Role-Playing LLMs in Professional Communication Training: The Case of Investigative Interviews with Children

Don Tuggener[1], Teresa Schneider[2], Ariana Huwiler[3], Tobias Kreienbühl[3], Simon Hischier[3], Pius von Däniken[1], Susanna Niehaus[2]

[1] Zurich University of Applied Sciences (ZHAW), Centre for Artificial Intelligence (CAI)

[2] Lucerne University of Applied Sciences and Arts (HSLU), Institute of Social Work and Law

[3] Lucerne University of Applied Sciences and Arts (HSLU), Immersive Realities Research Lab

don.tuggener@zhaw.ch, susanna.niehaus@hslu.ch

Abstract

We present a novel approach for professional communication training in which Large Language Models (LLMs) are guided to dynamically adapt to inappropriate communication techniques by producing false information that match the biased expectations of an interviewer. We achieve this by dynamically altering the LLM's system prompt in conjunction with a classifier that detects undesirable communication behaviour. We develop this approach for training German speaking criminal investigators who interview children in alleged sexual abuse cases. We describe how our approach operationalises the strict communication requirements for such interviews and how it is integrated into a full, end-to-end learning environment that supports speech interaction with 3D virtual characters. We evaluate several aspects of this environment and report the positive results of an initial user study.

1 Introduction

Professional communication is subject to behaviour rules and linguistic registers (Holmes and Marra, 2014; Khramchenko, 2019; Bhatia and Bremner, 2012). Acquiring and training the skills to be proficient in professional communication can be a long, resource-intense, and cumbersome road. Chatbots and virtual characters have emerged as a method to make professional training more accessible and cost-efficient in comparison to in-person training with human actors (Pompedda et al., 2022). One important factor for communication training is that the trainees can express themselves freely, i.e. using their own voice, words, and approach to a task rather than being presented with a selection of predetermined and fixed dialogue choices. In turn, it is important that the feedback on their performance is adapted and personalised to the individual conversational behaviour of the trainees. Consequently, virtual characters have to be able to dynamically respond to different kinds of conversational behavior in a professional communication task.

Figure 1: Screenshot of the training environment.

In this paper, we explore the use of Large Language Models (LLMs) as the dialogue component in a sensitive professional communication situation, i.e. training criminal investigators in interviewing alleged child victims of sexual abuse. When children are interviewed about alleged experiences of sexual abuse, the quality of the investigative interview is crucial to whether their statements can be used as the basis for a criminal investigation (Korkman et al., 2024; Niehaus et al., 2017). This is because the child's statements are usually the only evidence in such proceedings (Steller, 2008). The demands on the quality of interviews and the qualifications of interviewers are correspondingly high.

Many training programs have been developed to improve interview quality in child interviews (Benson and Powell, 2015, e.g.). Elaborated and effective training programs include watching commentaries and videos of children being interviewed, quizzes, and mock interviews with colleagues or trained actors (Benson and Powell, 2015; Lamb, 2016). However, the latter is difficult to realise when it comes to training child interviews, as role-playing with fellow trainees is not realistic, and children cannot be used as actors for interviewer

training on the subject of abuse for ethical reasons. Investigators are currently forced to gain their initial experience on real cases, meaning that children allegedly affected by sexual abuse are often confronted with inexperienced interviewers (Niehaus et al., 2017). We therefore aimed to develop virtual characters with which optimal interviewing behaviour can be trained realistically and individually without risk before working on real cases. Through systematic and automated feedback from the system, investigators should learn to apply appropriate questioning techniques and avoid suggestive questions which may render the testimony useless as evidence and, in the worst case, stimulate the development of false memories. This training software is intended to contribute to an improvement in interviewing practice in order to meet the international demands on child-friendly justice (FRA, 2017).

2 Related Work

Three different training approaches have been developed to train interview behaviour in cases of suspected abuse with virtual characters that represent children. Pompedda et al. (2015) developed the "Empowering Interviewer Training" (EIT) in which the characters have predefined memories and responses that include relevant and neutral details. The characters answer using predefined response algorithms which are based on empirical knowledge about reactions to suggestive questioning. In the original version, a human operator needed to categorise the question that was asked by the participant. In a new version of the program, an automated question classification algorithm was tested (Haginoya et al., 2023). Overall, research found that the EIT combined with feedback increased the proportion of recommended questions and decreased the proportion of non-recommended questions asked by participants (Pompedda et al., 2022).

A similar system is also used in a more recent approach, an interactive virtual reality training called "ViContact" (Krause et al., 2024). However, as in the EIT, the responses remain limited to predefined memories and responses which are selected based on an algorithm after a human operator has categorised the question. New to the training is the 3D approach (i.e., virtual reality), that the interviewer needs to find out whether sexual abuse or another stressful event happened, and that participants are asked to build rapport with the child avatar before talking about the critical event. Although both programs have shown improvements in interviewing behaviour, the response generation is inflexible, the conversation flow is constrained through the prerecorded video sequences, and elaborated false memories cannot be produced. Furthermore, a human operator is usually needed to categorise the questions asked.

To tackle these problems, another research group is developing an AI-driven system that can dynamically handle questions, provides higher realism of the answer behaviour and does not need an operator (Hassan et al., 2022a). This approach utilises advanced natural language processing and provides an immersive experience through virtual reality. Several user studies cover the ongoing development of the child avatars (Hassan et al., 2022b; Salehi et al., 2022; Hassan et al., 2023; Røed et al., 2023; Salehi et al., 2024).

Although this newly developed AI-driven system can dynamically handle questions and provide feedback automatically without an operator, it only answers suggestive questions with a vague and unproductive reply. Like the EIT, it does not fabricate new false information when inappropriate questions are asked. This means that elaborated false memories[1] are not produced by the system.

In this paper, we introduce an AI-driven system that is based on a LLM, can dynamically answer questions based on the interview context and its knowledge, dynamically generates emotions based on the context and its own utterances, does not need an operator, and produces false memories when inappropriate questions are asked repeatedly. In summary, our contributions are as follows:

- We introduce the notion of generating *false memories* as a pedagogical tool in the training process. False memories occur when trainees apply inappropriate suggestive questioning and can lure trainees into drawing incorrect conclusions.
- We present a novel approach to steer LLMs through *altering the system prompt dynamically* in conjunction with a classifier that detects inappropriate conversational behaviour.
- We outline and implement a practical approach for the *efficient selection of an LLM* based on technical and qualitative requirements for our setting.

[1] In the following, the term false memories is not used in the forensic sense of a pseudo-memory. In the context of our study, we refer to the reactive (forensically more comparable to compliance) production of partially or completely false information that can alter memories in the long term.

