# Safe Generative Chats in a WhatsApp Intelligent Tutoring System

Zachary Levonian[1,*], Owen Henkel[2]

[1]Digital Harbor Foundation, 1045 Light St, Baltimore, Maryland, USA
[2]University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, UK

**Abstract**

Large language models (LLMs) are flexible, personalizable, and available, which makes their use within Intelligent Tutoring Systems (ITSs) appealing. However, that flexibility creates risks: inaccuracies, harmful content, and non-curricular material. Ethically deploying LLM-backed ITS systems requires designing safeguards that ensure positive experiences for students. We describe the design of a conversational system integrated into an ITS, and our experience evaluating its safety with red-teaming, an in-classroom usability test, and field deployment. We present empirical data from more than 8,000 student conversations with this system, finding that GPT-3.5 rarely generates inappropriate messages. Comparatively more common is inappropriate messages from students, which prompts us to reason about safeguarding as a content moderation and classroom management problem. The student interaction behaviors we observe provide implications for designers—to focus on student inputs as a content moderation problem—and implications for researchers—to focus on subtle forms of bad content.

**Keywords**

large language models, intelligent tutoring systems, safety

## 1. Introduction

The capabilities of Large Language Models (LLMs) have led to a surge of interest in applying them to educational settings, including for automated tutoring, personalized learning, and adaptive assessment [1, 2]. A particularly promising application of LLMs is integration with Intelligent Tutoring Systems (ITSs), as they can combine the structured pedagogical processes and vetted curricula of ITSs and the flexibility and personalization enabled by conversational interfaces [3, 4, 5, 6].

Integrating LLMs into ITSs enables answering student questions, summarizing concepts, creating customized hints, and recontextualizing learning materials [7, 8, 3]. However, the use of LLMs in educational applications also raises concerns regarding potential risks, including the generation of toxic language, implicit biases, and inaccurate information, as well as inappropriate use by students [9, 10, 11, 12]. These risks become particularly important when designing educational applications that directly interact with students e.g. via a chat interface, necessitating a focus on the safety and accuracy of model-generated responses to students.

Recent advancements in LLMs have led to improvements in mitigating some of the most distressing behaviors of early generations, such as toxicity, wildly inaccurate information, and discussions of illegal or taboo topics [13, 14]. While this progress is welcome, it has also revealed a range of more subtle potential problems. For instance, small hallucinations (e.g., confusing "$2\pi r$" with "$\pi r^2$") may lead to persistent misconceptions [15]. Additionally, younger students might be more likely to anthropomorphize models and develop emotionally charged relationships with them [16, 17], and models tend to present an "average" view of the Anglophone internet which might not be appropriate in certain cultural contexts [18, 19].

Less discussed is how models should handle inappropriate or potentially offensive student inputs, as well as honest questions on politically or culturally sensitive topics. For example, if a student addresses an LLM application using profane language, should the model ignore the profanity and proceed, ask the student to stop using such language, or request that the student rephrase the question? Similarly, if a student asks an honest question about a potentially charged political topic (e.g., "Is it okay to get pregnant before you are married?"), should the model provide a standard "it depends" answer, ignore the question, or inform the student that they cannot discuss the topic?

Perhaps most seriously, if a student discloses some sort of trauma or abuse they have suffered, how should the model respond? While these are complex questions, they are also ones that teachers and tutors deal with regularly [20, 21, 22, 23]. Deciding how to respond to inappropriate or provocative student questions is a classic challenge of classroom management [24, 25], carefully choosing how to address and explain sensitive topics is a fraught area for nearly all teachers [26, 27], and handling sensitive student disclosures is such an important question that most school systems have codified mandatory reporting rules for teachers that specify which types of student disclosures must be reported to school leadership, mental health professionals, or law enforcement [28].

In this paper, we describe a system we designed for safeguarding student chats with an ITS and empirical data from a field deployment of that system with usage from more than 8,000 students. We formed a research collaboration with the developers of Rori, a WhatsApp-based chatbot math tutor. Rori is used primarily by low-income middle-school students in Sierra Leone, Liberia, Ghana, and Rwanda both in classroom settings and at home for math skills practice [29]. We designed a conversational experience for Rori's users that teaches them about growth mindset before they begin math skill practice.

Dinan et al. identify three broad safety issues in conversational systems: (a) instigator effects, in which the system generates harmful content, (b) yea-sayer effects, in which the system endorses or fails to object to harmful content, and (c) imposter effects, in which the system provides incorrect or harmful advice [30]. To safeguard students during that conversation, we designed a safety system consisting of

*Educational Datamining '24 Human-Centric eXplainable AI in Education and Leveraging Large Language Models for Next-Generation Educational Technologies Workshop Joint Proceedings, July 13, 2024, Atlanta, GA*

[*]Corresponding author.

