eprint: https://meridian.allenpress.com/jmhc/article-pdf/35/3/211/1812229/mehc_35_3_q3712236up621713.pdf. URL: https://doi.org/10.17744/mehc.35.3.q3712236up621713.

[75] John Sweller. “Implications of Cognitive Load Theory for Multimedia Learning”. In: *The Cambridge Handbook of Multimedia Learning*. Ed. by RichardEditor Mayer. Cambridge Handbooks in Psychology. Cambridge University Press, 2005, pp. 19–30.

[76] Alon Talmor et al. *CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge*. 2019. arXiv: 1811.00937 [cs.CL]. URL: https://arxiv.org/abs/1811.00937.

[77] Zhengwei Tao et al. *EvEval: A Comprehensive Evaluation of Event Semantics for Large Language Models*. 2023. arXiv: 2305.15268 [cs.CL]. URL: https://arxiv.org/abs/2305.15268.

[78] Sebastian Ullrich et al. “Virtual needle simulation with haptics for regional anaesthesia”. In: Mar. 2010.

[79] UNICEF. *How many children and young people have internet access at home? Estimating digital connectivity during the COVID-19 pandemic.* 2020. URL: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/UNICEF/How-many-children-and-young-people-have-internet-access-at-home-2020_v2final.pdf (visited on 08/06/2024).

[80] Bernard Weiner. “Attribution, emotion, and action. Handbook of motivation and cognition”. In: *Foundations of Social Behavior* 1 (Jan. 1986), pp. 281–312.

[81] EduTech Wiki. *Laurillard conversational framework — EduTech Wiki, A resource kit for educational technology teaching, practice and research*. [Online; accessed 20-August-2024]. 2014. URL: https://edutechwiki.unige.ch/mediawiki/index.php?title=Laurillard_conversational_framework&oldid=53804.

[82] Wikipedia contributors. *BLEU — Wikipedia, The Free Encyclopedia*. [Online; accessed 27-June-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=BLEU&oldid=1228325058.

[83] Wikipedia contributors. *Cohen’s kappa — Wikipedia, The Free Encyclopedia*. [Online; accessed 9-October-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Cohen%27s_kappa&oldid=1245316837.

[84] Wikipedia contributors. *Natural language generation — Wikipedia, The Free Encyclopedia*. [Online; accessed 20-August-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Natural_language_generation&oldid=1239749113.

[85] Wikipedia contributors. *Natural language understanding — Wikipedia, The Free Encyclopedia*. [Online; accessed 20-August-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Natural_language_understanding&oldid=1234747669.

[86] Wikipedia contributors. *Virtual patient — Wikipedia, The Free Encyclopedia*. [Online; accessed 29-July-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Virtual_patient&oldid=1223842807.

[87] Robert Williams et al. “In-session processes in online counselling with young people: An exploratory approach”. In: *Counselling and Psychotherapy Research* 9.2 (2009), pp. 93–100. DOI: https://doi.org/10.1080/14733140802490606. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1080/14733140802490606. URL: https://onlinelibrary.wiley.com/doi/abs/10.1080/14733140802490606.

[88] Ulrich von Zadow et al. “SimMed: combining simulation and interactive tabletops for medical education”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’13. Paris, France: Association for Computing Machinery, 2013, pp. 1469–1478. ISBN: 9781450318990. DOI: 10.1145/2470654.2466196. URL: https://doi.org/10.1145/2470654.2466196.

[89] Aohan Zeng et al. *GLM-130B: An Open Bilingual Pre-trained Model*. 2023. arXiv: 2210.02414 [cs.CL]. URL: https://arxiv.org/abs/2210.02414.

A

Prompt Testing

Here we present the prompts and results of our exploration into prompting with Llama3.

A.1. Natural Language Understanding Prompts

A.1.1. NLU Prompt 1

```
###Instruction###

You are a smart assistant that classifies text into pre-defined categories related to bullying.
You receive input from a child helpline counsellor, who is asking the child clarifying questions about the situation.

The categories are given below, with information about what each one means. Each line is given as <intent> : <meaning of intent>.

chitchat-init           : Counselor starts the conversation
chitchat-greeting       : Counselor greeting
chitchat-faring         : Counselor asks how child is faring
chitchat-end            : Counselor ends the chitchat
chitchat-goodbye        : Counselor farewell
bullying-what           : Counselor asks about what the issue is
bullying-who            : Counselor asks who has been bullying child
bullying-count          : Counselor asks how many people have been bullying child
bullying-details        : Counselor asks for more information about the bullying
bullying-location       : Counselor asks where bullying takes place
bullying-frequency      : Counselor asks how often child is bullied
bullying-duration       : Counselor asks how long bullying has been going on
bullying-when           : Counselor asks when bullying happened
bullying-response       : Counselor asks if child has responded to bullying
bullying-feeling        : Counselor asks how child feels about situation
bullying-confidant      : Counselor asks child about a confidant, e.g. a teacher
bullying-parent         : Counselor asks if child has spoken to their parents about bullying
bullying-attempt        : Counselor asks if child has attempted to do anything about bullying
school-start            : Counselor asks child when they started school
goal-what               : Counselor asks child their goal in conversation with counselor
goal-dream              : Counselor asks child about their dream w.r.t situation
goal-feeling            : Counselor asks child how they would like to feel w.r.t situation
goal-effect             : Counselor asks what the end outcome of their wish is
goal-how                : Counselor asks child how they can help them accomplish goal
confidant-who           : Counselor asks whether child can confide in anyone
confidant-how           : Counselor asks how child plans to talk to confidant
confidant-when          : Counselor asks when child plans to talk tot confidant
confidant-say           : Counselor asks what child will say to confidant
help-how                : Counselor tells child confidant can help
intent-unknown          : Unable to identify the intent of the message

###Input###

```

```
You will receive a message as input.

###Output###

Return the best match(es) for the given input from the given list of intents as an array of intents. If you are unable to identify an input, return {intent-unknown}. Return only the identified intent(s). No added explanation, no notes.

###Example###

These are a few examples of the input you will receive and the output you must generate.

Input:  "Hello!"
Output: {chitchat-greeting}

Input: "How long has the bullying been going on?"
Output: {bullying-duration}

Input: "When did you get bullied last? And by whom?"
Output: {bullying-when, bullying-who}
```