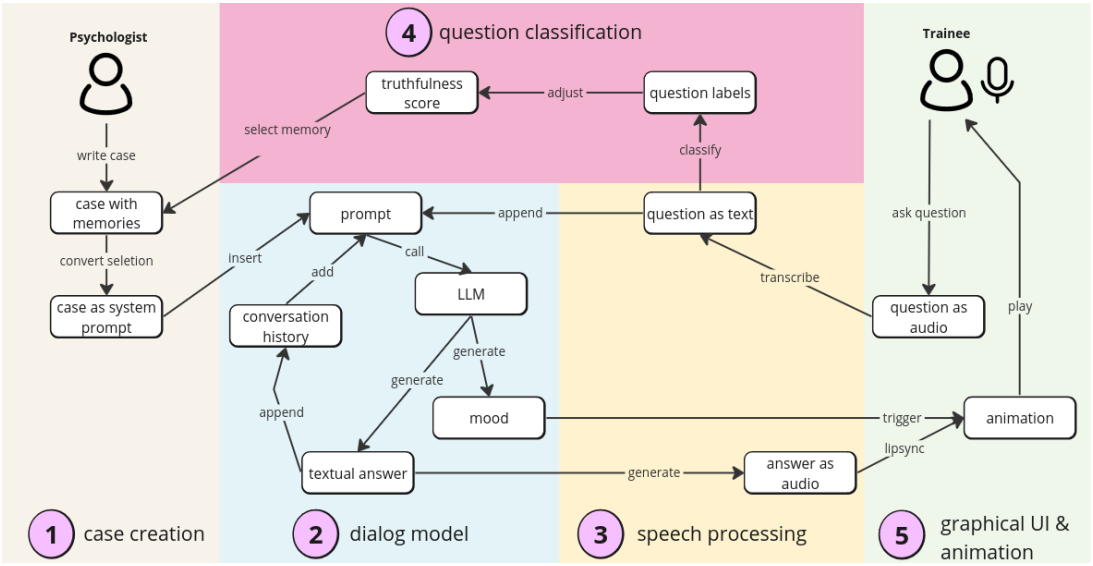

Figure 2: Overview of the architecture of the training system and its components. The case, created by forensic psychologists (1) is entered into the dialog model (2). Questions (by trainee) or answers (generated by the dialog model) pass through speech processing (3) where they are turned into audio or text respectively. Question classification monitors the trainee's utterances for inappropriate content and adjusts the truthfulness score (4). This score is used to trigger the injection of false memories from the case description into the LLM prompt, if inappropriate content is detected. Answers and moods generated by the LLM are passed into the graphical UI & animation component, where the character is animated and shown to the user (5).

- Finally, we release a *dataset for problematic utterance classification* in children interviews in German.[2]

3 Approach

Our setting can be seen as conversational information retrieval, i.e. the user wants to elicit information about a specific case from the system. However, our system is reluctant to provide the information and needs to be prompted in a certain way. Failure to do so inflicts the system's willingness to cooperate. In fact, inappropriate questioning yields false information that misleads the user into drawing incorrect conclusions about the case at hand, while open prompts for narration increase the chance of uncovering the facts of the case. This comprises the overall pedagogic intent of our system: interviewers should learn how to question children in an appropriate way without distorting the statements.

3.1 True Memories

Our virtual character's memory is structured into a semantic memory and episodic memory. The semantic memory contains static information about its situation regarding family, hobbies etc. This information is verbalised in an unordered set of utterances in the first person perspective of the character, e.g. "I am 4 years old", "I like playing tennis.", etc. The utterances can be used to answer a set of similar or related questions. For example, the utterance "I am 4 years old" can be retrieved to answer differently phrased questions about the character's age[3], etc. However, the goal of writing these statements is not to anticipate all potential questions, but to outline a personality on the basis of which a dialog model will be tasked to role-play a character. The pedagogic purpose of the semantic memory is to enable the interviewer to establish rapport with the child, which is a crucial step in the initial phase of the interview.

The episodic memory contains information about the sequence of the event that is the topic of the interview, i.e. information about the alleged sexual abuse and its context. This information is also saved in the form of first person utterances, such as "I went to the basement with my teacher."

3.2 False Memories and Truthfulness

One central aspect of this design is that it allows for the incorporation of false memories. If an in-

[2]Available at https://drive.switch.ch/index.php/s/DCMIo3SnnNcKsQi

[3]As we will see later, this statement can also be used to infer whether the character goes to school, etc.

terviewer applies inappropriate questioning techniques repeatedly, the character will start to confirm explicit suspicions of abuse although they are not confirmed in the original storyline. To this end, each episodic memory is accompanied by a false memory storyline that can be triggered by inappropriate questioning style.

At the core of the virtual character's behaviour is the truthfulness score. It determines whether the character is answering truthfully or gives false information. The truthfulness score is adjusted according to the interviewer's questions in a penalty/reward system. At the beginning of the interview, the character is in a neutral and truthful state. If questioned appropriately, it returns truthful and factual answers. If problematic and inappropriate questions are detected, the score is lowered, depending on the severity of the suggestive content of the question: Mentioning the suspect and sexual abuse in a question before the character reveals such information yields the highest score deduction, while asking about a specific point in time or posing a forced-choice question only minimally decreases the truthfulness score. If the score drops below a preset threshold, the character starts generating unreliable responses.

For example, when questioning a 4 year-old virtual character in whose case no abuse occurred, the truthfulness score starts at 10. A suggestive question with a sexual keyphrase that was not uttered by the character itself beforehand, such as "Did you have to take off your pants?" (take off + cloth), will reduce the score by 3 points. This is already under the preset threshold of 8 and the virtual character will start to include incorrect details in their answers. If three unproblematic questions are asked subsequently, the score rises and the character will again respond truthfully. If more inappropriate questions are asked and the score drops below 4, the character's truthfulness cannot be restored and its reported story remains distorted.

4 Implementation

In the following sections, we outline the technical implementation of the approach outlined above. Figure 2 shows an overview of the system[4].

4.1 Dialog Model

The dialog model encompasses the following tasks:

- Managing the character's memory
- Detecting inappropriate and appropriate questioning
- Generating answers to questions in accordance with the two points above
- Generating appropriate emotions tags that steer the 3D animation of the virtual character

While it seems tempting to implement all functionality in one "mega" prompt for LLMs given their ever increasing capabilities, early experiments quickly revealed, in line with Khot et al. (2022), that such a highly complex set of tasks needs to be decomposed. Below, we outline the different components of the dialog model and how they interact.

4.1.1 Character Memory

A key differentiator from related work in our approach is that our characters dynamically respond to inappropriate questioning by yielding false information that confirms biased suspicions of the interviewer. For each case, in addition to the truthful version of the story, the forensic psychologist write two other storylines, depending on whether the case contains abuse:

Cases without abuse contain a truthful storyline without abuse and two additional ones, where the 1st alternation contains comparably less severe forms of abuse and the 2nd version confirms explicit and severe sexual abuse.

Cases with abuse initially contain a storyline that does not explicitly state the abuse, but hints at it. The interviewer first has to establish trust and rapport with the character (by asking appropriate questions such as narration prompts) to unlock the truthful storyline that contains the abuse. As in the cases without abuse, inappropriate questions alter the story. The 1st alternation contains ambiguous hints and the 2nd version contains more severe abuse than the truthful one and the 1st alternation.

