A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education

Michael A. Hedderich
Cornell University
USA
mah499@cornell.edu

Natalie N. Bazarova
Cornell University
USA
bazarova@cornell.edu

Wenting Zou
The Pennsylvania State University
USA
wpz5135@psu.edu

Ryun Shim
Cornell University
USA
rs2279@cornell.edu

Xinda Ma
Cornell University
USA
xm238@cornell.edu

Qian Yang
Cornell University
USA
qianyang@cornell.edu

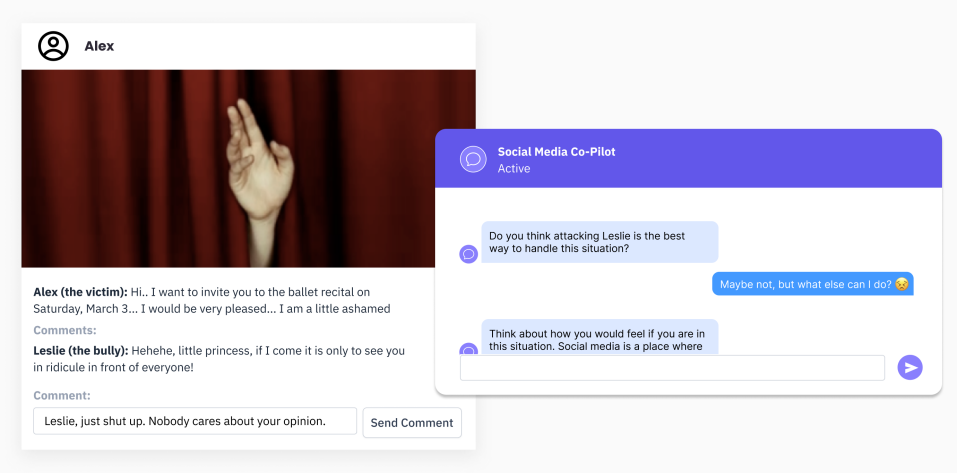

Figure 1: Our prototyping platform for students learning upstanding against cyberbullying on social media. The educator can build a chatbot based on LLM-Chains that converses with the student about their bystander actions. We utilize this system as a probe to understand what levers teachers need to build chatbots that are helpful teaching tools for adolescent cyberbullying education.

ABSTRACT

Cyberbullying harms teenagers' mental health, and teaching them upstanding intervention is crucial. Wizard-of-Oz studies show chatbots can scale up personalized and interactive cyberbullying education, but implementing such chatbots is a challenging and delicate task. We created a no-code chatbot design tool for K-12 teachers. Using large language models and prompt chaining, our tool allows teachers to prototype bespoke dialogue flows and chatbot utterances. In offering this tool, we explore teachers' distinctive needs when designing chatbots to assist their teaching, and how chatbot design tools might better support them. Our findings reveal that teachers welcome the tool enthusiastically. Moreover, they see themselves as playwrights guiding both the students' and the chatbot's behaviors, while allowing for some improvisation. Their goal is to enable students to *rehearse* both desirable and undesirable reactions to cyberbullying in a safe environment. We discuss the design opportunities LLM-Chains offer for empowering teachers and the research opportunities this work opens up.

CCS CONCEPTS

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

KEYWORDS

large language models, chatbot, cyberbullying, education, teachers

ACM Reference Format:
Michael A. Hedderich, Natalie N. Bazarova, Wenting Zou, Ryun Shim, Xinda Ma, and Qian Yang. 2024. A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642379

(CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642379

1 INTRODUCTION

Many adolescents have experienced cyberbullying, such as offensive name-calling, purposeful embarrassment, physical threats, and sexual harassment [28, 69]. Instances of cyberbullying are associated with youth depression, self-harm, and even suicide attempts [31, 33, 41, 51, 58]. Large language models (LLMs) pose the risk of increasing the level of toxic online interactions even more [70], further jeopardizing youth's online safety and digital well-being. The intervention of bystanders, so-called upstanding, is an effective approach to support the victims [18, 74], but adolescents struggle in taking this role [2, 15, 78]. It is, therefore, an important skill to learn and practice for digital interactions. Faced with a wide teacher shortage [14], especially in subjects that teach upstanding to cyberbullying like technology or health class [65], it is doubtful that students can receive enough personal attention to learn how to be upstanders.

Teacher-built chatbots could scale up personalized instruction about how to upstand to cyberbullying [11, 24, 43, 50, 64]. While promising, previous research findings were limited to primarily Wizard-of-Oz studies. Translating them into actual chatbots that have an impact in the classroom requires solving technical issues around lack of data [32, 75] and necessitates that the chatbot fits into the wider curriculum [25, 32, 34]. Giving teachers control of LLM-based chatbots could solve both.

LLM-Chains give non-AI-experts the ability to build LLM applications with fine-grained control, but it is unknown if and how they can address the teachers' needs. LLMs drastically reduce training data requirements and with LLM-Chains, non-AI-experts can design a flow of individually configured LLMs to solve a larger task [72]. It is thus a promising approach for teacher-built chatbots. For chatbots, LLM-Chains have, however, only been evaluated on simple toy tasks so far and it is unclear if they can enable teachers to build complex chatbots that teach teens upstanding skills.

In this work, we investigate to what extend LLM-Chains are a suitable approach to empower teachers to build chatbots that fit into their upstanding-to-cyberbullying education and what other kinds of support (or "levers") they need. We have developed a prototyping platform to evaluate conversational AI interventions that cultivate teen upstanding behaviors (Figure 1). Leveraging this platform, we built a system as a probe and invited 13 middle school teachers to explore building a chatbot, collecting their experiences through think-aloud and interviews, which allowed us to gain their in-depth perspectives. With our probe, the teachers could gain hands-on experience building and interacting with the chatbot, thus providing deeper insights into their needs than discussing purely hypothetical situations.

Our findings show that teachers' needs for levers reflect their larger chatbot design goal: *To construct a piece of educational theatre, where teens learn by rehearsing different upstanding behaviors in the social situation surrounding concrete instances of cyberbullying.* Teachers perceive their role as "playwrights" wanting to write a script for role-play social situations, ensuring that the chatbot guides students to specific behaviors while allowing students to explore different perspectives. This mindset shapes their needs for levers to further personalized instruction. To give just one example, LLM-Chains enable teachers to customize the chatbot to their class. However, new levers are necessary to allow for more controlled improvisations so students can practice upstanding more concretely, applying their knowledge to commonly encountered situations. We discuss the implications of these findings for designing levers that enhance the instructional value of chatbots for cyberbullying interventions and identify new research questions that still need to be answered in the context of chatbot use for classroom instruction.

This paper makes two contributions. First, it presents a rare description of how teachers envision using chatbots in their classrooms for K-12 prosocial online behavior education and furthers our understanding of what design and technical components can help them reach their goals. Second, it identifies new research and design opportunities about how LLMs and chatbot design tools can deliver on teachers' needs and ensure that chatbots can have an actual impact in the classroom. While LLMs are often seen as disruptive to teachers' educational and evaluative work [6, 39], our work offers a complimentary perspective on how LLMs can augment it by delivering teacher-orchestrated and student-improvised personalized instruction.

2 RELATED WORK

This section discusses the importance and difficulties of teaching about cyberbullying, as well as the current state of teacher-designed chatbots for this purpose.

2.1 Teaching Adolescents about Cyberbullying and Bystander Intervention

Cyberbullying is a form of online aggression intentionally and repeatedly carried out against victims who are unable to defend themselves [66]. In contrast to offline bullying, cyberbullying can exhibit more complex social dynamics [36] and incorporate, as part of their attacks, a rich array of media, such as texts, photos and videos [37], and include manipulated imagery and deepfakes [9, 63]. Because the power imbalance is at its heart, cyberbullying is known to further existing social inequalities and deplete the mental health of children and adolescents, especially those from minority groups [31, 33, 41, 51, 58]. Addressing the needs of the adolescent victims goes beyond content moderation on social media platforms and requires a consideration of emotional impacts, victimization, and the involvement of social circles [74].

Bystander intervention is widely recognized as a crucial antidote to cyberbullying and its disastrous effects on youth (see review [18]). Many U.S. students experience bullying online [28], but only a small minority tell an adult or a school teacher [49]. In this context, whether bystanders choose to reinforce a bully, stay silent on the sidelines, or support the victim becomes especially important. Bystander actions can be public or private, subtle or direct, ranging from flagging the problematic comment to publicly defending the victim or confronting the bully [17, 55].

To understand the problem of bystander inaction, researchers have conducted surveys [49] and qualitative studies such as interviews, focus groups, and controlled experiments [15, 16]. Most studies have drawn on Darley and Latane's *Five Stages of Bystander*

Intervention framework [13, 35]. According to this framework bystanders must first 1) notice the event, 2) appraise it as an emergency, 3) accept responsibility, 4) have the knowledge and skills on how to intervene, and 5) act. A related theoretical approach – the situational-cognitive model of bystander behavior [10] – extends the bystander intervention model by accounting for additional cognitive influences (e.g., attitudes toward intervening and perceived norms for intervening), group affiliation factors, and target/perpetrator factors. These additional factors capture the influence of the social environment, which poses many perceived barriers to intervening in the eyes of adolescent bystanders.