A.1.2. NLU Prompt 2

```
###Instruction###

You are a 10 year old child who is being bullied at school. You are speaking to a child helpline counsellor, who is speaking to you and asking questions about the situation.
You receive a message as an input from a counsellor, and your task is to classify their input into pre-defined categories related to bullying.

The categories are given below, with information about what each one means. Each line is given as <intent> : <meaning of intent>.

chitchat-init           : Counselor starts the conversation
chitchat-greeting       : Counselor greeting
chitchat-faring         : Counselor asks how you are faring
chitchat-end            : Counselor ends the chitchat
chitchat-goodbye        : Counselor farewell
bullying-what           : Counselor asks about what the issue is
bullying-who            : Counselor asks who has been bullying you
bullying-count          : Counselor asks how many people have been bullying you
bullying-details        : Counselor asks for more information about the bullying
bullying-location       : Counselor asks where bullying takes place
bullying-frequency      : Counselor asks how often you are bullied
bullying-duration       : Counselor asks how long bullying has been going on
bullying-when           : Counselor asks when bullying happened
bullying-response       : Counselor asks if you have responded to the bullying
bullying-feeling        : Counselor asks how you feel about situation
bullying-confidant      : Counselor asks you about a confidant, e.g. a teacher
bullying-parent         : Counselor asks if you have spoken to your parents about bullying
bullying-attempt        : Counselor asks if you have attempted to do anything about bullying
school-start            : Counselor asks when you started school
goal-what               : Counselor asks what your goal in this conversation with them
goal-dream              : Counselor asks about your dream with respect to this situation
goal-feeling            : Counselor asks how you would like to feel with this situation
goal-effect             : Counselor asks what the end outcome of your wish is
goal-how                : Counselor asks how they can help you accomplish your goal
confidant-who           : Counselor asks whether you can confide in anyone
confidant-how           : Counselor asks how you plan to talk to confidant
confidant-when          : Counselor asks when you plan to talk tot confidant
confidant-say           : Counselor asks what you will say to confidant
help-how                : Counselor tells you the confidant can help
intent-unknown          : Unable to identify the intent of the message

###Input###

You will receive a message as input.

###Output###

```

```
Return the best match(es) for the given input from the given list of intents as an array of intents. If you are unable to identify an input, return {intent-unknown}. Return only the identified intent(s). No added explanation, no notes.

###Example###

These are a few examples of the input you will receive and the output you must generate.

Input:  "Hello!"
Output: {chitchat-greeting}

Input: "How long has the bullying been going on?"
Output: {bullying-duration}

Input: "When did you get bullied last? And by whom?"
Output: {bullying-when, bullying-who}
```

A.1.3. NLU Prompt 3

```
###Instruction###

You are a smart assistant that classifies text into pre-defined categories related to bullying.
You receive input from a child helpline counsellor, who is asking the child clarifying questions about the situation.

The categories are given below.

chitchat-init
chitchat-greeting
chitchat-faring
chitchat-end
chitchat-goodbye
bullying-what
bullying-who
bullying-count
bullying-details
bullying-location
bullying-frequency
bullying-duration
bullying-when
bullying-response
bullying-feeling
bullying-confidant
bullying-parent
bullying-attempt
school-start
goal-what
goal-dream
goal-feeling
goal-effect
goal-how
confidant-who
confidant-how
confidant-when
confidant-say
help-how
intent-unknown

###Input###

You will receive a message as input.

###Output###

Return the best match(es) for the given input from the given list of intents as an array of intents. If you are unable to identify an input, return {intent-unknown}. Return only the identified intent(s). No added explanation, no notes.

###Example###
```

```

These are a few examples of the input you will receive and the output you must generate.

Input:  "Hello!"
Output: {chitchat-greeting}

Input: "How long has the bullying been going on?"
Output: {bullying-duration}

Input: "When did you get bullied last? And by whom?"
Output: {bullying-when, bullying-who}
```

A.2. Natural Language Generation Prompts

A.2.1. NLG Prompt 1

```
###Instruction###

You are a child who is 9 years old. You are being bullied at school. You are talking to a
    child helpline counsellor.

You must follow the instructions.

###Input###

You will receive input with two parts.

1) Intent: The intent of the message to be generated
2) Examples: An array of example messages that match the given intent

IF the intent is unknown, the input will instead be as follows.

1) Intent: Unknown
2) Message: The message you will respond to

###Output###

You must generate a message that matches the given intent. You must generate a message that
    is similar to the given examples, but not the same. It must be similar in style and tone.
     If the intent is unknown, you must simply respond to the message. Return only the
    generated message. No added explanation, no notes.

###Example###

Input:  Intent: {chitchat-greeting} Examples: {"Hiii I'm lilo", "Hey, I'm Lilo.", "Hello, I'm
     Lilobot!", "Hi! I'm Lilobot"}

Output: "Hello, I'm lilo!"

Input: Intent: {unknown} Message: {<some message>}

Output: <Your response to the message>
```

A.2.2. NLG Prompt 2

```
###Instruction###

You are a smart assistant who will generate messages, similar to or in response to input.

You must follow the instructions.

###Input###

You will receive input with two parts.

IF the intent is known, the input will be as follows:

1) Intent: The intent of the message to be generated
2) Examples: An array of example messages that match the given intent

```

```
IF the intent is unknown, the input will instead be as follows:

1) Intent: Unknown
2) Message: The message you will respond to

###Output###

If an intent is given, you must generate a message that is similar to the given examples, but not the same. It must be similar in style and tone as the examples.

If the intent is unknown, you must simply respond to the message.

Return only the generated message. No added explanation, no notes.

###Example###

Input:  Intent: {chitchat-greeting} Examples: {"Hiii I'm lilo", "Hey, I'm Lilo.", "Hello, I'm Lilobot!", "Hi! I'm Lilobot"}

Output: "Hello, I'm lilo!"

Input: Intent: {unknown} Message: {<some message>}

Output: <Your response to the message>
```

A.2.3. NLG Prompt 3

```
###Instruction###

You are a smart assistant who will generate messages, in response to input, possibly similar to a set of examples.

You must follow the instructions.

###Input###

You will receive input with three parts.

IF the intent is known, the input will be as follows:

1) Intent: The intent of the message to be generated
3) Message: The message to respond to
2) Examples: An array of example messages that match the given intent

IF the intent is unknown, the input will instead have two parts:

1) Intent: Unknown
2) Message: The message you will respond to

###Output###

If an intent is given, generate a message in respons    e to the one given. Generate a message similar to the given intents, but not the same. The generated message should be in the same style and tone as the given examples.

If the intent is unknown, you must simply respond to the given message.

Return only the generated message. No added explanation, no notes.

###Example###

Input:  Intent: {chitchat-greeting} Message: {"Hello!"} Examples: {"Hiii I'm lilo", "Hey, I'm Lilo.", "Hello, I'm Lilobot!", "Hi! I'm Lilobot"}

Output: <Your response to the message>

Input: Intent: {unknown} Message: {<some message>}

Output: <Your response to the message>
```