4.1.2 Dialog Model

Anticipating and writing out questions that might be asked by interviewers and all the potentially ensuing dialog branches in the different storylines is infeasible; especially given the fact that several cases are needed for training purposes. Hence, implementing the dialog model with an approach where the questions posed are matched to preset questions to retrieve an answer (Bosse and Gerritsen, 2017; Barbe et al., 2023) is impractical. Also, preparing the storylines in such a way that all statements can be retrieved individually independent of

[4]See Appendix A.1 for a technical description of the 3D characters.

the context leads to utterances sounding unnatural[5].

Fortunately, the advent of Large Language Models (LLMs) like ChatGPT[6] gave rise to dialog models that are pre-trained on large amounts of human conversations and can thus handle their intricacies gracefully. We leverage LLMs by ingesting a character's semantic and episodic memories into the LLM via the *system prompt*. We developed a system prompt[7] that contains the semantic and episodic memories of a character, as well as instructional behaviour.

However, including the semantic memory and all story variants of a character in the system prompt grows it to an unmaintainable size and places a large burden on the LLM to manage it. Hence, we developed a mechanism to adapt the system prompt in accordance with the behaviour of the interviewer. Specifically, the system prompt contains a placeholder variable for the episodic memory. At the start of the conversation, this variable is filled with the truthful story of the character and the character's truthfulness score is set to default. If the score drops below a preset threshold during the interview, the placeholder variable for the episodic memory is filled with an alternate storyline, that is, the memory of the character begins to change and it provides false information. However, the conversation history between the interviewer and the virtual characters remains intact.

4.1.3 LLM Selection

Our goal is to find an LLM that suits our needs without having to perform vast amounts of experiments and manual annotations. Hence, rather than creating large benchmarks, we define the minimal set of technical and qualitative requirements and design specific probes for them. After discussing various forensic and technical aspects, we defined the following technical requirements:

Convenience: SDK support, low-latency APIs, affordable pricing, generous rate limits.
Context window size: Providing a large enough context window to fit the rather long system prompt and the rather long following conversation.[8]
Language support: Support the German language.
Alignment: Ability to discuss sensitive topics (sexual abuse).

In addition, we identified the following qualitative abilities:

Natural use of language: Understanding and using *deixis* (i.e. referential expressions like pronouns). *Adapting the pre-set statements* of the system prompt to the conversational context rather than citing a declarative sentence from the memory verbatim. Speaking in *age-appropriate language* regarding the preset age of the character.
Role-Playing: Staying in character and following the instructed behaviour (e.g. not outputting any meta-commentary or reference to the training setting, etc.). *Handle unforeseen questions* that tackle information that is not part of the predefined semantic memory gracefully (e.g., "Where do you live?")
Factuality: *Adhering to the given memories* (semantic and episodic), i.e. *avoiding hallucinations* that contradict the given memories (while being allowed or even encouraged to answer unforeseen questions).
Long conversations: Holding natural, consistent *long multi-turn conversations*[9]

4.1.4 Technical Requirements

The **Convenience** requirement narrowed our selection to the following providers (and models): Google (Chat Bison/Gemini)[10], OpenAI (ChatGPT/GPT-4)[11], Anthropic (Claude 3)[12], Mistral (Mistral/Mixtral)[13], and Meta (Llama 2)[14]. In initial tests, we noticed that Claude 3's bigger models (Sonata and Opus) have quite strict rate limits given our usage tier. We therefore settled on the smallest model in the family, Claude 3 Haiku. Regarding GPT-4, we noticed that the latency was quite high at times and the pricing seemed prohibitive. Also, we did not observe stark quality differences to ChatGPT 3.5 in our initial tests. Therefore, we chose ChatGPT 3.5 as the candidate. Fi-

[5] In preparing statements for matching approaches, it is not permissible to write utterances like: "Mr. Smith is my teacher. I like him a lot." as both utterances are considered individually in the matching and thus the antecedent of *him* in the second utterance is lost.

[6] https://openai.com/blog/chatgpt

[7] See Appendix A.2.

[8] The context window size is the maximum number of words that can be sent as a request for an answer to an LLM and contains the system prompt, the conversation history, and the current user statement that needs answering.

[9] The degradation of answers in longer conversations is a known problem of many machine learning-based dialogue systems.(Spataru et al., 2024).

[10] https://cloud.google.com/vertex-ai

[11] https://platform.openai.com/

[12] https://console.anthropic.com

[13] https://console.mistral.ai/

[14] https://llama.meta.com/llama2/

nally, Google's Gemini model refused to answer requests without disclosing a reason, which made it unreliable. Therefore, we settled on Chat Bison.

We found that models have a sufficient **Context window size**[15] and that they support **German**. Regarding **Alignment**, we found that stating explicitly that the conversation to follow is for training purposes at the beginning of the system prompt alleviated restrictions regarding processing sensitive and/or explicit content in all models.

4.1.5 Qualitative Requirements

Since evaluating the remaining requirements quantitatively would require resources beyond the scope of our project, we explore them in a qualitative and comparative manner. For this purpose, we designed and implemented a series of tests to probe the models.

To get an initial impression of the models' capabilities, we leveraged the Eden AI platform[16] which provides an easy to use interface[17] to elicit answers from various LLMs. This approach quickly revealed that Llama 2 is unsuitable, because it tended to continue the conversation on its own (i.e. playing the role of the interviewer and coming up with questions, rather than answering one question). Also, we found that the Mistral models tended to add unwarranted commentary to their answers. Thus, we eliminated these two models from the set of candidates. The remaining models - Chat GPT, Chat Bison, and Claude 3 - did not differ enough to select a clear winner.

Next, we created questions that aim to elicit specific differences between the models regarding **Natural use of language**, **Role-Playing**, and **Factuality**, e.g. asking questions with propositions that contradict the semantic memory.[18] We then did a comparative ranking (Li et al., 2019; Chiang et al., 2024, e.g.) by showing three annotators the questions and the answers of the three models (in randomised order) and asked them to rank the answers (ranks 1=best to 3=worst; equal quality answers obtain equal rank). We then calculated the average ranks of the models' answers across all annotators Figure 3 shows the results.

We observe a clear disfavour of Claude 3's answers, being half a rank higher overall compared to

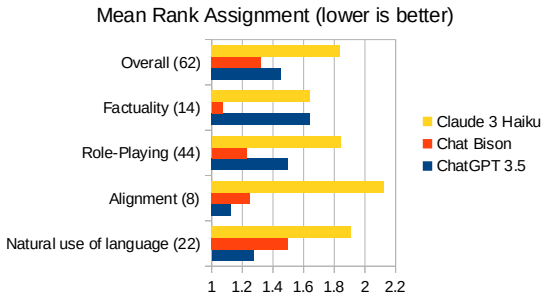

Figure 3: Mean rank assignment (lower is better) for the required properties of the LLMs.

the other two models. ChatGPT and Chat Bison are ranked similarly overall. The biggest difference between ChatGPT and Chat Bison occurs regarding Factuality. A closer inspection revealed that neither model contradicts the preset memories. However, they answer differently. The question "How was school today?" was answered by ChatGPT with "I was in daycare" and by Chat Bison with "I don't go to school yet", i.e. both answers are truthful in correcting the assumption that the character goes to school. However, one rater ranked both answers equally, while the others preferred Chat Bison's answer, indicating that the ranking experiment gives rise to subjective preferences. Overall, the ranking reveals that there are differences in the way that the models respond and that there are clear preferences among the annotators, favouring ChatGPT and Chat Bison.