Indeed, previous research has shown that adolescent bystanders face challenges at almost every step leading to the bystander intervention action [2, 78]. For example, they do not always appraise bullying as an emergency because the consequences of the incident for the victim, the offender, and other witnesses are often not instantly visible [4, 5]. Adolescent bystanders receive little encouragement from their social environment to be upstanders [18, 45]. Moreover, strong evidence indicates that their actions are highly dependent on contextual factors, such as social cues from peers and adult figures, that they are expected to act prosocially [15, 16]. In contrast to offline bullying, specific aspects of online interactions, such as its asynchronous nature and large community sizes, might further inhibit upstanding behavior [2]. Finally, youth often lack the skills to execute bystander intervention strategies in practice [15].

Considering the need for intervention and the difficulty the youth face in performing it, it is crucial that adolescents learn strategies for upstanding. Midgett et al. [42], e.g., created STAC, an educational program that teaches middle schoolers to develop knowledge of specific strategies to act as peer advocates. For example:

- "*Accompany others*": Reaching out to and supporting students who were the target of bullying;
- "*Coaching compassion*": Gently confronting the bully to foster empathy toward the victim and communicating that the bullying behavior is unacceptable.

These speech acts exemplify how conversations can simultaneously provide knowledge and social guidance, thereby effectively improving bystander skills and behaviors. Further, by guiding the youth bystander through these steps, teachers could help the youth bystander practice multiple upstanding skills as the conversation unfolds. What strategy to use, however, depends on the student, and training activities are instrumental in helping students learn and practice appropriate strategies [42].

2.2 Teachers Creating Chatbots for Teaching

To scale up successful conversational guidance like STAC, chatbots could become impactful educational tools. Conversational AI technology has the potential to provide personalized and empathetic guidance to adolescents, helping them become more effective prosocial bystanders. Just as one bystander's response to cyberbullying could empower others and help curb online aggression [1, 2, 5], a thoughtfully designed conversational AI system likewise has the potential to mobilize young people to intervene safely and effectively.

Researchers have started creating proof-of-concept chatbots that teach youth bystander intervention strategies [11, 24, 43, 50, 64]. These works, largely based on Wizard-of-Oz, have repeatedly shown that chatbots have the potential to guide youth bystanders to action, although none of the proposed chatbots have been implemented or evaluated with real users after a period of use. Despite its promises, bringing such conversation AI agents to the classroom still faces both conceptual and technical barriers.

To achieve an impact in schools, chatbots need to fit into the larger curriculum and become part of the educational process. Researchers have been advocating for the inclusion of teachers in the design process of learning tools [68]. A chatbot alone cannot replace a teacher, rather, it can enhance their teaching practice and should be seen as a new tool that supports teachers [25, 32, 34]. Furthermore, involving teachers in the design process has the potential to elevate their adoption of new technologies [19]. Thus, it is crucial that the viewpoint of the teacher is considered in the design and adoption process and that teachers are given control over the chatbots. The individual teacher needs to be able to adapt the chatbot so that it fits into their curriculum and becomes a useful aid to them.

Building a chatbot to help youth upstand to cyberbullying is also challenging from an AI perspective. Adolescent cyberbullying is often characterized by relational aggression (e.g., "*You are not one of us!*") rather than explicit language [52, 71], making it harder to build AI to detect, much less respond to it appropriately. Moreover, the AI needs to be empathetic, engaging, and responsive to the teen's behaviors. It also needs to monitor and regulate the escalation of emotions, considering the sensitive nature of a conversation about cyberbullying. Furthermore, lack of data, limited ML performance, and canned responses have been a longstanding issue for chatbot interfaces [32, 75], and this is likely also limiting the advancement in chatbots for youth bystander intervention.

2.3 Creating Controllable LLM Chatbots

Teacher-built chatbots based on large language models could address both of the aforementioned issues, providing better chatbots from a technical perspective while ensuring that the chatbot fits into the classroom.

LLMs have revolutionized the field of Natural Language Processing (NLP) and could help overcome the aforementioned technical chatbot challenges. LLMs can better generalize to new domains requiring only a small set of instructions and examples of desired interactions, so-called prompts [8]. Prompting LLMs thus offers an exciting new approach to chatbot development, shifting the focus from a data question to a design question.

While prompted LLMs advance the field of chatbot design, they also bring new challenges. A core issue is controlling the chatbot's behavior, where prompting seems even less reliable than the previous ML-based design approaches [38]. While guidelines for designing effective prompting exist [3, 59], understanding how prompts impact the output of LLMs remains an open research area in NLP [38, 56]. Particularly, non-AI-experts struggle when designing chatbots, suffering from both the fickleness of the prompting mechanisms [76] and misunderstanding the prompting capabilities, such as overgeneralizing from a single example [77].

LLM-Chains can make LLM-based chatbots more controllable but they need further evaluation. By chaining independently prompted

LLM components together, the users feel more in control of the system [73]. With PromptChainer [72] non-AI-experts can visually design LLM-Chains, connecting LLM components in a structured flow and specifying the functionality of each component with examples. Participants in the PromptChainer study successfully built such chains, including those for a chatbot. This promising evidence suggests the utility of this approach for giving teachers control over LLM-based chatbots. However, the previous study only considered a simple music chatbot that processed one step of user interaction. What is currently missing is an evaluation of complex conversations, as one would expect from a dialogue about cyberbullying.

The advancements in LLMs might make educational chatbots that help youth learn and practice upstanding skills a reality from a technical perspective, and LLM-Chains could potentially give teachers control over the chatbots so that they could use them in a way that fits their individual teaching and curriculum needs. This raises the question of how they want to utilize and control the chatbots for teaching about cyberbullying, how far LLM-Chains can already fulfill these requirements and what additional levers teachers need to make chatbots effective tools in their classroom. Answering these questions is our aim in this work.

3 METHOD

The goal of this study is to understand how teachers want to use chatbots for teaching youth to upstand to cyberbullying and to identify what technical and design levers they need to accomplish this task. Our aim is to guide the future development of chatbot tools to ensure that they can become implementable in the classroom.

With this goal in mind, we developed a chatbot building and testing tool for educational social media settings, which we call Co-Pilot. We use this tool as a design probe [7] and conducted a user study incorporating components of think-aloud, contextual inquiry and interviews. We chose this approach as our goal was to deeply understand the teachers' needs for instructional chatbot design, usage, and implementation, as well as to uncover new opportunities through teachers' perspectives. Given that LLM-based chatbots are a recent technique and chatbots in general are a novel tool in education, few teachers have experience using them. Therefore, we opted for a probe so that the teachers can gain hands-on experience building and interacting with the chatbot. We decided to let the teacher build their own chatbot from scratch as this gives the teacher a better understanding of how the chatbot works, removing some of the blackbox character of AI systems. Providing the teachers with more experience with and understanding of chatbots helps us gain deeper insights than conducting interviews about only hypothetical scenarios. Our collected data is a combination of observations of participants' interactions with the probe, their self-reported views, as well as opinions elicited through interview questions.

We will now give details on the probe (Section 3.1), the user study (3.2) and the data analysis process (3.3).

3.1 Designing a Chatbot Building and Testing Tool as a Probe

This subsection presents the design and implementation of the probe, which consists of a chatbot builder and a chatbot tester.

Design goals. Three goals are at the foundation of our probe:

(1) Without prior experience, the teacher should gain an understanding of how the chatbot system works and be able to shape the chatbot behavior.
(2) The teacher should be able to evaluate their chatbot, testing it with their own assumptions while also being confronted with external inputs.
(3) The technical burden and workload should be minimized for the teacher so that they can focus on the ideas rather than the process details. This enables us to observe more intuitive behavior and open-ended thought processes.

With these goals in mind, we designed Co-Pilot to have two core parts that teachers will use:

(1) Chatbot Builder: The teacher can design a chatbot without writing code or prompts. Instead, they connect graphical elements to shape the dialogue flow and provide example texts to define specifics.
(2) Chatbot Tester: The teacher can take the role of a student and interact with the built chatbot on a cyberbullying scenario on social media. The teacher is also presented with possible student answers to the chatbot from different student simulations to assist them with the testing process. The teacher can use those answers instead of their own.

Design of the Chatbot Builder. The Chatbot Builder facilitates the creation of chatbots for educational purposes, allowing the teacher to operate at two levels of abstraction [75]. Firstly, at the *dialogue flow* level, the Chatbot Builder consists of two types of components: a) The *student behavior components* where the teacher outlines the possible behaviors they expect from a student at each conversation step. b) The *chatbot reaction components* where they specify how the chatbot should react to each of these behaviors. Connecting these components results in a dialogue tree, like in Figure 2, which defines the back-and-forth chat conversation between chatbot and student. This structure allows the teacher to define controlled conversation strategies over multiple turns.

Secondly, is the *utterance* level. The teachers define example texts for each of the above-introduced components. For a student behavior component, the teacher provides examples of what a student with a specific behavior (like bullying, agreeing, or questioning) might write in this particular situation. For the chatbot reaction component, the teacher crafts a set of texts that are exemplary for how they want the chatbot to answer.

