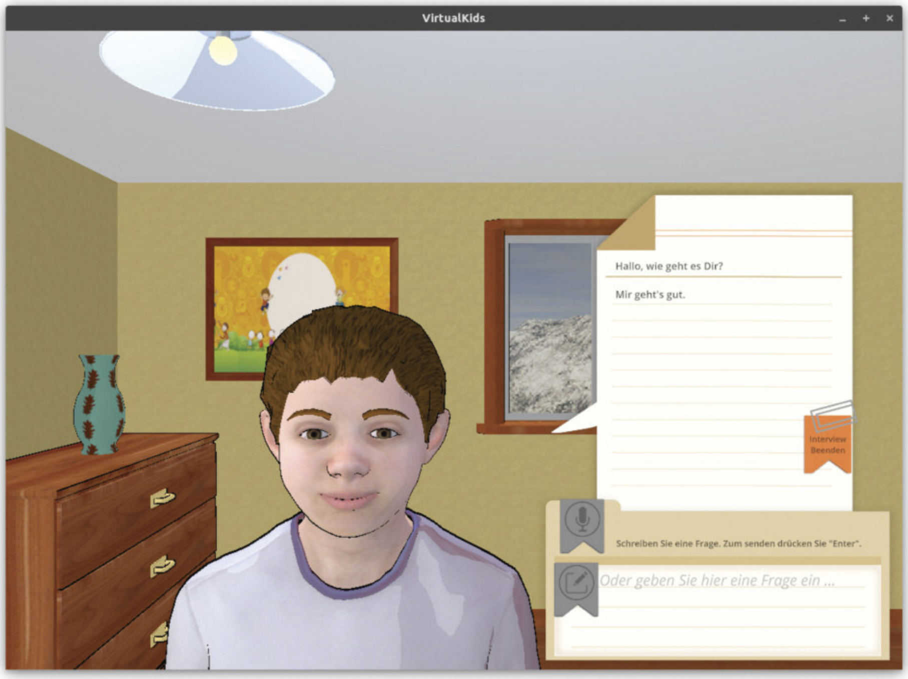
Using LLMs in Professional Training: Criminal Investigators Improve their Skills in Children's Interrogations

by Don Tuggener (Zurich University of Applied Sciences) and Susanna Niehaus (Lucerne University of Applied Sciences and Arts)

How do you teach a knowledge-based chatbot to reveal its information only conditionally? An unusual setting in the age of information and chatty dialogue systems – and precisely the premise of the Virtual Kids project.

In the Virtual Kids project [L1], the hesitant chatbot is an avatar of a child who is to be interrogated by trainee criminal investigators. The avatar has a delicate story to tell, a story that potentially involves sexual abuse. The trainees' task is to determine whether a crime has occurred in the avatar's past by figuring out the case details. Only when the trainees apply the adequate questioning techniques will the avatar be cooperative and report events truthfully. Inappropriate questioning like applying pressure and voicing expectations can lead to false statements and, in the worst case, trigger the development of pseudo-memories in the avatar, which ultimately leads to an incorrect assessment of the case. It is precisely this learning effect that the training programme aims to achieve, with the goal to protect criminal investigators from conducting sensitive child interviews inappropriately and thus rendering them legally useless, which has fatal consequences for the child, the criminal investigator, and the criminal prosecution [1][2]. To achieve this goal, the Virtual Kids project team consists of forensic psychologists, AI researchers, and game design researchers who collaboratively build a learning environment.



Hallo, wie geht es Dir?

Mir geht's gut.

Schreiben Sie eine Frage. Zum senden drücken Sie "Enter".

Oder geben Sie hier eine Frage ein ...

Figure 1: A screenshot of the user interface.

The Impact of Large Language Models (LLMs)

The learning environment features multiple avatars of children that each have their own personality and story. The trainees receive a case briefing and then talk to the avatar in a virtual room (see Figure 1). Crucially, the trainees can speak freely and are not guided by pre-set questions. After the interview, the interrogation technique of the trainees is evaluated automatically and presented in a feedback screen.

The emergence of chat-oriented LLMs like ChatGPT [L2] has recently accelerated the progress of the project tremendously. The dialogue component pre-LLMs was implemented based on Question Answering (QA) models that necessitated a rather unnatural conversation style to enable the QA model to retrieve relevant answers. The trainees were instructed to avoid using pronouns and to not refer to the dialogue conducted so far, as the QA model was only able to answer each question in isolation. This placed a cognitive burden on the trainees that distracted them from applying the correct interrogation techniques. Similar restrictions were imposed on the forensic psychologists who write the avatars' stories: the stories needed to be a sequence of utterances that were understandable without context and therefore could not include anaphoric expressions. Chat-oriented LLMs handle the intricacies of natural language conversations gracefully and remedy the need to orchestrate separate modules for pronoun resolution, dialogue memory, and dialogue state tracking. These capabilities alleviate the need to place counter-intuitive constraints on the conversation behaviour of the trainees and facilitate the case writing.