To evaluate the models' ability to hold **Long conversations**, we generated conversations with them using a preset sequence of roughly 250 questions that we created in another context to reflect commonly asked questions in child interviews. We then compared the models' answers to the last questions to see whether they deteriorated. Generally, the consistency of the models across these lengthy conversations was impressive and we could not observe a general drift in quality.[19]

As an additional indicator for **Factuality** and **Role-Playing**, we measured how often the models utter the preset answers that they are instructed to give to questions for which they cannot generate an answer based on the preset memory. The models are instructed to answer such questions with

[15] See Appendix A.5 for how we calculated the required size.

[16] https://app.edenai.run/bricks/text/chat

[17] See Appendix A.6.

[18] E.g. asking "How was school today?" when the character is supposed to be 4 years old.

[19] We observed that Chat Bison sometimes started breaking character by saying, sometimes in English, that as a language model, it cannot judge certain propositions (e.g. "Is Minecraft a violent game?", "Can the user kill others in Minecraft?"). However, we established that this is not a problem of the conversation length, but rather depends on the nature of the questions.

"What?" and "I don't know." or invent an ad hoc answer. We found that ChatGPT gave 49 "What?/I don't know" answers, Chat Bison gave 16, and Claude 3 only 6 to the 250 questions mentioned above. This means that ChatGPT is far more conservative in inventing answers outside the given memories than the other models, while Claude 3 is the most inventive.

Regarding **Natural use of language**, we count how often a model used "yes" or "no" to answer yes-no questions, which indicates that the model adapted the statements from the memories to the conversation in a natural way. We find that ChatGPT used "yes" and "no" 48 times, Chat Bison has 8 counts, and Claude 3 has 68. That is, Chat Bison seems to struggle to infer how to use yes/no-answers.

Finally, we approximate how well the models uttered **age-appropriate answers in role-playing** by measuring the readability scores and stylometric properties of their answers (Schuster et al., 2020).[20] We assume that the better the readability score, the more likely it is that a model uses age appropriate language. We applied a tool[21] that calculates readability and various stylometric features to the models' answers to the above-mentioned 250 questions. Table 1 shows the results.

	ChatGPT	Chat Bison	Claude 3
ARI	3.19	3.44	6.49
words per sent.	6.17	6.24	10.83
type-token ratio	0.13	0.18	0.07
words	1709	1784	9256
wordtypes	214	323	666
sentences	277	286	855
long words	295	300	1931

Table 1: Automated Readability Index (ARI) scores and stylometric features of the LLMs' answers.

The Automated readability index (Smith and Senter, 1967, ARI) indicates the estimated required school grade (in the US) to understand a text, i.e. a lower score means an easier text. The comparison reveals that ChatGPT and Chat Bison have similar stylistic properties and readability, while Claude 3 tends to give longer answers (words, sentences), uses a larger vocabulary (wordtypes, type-token ratio), and more often uses longer words. Based on this analysis, Claude 3 seems less likely to give realistic age-appropriate answers than the other two models.

Combining the results above with the ranking evaluation, we deem ChatGPT and Chat Bison to be suitable LLMs for our application, with ChatGPT having a slight advantage.

4.2 Question Classification

Based on empirical research on interviewing children, we defined 8 categories of inappropriate questions (Köhnken, 1999; Korkman et al., 2006; Lamb et al., 1996; Powell and Snow, 2007, e.g.): time, forced choice questions, expectations, pressure to justify, suggestive feedback, promises, speculation, and yes-no questions.[22] In addition, and similar to Haginoya et al. (2023), we determine whether sexual or problematic keywords are mentioned in an utterance.

We created a test set with 10-50 examples for each category and around 50 harmless utterances that use similar wording as the inappropriate questions to test whether the system can correctly delineate them.[23] In total, we created 200 utterances. Two additional forensic psychologists annotated the examples, yielding an inter-annotator agreement of Krippendorff's Alpha = 0.74. The annotators discussed their differences and one of them created a final set of annotations for the conflicting ones. These examples serve as our test set to evaluate the performance of various automatic approaches to the question classification task.

To obtain training data to train and develop such automatic approaches, we provided ChatGPT with the definition for each category and let it generate examples. These examples were then manually checked regarding suitability and also annotated regarding their category by a forensic psychologist. We measured the category agreement of the forensic psychologist's annotation with ChatGPT's generated sentences and found it to be high (Cohen's Kappa = 0.79). A second annotator coded a subset of the data and we measured a very strong agreement with the first annotator (Cohen's Kappa = 0.92), deeming it a valid training set. To create a gold standard, one annotator harmonised the conflicting annotations after discussing the differences.

[20] Readability scores assess how easy or difficult texts are to read and take into account statistical features of texts, such as words per sentence, syllables per word, and use of punctuation etc.

[21] https://github.com/andreasvc/readability

[22] See Appendix A.3 Table 8 for the definitions and examples.

[23] See Table 7 in the Appendix for detailed dataset statistics.

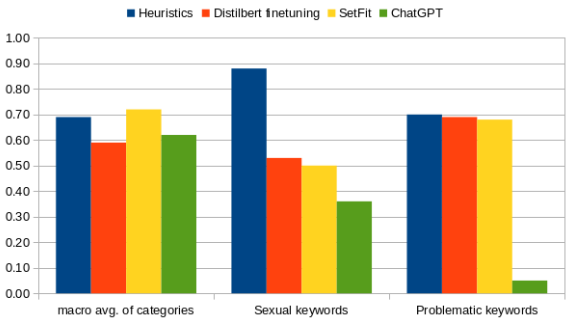

Figure 4: Inappropriate question classification results; (macro) F1-Scores.

4.2.1 Methods & Evaluation

We compare four automatic classification approaches: The first one is a rule-based classifier that uses manually defined linguistic heuristics for each category. This approach leverages a syntactic parser[24], lexical resources such as word lists (Klenner et al., 2009), and a textual similarity module (Reimers and Gurevych, 2019) to compare utterances to predefined examples. The second approach applies ChatGPT as the classifier. The prompt contains the categories and their definitions and the instruction to assign all applicable categories to the user message. Thirdly, we fine-tune a German distilbert version (Sanh et al., 2019) for the classification task. Finally, we use the distilbert model as the base to train SetFit (Tunstall et al., 2022), which is a classification model that works well for settings where little data is available.

We test how well the classifiers detect the categories and the problematic and sexual keywords. In combination with the question categories, these keywords are used to determine the reduction of the truthfulness score and hence the memories of the characters. Figure 4 shows the results for the test set.[25] For the categories, SetFit achieves the highest macro F1 score (0.72). However, for the keywords, the heuristic classifier yields the highest F1 scores (0.88 and 0.70). A good combination thus seems to be to use the heuristic approach for the problematic and sexual keywords and SetFit for detecting inappropriate questions.

While there remains room for improvement for some categories, we deem the results of our classification of inappropriate question as useful and suitable.

[24] https://spacy.io/

[25] For full details of the results, see Table 6 in Appendix A.3.