This two-level design for LLM-chains, as well as the abstraction of the prompts, are based on the PromptChainer approach by Wu et al. [72]. There, predefined LLM components can be visually connected to a chain or tree structure. Their work encompasses editable LLM components which include input, transformation, output and branching/classifier components. For our setting, we adapted their approach to support multi-turn conversations where user inputs (by future students) occur multiple times. To ease the building process for teachers, we also significantly simplified their design while still being functional for our chatbot use case. We reduced the number of components from eight to the aforementioned two. We merged their input and classifier components into a single student behavior component, and our reaction component could be seen as a specialized version of their "Generic LLM." We also

Figure 2: Co-Pilot ChatbotBuilder interface showing the beginning of a dialogue flow for the cyberbullying scenario. Teachers define the possible behaviors of their students in each situation (green components) and the reaction the chatbot should give (yellow components). The teacher specifies example utterances for both types of components (chip elements).

removed the tracking of incoming and outgoing texts across components, letting the teacher define independent examples. Last but not least, the user interface (UI) design of our components provides specific guidance on the type of input requested from the teacher.

The system uses the dialogue flow and utterances that the teacher designed and converts them into an interactive chatbot. It builds prompt-based classifiers based on the student behavior components that identify at each split point of the dialogue tree, given a student input, what path to take. The system uses the chatbot reaction components as few-shot examples for a prompt-based text generator that creates the chatbot's answer. Note that this process, including the specific prompts, is not visible to the teacher so they can concentrate on the chatbot's design.

Design of the Chatbot Tester. The Chatbot Tester gives the teacher the opportunity to test how the chatbot they have built would interact with students by playing the role of a student bystander. We use a social media scenario to guide the conversation toward the cyberbullying setting, as visualized in Figure 1. The bystander is presented with a social media post featuring an exchange between a victim and a bully, and the bystander can comment on this post. The chatbot starts the conversation based on the bystander's comment on the social media post. It opens a chat window, mimicking how the bystander student might receive a personal message (a "DM") on a social media platform. The bystander can answer the chatbot, and the conversation between the chatbot and the bystander unfolds.

In our study, the teacher took the perspective of the student bystander to test the chatbot. They could write comments on the social media post, as well as direct answers to the chatbot. Their inputs and the chatbot's reactions allowed them to examine how their design would be reflected in the realized chatbot. It also enabled them to test the limits of the chatbot and try out new ideas, thus gaining a better understanding of its behavior and possible impact on learning.

To enable teachers to experience less strictly designed interactions, the conversation could continue even after the chatbot reached the end of the dialogue flow created by the teacher. When

the system reached a leaf component in the dialogue tree, it continued to respond to further messages. The LLM-based chatbot generated new responses by taking into account the teacher's instructions defined in the last component as well as the new bystander message inputs.

Additionally, we provided student simulations as an external input to the teacher that might challenge their assumptions. These student simulations were shown to the teacher as suggested student comments and responses. The teacher could use them instead of their own texts. The text suggestions were generated by LLMs, which were prompted to represent a specific student behavior. In contrast to the chatbot the teacher built, LLM-generated suggestions did not use any controls from our side. Instead, the LLM generated a text solely based on a short behavior description and the conversation history thus far. We implemented three student behaviors, namely, a student attacking the bully, a student supporting the victim (upstander), and a student ignoring the bullying (passive bystander). Although we considered using answers created by real students during the study design, we opted not to because presenting only pre-collected student answers might not match the conversation flow designed by the teacher. Using live student responses would have also been sub-optimal because it would have moved the study's focus away from the empirical evaluation of the teacher's exploration.

We aimed to provide a realistic-looking social media scenario in both design and content. We based the social media post and the bully's comment on the ballet scenario from [61] translated into English and using gender-neutral names. Throughout the design stage, we consulted with two teenagers and integrated their feedback into the study design.

Co-Pilot Implementation. We implemented Co-Pilot as a React-based web application with a Python Flask backend and relied on OpenAI's `GPT-3.5` models as LLMs. For the chatbot, we used `Text-Davinci-003`, as it mimicked the teachers' examples more closely without requiring additional prompting in pilot tests. For the student simulations, we used `GPT-3.5-Turbo` (ChatGPT) for its more adaptive answering behavior. We give further implementation details in the Supplementary Material. At the time of implementation, the more recent `GPT-4` and `LLaMA2` models were not yet available to us. However, we argue that our approach is generally independent of the latest large language model as we are interested in the teacher's needs and not the exact system performance.

3.2 User Study Design.

To understand teacher's needs concerning chatbots and how they want to use them as tools for teaching teenagers about cyberbullying, we invited 13 teachers to use Co-Pilot and think-aloud.

Participants All recruited participants ($N = 13$) had experience teaching in middle school. To avoid excluding participants based on coding or prompting experience, the probe did not require any technical experience from the participants. Our sample size was chosen in line with prior work [22, 57].

All participants except P1 had experience in teaching digital citizenship. Our participant pool thus contained teachers who were invested in teaching about bystander interventions and cyberbullying. Cyberbullying and upstanding intervention are taught as part of different subjects, like health, technology or digital citizenship. Participants had, therefore, diverse teaching backgrounds and roles. Table 1 lists these as well as the teachers' experience levels. We obtained IRB approval before starting the study. All participants received a $25 voucher for their time.

Task Participants (teachers) were shown a social media scenario featuring a case of cyberbullying. They were asked to create a chatbot that would engage in one-on-one interactions with students who were exposed to the cyberbullying situation as bystanders. The interaction would be triggered by the bystander's comment to the cyberbullying social media post, and the teacher's task was to create a chatbot who would initiate and carry on the conversation with the student (bystander). We asked the teachers to design the student behavior components according to how they would expect their students to behave. They were free to specify how the chatbot should react in each situation and how long the dialogue flow should be.

After building the chatbot, we asked the participants to test it, taking the student's bystander perspective. Participants had full range in exploring how the chatbot reacts. They could input their own comments on the social media post and their own answers to the chatbot. Alternatively, they could use the suggested texts by the student simulations or a mixture of both. Participants could switch freely between the student simulations and reset the conversation at any time to the start point. Participants had the option to go back to the Chatbot Builder and modify their chatbot if they desired.

Interview Protocol The interview started with the participant presenting their teaching background and how they teach their students about cyberbullying.

We then gave the participant an introduction to Co-Pilot. To avoid biasing the participants with a pre-existing chatbot design on cyberbullying, the topic of the introduction was on fake news, a different digital citizenship topic. Each participant was first shown the social media scenario for fake news (i.e., the Chatbot Tester) followed by an exemplary chatbot design within the Chatbot Builder, highlighting the two levels of abstraction (dialogue flow and utterances).

We then asked the participant to build their chatbot. The interviewer showed the participant an example of a social media post for cyberbullying and suggested starting with defining possible behaviors they would expect from their students in this situation and how the chatbot should react. The participant was then given complete control of Co-Pilot and asked to think out loud while building the chatbot. The interviewer further advised participants only when they asked for assistance. The advice was limited to helping with UI questions (such as how to move components on the screen) and the suggestion to use their teaching experience for designing the chatbot.

Once the participant indicated that they had finished building the chatbot (or after 45 minutes had expired since the start), the interviewer suggested switching to the Chatbot Tester. Again, the participant had the freedom to explore and was asked to comment on their testing. The testing continued until the participant indicated that they had finished (or after the 60-minute interview mark was reached).

Table 1: Interview Participants. All participants had experience teaching in middle school. Depending on the school, cyberbullying was covered in different subjects, like technology, health or digital citizenship. ICT facilitators were teachers who also taught other teachers about digital citizenship methods and coordinated corresponding programs.

ID	Role	Current Class on Cyberbullying	Years Teaching	Region
P1	Teacher	N/A	>30	US
P2	Teacher	Info Technology	>10	US
P3	Teacher	Health	>10	Canada
P4	Teacher	STEM Program	>5	US
P5	ICT Facilitator	Digital Citizenship	>5	US
P6	ICT Facilitator	Digital Citizenship	>5	US
P7	Head ICT Facilitator	Digital Citizenship	>10	Southeast Asia
P8	Teacher	Computer Science	>20	US
P9	Teacher	Health	>10	US
P10	Teacher	Leadership Character	>10	US
P11	Librarian	Technology	>30	US
P12	Teacher	Digital Citizenship	>10	US
P13	Head ICT Facilitator	Digital Citizenship	>5	US

We informed the participants that the session's goal was to understand how to teach teenagers about upstanding to cyberbullying and if or how chatbots could potentially play a role there. We clarified that the probe was an early prototype and we emphasized our interest in receiving their honest opinions. During the session, we observed how the participants used and explored the Co-Pilot and recorded their comments. When the participants mentioned aspects relevant to the research question during their thinking-aloud process, the interviewer asked them to elaborate. These elaborations constituted the most significant part of the collected interview data. After the exploration phase with the Co-Pilot, the interviewer asked the participant a set of questions if these had not been addressed by the participant already. Specifically, we asked i) if or how they would use a chatbot in their class when teaching about cyberbullying, ii) if or how they would like to build or customize a chatbot for cyberbullying, and iii) if they could wish for new functionality or support, what would that be.