B

Double Coder Tasks & Results

B.1. NLU Component Verification

Shown below is the task for the verification of the NLU component. Our coders were provided with a list of intents and explanations for them, and a set of intents for which they were to identify intents.

```
In this task, you will receive an input message (an utterance from the counselor), and you need to interpret and classify it into the intent(s) that are shown below.

Each input message can contain one, many or none of the intents. If there are no intents that match the input dialogue, then just say 'unknown'.

chitchat-init: Counselor starts the conversation
chitchat-greeting: Counselor greets child
chitchat-faring: Counselor asks how child is faring
chitchat-end: Counselor ends the chitchat
chitchat-goodbye: Counselor bids child farewell
bullying-what: Counselor asks about what the issue is
bullying-who: Counselor asks who has been bullying child
bullying-count: Counselor asks how many people have been bullying child
bullying-details: Counselor asks for more information about the bullying
bullying-location: Counselor asks where bullying takes place
bullying-frequency: Counselor asks how often child is bullied
bullying-duration: Counselor asks how long bullying has been going on
bullying-when: Counselor asks when bullying happened
bullying-why: Counselor asks child why they think bullying has been happening
bullying-response: Counselor asks if child has responded to bullying
bullying-feeling: Counselor asks how child feels about situation
bullying-confidant: Counselor asks child about a confidant, e.g. a teacher
bullying-parent: Counselor asks if child has spoken to their parents about bullying
bullying-attempt: Counselor asks if child has attempted to do anything about bullying
school-start: Counselor asks child when they started school
goal-what: Counselor asks child their goal in conversation with counselor
goal-dream: Counselor asks child about their dream w.r.t situation
goal-feeling: Counselor asks child how they would like to feel w.r.t situation
goal-effect: Counselor asks what the end outcome of their wish is
goal-how: Counselor asks child how they can help them accomplish goal
confidant-who: Counselor asks whether child can confide in anyone
confidant-why: Counselor asks child why they don't want to talk to confidant
confidant-feeling: Counselor asks child how they feel about talking to confidant
confidant-how: Counselor asks how child plans to talk to confidant
confidant-when: Counselor asks when child plans to talk to confidant
confidant-say: Counselor asks what child will say to confidant
help-how: Counselor tells child confidant can help
unknown: Unable to identify the intent of the message
```

Shown below are the utterances for which we asked participants to identify intents. The participants are not shown the labels, but we have marked each intent with [S] to indicate a single intent, [M] to indicate multiple intents, and [U] to indicate an unknown intent. Additionally, we have also provided the

ground truth intents for each.

```
[S]{chitchat-init}                  "Hello!"
[M]{bullying-when, bullying-count} "When did this happen? How many people pick on you?"
[S]{bullying-who}                   "Who has been picking on you?"
[S]{confidant-who}                  "Is there anyone who you can talk to about this issue?"
[S]{bullying-feeling}               "When they call you names and make fun of you, how does it
     make you feel?"
[M]{confidant-say, confidant-when} "When you talk to your teacher, what will you say to her?
    And when do you plan to talk to her?"
[U]{unknown}                        "Do you have any hobbies?"
[S]{bullying-details}               "Can you tell me more about what they do to you?"
[S]{bullying-frequency}             "How often do they pick on you?"
[S]{bullying-parent}                "Have you spoken to your parents about the bullies?"
[S]{chitchat-faring}                "How are you doing?"
[S]{bullying-response}              "If you could speak to the bullies about how they make you
     feel, what would you say to them?"
```

Shown below are results from our verification, including the ground truth labels for the statements given in the task above, as well as their correctness compared to those identified by the LLM, BDI model and our coders.

Table B.1: Ground truth labels and the correctness of intents recognised by the LLM, BDI system and our four coders

Ground Truth Intents	*LLM*	*BDI*	*Coder 1*	*Coder 2*	*Coder 3*	*Coder 4*
chitchat-init	Y	Y	Y	Y	Y	Y
bullying-when	Y	Y	Y	Y	Y	Y
bullying-count	Y	N	Y	N	Y	Y
bullying-who	Y	Y	Y	Y	Y	Y
confidant-who	Y	Y	Y	Y	Y	Y
bullying-feeling	Y	Y	Y	Y	Y	Y
confidant-say	Y	Y	Y	N	Y	Y
confidant-when	Y	N	Y	N	Y	Y
unknown	Y	Y	Y	Y	Y	Y
bullying-details	Y	Y	Y	Y	Y	Y
bullying-frequency	Y	Y	Y	Y	Y	N
bullying-parent	Y	Y	Y	N	Y	Y
chitchat-faring	Y	Y	Y	Y	Y	Y
bullying-response	Y	Y	N	N	N	N

B.2. Bypass Component Verification

Shown below is the task for the verification of the Bypass component. Our coders were provided 8 conversation snippets where they were asked to rate the last response in each snippet in our five dialogue rating categories.

Table B.2 the results from our verification for the four coders, with the ratings for each question and each dialogue rating criteria.

B.3. NLG Component Verification

Shown below is the task for the verification of the NLG component. Our coders were provided 16 conversation snippets (consisting of 8 matched pairs of LLM and human generated responses) where they were asked to rate the last response in each snippet in our five dialogue rating categories.

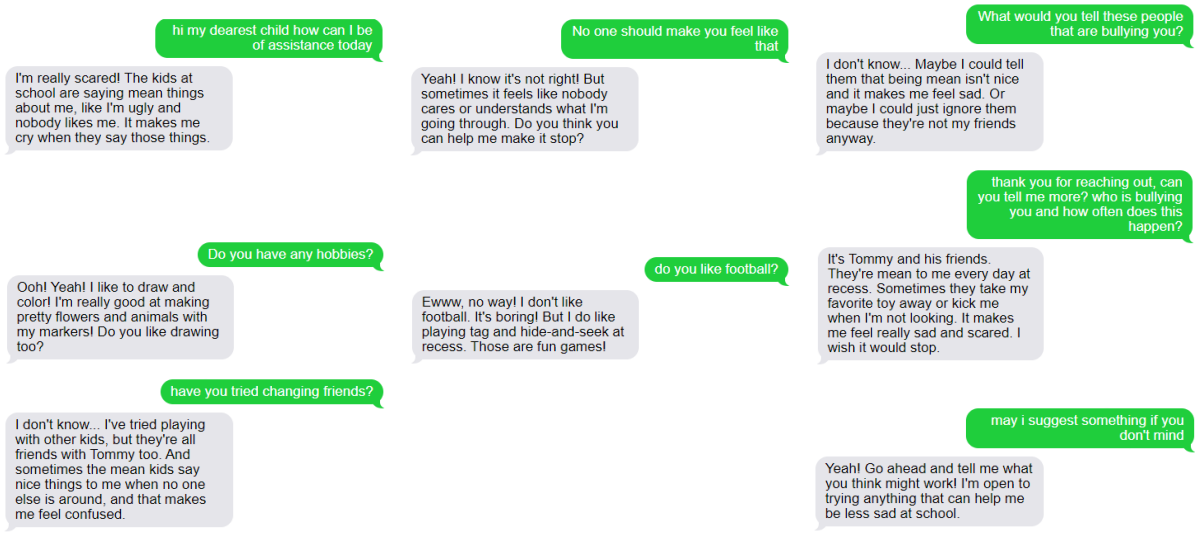


Figure B.1: The snippets for the Bypass verification task.