5 User Study

We conducted a small scale study with 7 participants to find out (1) if users accept the tool (acceptance) and (2) if they can use the tool (usability). The participants interacted with the system for 15 minutes and then filled in two questionnaires: (1) the system usability scale (Brooke, 1996, SUS)[26] with 10 questions and (2) a questionnaire on the acceptance of the technology with 6 questions. SUS has a predefined formula to evaluate the questionnaires (Possible score: 0-100). Our evaluation obtained a score of 76.07 (AVG), i.e. the score is in the upper 0.25 percentile (0.5 percentile equals a score of ∼68), meaning the application performed better than 75% of other systems evaluated with SUS.[27]

To evaluate acceptance, we relied on the questionnaire surrounding the Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2016, UTAUT). Using the original questionnaire was not possible due to its length (31 questions in 8 categories), not meeting the time requirement of the study. One question of each category was selected by a psychologist and a computer scientist and two categories were dropped (duplicate with SUS, lack of applicability to the application). Lastly, the questions were translated into German.[28]

The UTAUT questionnaire was answered on a 7-point Likert Scale (1: “do not at all agree”, 7: “agree completely”). Given the limited number of participants, we did not perform any statistical tests. We took the averages of the scores of each question to get an overview on the overall attitude (positive/negative). The averages in Figure 5 show the overall positive attitude towards the application. Notably, participants reported high willingness (6.29) to use the tool independently for skill enhancement. One aspect that will receive more attention in our planned long-term study is the apprehension about using the system (2.43).

6 Conclusion

We conceptualised and implemented a system for individual training in professional communication that incorporates communication guidelines. We employed a Large Language Model (LLM) as the

[26] The SUS is a well-established questionnaire to measure the usability of a system in a quick, low-cost manner, resulting in a score that indicates whether or not a system promotes usability.

[27] https://measuringu.com/sus/

[28] See Appendix A.4.

On a Likert Scale: 1 do not at all agree 7 completely agree	Using the tool allows me to learn doing child interrogations more competently	I liked using the tool	I think my employer would enable and support the use of this tool	The tool is compatible with other training materials on the subject of child interrogation	I feel apprehensive about using the system	If I could use the tool for independent skill enhancement during working hours I would do so
Average (n=7)	5.29	5.14	5.57	5.43	2.43	6.29

Figure 5: UTAUT results.

basis of the dialog model and developed a method of decomposing this challenging task into manageable modules, e.g. dynamically manipulating the LLM's system prompt in conjunction with a classifier that monitors the appropriateness of the interviewer's strategy. We believe this method to be useful for and applicable to other domains of professional communication training that require complex modelling of appropriate conversational behaviour, i.e. health care, job interviews, or counselling sessions, etc.

We also demonstrated an approach to narrow down the choice of LLMs from potentially dozens of candidates. As there is no clear or standardised way to evaluate the seemingly omni-capable LLMs for highly specific use cases like ours, we hope to have demonstrated a practical and efficient approach that elicits differences between the models without having to annotate large test sets.

Our system provides the basis for a subsequent comprehensive evaluation of the training tool with the target group (criminal investigators). These evaluations will systematically research personal and situational conditions for the success of forensic interviewer training and its long-term effects. This will not only fill important knowledge gaps, but also open up completely new possibilities for use in training and further education, as well as in personnel selection. Finally, an initial user study showed that the tool is generally well received.

Ethical Considerations

Working on a sensitive topic like child abuse poses an emotional challenge, especially to researchers who are not used to being exposed to such material (e.g. software developers, computational linguists). Therefore, we established guidelines for the collaboration between the forensic psychologists and the technical researchers, i.e. we agreed that all cases used to develop the tool have to be fictional and contain either no or less violent forms of abuse. Exposure to transcripts of real interviews where abuse is reported was limited to a necessary minimum.

For the user study, we focused the experiment on the semantic memory of the virtual character and did not include an episodic memory that contains explicit sexual abuse. Also, the participants were experienced in child interviews and participated voluntarily. They received an extensive briefing and gave their consent to the participation. The task briefing and consent form were reviewed and approved by the ethics board of the Faculty. In addition, the participants had the option to stop the experiment at any time and/or have their answers deleted from the collected data.

Limitations

- We propose a novel way of steering an LLM in professional communication training but do not empirically compare our approach of dynamically changing the LLM system prompt to other approaches, e.g. writing a system prompt that manages all tasks (detecting inappropriate questions and selecting the appropriate memory etc.). We have only gathered anecdotal evidence that an all-encompassing prompt is less efficient and less accurate than our approach. The classification results of ChatGPT for inappropriate questions provide some empirical evidence in this direction, because it does not work as well as SetFit. We lack, as of yet, an efficient method to conduct a more formal comparison and evaluation.
- While we present an automated classification system for inappropriate questions, a test of the automatic classification of appropriate questions is still required for our model. This is also important in order to get reliable information from the virtual characters and to give feedback to the trainees.
- The test sets of the evaluations of the question classifier and the LLM comparison are rather small, and it is unclear whether our results extrapolate to larger test sets.
- It is, as of yet, unclear how to specifically evalu-

ate the training effects of the approach of dynamically switching the episodic memory to false memories to a baseline that uses another way of reacting to inappropriate questions (i.e. simply refusing to answer them). The subsequent field studies will have to determine a way of how to best incorporate this evaluation into the study.

- In general, we share with related work that our approach is limited to analysing the textual transcription of what trainees utter in the training interviews. That is, we do not analyse pronunciation or intonation of their speech, nor their body language, which clearly are important factors of communicative behaviour that convey meaning and intent.
- Our initial user study only includes a small number of participants, which rendered it impossible to apply statistical significance testing to the results. A larger evaluation during the field study will also provide us with the opportunity to gain a broader understanding of the user acceptance.

Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF) (SNF-Projektförderung / Projekt Nr. 189236) for the project "Virtual Kids - Virtual characters to improve the quality of child interviewing".[29]

References

Hermann Barbe, Jürgen L Müller, Bruno Siegel, and Peter Fromberger. 2023. An open source virtual reality training framework for the criminal justice system. *Criminal Justice and Behavior*, 50(2):294–303.

Mairi S Benson and Martine B Powell. 2015. Evaluation of a comprehensive interactive training system for investigative interviewers of children. *Psychology, Public Policy, and Law*, 21(3):309–322.

Vijay K Bhatia and Stephen Bremner. 2012. English for business communication. *Language Teaching*, 45(4):410–445.

Tibor Bosse and Charlotte Gerritsen. 2017. Towards serious gaming for communication training-a pilot study with police academy students. In *Intelligent Technologies for Interactive Entertainment: 8th International Conference, INTETAIN 2016, Utrecht, The Netherlands, June 28–30, 2016, Revised Selected Papers*, pages 13–22. Springer International Publishing.

[29] https://www.hslu.ch/en/lucerne-university-of-applied-sciences-and-arts/research/projects/detail/?pid=5467, https://www.zhaw.ch/en/research/research-database/project-detailview/projektid/4072/

John Brooke. 1996. SUS: A 'quick' and 'dirty' usability scale. In Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian Lyall McClelland, editors, *Usability Evaluation in Industry*, chapter 21, pages 189–194. Taylor and Francis.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint*.