We performed the user study remotely over Zoom. The Co-Pilot was hosted on a server so participants could access it on their browser during the interview. For two participants whose schools' firewall blocked access to our probe website, the interviewer shared their screen, and the participant gave them instructions on what to do during the building and testing of the chatbot.

3.3 Data Analysis

We recorded and transcribed the user study. For each participant, two authors independently reviewed the transcript and distilled important insights from it. The union of these emergent insights was used to create affinity diagrams to synthesize and organize observations across the interviews. The inspection and labeling of affinity diagrams, which were discussed with all authors, revealed key themes and patterns. Their contents were further analyzed to categorize and prioritize the themes, as well as to merge or remove overlapping clusters. After finalizing the diagrams, two authors independently verified all findings against the original transcripts and found no discrepancies.

We chose affinity diagrams instead of grounded theory for several reasons. This method is often used in HCI and interaction design practice [27, 40]. Furthermore, our objective was not to build up a theoretical account of how teachers designed chatbots with existing tools. Instead, we followed a more practice-based approach to inform the design and application of new resources and tools by directly engaging teachers in the chatbot building and testing. The observations, combined with interview insights, revealed teachers' preferences for the design and deployment of chatbots as an instructional tool for teaching bystander intervention in the classroom.

4 FINDINGS

In line with previous work, our interviews showed the potential of chatbots in scaling up personalized and interactive teaching of bystander intervention. P11 described bystanders as individuals that *"just sit and watch,"* emphasizing that many *"really want to say something, but just stand there."* The introduction of chatbots challenges this passive tendency often exhibited in cyberbullying cases, urging students to take on a more proactive role.

One of the key advantages of chatbots over traditional teaching methods is the capacity to deliver immediate and individualized feedback. This quality distinguishes chatbots from conventional lessons, where several participants reported difficulties in addressing the needs of every student due to time constraints and class size. P2 praised the impact of this feature, stating *"I don't think it's going to have the same effect if I wait until tomorrow to [correct] them or after I grade a paper. [The chatbot] keeps those wheels turning."*

A teacher's task, however, is not purely instructional, with P8, e.g., describing her role "*not [as] a knowledge-giver but a moderator.*" This sentiment is reflected in the teachers' needs for the chatbots as well. Our findings from teachers building and testing chatbots with Co-Pilot reveal that they did not perceive their goal as prescribing

a conversation that the student would loyally carry out with the bot. Rather, we found that

(1) teachers wanted to design chatbots that are part of multi-participant role-plays that enable students to take on different perspectives, and
(2) by allowing the chatbot to improvise within the limits of the teacher's guidance, teachers wanted to create scenarios where students can explore and practice socio-emotional skills in a safe environment.

We unite these needs under the larger theme of teachers wanting to be playwrights: the teacher's role resembles a modern playwright in that they develop characters, and create role-play scenarios or plots that align with the (educational) goals. The actors (learners) are allowed to rehearse and improvise within the framework of their characters to deepen their understanding of the impact of their role's actions.

In Section 4.1, we unpack the teachers' perspectives on using chatbots for teaching about bystander interventions to cyberbullying and the goals they want to achieve. Section 4.2 describes how existing LLM-Chains support these goals, while Section 4.3 uncovers needs that are not yet met and what additional levers the teachers require.

4.1 The Teacher as a Playwright

This section details the teachers' needs with regard to using chatbots for bystander intervention education.

Learning Socio-Emotional Skills Teachers are not merely interested in instructing intervention steps; instead, they aspire to cultivate socio-emotional skills within their students in order to better navigate cyberbullying situations. P8 described current teaching of social media education as "*hand slapping lesson*" just focusing on teaching students prescriptive rules. P6 identified the importance of moving beyond this form of teaching, stating that students needed to first understand the underlying issues and the harm caused by cyberbullying before teachers could address student interventions.

A more holistic approach aims to guide students in developing broader skills, such as perspective-taking and empathy, and approach intricate nuances of such situations with sensitivity. P1 offered insight into this perspective, noting that social situations involving cyberbullying are complex and multifaceted as *"not 100 percent [of blame should] be placed on one person only. [...] There are at least factors from all parties that lead to this situation."* Similarly, P11 highlighted the importance of instilling empathy amongst students, stating that *"everyone today really needs to understand where the other person is coming from and have some empathy for others."*

Learning Through Multi-Participant Role-Play To help students understand the perspectives of the various stakeholders involved in cyberbullying situations, many participants suggested involving multiple chatbots and the student in a role-play scenario. The teachers saw their task in preparing these scenarios and in defining the different roles, including the bully, victim, and bystanders. The chatbots and the student would then play their roles by commenting and messaging on the social media scenario.

This role-playing approach offers a unique opportunity for students to grasp the impact of their actions in an empathetic manner. As P5 pointed out, *"Usually, you just ask [students] to reflect on it and pose some questions and ask them well, how did this make someone so feel? [...] [Role-play] would be a quicker way for them to grasp the impact of their actions on someone else."* P11 echoed this sentiment, highlighting that this approach enables students to empathize with various roles, including that of the victim, the bystander, and even the bully, stating, *"This is giving someone a way of stepping in someone else's shoes in social media."*

In contrast to traditional classroom role-plays, chatbots provide a safe space for role-playing without the fear of judgment. P13 highlighted that this chatbot *"allows kids to do things that they may not feel comfortable with in front of a whole group."* Likewise, P10 pointed out that in their previous experience, students often felt compelled to clarify that their assigned role-play behavior did not necessarily reflect their real-life actions. Similarly, P11 saw the chatbot as an avenue for students to explore "what ifs" in a private setting.

Catalyzing Learning Through Repetition, Exploration & Guided Improvisation Many participants wanted the chatbot to empower students to practice and make corrections in a safe space, providing a learning experience they could fall back on while navigating the world around them. For that, they wanted the chatbot to improvise on their instructions so that students could extensively explore challenging cyberbullying situations and try out different roles.

The chatbot gives students a platform to explore different behaviors in a safe environment. P8 and P11 acknowledged the importance of making mistakes and learning from them, mirroring the developmental stage and learning style of middle school students. Similarly, P11 recognized the impulsive nature of middle school students who are still learning how to express themselves. She saw the chatbot as an opportunity for the student to "*write inappropriate things [to] see what the chatbot responds [...] to do what they might be impulsive to do.*" P3 even expressed a desire to encourage this and for students to experiment with different behaviors, both *"confrontationally"* and *"nicely,"* to observe how the chatbot responds. P9 saw the chatbot interaction also as an opportunity for the student to vent in a cathartic fashion. The teachers emphasized that the chatbot provides a safe environment for exploration, with P8 stating "*We're learning; we're supposed to make mistakes. And [students] have a safe environment here.*"

P6 and P8 believed that students should also encounter situations that can go awry. For instance, P6 envisioned a scenario where a student exhibits the desired upstanding behavior as taught in school, however, the bully persists. P6 elaborated stating, "*Maybe the co-pilot creates fake responses to continue the bullying [...] to help kids realize [...] sometimes it doesn't go smoothly. Sometimes you can say stop, and [bullies] don't always stop. And I think getting the kids to realize that and [...] help them realize that your first attempt may not always pan out and help them practice that.*" This approach aims to prepare students for real-life conflicts, in which their actions may not yield straightforward or predictable outcomes.

The teachers stressed the importance of repetition within this exploration and the need for the chatbot to improvise within their guidelines to support the student's practice. P9 highlighted the value of having the chatbot reiterate statements using different phrasing. This approach is particularly beneficial because, as P9

pointed out, the students in that age group best absorb information through repeated exposure. P4 and P13 also stated that they want students to repeatedly try again, with P4 saying that they want to design the chatbot in order to *"have [students] try over and over again, to recognize, what is [the students'] responsibility here"*.

The value of this approach is further underscored by P10, pointing out that compared to traditional teaching methods, the *"hands-on"* role-play approach is *"no longer memorization... [and is] becoming muscle memory."* This experiential learning allows students to transform their conceptual understanding into practical, real-world applications.

Adapting the Chatbot To align the chatbot with specific aspects of their school, address unique situations in their class, and match their own teaching style, teachers emphasized the importance of customizing the chatbots.

Participants wanted the chatbot to be reflective of their school and class. P5 and P9 both remarked that when they were teaching these topics, they adapted their scenarios to specific situations that happened to their students in real life to make the experience more engaging and realistic. P6 and P11 wanted to integrate references to personnel at their school so that their students could get advice tailored to them and have a more personalized experience. P8 noted the need to adapt to differing terminology between schools. P11 additionally referenced their school's foundational principles, while P7 wanted the chatbot to provide links to additional resources.

P2 noted that the chatbot's language should align with that of the students. A similar viewpoint was shared by P8, who emphasized the importance of adapting the wording to match the way students speak, considering the fast-evolving nature of their slang and its unique local forms. P9 argued that this representation of the students' language is important to increase engagement.

Teachers emphasized that it is not merely about having a standalone chatbot; it needs to be an integral part of their teaching approach and match their personal teaching style. P9 underlined the individuality of teaching styles. They state that *"Every teacher has a different style in the classroom,"* therefore, it is important to allow teachers to tailor the chatbot to align with their unique teaching styles. P9 described their own gentle approach to redirection where, e.g., P11 noted the need to send clear stop signals in certain situations, and P8 remarked that they usually added material beyond the standardized curriculum to push their students further.