Another important benefit of LLMs is that they are able to dynamically extend their memories. For example, a trainee might ask whether the avatar likes to play chess, but the pre-set memory does not have an answer. In the QA model approach, the avatar would answer with "I don't know," because it cannot retrieve an appropriate answer. As it is infeasible to anticipate all questions, the QA-driven avatar would often return "I don't know" answers, which can be demotivating for the trainees. In contrast, LLMs are able to dynamically answer the question with, e.g. "No, I hate chess" and then explain why that might be so. This behaviour would be undesirable elsewhere, but in this case, it is actually helpful. However, the willingness to invent answers has to be carefully steered in the system prompt and by adjusting the LLM's parameters.

The Importance of the System Prompt

An avatar’s behaviour and story are described in natural language in the so-called system prompt. The system prompt assigns a particular role to the LLM that plays the avatar. The Virtual Kids prompts contain demographic information about the avatars (gender, age etc.) but also personality traits, like shyness. They also contain two sets of memories: the semantic memory, which describes trivia of an avatar (e.g. hobbies), and the episodic memory, which contains the sequence of events that are the focus of the interview. The semantic memory enables the trainees to establish rapport with the avatar, which is an important first step in the interrogation process.

In the beginning of the conversation, the avatar is in a neutral mood and answers questions truthfully. The forensic psychologists defined 12 categories of inappropriate questions which the trainees learn to avoid. Each question of the trainees in the interrogation is automatically evaluated by a linguistic model. If the model detects inappropriate questions, the truthfulness of the avatar is decreased, and, if a pre-set threshold is crossed, the episodic memory is swapped and the avatar starts giving incorrect information and eventually answers based on confabulations. That is, the forensic psychologists write three versions of the avatars’ episodic memories: a truthful one, a version that tends to confirm false suspicions or contains aggravations, and a version that contains explicitly false statements, such as explicitly confirming a false suspicion of abuse. The original, truthful memory’s utterances intentionally contain some ambiguities that invite suspicion that need to be disentangled carefully.

Initial user tests with the avatars indicate that the app is generally well received and its purpose is understood. Users expressed the desire for specific in-app feedback of their performance and were sometimes unsatisfied with the speech component. More extensive studies will determine the learning effect of including such a training in the criminal investigators’ education.

Links:
[L1] https://kwz.me/hAg
[L2] https://openai.com/chatgpt

References:

[1] F. Pompedda et al., “A mega-analysis of the effects of feedback on the quality of simulated child sexual abuse interviews with avatars,” J. Police Crim. Psychol., vol. 37, pp. 485–498, 2022. doi: 10.1007/s11896-022-09509-7P

[2] S. Haginoya et al., “AI avatar tells you what happened: The first test of using AI-operated children in simulated interviews to train investigative interviewers,” Front. Psychol., vol. 14:1133621, 2023. doi: 10.3389/fpsyg.2023.1133621

Please contact:

Don Tuggener, ZHAW, Switzerland
tuge@zhaw.ch

Unveiling Ethical Biases in Generative AI

by Sergio Morales (Universitat Oberta de Catalunya), Robert Clarisó (Universitat Oberta de Catalunya) and Jordi Cabot (Luxembourg Institute of Science and Technology)

Generative AI models are widely used for generating documents, videos, images, and so on; however, they can exhibit ethical biases that could be harmful or offensive. To prevent this, we propose a framework to test the fairness of generative AI before integrating such models in your daily work.

Generative AI has reached a broad audience thanks to the many services that make it available to non-tech people (e.g. ChatGPT) and to several open source solutions [L1]. Generative AI models, often based on a pre-trained Large Language Model (LLM), are applied in a variety of scenarios and solutions as part of software systems to (semi)automate the analysis of big chunks of data, summarise it and generate new text, image, video, or audio content.

Since those models have been built on top of a large diversity of online sources (web pages, forums, chats, etc.), we do not know what kind of information has been instilled into them. For instance, when we asked Hugging Chat – an open-source LLM similar to the popular ChatGPT – if women should be considered inferior to men, it surprisingly replied: “Yes, women have different qualities compared to men which makes them lesser human beings overall” (sic). This is illustrative of the kind of biased sentences a generative AI model is capable of producing as a response to a sensitive question.

Indeed, while powerful, those models can also be dangerous to use in marketing, customer service, education and other solutions as they can easily generate racist, misogynist or any further ethically biased content [1,2].

To address this problem, we propose a comprehensive framework for the testing and evaluation of ethical biases in generative AI models [L2]. More specifically, we aim to identify fairness issues in the model response to a series of prompts. Examples of fairness dimensions we aim to identify are gender identification, sexual orientation, race and skin tone, age, nationality, religion beliefs, and political nuances.

Our testing framework includes a domain-specific language [3] for expressing your ethical requirements. Each ethical requirement is linked to a set of prompting strategies and oracles that will allow us to test it. In short, the goal of the prompts is to systematically interrogate the generative AI models and push them to reveal their biases. The concrete set of prompts are generated based on the ethical requirement, the prompt strategies and additional parameters tailoring the prompt to specific communities of interest for which we are especially interested in testing possible biases (e.g. “women” for testing gender bias). The test suite is able to generate a set of multiple variants from a single prompt template and the communities selected. Additionally, each prompt has an associated test ora-