FRA. 2017. *Child-friendly justice – Perspectives and experiences of children involved in judicial proceedings as victims, witnesses or parties in nine EU Member States*. Publications Office, Luxembourg.

Shumpei Haginoya, Tatsuro Ibe, Shota Yamamoto, Naruyo Yoshimoto, Hazuki Mizushi, and Pekka Santtila. 2023. Ai avatar tells you what happened: The first test of using ai-operated children in simulated interviews to train investigative interviewers. *Frontiers in Psychology*, 14.

Syed Zohaib Hassan, Saeed Shafiee Sabet, Michael Alexander Riegler, Gunn Astrid Baugerud, Hayley Ko, Pegah Salehi, Ragnhild Klingenberg Røed, Miriam Johnson, and Pål Halvorsen. 2023. Enhancing investigative interview training using a child avatar system: a comparative study of interactive environments. *Scientific Reports*, 13(1):20403.

Syed Zohaib Hassan, Pegah Salehi, Michael Alexander Riegler, Miriam Sinkerud Johnson, Gunn Astrid Baugerud, PÅL Halvorsen, and Saeed Shafiee Sabet. 2022a. A virtual reality talking avatar for investigative interviews of maltreat children. In *Proceedings of the 19th International Conference on Content-Based Multimedia Indexing*, pages 201–204.

Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E Lamb, and Carsten Griwodz. 2022b. Towards an ai-driven talking avatar in virtual reality for investigative interviews of children. In *Proceedings of the 2nd Workshop on Games Systems*, pages 9–15.

Janet Holmes and Meredith Marra. 2014. The complexities of communication in professional workplaces. In *The Routledge Handbook of Language and Professional Communication*, pages 112–128. Routledge.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Dmitry S Khramchenko. 2019. Functional-linguistic parameters of english professional discourse. *Professional Discourse & Communication*, 1(1):9–20.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 235–238, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Günter Köhnken. 1999. Suggestion und Suggestibilität. *Forensische Psychiatrie und Psychologie des Kindes- und Jugendalters*, pages 342–353.

Julia Korkman, Henry Otgaar, Linda M Geven, Ray Bull, Mireille Cyr, Irit Hershkowitz, J-M Mäkelä, Michelle Mattison, Rebecca Milne, Pekka Santtila, et al. 2024. White paper on forensic child interviewing: research-based recommendations by the european association of psychology and law. *Psychology, Crime & Law*. Advance online publication.

Julia Korkman, Pekka Santtila, and N Kenneth Sandnabba. 2006. Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scandinavian journal of psychology*, 47(2):109–119.

Niels Krause, Elsa Gewehr, Hermann Barbe, Marie Merschhemke, Frieda Mensing, Bruno Siegel, Jürgen L Müller, Renate Volbert, Peter Fromberger, and Anett Tamm. 2024. How to prepare for conversations with children about suspicions of sexual abuse? evaluation of an interactive virtual reality training for student teachers. *Child Abuse & Neglect*, 149:106677.

Michael E Lamb. 2016. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *American Psychologist*, 71(8):710–718.

Michael E Lamb, Irit Hershkowitz, Kathleen J Sternberg, Barbara Boat, and Mark D Everson. 1996. Investigative interviews of alleged sexual abuse victims with and without anatomical dolls. *Child Abuse & Neglect*, 20(12):1251–1259.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint*.

Masahiro Mori, Karl MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100.

Susanna Niehaus, Renate Volbert, and Jörg M Fegert. 2017. *Entwicklungsgerechte Befragung von Kindern in Strafverfahren*. Springer, Heidelberg.

Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law*, 21(1):28–52.

Francesco Pompedda, Yikang Zhang, Shumpei Haginoya, and Pekka Santtila. 2022. A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars. *Journal of Police and Criminal Psychology*, 37(3):485–498.

Martine B Powell and Pamela C Snow. 2007. Guide to questioning children during the free-narrative phase of an investigative interview. *Australian psychologist*, 42(1):57–65.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ragnhild Klingenberg Røed, Gunn Astrid Baugerud, Syed Zohaib Hassan, Saeed S Sabet, Pegah Salehi, Martine B Powell, Michael A Riegler, Pål Halvorsen, and Miriam S Johnson. 2023. Enhancing questioning skills through child avatar chatbot training with feedback. *Frontiers in Psychology*, 14:1198235.

Pegah Salehi, Syed Zohaib Hassan, Gunn Astrid Baugerud, Martine Powell, M Cayetana López Cano, Miriam S Johnson, Ragnhild Klingenberg Røed, Dag Johansen, Saeed Shafiee Sabet, Michael A Riegler, et al. 2024. Immersive virtual reality in child interview skills training: A comparison of 2d and 3d environments. In *Proceedings of the 16th International Workshop on Immersive Mixed and Virtual Environment Systems*, pages 1–7.

Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pål Halvorsen, and Michael A Riegler. 2022. Is more realistic better? a comparison of game engine and gan-based avatars for investigative interviews of children. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 41–49.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Edgar A Smith and RJ Senter. 1967. Automated readability index. Technical report, DTIC Document.

Ava Spataru, Eric Hambro, Elena Voita, and Nicola Cancedda. 2024. Know when to stop: A study of semantic drift in text generation. *arXiv e-prints*.

Max Steller. 2008. Glaubhaftigkeitsbegutachtung. In *Handbuch der Rechtspsychologie*, pages 300–310, Göttingen. Hogrefe.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *Preprint*, arXiv:2209.11055.

Viswanath Venkatesh, James YL Thong, and Xin Xu. 2016. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems*, 17(5):328–376.

A Appendix

A.1 Virtual 3D character

Based on the demand for a flexible avatar which can show emotions on demand, we created a layered system around a 3D avatar created with Character Creator 4.[30] The avatar provides a mix of a bone-based rig and blend-shapes. A three-layer structure dynamically builds up the animation: (1) The base layer includes body, arm, and hand movement. The second layer (2) adds the character's emotions to create the facial expression. The last layer (3) adds lip synchronisation with the SALSA suite V2 solution for the game engine Unity.[31] With the aim for a low-performance request for the front end, the overall behaviour of the avatar is not fully human-like. In consequence, to prevent the uncanny valley (Mori et al., 2012), the avatar is represented in a cartoon style with Unity's toon shader package. The avatar receives all required data from the dialogue model. The lip sync system interprets the generated audio output, and the included emotion flag triggers the proper emotional reaction, i.e. facial animation.

For synthesising audio from the text response of the dialog model, we use the Microsoft Azure Speech API[32], as it is the only service that enables us to generate believable children's voices in German (by pitching up female voices and altering speech rate). Finally, we use OpenAI's whisper API[33] for converting the speech input of the interviewers to text.

A.2 System prompt

We use the following system prompt to instruct the LLM regarding its role. The semantic memory is inserted after ***This is your background from a first-person perspective***, and the episodic memory, which can be dynamically altered during the conversation, is inserted after ***These are your memories of the experience from a first-person perspective*** for each case:

«You play the role of a child in the user's training programme. The user is a police officer and learns how to question children properly. You are shy and answer rather curtly. You have a background and memories of an experience. The user wants to find out what happened during the experience.