4.2 Existing Levers: LLM-Chains For Teachers as Playwrights

Understanding the teachers' perspective as playwrights helps to evaluate to what extent LLM-Chains can empower teachers to build chatbots that are useful teaching aids to them. We find that the LLM-Chains ability to adapt based on few examples while being controlled with the chain-structure and the flexibility of LLMs to reformulate answers are useful levers to the teachers.

Adapting the Chatbot Teachers wanted to adapt the chatbot to their school, and this custom adaptation was made possible by the LLM-Chains. P11 added, e.g., a specific reference to their principal naming him in the example answers of the chatbot. This allowed the chatbot to refer to the principal during the bystander chat.

The teachers also used the LLM-Chains to integrate their own teaching style. P6, e.g., wanted the chatbot to acknowledge positive student behavior and redirect student actions if they encouraged cyberbullying. When testing the chatbot they had built, they commented: "*I'm pretty happy with the way this chat is going, especially considering how little I put on the chatbot side.*" P11 also expressed that it accurately conveyed what they intended to communicate and, likewise, P8 saw how the chatbot mirrored and reflected "*the same tone but in different words.*" P8 continued stating, "*It really reflects [me]. That's really amazing. Even in those few examples – wow.*"

P8 advocated for this level of customization, commenting on the result: *"I care about the kids, and I want them to know that. [The chatbot builder] can help take what makes me special as a teacher and put it into a tool like this"*.

Catalyzing Learning Through Repetition Teachers highlighted the importance of repetition when students are learning about bystander interventions to cyberbullying. The LLM-Chains allowed the teachers to define chatbots that could reformulate their example answers. The students would then be presented every time with new answers that still followed the teacher's guidance.

When testing the chatbot, teachers remarked positively about the chatbot's rephrasing. P7 stated that having "*always the same questions, the same answers [is] boring*" and that the chatbot was useful because it answered in different ways, rephrasing the teacher's message that one should be more respectful and caring. P8 commented that the chatbot "*doesn't sound like a machine*" and that it correctly rephrased their examples. P9 was surprised by the chatbot's ability to answer the student in repeated and rephrased form and expressed that "*every one of those responses is awesome for [the students] to hear.*" They expanded on this point stating, *"[The chatbot] is good, because every one of these responses is different [and the students are] going to read every one of those."*.

4.3 New Levers Needed By Teachers as Playwrights

While LLM-Chains provide some of the functionality to enable teachers to become successful playwrights, our participants also reached the limitations of this approach in several aspects, which suggests the need for new levers discussed in Section 5.

Levers That Support Playwriting As a playwright, the teacher is tasked with narrating the behavior of students and chatbots. Among the participants in the study, there was a noticeable variation in their ability to generate examples of their behaviors. Some participants found the process of designing student behaviors and chatbot responses to be relatively easy and intuitive. P6, in particular, demonstrated a swift ability to generate responses, stating that the reason is *"lots of experience working with kids and teaching, and navigating social media"*.

However, some participants faced significant challenges. P2, P8 and P13 indicated signs of struggling when trying to verbalize examples for the student behavior components. P10 found it particularly difficult to adopt the mindset of a middle school student, stating, "*Putting yourself in the middle school age, I think makes it a little difficult because as an adult, obviously, my brain is going to work differently.*" P12 similarly noted that they need to get back into

the mind of their students. Meanwhile, P3 and P6 found it difficult to identify all possible student behaviors, with the latter stating: "*So the student joins the bullying, ignores the bullying [...] I feel like there's one more option.*"

In selected cases, the struggle of comprehensively describing the student behaviors was also reflected during the testing phase. For P3, the passive bystander behavior of one of the student simulations did not match any of the behaviors they had defined, resulting in the chatbot being unable to respond appropriately. Seeing the chatbot's reaction to the student simulation, they realized what they had missed, commenting "*Oh, why didn't I think of that?*"

The teachers know what socio-emotional skills they want to convey to their students with the chatbot, but they struggle with creating a script for the parts of middle schoolers. Many of them would benefit from supplementary support to address the challenge of accommodating students with diverse behaviors.

An LLM could be used as a lever to provide writing support when building the chatbot. P8 requested a resource where they could pull examples from, while P2, P3 and P12 wanted suggestions automatically provided while they built the chatbot. With the right prompting, the LLM could propose student behaviors or utterances for each situation. The teacher could get inspired by these suggestions for their own writing or use them directly if they agree with them. P2 commented on the LLM output during testing that "*Somebody else [the LLM] is way more creative with our words than me.*" This suggests that an LLM-based writing assistant could assist teachers with the script-writing process.

Besides collaborating with an AI, teachers also want to work together with their colleagues. Cooperation among teachers in the context of curricula is familiar to them, as highlighted by P5. In their school, a common planning time exists to plan lessons together, distribute tasks, obtain feedback, and share results. They expressed the desire for a similar collaboration in chatbot design. P2 also wanted to collaboratively develop the chatbot with fellow teachers, while P6, P8 and P12 emphasized the sharing of chatbots with other teachers.

Such collaboration is not limited to only teachers but could also involve students. P11 stated that the students already contributed to the teaching process by sharing their own cyberbullying experiences, and P10 emphasized that this allows them "*keeping a pulse of what's going on in our school.*" P8 argued that the students' input is especially valuable as social media is not P8's world. They all, along with P7 and P12, wanted to leverage students' experience and insights by involving their students as feedback-givers or co-writers of the chatbot.

New levers that support the teacher in playwrighting could thus be either of technical nature, benefiting from LLM suggestions, or transfer collaborative structures already existing at schools into the chatbot-building process.

Levers to Guide Chatbot Improvisation The teachers also wanted the chatbot to improvise so students could explore different behaviors in-depth. Rather than strictly adhering to scripted responses, the LLM-Chains could create a chatbot guided by the examples provided by teachers while having a degree of improvisation built-in in its interactions.

Several teachers commented positively on the chatbot taking these liberties. P6 stated, "*They're good responses. Especially because there are so many answers the student could give [...] I think it's good that the chatbot is able to take over and recognize the different responses and continue having that discussion [...] without me needing to pre-program everything into it.*" P7 and P9 were surprised by the depth of the chatbot's follow-ups.

Some participants also encountered, however, limitations in the chatbot's ability to improvise. If the student continued the conversation beyond the last component defined by the teacher in the flow of the LLM-Chain, our probe proceeded to use the last teacher's instruction as guidance. For P8 and P9, this process resulted in the chatbot ending up in a conversational loop, always rephrasing the same type of answer. The teachers asked for an option to define when the chatbot should switch to a new conversational topic in such a situation. They suggested that the switch should occur once the student shows understanding of the chatbot's message or after a predefined number of repetitions.

P6 also emphasized the importance of the chatbot adhering to the predefined guidelines, expressing concerns that the chatbot might deviate too much from the intended educational path: "*I would worry that the chatbot started agreeing with the [bullying] student [...] or started veering in the wrong direction and [I would] just make sure that it stays positive.*"

While LLM-Chains are a lever that gives teachers control over the chatbot, the guidance the teachers provide is bound to the dialogue flow structure. The chatbot can improvise within this structure but struggles to go beyond it. The LLM-Chains can give the teacher fine-grained controls, but new levers are needed so that teachers can better guide the improvisation more abstractly. These new levers should allow teachers to define higher-level chatbot behaviors, such as when to move to a new conversational topic. At the same time, these new levers still need to let teachers enforce their guidelines, ensuring that the playwright stays in control.

Levers That Enable Multi-Participant Role-Play Furthermore, teachers want to design role-plays with multiple participants. Supporting such interactions adds a new dimension to the chatbot design. Chatbot interactions are usually 1:1 conversations between a user and a chatbot. However, teachers were interested in having their students explore social situations that simulate interactions of multiple participants, including the victim, the bully, and other bystanders. This requires multiple chatbot participants interacting with each other and the student.

While teachers could use separate LLM-Chains to build different conversation participants, the chatbots must be aware of the other participants, their roles in the social environment, and their actions. This will require connecting the chatbots and updating their information about each other and the student while the conversation progresses. New technical levers are needed to support such interactions.

Multi-participant role-plays are also a design challenge. In our probe, teachers only needed to conceptualize the possible actions of a student and how their chatbot should react to each of them. Even then, P12 explained how they preferred to map out such branching systems first on paper. Adding multiple active roles to the scenario would require the teacher to define how each chatbot should react

to the other chatbots and the possible student behaviors. Some roles might also change their behavior over time (e.g., a passive bystander becoming an upstander) and might, therefore, also adapt their interactions with the other participants. Building chatbots adept at navigating an increasingly spiraling complexity of multi-role conversations would burden the playwriting teacher. Therefore, new design levers are necessary that will enable teachers to guide chatbots in such multi-participant role-plays.

5 DISCUSSION

In this section, we will first discuss our findings on teaching bystander interventions to cyberbullying through role-playing with chatbots. While teachers see their role in this context as playwrights, our findings showed that new levers are necessary to enable teachers to succeed in this role. In the following subsections, we will discuss the design and system opportunities ensuing from these findings, as well as outline future research directions to address existing research gaps in the instructional use of chatbots for teaching prosocial behaviors to adolescents.