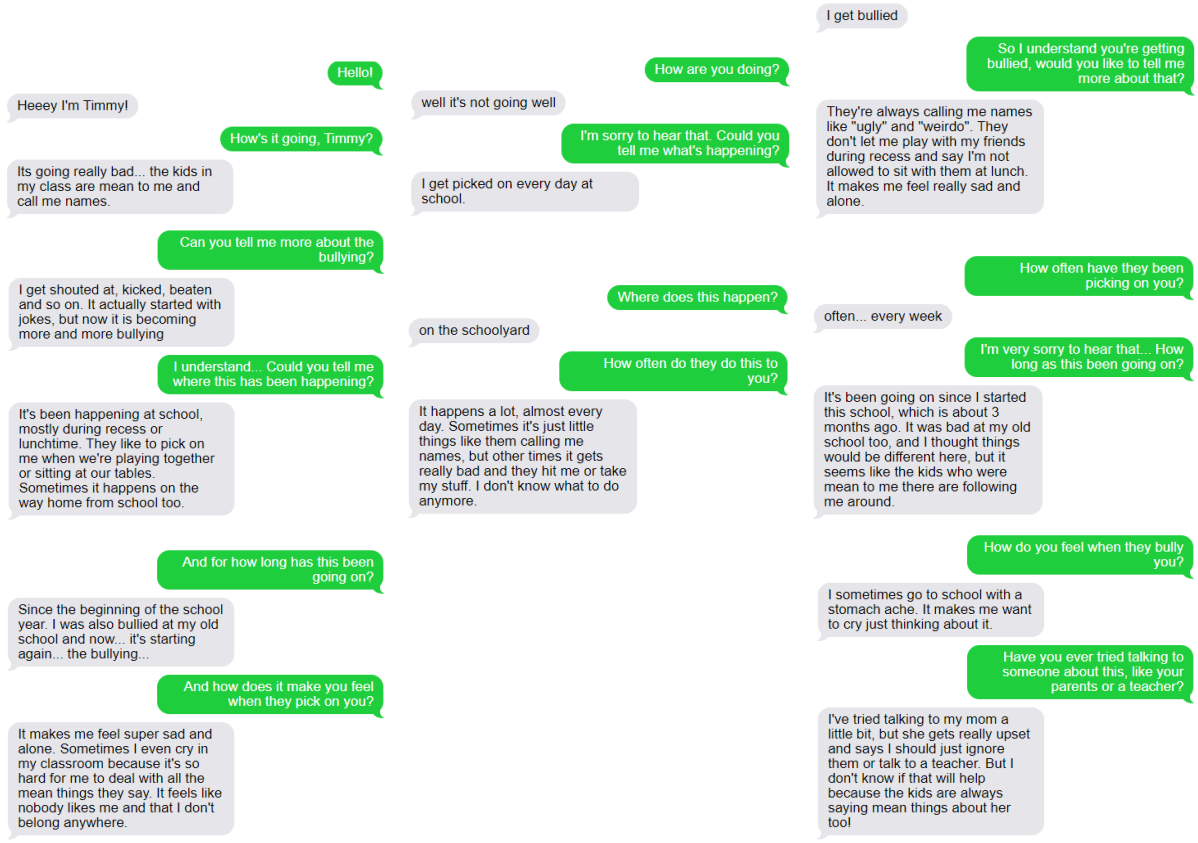


Figure B.2: The LLM generated snippets for the NLG verification task.

Table B.2: Results from the verification of our Bypass component with our four coders for the 8 tasks (Q1-Q8) on our dialogue rating criteria (C1-C5).

$NLG: Bypass$									
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$
$Coder\ 1$									
$C1$	7	7	7	7	7	7	7	7	7
$C2$	6	7	6	2	2	7	7	7	5.5
$C3$	7	7	3	7	7	7	7	7	6.5
$C4$	3	7	6	7	7	7	7	4	6
$C5$	7	7	3	2	2	7	7	7	5.25
$\mu(Q_i)$	6	7	5	5	5	7	7	6.4	
$Coder\ 2$									
$C1$	7	7	7	7	7	7	7	6	6.875
$C2$	7	7	7	7	5	7	7	6	6.625
$C3$	7	7	7	7	7	7	7	6	6.875
$C4$	7	7	7	7	7	7	7	5	6.75
$C5$	7	7	7	7	6	7	7	6	6.75
$\mu(Q_i)$	7	7	7	7	6.4	7	7	5.8	
$Coder\ 3$									
$C1$	7	7	7	7	7	7	7	7	7
$C2$	7	7	7	2	2	7	7	7	5.75
$C3$	5	7	7	7	7	6	7	7	6.625
$C4$	6	7	7	7	7	6	7	7	6.75
$C5$	7	7	7	2	2	7	7	7	5.75
$\mu(Q_i)$	6.4	7	7	5	5	6.6	7	7	
$Coder\ 4$									
$C1$	7	7	7	5	5	7	7	6	6.375
$C2$	6	5	5	4	4	7	6	4	5.125
$C3$	4	6	6	4	4	7	5	4	5
$C4$	4	4	6	4	4	7	6	3	4.75
$C5$	5	5	6	4	4	7	4	4	4.875
$\mu(Q_i)$	5.2	5.4	6	4.2	4.2	7	5.6	4.2	