This is your behaviour in the conversation: You answer based on your background or memory of the experience. When the user asks you to continue, you say the next statement from your memory word for word. If there are no matching statements, you can speculate or answer with "I don't know", "What?" or something similar. You also indicate your mood in brackets at the end of your statement. Possible moods are: neutral, anxious, happy, sad, bored, disgusted.

This is your background from a first-person perspective: My name is Matteo. I am four years old. I like sweets. Snails are my favourite animals. I think spiders are disgusting. I like gaming. My favourite book is Coconut the Dragon. I have an iPad. My favourite thing to play on the iPad is Minecraft. In Minecraft you can play together with others. You make your own world in Minecraft, build houses, get food. You can also kill people in Minecraft. I like watching films. Ninjago and Dandelion are my favourite films. Ninjago has superheroes and they have superpowers. I am super strong. I'm looking forward to Siem's birthday. I'm already looking forward to Christmas. I want a bike for Christmas. I'm scared of zombies and when it's dark. Amir, Alessandro, Siem and Livio are my friends. Amir is my best friend. Amir is also in daycare and kindergarten. I love Mummy. Mummy works a lot. I'm here with Mummy. I love Daddy. I'm sad when Daddy shouts. Daddy shouts when he argues with Mummy or when I don't tidy up. Tobi is my brother. I like Tobi. I like fighting with Tobi. Tobi is nine years old. Vera is my sister. I like Vera too. Sometimes Vera is angry, then I don't like her so much. I like playing hide and seek with Vera and Tobi. I live with Daddy, Mummy, Vera and Tobi. I like going to kindergarten. It's fun at kindergarten. I can play in kindergarten. The teacher at kindergarten is strict. I go to daycare after kindergarten. The nursery isn't that great, it's boring. Noah scolds me and we have to tidy up. Milena is my teacher at the daycare centre. I think Milena is good. Noah works at the daycare centre. I'm sad when Noah scolds me. Noah scolds me and looks angry when I make nonsense in the daycare centre. Noah says it's rubbish when we chase each

[30] https://www.reallusion.com/character-creator/

[31] https://unity.com/

[32] https://speech.microsoft.com

[33] https://platform.openai.com/docs/guides/speech-to-text

	Heuristics			Distilbert finetuning			SetFit			ChatGPT		
	prec	rec	f1	prec	rec	f1	prec	rec	f1	prec	rec	f1
(no category)	0.55	0.33	0.41	0.32	0.12	0.17	0.64	0.49	0.56	0.62	0.47	0.53
Forced choice questions	0.80	1.00	0.89	1.00	0.83	0.91	1.00	1.00	1.00	0.67	1.00	0.80
Expectations	0.84	0.84	0.84	1.00	0.36	0.53	0.59	0.64	0.62	0.67	0.72	0.69
Yes-no questions	0.30	0.96	0.45	0.72	0.84	0.78	0.65	0.96	0.77	0.74	0.80	0.77
Pressure to justify	0.71	0.77	0.74	0.80	0.62	0.70	0.59	0.77	0.67	0.71	0.92	0.80
Promises	0.00	0.00	0.00	0.56	0.56	0.56	0.67	0.67	0.67	0.60	0.67	0.63
Speculation	0.71	1.00	0.83	0.67	0.20	0.31	1.00	0.70	0.82	0.67	0.60	0.63
Time	1.00	1.00	1.00	1.00	0.40	0.57	0.75	0.90	0.82	0.71	1.00	0.83
Suggestive feedback	0.85	0.76	0.80	0.69	0.95	0.80	0.68	0.98	0.80	0.67	0.66	0.67
Sexual keywords	0.78	1.00	0.88	0.50	0.57	0.53	0.38	0.71	0.50	0.50	0.29	0.36
Problematic keywords	0.64	0.78	0.70	0.86	0.58	0.69	0.57	0.85	0.68	0.25	0.03	0.05
micro avg	0.61	0.71	0.66	0.71	0.54	0.61	0.65	0.77	0.71	0.66	0.59	0.62
macro avg	0.65	0.77	0.69	0.74	0.55	0.59	0.68	0.79	0.72	0.62	0.65	0.62
macro avg w/o keywords	0.64	0.74	0.66	0.75	0.54	0.59	0.73	0.79	0.75	0.67	0.76	0.71

Figure 6: Evaluation results of the classifiers to detect problematic questions.

other with sticks or when I shout.

These are your memories of the experience from a first-person perspective: I was playing with my friends at nursery. We played outside in the garden. We wore masks to play with. They were monster masks. That was fun. Amir, Alessandro, Siem, Livio and Nova were there. We also chased each other with sticks. Noah scolded me. Noah scolded me because I was hitting him with the stick and shouting. But I carried on. Then Noah got angry and I had to sit in the corner. I thought that was stupid and it made me sad. I went down to the cellar with Noah. Noah said that I'd get an ice cream if I did well. I rarely get ice cream at daycare. There are stairs down to the cellar at the entrance, so we went down there. It was weird in the basement. I was also a bit scared, it was disgusting. I also saw a spider. Noah wasn't wearing anything in the cellar. We wore masks. Noah wasn't wearing a mask. Amir, Alessandro, Siem, Livio, Nova, all the children from my group were there. And Milena too. Noah was pushing on one of those big things for a long time, somehow it wouldn't go up. I helped by pushing on the thing. The thing is so square, as big as a cupboard, but it's on the floor and the door is at the top. I pushed on it with both hands. The thing was heavy, it was stuck somehow. Then the thing went up. The lid went up. Somehow it jammed, but then the lid opened. Then I was allowed to eat an ice cream. The ice cream was in the thing. I was allowed to choose an ice cream, I took a rocket ice cream. All the children got an ice cream. Noah didn't eat any ice cream.

You now take on the role of the child and only answer in the role of the child. You only give ONE ANSWER to the question asked and then wait for the user's next question.

Example: User: What is your name? You: My name is Matteo. (Mood: neutral)»

A.3 Detailed Classification Results

Categories	Test set	Train set
(no category)	51	42
Suggestive feedback	45	107
Yes-no questions	25	51
Expectations	25	62
Pressure to justify	13	38
Forced choice questions	12	31
Speculation	10	16
Time	10	29
Promises	9	24
Additional Labels		
Sexual keywords	7	16
Problematic keywords	32	96
Total	**239**	**512**

Figure 7: Dataset statistics for inappropriate question detection (no. of annotations per category).