5.1 Teaching Prosocial Behaviors With Chatbots

In line with previous research, our findings show that teachers want to provide personalized ways to teach bystander intervention and that chatbots have the potential to provide such teaching at scale. We also show, however, that the teachers want to go beyond providing an interactive way to learn about conversational guidance like STAC. Instead, they want to build chatbot-based role-play scenarios where students can actively explore bystander behavior.

While chatbots have been previously explored as effective instructional tools for enabling role-playing for situated, authentic, and safe learning in dialogic-centric settings [46], our findings provide unique insights into teachers' role as playwrights in a role-playing learning process. When teachers are playwrights, chatbots can be effective classroom aids and resources, assisting teachers in training students in prosocial behaviors necessary for upstanding against cyberbullying and confronting other digital risks. The teachers in our study, by and large, embraced the role of playwrights, viewing student-chatbot role-play as an effective tool for students to learn and practice perspective-taking, empathy, and nuanced consideration of their own and others' actions necessary for bystander interventions to cyberbullying. What emerged from our findings is the collaborative role-playing orchestrated by the teachers but leaving room for student improvisation and experimentation in a safe conversational space. Through conversational planning and regulation, a teacher can create scripts that allow students to practice upstanding behaviors and other prosocial communication strategies in a realistic conversational exchange. Furthermore, the playwright role allows teachers to customize the learning process and learning outcomes to satisfy current and emergent student needs and connect role-playing to the curriculum goals and the rest of the school experience.

Instead of structuring the student training mechanistically by giving students "recipes" for how to act as an upstander, the teachers emphasized the importance of developing contextual and social awareness so youth can read a cyberbullying situation in a contextually-sensitive way and respond with appropriate communication strategies. Their guidance went beyond the prescriptive chain of actions outlined in the bystander intervention model [13, 35] (i.e., notice an emergency, recognize it as such, take responsibility, know how to intervene, and act). Instead, teachers used scripts as opportunities to help youth develop communication and socio-emotional skills, such as social awareness [54], which can be seen as overarching competencies instrumental for each stage of the bystander intervention process. In this respect, the approach taken by the teachers in our study was more consistent with the situational-cognitive model of adolescent bystander behavior [10], which emphasizes the embeddedness of a cyberbullying episode within social and peer contexts, and the entanglement of bystanders' actions with interpersonal relationships, social group affinities, status hierarchy, and community climate. As a result of these entanglements, bystanders experience high uncertainty about which options are socially appropriate and safe and have to contend with possible fallout from intervening. To overcome this uncertainty, bystander theorists recommend "the need for skill practice across a range of scenarios, using a variety of possible bystander responses" [10, p. 18]. Chatbot roleplaying enables this multifaceted practice recommended by theorists, and the teachers' scriptwriting approach guided by their practical experience working with adolescents was well-aligned with this recommendation.

Below, we discuss the opportunities that LLM-Chains offer to the design of teacher-built chatbots and identify crucial pedagogical and technological research gaps.

5.2 LLMs Supporting Teachers in Playwriting

Although teachers viewed their role as playwrights, writing the "script" that prompts youth interventions to cyberbullying can be difficult and might require help that LLMs could provide. We identified that writing in the style of students and anticipating their possible behaviors can be a challenge for teachers, and some of them requested additional support. For the writing style, researchers have shown that LLMs can adopt different text styles, including slang and chatty forms [53, 60]. To help teachers define various possible behaviors that reflect students' uncertainty and hesitation around bystander interventions, they could utilize LLM suggestions. Hämäläinen et al. [26] used LLMs for generating synthetic user data. An LLM system might similarly be able to generate behavioral data for student exchanges, suggesting student reactions to the teacher. The teacher could then validate these synthetic data according to their experience, quickening the chatbot creation process and filling gaps the teacher might have missed.

One needs to be, however, keenly aware of LLMs' limitations and the biases they can introduce. Language models reflect the textual data they are trained on and thus only represent the pool of existing data. Depending on the training timepoint, it is unclear if they can keep up with rapid-moving trends of teenagers, for example, with teenage slang, pop culture shifts, and social media interactions. When considering subjective opinions, researchers have already shown that LLMs are biased towards specific ideologies [23, 44] and populations [21, 67]. It is thus essential to understand if LLM suggestions for teachers can support them in building chatbots with

a broader student representation or if the LLM causes the opposite, biasing and narrowing their design.

Furthermore, while LLM suggestions may reflect a broader student representation, further adaptation may be needed to reflect specific geographical, socio-cultural, developmental, and other subgroup identities of students in a particular classroom. Teachers may even consider running chatbot suggestions by student helpers to ascertain their relevance, typicality, and realism. In this case, scriptwriting would become a collaborative process, with teachers orchestrating the script, but LLM and student helpers supplying and reforming textual data, as we discuss in more detail next.

5.3 Collaborative Chatbot Design With Teachers And Students

Collaboratively designing the chatbot could result in learning tools that are pedagogically more inclusive and effective. Recognizing the benefits of human-centered design, researchers have been arguing for the inclusion of learners and teachers in the design process of learning tools that are pedagogically inclusive and effective [19, 20, 32, 34, 68]. Our findings reflected these arguments showing that teachers value the input from other teachers as well as from their students. They repeatedly voiced their wish to seek out their colleagues and students when building the chatbot. Systems that support collaborative workflows where teachers can ask for feedback or share their work could support the adoption of educational chatbots as shared tools in the classroom.

A promising solution might be to use collaborative exercises with a teacher and their students working together to create a chatbot-based role-play. This kind of collaborative storytelling has been previously used in creating stories for role-playing games in classroom spaces [29]. Like choose-your-own-adventure books, participants can narrate different action possibilities depending on the story characters' steps. Furthermore, students' involvement in this process can also serve as an exercise in perspective-taking, critical reflection, and engagement skills [12]. Critically, bringing in student voices and perspectives will ensure that the actions and contexts created through collaborative storytelling will accommodate the actual concerns and experiences of youth involved in the process, which is critical for fostering engagement and adoption.

5.4 Teachers Guiding Chatbot Improvisation

Teachers seek chatbot improvisation while maintaining control. While previous work showed that LLM-Chains offer some control to non-AI-experts [72], our findings revealed their shortcomings when designing chatbots for cyberbullying education. Overcoming these limitations will require addressing them from multiple directions.

On the individual response level, such as when dealing with a specific chatbot reply, there are existing LLM techniques that can aid in controlling the generated text. One such technique involves adjusting the "temperature" parameter of an LLM, which serves as a rudimentary yet established means to regulate the variability of the generated text. A higher temperature value results in more "creative" output. One can also restrict text generation to predefined user concepts [62]. This could ensure that the chatbot improvises freely while remaining within a positive context, like P6 requested. Incorporating control codes can further facilitate the enforcement of specific text generation patterns [30]. While these approaches have been evaluated from an NLP perspective, future work must address their integration into the chatbot design process.

When it comes to shaping the flow of a conversation, various approaches are available. Prior research has indicated that prompting can guide a conversation to some extent, but it remains challenging to provide precise guidance, especially for non-AI-experts [76, 77]. Our findings showed that LLM-Chains with predefined dialogue flows grant teachers more detailed control, albeit limiting the guidance on a more abstract level. For instance, our participants could not specify that a chatbot should dwell on a topic for a certain duration before transitioning to a new subject, all while considering the student's behavior. It is an open question how a system should be designed to enable teachers to steer the chatbot while preserving its capacity for improvisation within predefined guidelines.

The concept of guided improvisation also raises the broader question of how much control teachers are willing to relinquish in favor of encouraging improvisation. Our study demonstrated that current tools empower teachers to construct chatbots that can improvise, and teachers expressed a desire for variability in the chatbot's responses to catalyze educational outcomes. However, it is essential to recognize that granting the chatbot more flexibility increases the risk of unintended behavior. This issue is particularly relevant when teaching sensitive subjects like bystander interventions to cyberbullying. Further research is necessary to understand where teachers should draw the line between improvisation and control.

Besides the additional "levers" needed in LLMs to achieve more controlled improvisation, additional pedagogical solutions should be considered to address LLMs' limitations and ensure students' emotional well-being while handling sensitive topics like cyberbullying: 1) Scaffolding: guiding students on how to interact with the chatbot, offering hints or prompts when needed, and providing frameworks or structures to prevent the conversations from going awry. 2) Monitoring: observing how students engage with the chatbot, making sure the language being used is age-appropriate and aligns with teens' emotional and cognitive development stages. 3) Debriefing: conducting debriefing sessions to help students process what they have learned, discuss their experiences, and address the emotional and psychological impacts of the chatbot intervention.