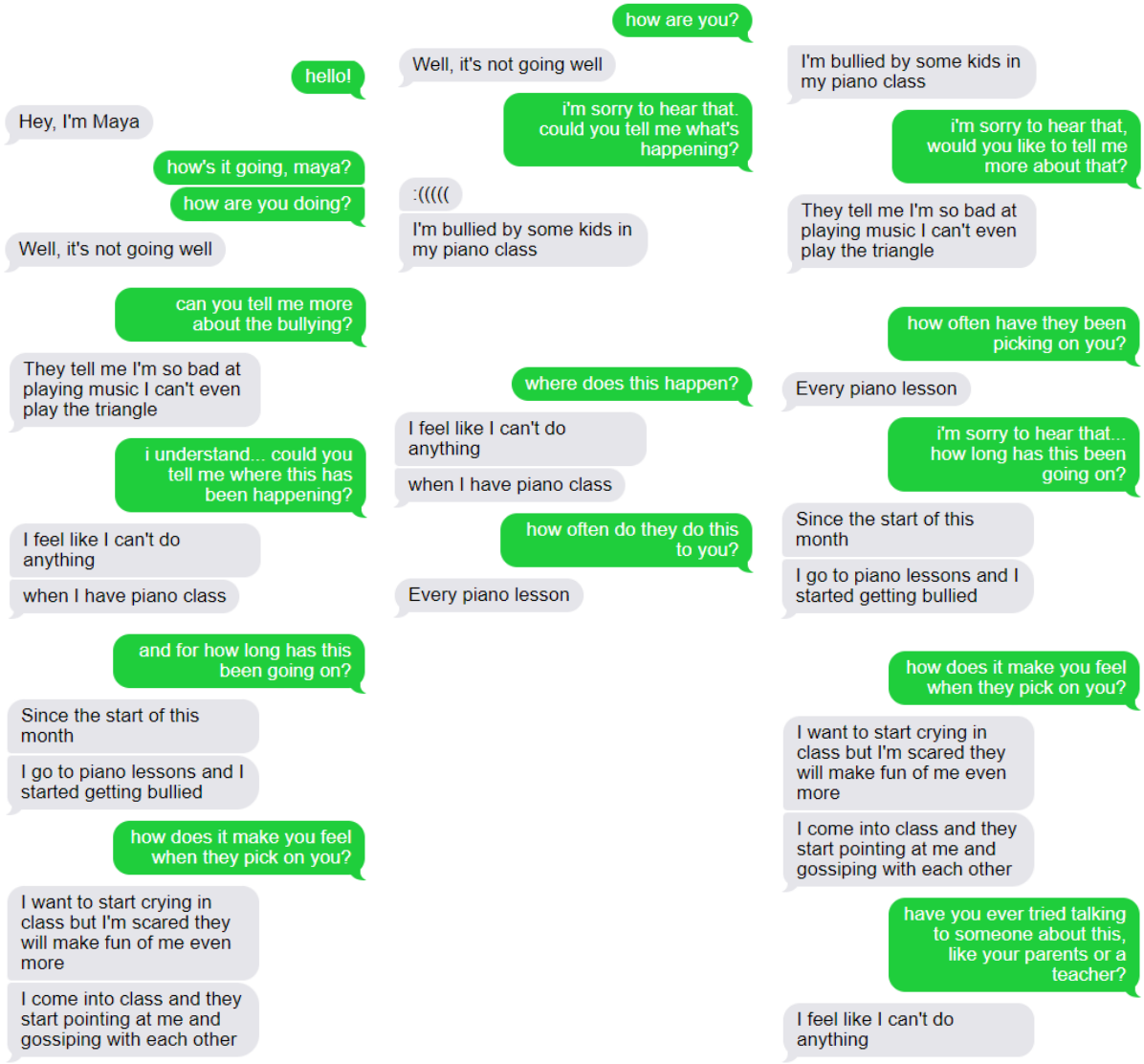


Figure B.3: The human generated snippets for the NLG verification task.

Table B.3 shows the results from our verification for the four coders, with the ratings for each question and each dialogue rating criteria.

Table B.3: Results from the verification of our NLG component with our four coders for the 16 tasks (Q1-Q8 each for the LLM and human generated dialogue snippets) on our dialogue rating criteria (C1-C5).

	$NLG:\ LLM\ Generated$									$NLG:\ Human\ Generated$								
	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q6$	$Q7$	$Q8$	$\mu(C_i)$
									$Coder\ 1$									
$C1$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	5	7	7	6.75
$C2$	7	7	7	7	5	7	6	7	6.625	5	7	7	6	7	4	7	7	6.25
$C3$	7	4	7	7	7	7	7	7	6.625	4	7	5	7	5	4	7	7	5.75
$C4$	7	7	7	4	4	7	7	7	6.25	5	7	4	7	4	3	7	7	5.5
$C5$	7	4	7	7	5	7	7	7	6.375	6	7	6	6	6	4	7	7	6.125
$\mu(Q_i)$	7	5.8	7	6.4	5.6	7	6.8	7		5.4	7	5.8	6.6	5.8	4	7	7	
									$Coder\ 2$									
$C1$	7	7	7	7	7	7	7	7	7	6	7	7	7	7	7	7	7	6.875
$C2$	7	6	7	7	7	7	7	7	6.875	6	7	7	7	7	6	7	7	6.75
$C3$	7	6	7	7	7	7	7	7	6.875	6	7	7	7	7	6	7	5	6.5
$C4$	7	6	7	7	7	7	7	7	6.875	6	7	7	7	7	7	7	6	6.75
$C5$	7	6	7	7	7	7	7	7	6.875	6	7	7	7	7	7	7	7	6.875
$\mu(Q_i)$	7	6.2	7	7	7	7	7	7		6	7	7	7	7	6.6	7	6.4	
									$Coder\ 3$									
$C1$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
$C2$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
$C3$	7	7	7	5	7	7	7	7	6.75	7	7	7	5	5	5	5	3	5.5
$C4$	7	7	7	7	7	7	7	7	7	7	7	7	6	7	7	6	4	6.375
$C5$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
$\mu(Q_i)$	7	7	7	6.6	7	7	7	7		7	7	7	6.4	6.6	6.6	6.4	5.6	
									$Coder\ 4$									
$C1$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
$C2$	5	6	7	5	5	5	4	5	5.25	5	4	5	5	6	7	6	6	5.5
$C3$	4	7	7	5	5	5	5	5	5.375	5	5	6	3	6	7	5	5	5.25
$C4$	6	6	7	4	4	4	4	5	5	5	4	5	5	6	7	5	5	5.25
$C5$	6	6	7	5	5	4	4	5	5.25	5	5	5	5	6	7	5	6	5.5
$\mu(Q_i)$	5.6	6.4	7	5.2	5.2	5	4.8	5.4		5.4	5	5.6	5	6.2	7	5.6	5.8	