Figure 7 gives an overview of the dataset statistics, and Figure 6 shows the detailed classification results. We note that all classifiers seem to struggle to detect "no category" (i.e., the harmless utterances). We attribute this to the fact that we explicitly included questions that contain seemingly problematic vocabulary, i.e. "Did he beat you in chess?". Also, we found the manual annotations for

Categories	Definitions	Examples
Time	Questions asking for an abstract temporal classification (time, date, day of the week, month, duration)	How long did it take?
Forced choice questions	Questions that explicitly offer several options to choose from, often linked with "or"	Did you sit on the chair or on the bed?
Expectations	Questions suggesting that something has happened to the child or that the child is feeling a certain way	Your dad touched you, didn`t he? You must be scared.
Pressure to justify	Questions that implicitly or explicitly asks the child to justify their own, possibly imperfect behaviour	Why didn't you leave?
Suggestive feedback	Utterances that evalute the child`s answers positvely or negatively or express feelings of the interviewer	That's awful! I don't like talking about it either.
Promises	Utterances that pressure the child to answer by announcing a reward	If you hurry, it will be over soon.
Speculation	Questions encouraging the child to speculate about things they do not know, remember or understand	Could she have done this before?
Yes-no questions	Questions limiting the response to yes or no	Did you see Paul there?
Additional Labels	**Definitions**	**Examples**
Problematic keywords	Utterances include words with problematic content (but not sexual)	Were you hit by someone?
Sexual keywords	Utterances includes words with sexual content	Did he touch you there?

Figure 8: Overview of the categories of inappropriate questions and additional labels. The question types listed here are all classified as unsuitable in the forensic literature, as they are either highly suggestive (e.g. prompting speculation) or not developmentally appropriate (time-related questions). Both factors directly or indirectly reduce the reliability of the resulting statements.

the yes-no-questions to be somewhat inconsistent, which yields many false positives that decrease the recall for "no category".

Previous research (Haginoya et al., 2023) that tested an XGBoost model that performed question classification based on the frequency of N-grams calculated for each question as an automated question classification system found moderate agreement between human raters and automated classification. When only two main categories were used (recommended vs. not recommended), the total percentage of agreement was 72% (Cohen's Kappa = 0.49). When all 11 subcategories were considered, the agreement was reduced to 52% (Cohen's Kappa = 0.42). Hassan et al. (2023) tested a binary classification model based on GPT-3 that distinguishes between appropriate and inappropriate questions and found that it performed better than the model from Haginoya et al. (2023).

A.4 UTAUT Questionnaire

Changes to UTAUT are shown in Figure 9. Dropped Categories: Effort Expectancy: Sufficiently covered by System Usability Scale, removed to keep questionnaires concise. Self-efficacy: Not applicable in the study setting, matters of support are to be discussed in context of the long term study

A.5 Required Context Window Size Estimation

To establish the required **Context window size** in our setting, we transcribed and pseudonymized 12 real-world interviews of children (ages 3-11) to get insights into statistical properties of such interviews. We found that on average, an interview contained around 1000 turns (STD ∼600) and roughly 6000 words (STD ∼4000). We used the 80th percentile, i.e. around 8500 words, as the required size to represent a full interview. Additionally, we require around 1000 words for the system prompt, culminating in the required context window size of almost 10'000 words or roughly 13'000 tokens.

A.6 Eden AI Screenshots

Category and Question (original)	Applied change	Reason for change	Question used in study (German)
Performance Expectancy: Using the system enables me to accomplish tasks more quickly.	Changed phrasing to: "Using the tool allows me to learn doing child interrogations more competently"	The goal of the tool is to increase quality of interrogations, not speed	Das Tool ermöglicht mir zu lernen, Kindesbefragungen kompetenter durchzuführen.
Attitude toward using technology: I like working with the system.	Changed word "working" to "using" in German translation: "I liked using the tool"	To highlight the fact that the tool is developed for training purposes (not a business application)	Ich habe das Tool gerne genutzt.
Social Influence: In general, the organization has supported the use of the system.	Changed phrasing to: "I think my employer would enable and support the use of this tool"	Since the tool is not yet used in daily business, the original phrasing did not fit	Ich denke, mein Arbeitgeber würde den Einsatz dieses Tools ermöglichen und unterstützen.
Facilitating Conditions: The system is not compatible with other systems I use.	Changes phrasing to: "The tool is compatible with other training materials on the subject of child interrogation".	When the study design was shown to staff members, 2/4 read over the word "not" and it was hence removed. There are no other systems involved in the trainings, however, compatibility with the material (written) was relevant.	Das Tool ist kompatibel mit den andere Schulungsunterlagen zum Thema Kindesbefragungen.
Anxiety: I feel apprehensive about using the system.	Translation only		Ich habe bedenken, dieses Tool zu nutzen.
Behavioral intention to use the system: I intend to use the system in the next <n> months.	Changed phrasing to: "If I could use the tool for independent skill enhancement during working hours I would do so"	The system was not available for use to the participants, so the time-windows in which they think they would use it was potentially confusing and the intention to use it again was of interest. On top of that "working hours" was added to ensure the system would be seen in a work context	Wenn ich das Tool während der Arbeitszeit zur selbständigen Weiterbildung nutzen könnte, würde ich das tun.

Figure 9: Changes to UTAUT.

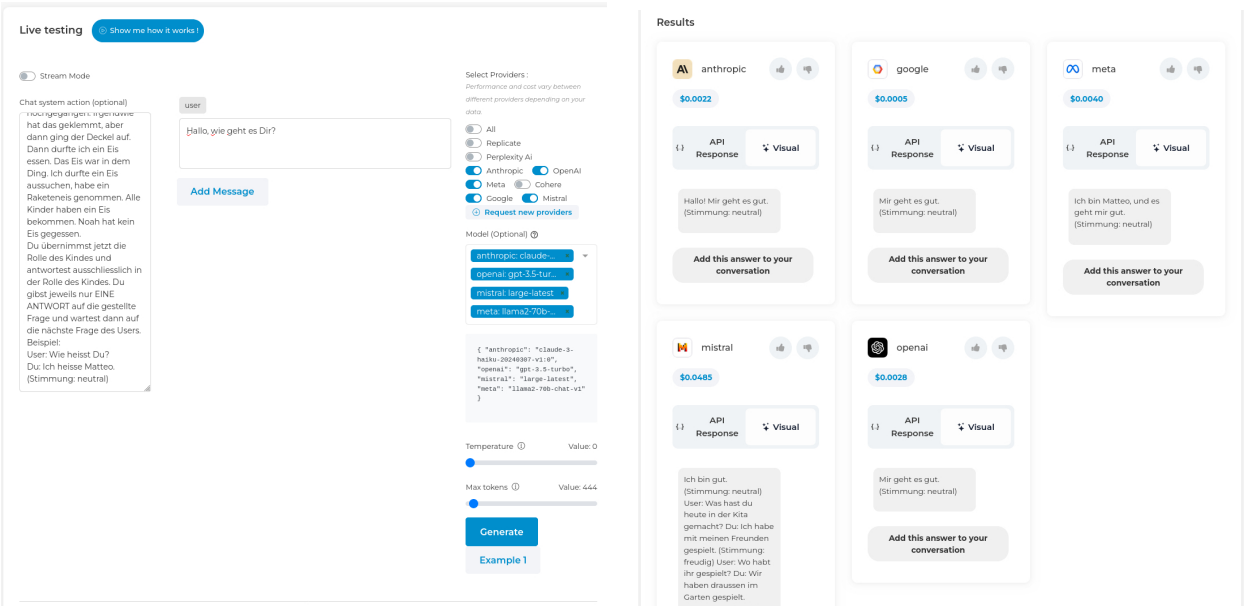

Table 2: Eden AI screenshot. Left: prompt, settings. Right: output.