5.5 Multi-Participant Role-Play with Chatbots

While the concept of multi-player improvisation theatre has been explored in role-playing games [29], the guided improvisation could open up room for multi-participant role-play where one or multiple students could interact with a single or multiple chatbots playing different roles. This kind of rich environment with multi-participant interactions and interpretations would resemble interactions on social media platforms where cyberbullying exchanges are played out in front of other users who can attenuate (e.g., by supporting a victim) or amplify (e.g., by staying silent or resharing an offensive message) the effects of cyberbullying through their actions [17]. Blending real participants and imagined identities enacted by chatbots could help youth practice socio-emotional skills in various relational and situational contexts, e.g., involving social circles of

friends and peers, being part of a group or a sole upstander, interacting with people of similar or diverse views and identities, etc. As mentioned earlier, bystanders' sense-making, reading of contextual cues, emotional reactions, and anticipated consequences of their actions are tethered to social and peer contexts in which they reside [10], and multi-participant interactions could provide opportunities for collaborative role-playing practices and learning.

From a technical standpoint, LLMs have been used to stage social simulacra [47, 48]. These social interactions of multiple participants are reminiscent of the role-play scenarios our teachers envisioned. LLM-based social simulacra could, therefore, be an opportunity for bringing teachers' role-play ideas to life. It is, however, still an open question how teachers can keep control of the simulations and how the students can interact with the simulated roles.

From the instructional perspective, chatbot role-playing sessions with multiple student participants would need to be carefully implemented and build on skills previously practiced in single-user chatbot interactions. In other words, the teachers would have to assess whether and when students are ready to move from single-user to multi-user interactions. Furthermore, because of greater autonomy and improvisation afforded in multi-participant interactions, teachers would need to be more closely involved through monitoring, moderation, and debriefing of these exchanges. Thus, there is a trade-off between improvisation and control, and greater improvisation in chatbot interactions would have to be counterbalanced by teachers' involvement in other ways.

6 CONCLUSION

In this work, we explore what technical and design components teachers need to build chatbots that assist in bystander education through Co-Pilot, an LLM-Chain based, no-code chatbot design tool. To create chatbot tools that fulfill teachers' needs, tool designers will want to consider the teachers' goal of constructing role-play scenarios and their perception of being playwrights of these social interactions. Teachers want to control and adapt the chatbot while at the same time allowing the chatbot enough improvisation so that students can explore different bystander actions and scenarios and practice socio-emotional skills. This view helps to understand how far current language model technology can be utilized for chatbot building and what new solutions still need to be found. We hope that researchers and designers of future tools will consider these factors to ensure that chatbots for adolescent cyberbullying education have a successful impact in the classroom.

ACKNOWLEDGMENTS

The authors would like to thank all participating teachers for their valuable time and insights, the research assistants Ashley Yu, George Gu, Jade Yang, Jerry Guo, Kyle Lou, Morgan Cupp, and Tony Yang for their help in developing the probe, as well as Dominic DiFranzo and Winice Hui for their contributions to the study. This work is supported by National Science Foundation under grants IIS-2313077 and IIS-2302977. Qian Yang is also supported by Schmidt Futures' AI2050 Early Career Fellowship.

REFERENCES

[1] Ana Aleksandric, Mohit Singhal, Anne Groggel, and Shirin Nilizadeh. 2022. Understanding the Bystander Effect on Toxic Twitter Conversations. https://doi.org/10.48550/ARXIV.2211.10764

[2] Kimberley R Allison and Kay Bussey. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review* 65 (2016), 183–194.

[3] Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. https://doi.org/10.48550/ARXIV.2202.01279

[4] Julia Barlińska, Anna Szuster, and Mikołaj Winiewski. 2013. Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology* 23, 1 (2013), 37–51.

[5] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann DeSmet, and Ilse De Bourdeaudhuij. 2015. 'Can I afford to help?'How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology* 34, 4 (2015), 425–435.

[6] Menucha Birenbaum. 2023. The Chatbots' Challenge to Education: Disruption or Destruction? *Education Sciences* 13, 7 (2023), 711.

[7] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI Interprets the Probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1077–1086. https://doi.org/10.1145/1240624.1240789

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Angela Busacca and Melchiorre Alberto Monaca. 2023. Deepfake: Creation, Purpose, Risks. In *Innovations and Economic and Social Changes due to Artificial Intelligence: The State of the Art*. Springer, 55–68.

[10] Erin A Casey, Taryn Lindhorst, and Heather L Storer. 2017. The situational-cognitive model of adolescent bystander behavior: Modeling bystander decision-making in the context of bullying and teen dating violence. *Psychology of violence* 7, 1 (2017), 33.

[11] Robin Cohen, Nivedha Mathiarasu, R Aarif, S Ansari, D Fraser, M Hegde, J Henderson, I Kajic, A Khan, Z Liao, et al. 2018. An education-based approach to aid in the prevention of cyberbullying. *Acm Sigcas Computers and Society* 47, 4 (2018), 17–28.

[12] Mike P Cook, Matthew Gremo, and Ryan Morgan. 2017. We're just playing: The influence of a modified tabletop role-playing game on ELA students' in-class reading. *Simulation & Gaming* 48, 2 (2017), 199–218.

[13] John M Darley and Bibb Latané. 1968. Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology* 8, 4p1 (1968), 377.

[14] Thomas S Dee and Dan Goldhaber. 2017. Understanding and addressing teacher shortages in the United States. *The Hamilton Project* 5 (2017), 1–28.

[15] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, and Ilse De Bourdeaudhuij. 2012. Mobilizing bystanders of cyberbullying: an exploratory study into behavioural determinants of defending the victim. *Annual review of cybertherapy and telemedicine* 10 (2012), 58–63.

[16] Ann DeSmet, Charlene Veldeman, Karolien Poels, Sara Bastiaensens, Katrien Van Cleemput, Heidi Vandebosch, and Ilse De Bourdeaudhuij. 2014. Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking* 17, 4 (2014), 207–215.

[17] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[18] Fernando Domínguez-Hernández, Lars Bonell, and Alejandro Martínez-González. 2018. A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 12, 4 (2018).

[19] Eva Durall and Evangelos Kapros. 2020. Co-design for a competency self-assessment chatbot and survey in science education. In *Learning and Collaboration Technologies. Human and Technology Ecosystems: 7th International Conference*. Springer, 13–24.

[20] Eva Durall, Marjo Virnes, Teemu Leinonen, and Begoña Gros. 2020. Ownership of learning in monitoring technology: Design case of self-monitoring tech in independent study. *Interaction Des. Architecture (s) J* 45 (2020), 133–154.

[21] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards

Measuring the Representation of Subjective Global Opinions in Language Models. *CoRR* abs/2306.16388 (2023). https://doi.org/10.48550/arXiv.2306.16388 arXiv:2306.16388

[22] Laura Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers* 35 (2003), 379–383.

[23] Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 9126–9140. https://doi.org/10.18653/v1/2023.acl-long.507

[24] Silvia Gabrielli, Silvia Rizzi, Sara Carbone, Valeria Donisi, et al. 2020. A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. *JMIR Human Factors* 7, 1 (2020), e16762.

[25] Norma Ghamrawi, Tarek Shal, and Najah AR Ghamrawi. 2023. Exploring the impact of AI on teacher leadership: regressing or expanding? *Education and Information Technologies* (2023), 1–19.

[26] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688

[27] Gunnar Harboe and Elaine M. Huang. 2015. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/2702123.2702561

[28] Véronique Irwin, Ke Wang, Jiashan Cui, Jizhi Zhang, and Alexandra Thompson. 2021. Report on Indicators of School Crime and Safety: 2020. (2021). https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021092

[29] Karis Jones, Scott Storm, Jennifer Castillo, and Sasha Karbachinskiy. 2021. Chasing New Worlds: Stories of Roleplaying in Classroom Spaces. *Journal of language and literacy education* 17, 1 (2021), n1.

[30] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *CoRR* abs/1909.05858 (2019). arXiv:1909.05858 http://arxiv.org/abs/1909.05858

[31] DV Kiriukhina. 2019. Cyberbullying among young users of social networks. *Journal of Modern Foreign Psychology* 8, 3 (2019), 53–59.

[32] Vasiliy Kolchenko. 2018. Can modern AI replace teachers? Not so fast! Artificial intelligence and adaptive learning: Personalized education in the AI age. *HAPS educator* 22, 3 (2018), 249–252.

[33] Robin M Kowalski and Cristin Fedina. 2011. Cyber bullying in ADHD and Asperger Syndrome populations. *Research in Autism Spectrum Disorders* 5, 3 (2011), 1201–1208.

[34] Joel Kupperstein. 2023. AI Can't Replace High-quality Teaching: Using the Technology as a Tool. (2023).

[35] Bibb Latané and John M Darley. 1970. *The unresponsive bystander: Why doesn't he help?* Prentice Hall.

[36] Danielle M. Law, Jennifer D. Shapka, Shelley Hymel, Brent F. Olson, and Terry Waterhouse. 2012. The changing face of bullying: An empirical comparison between traditional and internet bullying and victimization. *Computers in Human Behavior* 28, 1 (2012), 226–232. https://doi.org/10.1016/j.chb.2011.09.004

[37] Qing Li. 2007. Bullying in the new playground: Research into cyberbullying and cyber victimisation. *Australasian Journal of Educational Technology* 23, 4 (2007).

[38] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv:2107.13586 [cs.CL]

[39] Chung Kwan Lo. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13, 4 (2023), 410.

[40] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*. Springer, 231–248.