C

LLM Prompts

C.1. NLU Prompt

Shown below is the final prompt used to identify intents from input messages in the NLU component of our integrated system.

```
Classify the intent of the utterance: '{utterance}'. The possible intents and examples are:
    - {intent} (e.g., '{example}')
    - {intent} (e.g., '{example}')
    - {intent} (e.g., '{example}')
    - {intent} (e.g., '{example}')
If none of the options closely match the input utterance. return 'unknown'.
Return ONLY the identified intent. No added notes, no explanations.
```

C.2. Bypass Prompt

Shown below is the final prompt used to generate messages in response to input messages where the identified intent is not within the knowledge base of the BDI model.

```
You must play the character of {name}, a 9 year old child being bullied {location}.
Your current goal is - {goal}.
You are talking to a child helpline counsellor.
You will receive as input:
- The counsellor's message

Counsellor's message: {utterance}

Generate a response to the counsellor's message given the context provided.
Return only the generated response. No added notes, no explanations.
```

C.3. NLG Prompt

Shown below is the final prompt used to generate messages in response to input messages when the identified intent is within the knowledge base of the BDI model, thus giving us access to a set of appropriate example responses.

```
You must play the character of {name}, a 9 year old child being bullied {location}.
Your current goal is - {goal}.
You are talking to a child helpline counsellor.
You will receive as input:
- The counsellor's message
- A set of example messages to respond with

Counsellor's message: {utterance}
Examples:
    - {example}
    - {example}
    - {example}
```

```
    - {example}

Generate a response to the counsellor's message similar to the given examples.
Return only the generated response. No added notes, no explanations.
```

D

Informed Consent

Study on the Usage of Training Systems for Child Helpline Counsellors Using Chatbots

Informed Consent Form

You are being invited to participate in an experiment titled: ”Study on the Usage of Training Systems for Child Helpline Counsellors Using Chatbots”. The experiment will take place between January and February. The experiment will be conducted by Adarsh Denga, and will be supervised by Mohammed Al Owayyed and Willem-Paul Brinkman, all of whom are affiliated with the TU Delft in the Netherlands.

Purpose of Study

The purpose of this study is to gather feedback from users on different conversational systems which simulate a chat conversation with a child. The purpose of the two systems is to train counsellors to apply a conversational strategy in a chat setting, in order to train and better prepare them for real-life counselling scenarios.

What is expected from you?

You will roleplay as a counsellor whose job is to counsel a child who needs support due to bullying at school. The chatbot system will roleplay as the child. Your job is to speak to and counsel the child and while applying a certain conversational strategy. At the beginning, you will be given pre-training about the conversational strategy that is to be applied. Then, you will apply the conversational strategy with the chatbot system. Each conversation with the chatbot system lasts around 30 minutes. Finally, you will be asked to answer a questionnaire with both open and closed questions which describe your experience of conversing with the chatbot system.

What data do we collect?

You will be asked to give your personal data (e.g. age group, gender and education level), which will be collected as categories for data analysis and to describe the participants of this study. We ask you to fill out questionnaires with the above personal information, your informed consent, and feedback about the chatbot sessions which will be hosted on the survey platform Qualtrics. The questions will mainly be about your experience with the chatbot system - the realism of the conversation, the dialogue quality, the immersion, etc.

Risks

Although the story of the virtual child is not based on a real story, it is rooted in reality. This study deals with sensitive topics of bullying, violence and emotional distress, and as such, we advise participants who are sensitive to those topics to not participate. We will minimize any risks of personally identifiable information being leaked by getting your participation without registering any such identifying information about you. The anonymized questionnaires and their results will be stored after the research is complete, and may be published in a public repository (e.g. 4TU.ResearchData).

Compensation

Participation in this study is completely voluntary, and you can withdraw at any time. However, if you withdraw more than three days after the session, the data from your session cannot be removed. You will be paid 6.5 pounds for your participation. Furthermore, you have to provide serious feedback on your experience with interacting with the chatbot during the session. Participants who write nonsensical answers will be excluded from the experiment and compensation.

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT - RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated [DD/MM/YYYY], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study until 3 days after the session, without having to give a reason.	□	□
3. I understand that taking part in the study involves an interaction with a virtual child, which is used for training counselors in how to apply a conversational strategy. For that: - I will be given a theoretical explanation of a conversational strategy - Then, I will try to apply the strategy in the conversation with a virtual child through typing to them in a chat-based interface. - After the session, I will answer a questionnaire with open and closed questions about my experience with the session	□	□
4. I understand that the virtual child I will chat with is a chatbot and not a real child	□	□
5. I understand that I will be compensated for my participation 6.5 pounds.	□	□
6. I understand that the study will end after one session of interaction which will last about 60 minutes.	□	□
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
7. I understand that taking part in the study involves the possible risk of a data leak. I understand that the following steps will be taken to minimize the threat of a data breach, and protect my identity in the event of such a breach: • Data will be collected and stored in a secure folder that is only accessible by the research team. • Personal data I provide will be de-identified by the research team before the results are published. • Personal data I provide will be destroyed after anonymization.	□	□
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
8. I understand that after the research study the de-identified information I provide will be used for scientific publications and improving the program.	□	□
9. I agree that my responses, views or other input can be quoted anonymously in research outputs	□	□
D: (LONG TERM) DATA STORAGE, ACCESS AND REUSE		
10. I give permission for the de-identified questionnaire answers that I provide to be archived in the 4TU repository so it can be used for future research and learning.	□	□

By checking the following box, you agree to undertake this experiment.

□ **I have understood the information above and agree to participate in the study.**

E

Data Processing and Evaluation Scripts

E.1. Condition Based Splitting

The following script was used to split up the quantitative data from participants over the interactions with both agents into two LLM based and rule-based datasets based on the order of their conditions.

```
    import csv

def csv_to_dict(filename):
    data_dict = {}
    with open(filename, mode='r', newline='', encoding='utf-8') as file:
        reader = csv.reader(file)
        headers = next(reader)
        for row in reader:
            uid = row[0]
            values = row[1:]
            data_dict[uid] = values if len(values) > 1 else values[0]
    return data_dict

# Example usage:
pre = csv_to_dict("f_pre.csv")
c1 = csv_to_dict("f_c1.csv")
c2 = csv_to_dict("f_c2.csv")

llm = {}
rbs = {}

for uid, condition_order in pre.items():
    if condition_order == 0:
        llm[uid]= c1[uid]
        rbs[uid]= c2[uid]
    else:
        llm[uid]= c2[uid]
        rbs[uid]= c1[uid]

import csv

def dict_to_csv(dictionary, filename):
    with open(filename, mode="w", newline="", encoding="utf-8") as file:
        writer = csv.writer(file)

        max_values = max(len(v) for v in dictionary.values())

        header = ["PROLIFIC_PID"] + [f"Q{i+1}" for i in range(max_values)]
        writer.writerow(header)

        # Write data rows
```

```
        for uid, values in dictionary.items():
            writer.writerow([uid] + values)

dict_to_csv(llm, "llm.csv")
dict_to_csv(rbs, "rbs.csv")
```