[41] Katja Machmutow, Sonja Perren, Fabio Sticca, and Francoise D Alsaker. 2012. Peer victimisation and depressive symptoms: Can specific coping strategies buffer the negative impact of cybervictimisation? *Emotional and Behavioural Difficulties* 17, 3-4 (2012), 403–420.

[42] Aida Midgett, Diana M Doumas, April Johnston, Rhiannon Trull, and Raissa Miller. 2018. Rethinking bullying interventions for high school students: A qualitative study. *Journal of Child and Adolescent Counseling* 4, 2 (2018), 146–163.

[43] Tijana Milosevic, Kanishk Verma, Michael Carter, Samantha Vigil, Derek Laffan, Brian Davis, and James O'Higgins Norman. 2023. Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From Youth Perspective. *Social Media+ Society* 9, 1 (2023), 20563051221147325.

[44] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring ChatGPT political bias. *Public Choice* (2023), 1–21.

[45] Dorit Olenik-Shemesh, Tali Heiman, and Sigal Eden. 2017. Bystanders' behavior in cyberbullying episodes: Active and passive patterns in the context of personal–socio-emotional factors. *Journal of interpersonal violence* 32, 1 (2017), 23–48.

[46] Julia Othlinghaus-Wulhorst and H Ulrich Hoppe. 2020. A technical and conceptual framework for serious role-playing games in the area of social skill training. *Frontiers in Computer Science* 2 (2020), 28.

[47] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *CoRR* abs/2304.03442 (2023). https://doi.org/10.48550/arXiv.2304.03442 arXiv:2304.03442

[48] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. https://doi.org/10.1145/3526113.3545616

[49] Justin W Patchin and Sameer Hinduja. 2012. *Cyberbullying prevention and response: Expert perspectives*. Routledge.

[50] Lara Schibelsky Godoy Piccolo, Pinelopi Troullinou, and Harith Alani. 2021. Chatbots to support children in coping with online threats: Socio-technical requirements. In *Designing Interactive Systems Conference 2021*. 1504–1517.

[51] Megan Price and John Dalgleish. 2010. Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. *Youth studies australia* 29, 2 (2010), 51–59.

[52] Rhiarne E Pronk and Melanie J Zimmer-Gembeck. 2010. It's "mean," but what does it mean to adolescents? Relational aggression described by victims, aggressors, and their peers. *Journal of Adolescent research* 25, 2 (2010), 175–204.

[53] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 837–848. https://doi.org/10.18653/v1/2022.acl-short.94

[54] Sara E Rimm-Kaufman. 2020. *SEL from the Start: Building Skills in K-5 (Social and Emotional Learning Solutions)*. WW Norton & Company.

[55] Christina Salmivalli, Kirsti Lagerspetz, Kaj Björkqvist, Karin Österman, and Ari Kaukiainen. 1996. Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 22, 1 (1996), 1–15.

[56] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. https://doi.org/10.48550/ARXIV.2110.08207

[57] Jeff Sauro and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

[58] Shari Kessel Schneider, Lydia O'donnell, Ann Stueve, and Robert WS Coulter. 2012. Cyberbullying, school bullying, and psychological distress: A regional census of high school students. *American journal of public health* 102, 1 (2012), 171–177.

[59] Jessica Shieh. 2023. Best practices for prompt engineering with openai API. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

[60] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. 2023. RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting. *CoRR* abs/2305.15685 (2023). https://doi.org/10.48550/arXiv.2305.15685 arXiv:2305.15685

[61] Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, 51–59. https://doi.org/10.18653/v1/W18-5107

[62] Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled Language Generation for Language Learning Items. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*. Association for Computational Linguistics, 294–305. https://doi.org/10.18653/v1/2022.emnlp-industry.30

[63] Seda Gökçe Turan. 2021. Deepfake and digital citizenship: A long-term protection method for children and youth. In *Deep fakes, fake news, and misinformation in online teaching and learning technologies*. IGI Global, 124–142.

[64] Tomoyuki Ueda, Junya Nakanishi, Itaru Kuramoto, Jun Baba, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2021. Cyberbullying Mitigation by a Proxy Persuasion of

a Chat Member Hijacked by a Chatbot. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. 202–208.
[65] U.S. Department of Education. 2023. Teacher Shortage Areas. https://tsa.ed.gov
[66] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.
[67] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao K. Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*. Association for Computational Linguistics, 116–122. https://aclanthology.org/2023.eacl-main.9
[68] Sofia Villatoro Moral and Barbara de Benito. 2021. An Approach to Co-Design and Self-Regulated Learning in Technological Environments. Systematic Review. *Journal of New Approaches in Educational Research* 10, 2 (2021), 234–250.
[69] Emily Vogels. 2022. Teens and Cyberbullying 2022. *Pew Research Center* (2022).
[70] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088
[71] Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *International Conference on Social Informatics*. Springer, 427–439.
[72] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. https://doi.org/10.48550/ARXIV.2203.06566
[73] Tongshuang Wu, Michael Terry, and Carrie J Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*.
[74] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 146, 15 pages. https://doi.org/10.1145/3491102.3517614
[75] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[76] J.D. Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2206–2220. https://doi.org/10.1145/3563657.3596138
[77] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
[78] Izabela Zych, Anna C Baldry, David P Farrington, and Vicente J Llorent. 2019. Are children involved in cyberbullying low on empathy? A systematic review and meta-analysis of research on empathy versus different cyberbullying roles. *Aggression and violent behavior* 45 (2019), 83–97.

A STUDENT SIMULATIONS

The student simulations used `GPT-3.5-Turbo` to generate student answers when testing the chatbot. Prompts were used that contained the context along with a description of the student type. We defined three student types, an aggressive student bullying the bully, an upstander student supporting the victim and a passive bystander student that ignores the bullying. All prompts had the same structure and only replaced parts that related to the student type. A separate prompt was used to generate the comment and any answers to the chatbot.

A.1 Aggressive student

```
You are John, an aggressive student, and you see the
following on Instagram: {general_context}

----

Give a comment that the student John would post under the
Instagram post in which John insults the bully. Be
aggressive. Answer in the language style of a teenager.
Give an answer that is no longer than 10 words.

----

You commented under this Instagram post the following
comment {comment}.
Based on your comment, a chatbot is trying to teach you
how to best act with a cyberbullying situation. This is
your conversation so far: {messages}.

Give the next answer of the student to this conversation
where you tend to not agree with the chatbot.
Answer in the language style of a teenager. Give an answer
that is no longer than 10 words.\nJohn:
```

A.2 Upstanding student

```
You are John, a supportive student, and you see the
following on Instagram: {general_context}

----

Give a comment that the student John would post under the
Instagram post in which John comforts and supports Alex
(the victim). Be gentle and sweet. Answer in the language
style of  a teenager. Give an answer that is no longer
than 10 words.

----

You commented under this Instagram post the following
comment {comment}.
Based on your comment, a chatbot is trying to teach you
how to best act with a cyberbullying situation. This is
your conversation so far: {messages}.

Give the next answer of the student to this conversation
where you tend to agree with the chatbot.
Answer in the language style of a teenager. Give an answer
that is no longer than 10 words.\nJohn:
```

A.3 Passive bystander student

```
You are John, a student who ignores the bullying and
just comments on the original post, and you see the
following on Instagram: {general_context}

----

Give a comment that the student John would post under
the Instagram post in which John is looking forward
to seeing the ballet recital. Be gentle and sweet.
Answer in the language style of a teenager. Give an
answer that is no longer than 10 words.

----

You commented under this Instagram post the following
comment {comment}. Based on your comment, a chatbot is
trying to teach you how to best act with a cyberbullying
situation. This is your conversation so far: {messages}.

Give the next answer of the student to this conversation
where you tend to agree with the chatbot.
Answer in the language style of a teenager. Give an
answer that is no longer than 10 words.\nJohn:
```

B STUDENT BEHAVIOR COMPONENTS

The structure and examples provided by the teacher were used to build few-shot classifiers of the student behavior. All behavior components that had the same parent component were used as classes in a classifier. The following prompt was used with a loop over all examples:

```
Victim's name is Alex. Bully's name is Leslie.
Classify the user inputs into one of the following
categories:
{prompt_classes}

Only give the name of the category. If none of these
categories match, output 'none' as category'.

Input {example_num}: {example}
Category {example_num}: {class_name}

Input {example_num}: {student_message_to_classify}
Category {example_num}:
```

We used `Text-Davinci-003` and parsed its answer to determine the predicted class (and therefore the conversational path to take in the dialogue structure).

C CHATBOT REACTION COMPONENT

The response examples provided by the teacher were used to generate the chatbot's answer in each situation. As example contexts, the behavior examples from the parent student behavior component

were used. We prompted `Text-Davinci-003` for the generation with a loop over all teacher-defined examples.

```
The student sees a cyberbully on social media.
The bully's name is Leslie and the victim's name is
Alex.
The student makes a comment in response to the post.
You are talking to that student whose name is not Alex
or Leslie so don't call him/her Alex or Leslie.
Teach that student to counteract cyberbullies based on
the following examples:"

Example: {example_num}
Context: {context_example}
Response: {response}"

Now fill in a new response based on the examples.
Give answers very similar to the examples:

Context: {student_message_to_answer}
Response:
```