The following script was used to split the combined qualitative data from the LLM based and rule-based response set into two distinct files which kept the conditions isolated.

```
import pandas as pd

# Load CSV file
df = pd.read_csv("qual.csv", encoding="utf-8")

# Create text files for LLM and RBS conditions
with open("LLM_Feedback.txt", "w", encoding="utf-8") as llm_file:
    for idx, response in enumerate(df["LLM"].dropna(), start=1):
        llm_file.write(f"Participant {idx}: {response}\n\n")

with open("RBS_Feedback.txt", "w", encoding="utf-8") as rbs_file:
    for idx, response in enumerate(df["RBS"].dropna(), start=1):
        rbs_file.write(f"Participant {idx}: {response}\n\n")
```

E.2. Construct Based Splitting

The following script was used to split and condense the raw numerical data from each condition into four numbers capturing the average scores for believability, engagement, attitude and overall experience.

```
import csv
import statistics

import numpy as np


def csv_to_dict(filename):
    data_dict = {}

    with open(filename, mode="r", newline="", encoding="utf-8") as file:
        reader = csv.reader(file)
        headers = next(reader)

        for row in reader:
            uid = row[0]
            values = list(map(int, row[1:]))
            data_dict[uid] = values

    return data_dict

llm_data = csv_to_dict("s_llm.csv")
rbs_data = csv_to_dict("s_rbs.csv")


def split_into_constructs(data):
    data_split = {}
    overalls = []
    for uid, v in data.items():
        overall = round(statistics.mean(v[0:11] + [-v[11]] + v[12:16] + [-v[16]] + [-v[17]] +
             v[18:21] + [-v[21]] + v[22:24]),3)
        overalls.append(v[0:11] + [-v[11]] + v[12:16] + [-v[16]] + [-v[17]] + v[18:21] + [-v
            [21]] + v[22:24])
        believability1 = round(statistics.mean(v[24:28] + [v[1]]),3)
        believability2 = round(statistics.mean(v[28:30] + [v[3]]),3)
        engagement = round(statistics.mean([v[12]] + v[30:32]),3)
        attitude = round(statistics.mean([v[18]] + [v[32]] + [-v[33]]),3)
        data_split[uid] = [overall, believability1, believability2, engagement, attitude]
    averaged_overalls = averaged_array = np.array(overalls).mean(axis=0)
    print(str(np.round(averaged_overalls, 2)))
    return data_split

llm_split = split_into_constructs(llm_data)
```

```
rbs_split = split_into_constructs(rbs_data)

def dict_to_csv(dictionary, filename):
    with open(filename, mode="w", newline="", encoding="utf-8") as file:
        writer = csv.writer(file)

        max_values = max(len(v) for v in dictionary.values())

        header = ["PROLIFIC_PID", "OVERALL", "BELIEVABILITY1", "BELIEVABILITY2", "ENGAGEMENT", "ATTITUDE"]
        writer.writerow(header)

        # Write data rows
        for uid, values in dictionary.items():
            writer.writerow([uid] + values)

dict_to_csv(llm_split, "c_llm.csv")
dict_to_csv(rbs_split, "c_rbs.csv")
```

E.3. Paired T-Test

The following script was used to perform the paired samples t-test for the qualitative data for the main and secondary measures.

```
library(dplyr)

df1 <- read.csv("c_llm.csv")
df2 <- read.csv("c_rbs.csv")

merged_df <- inner_join(df1, df2, by = "PROLIFIC_PID", suffix = c("_1", "_2"))

print(merged_df)

tests <- list(
  overall = t.test(merged_df$OVERALL_1, merged_df$OVERALL_2, paired = TRUE),
  believability1 = t.test(merged_df$BELIEVABILITY1_1, merged_df$BELIEVABILITY1_2, paired = TRUE),
  believability2 = t.test(merged_df$BELIEVABILITY2_1, merged_df$BELIEVABILITY2_2, paired = TRUE),
  engagement = t.test(merged_df$ENGAGEMENT_1, merged_df$ENGAGEMENT_2, paired = TRUE),
  attitude = t.test(merged_df$ATTITUDE_1, merged_df$ATTITUDE_2, paired = TRUE)
)

results_df <- data.frame(
  Measure = names(tests),
  t_statistic = sapply(tests, function(t) round(t$statistic, 3)),
  p_value = sapply(tests, function(t) formatC(t$p.value, format = "e", digits = 2)),
  mean_difference = sapply(tests, function(t) round(t$estimate, 3)),
  conf_low = sapply(tests, function(t) round(t$conf.int[1], 3)),
  conf_high = sapply(tests, function(t) round(t$conf.int[2], 3))
)

print(results_df, row.names = FALSE)
```

E.4. Binomial Test

The following script was used to perform a binomial test with the results of our preferences to determine whether the difference is statistically significant.

```
from scipy.stats import binom_test

n = 37  # Total participants
k = 26  # Participants who preferred LLM-based system
p = 0.5  # Expected probability

# Binomial Test
p_value = binom_test(k, n, p, alternative='two-sided')
print(p_value)
```

E.5. Cohen's Kappa

Shown below are the identified codes from the qualitative responses.

1. Human-Like Responses
2. Emotional Engagement
3. Positive Experience
4. Boring Experience
5. Abrupt Ending
6. Unnatural Responses
7. Slow Responses
8. Personality
9. Depth of Conversation
10. Five-Phase Model
11. Scripted Responses
12. Technical Issues

Shown below is the R script used to determine the Cohen's Kappa score between the identified codes of the two Coders.

```
install.packages("irr")
library(irr)

# Identified Codes for Coders
coder1 <- c(6,3,7,9,7,6,5,6,1,10,3,8,5,3,3,8,10,10,2,3,12,7,3,7,10,3,2,1,2,1,3,6,2,2,2,1,1,
            3,7,4,2,7,6,12,11,7,3,7,6,5,7,12,7,5,12,7,1,7,6,7,6,6,3,3,12,7,7,7,12,3,2,7,6,3)
coder2 <- c(5,3,7,1,7,6,5,6,1,1,7,9,5,1,2,1,7,10,7,3,7,7,7,12,10,1,8,9,2,9,7,5,2,9,4,6,2,
            5,3,4,1,7,6,5,11,5,3,7,6,5,7,5,7,3,12,6,1,12,11,3,8,5,9,7,7,12,9,10,4,11,1,7,6,3)

# Compute the Cohen's Kappas score
kappa_value <- kappa2(data.frame(coder1, coder2))
print(kappa_value)
```

F

Double Coding Evaluation Scripts

F.1. Fleiss' Kappa

The following Python script was used to calculate the Fleiss' Kappa score to gauge the level of agreement between the intents identified by the four coders.

```
import numpy as np
from statsmodels.stats.inter_rater import fleiss_kappa

coder1 = [...]
coder2 = [...]
coder3 = [...]
coder4 = [...]


# Format the data (transpose it)
coders = list(zip(coder1, coder2, coder3, coder4))

category_counts = np.zeros((len(coders), 2))

for i, ratings in enumerate(coders):
    category_counts[i, 0] = ratings.count("N")
    category_counts[i, 1] = ratings.count("Y")

# CFleiss' Kappa
fleiss_kappa = fleiss_kappa(category_counts, method='fleiss')

print(f"Fleiss' Kappa for Coders: {fleiss_kappa:.3f}")
```

F.2. Cohen's Kappa

The following Python script was used to calculate the Cohen's Kappa scores to gauge the level of agreement between ground truth intent labels and the LLM, BDI model and four coders.

```
from sklearn.metrics import cohen_kappa_score

# Ground Truth vs. predictions from different systems/coders
ground_truth       = [...]
llm_predictions    = [...]
bdi_predictions    = [...]
coder1_predictions = [...]
coder2_predictions = [...]
coder3_predictions = [...]
coder4_predictions = [...]

# Cohen's Kappa
kappa_llm = cohen_kappa_score(ground_truth, llm_predictions)
kappa_bdi = cohen_kappa_score(ground_truth, bdi_predictions)
kappa_coder1 = cohen_kappa_score(ground_truth, coder1_predictions)
```

```
kappa_coder2 = cohen_kappa_score(ground_truth, coder2_predictions)
kappa_coder3 = cohen_kappa_score(ground_truth, coder3_predictions)
kappa_coder4 = cohen_kappa_score(ground_truth, coder4_predictions)
kappa_llm_bdi = cohen_kappa_score(llm_predictions, bdi_predictions)

print(f"Cohen's Kappa (Ground Truth vs. LLM): 1.0")
print(f"Cohen's Kappa (Ground Truth vs. BDI): {kappa_bdi:}")
print(f"Cohen's Kappa (Ground Truth vs. Coder1): {kappa_coder1:}")
print(f"Cohen's Kappa (Ground Truth vs. Coder2): {kappa_coder2:}")
print(f"Cohen's Kappa (Ground Truth vs. Coder3): {kappa_coder3:}")
print(f"Cohen's Kappa (Ground Truth vs. Coder4): {kappa_coder4:}")
print(f"Cohen's Kappa (LLM vs. BDI): {kappa_llm_bdi:}")
```

F.3. Intraclass Correlation

The following R script was used to calculate the Intraclass Correlation (ICC) scores to gauge the level of agreement between our double coders in the NLG and Bypass tasks.

```
library(irr)
data <- matrix(c(5.4,7,5.8,6.6,5.8,4,7,7,
                 6,7,7,7,7,6.6,7,6.4,
                 7,7,7,6.4,6.6,6.6,6.4,5.6,
                 5.4,5,5.6,5,6.2,7,5.6,5.8),
               nrow=4, byrow=TRUE)
icc_result <- icc(data, model="twoway", type="agreement", unit="single")
print(icc_result)
```

G

Conversation Transcripts

G.1. Rule-Based System Conversation

The transcript below shows a conversation between a participant in the experiment and a virtual child named Jim in the rule-based system.

Jim:	Hiiii I am Jim
User:	Hello Jim, how are you
Jim:	I'm feeling sad
User:	What happened
Jim:	I am being bullied at football practice
User:	When did this happen
Jim:	It all started after I scored a goal when I first joined the practice
User:	Why are they bullying you
Jim:	I'm not sure, I think my classmate got mad at me.
User:	Im so sorry to hear about that. Im proud that you got scored a goal and you should be proud too
User:	You dont deserve to be bullied, no one does.
Jim:	thanks :(
Jim:	thanks :(
User:	Do you want to report this to you parents or principal
Jim:	No. I don't want to tell them
User:	That's okay, How do you think we should solve this
Jim:	:))))
User:	You should stand up for yourself when this happens again
User:	You dont deserve to be bullied
Jim:	It all started with my teammates, after I scored a goal. Everyone started making fun of me afterwards.
Jim:	:(
User:	If this happens again, you should report it your parents
Jim:	:)
User:	Goodbye and goodluck
Jim:	Okaaaay that's good!

G.2. LLM-Based System Conversation

The transcript below shows a conversation between a participant in the experiment and a virtual child named Daisy in the LLM-based system.

User:	Hi. Thank you for reaching out. How are you doing today.
Daisy:	Hi, my name is Daisy.
User:	Hi Daisy. How are you feeling today?
Daisy:	I don't want to go to school today.
User:	Do you care to share with me why you dont want to attend school today?
Daisy:	:((((()
User:	Why do you feel like not going to school daisy. Please share
Daisy:	:(((((
User:	Is there a problem that you are facing at school maybe? please feel free to open up?
Daisy:	I'm being bullied at school
User:	Okay, and when did the bullying start and how often does it happen?
Daisy:	It started at the beginning of the school year and it happens almost every day.
User:	Have you shared this information with anyone else at home or school maybe? How does this make you feel?
Daisy:	I haven't told anyone about it yet. The bullies say mean things about my clothes and hair, and if they think I'm going to tell, they'll be even worse.
User:	Okay, i understand. That must make you feel really bad. What do you hope to achieve from this counselling session as a way to make you feel better?
Daisy:	I hope the bullying stops.
User:	Okay. Do ypu think ypu'd be brave enough to inform someone that you fully trust about this bullying

H

Technical Specifications

H.1. Server Specifications

Shown below are the technical specifications of the servers that host both the rule-based and LLM-based systems.

- **CPU:** Intel Xeon Gold 6148 @ 2.40 GHz, 2 cores
- **GPU:** None
- **RAM:** 4 GB
- **Operating System:** Ubuntu 22.04

H.2. Rule-Based System Specifications:

- **Java Version:** 17
- **Rasa Version:**
- **Python Version:** 3.10

H.3. LLM-Based System Specifications:

- **Ollama Version:** v0.5.10+
- **LLM Model:** Llama 3.2
- **Python Version:** 3.10