✉ zach@levi.digitalharbor.org (Z. Levonian); owen.henkel@education.ox.ac.uk (O. Henkel)
🌐 https://levon003.github.io/ (Z. Levonian)
iD 0000-0002-8932-1489 (Z. Levonian); 0009-0001-8850-067X (O. Henkel)

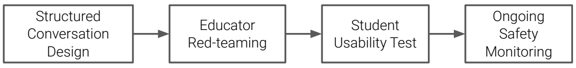


**Figure 1:** Designing for safety: our process.

filters and corresponding actions when messages are flagged by those filters. Across two studies, we find no evidence of instigator or imposter effects but limited evidence of the yea-sayer effect.

In study one, we assess the usability and ethical acceptability of the system in the classroom. In study two, we deploy the system for use by students at home. The empirical evidence from these two studies provides implications for designers—to focus on student inputs as a content moderation problem—and implications for researchers—to focus on subtle forms of bad content.

## 2. System Design

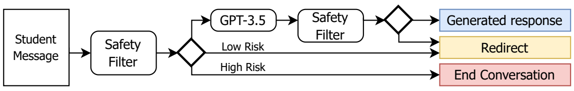
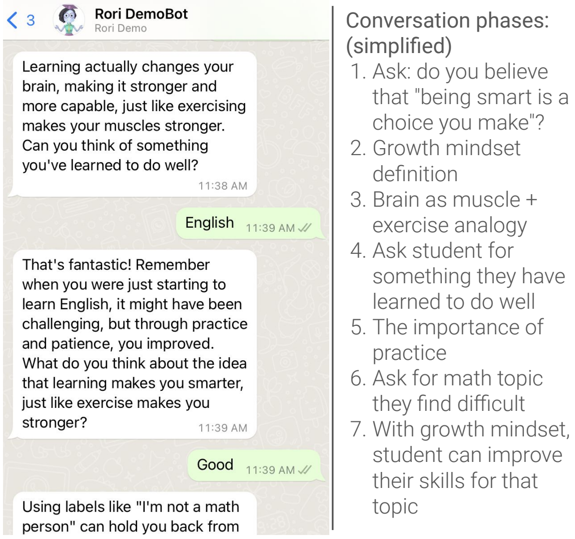


**Figure 2:** The generative chat moderation system.

To design a safe generative chat experience, we implemented a system on the basis of educator feedback and through multiple phases of evaluation as shown in Figure 1.

**Designing a semi-structured conversation.** We chose to implement a generative chat for encouraging a growth mindset, an approach linked to positive educational outcomes, including in mobile learning contexts [31, 32, 33]. We used a prompting approach that moves the conversation through multiple phases: introducing the concept of a growth mindset, asking the student to reflect on a time that practice has helped them, and identifying a specific math skill that they want to practice. The system initiates the conversation with the message “Do you agree with the statement ‘Being smart is a choice you make, not the way you are’?” and moves the student through various conversational phases, as shown in Figure 3. During the conversation, we detect standard navigation keywords (e.g. “menu”) to navigate away from the conversation and on to math skills practice. We limited the total conversation length—a max of 8 turns during the usability test and 10 during the field deployment—to decrease the chance of major digressions and to reduce any student frustration. By designing the conversation as system-initiated rather than student-initiated and ending each system message with a question, we provide structure that keeps the conversation flowing and focused on growth mindset.

**Designing safety guardrails.** To ensure students have a safe experience during the conversation, we implemented guardrails that would redirect or end the conversation. Each student and system message is passed through a safety filter that determines how the system will respond to the student. Figure 2 demonstrates the final design. The safety filter consists of (1) a word list and (2) a statistical moderation model. The word list—consisting only of unambiguous curse words—is applied first. While a word list is rigid and inflexible, we chose to include it because it is easier for educators and parents to reason about than a statistical model [34].



Learning actually changes your brain, making it stronger and more capable, just like exercising makes your muscles stronger. Can you think of something you've learned to do well? 11:38 AM

English 11:39 AM

That's fantastic! Remember when you were just starting to learn English, it might have been challenging, but through practice and patience, you improved. What do you think about the idea that learning makes you smarter, just like exercise makes you stronger? 11:39 AM

Good 11:39 AM

Using labels like "I'm not a math person" can hold you back from

Conversation phases: (simplified)

1. Ask: do you believe that "being smart is a choice you make"?
2. Growth mindset definition
3. Brain as muscle + exercise analogy
4. Ask student for something they have learned to do well
5. The importance of practice
6. Ask for math topic they find difficult
7. With growth mindset, student can improve their skills for that topic

**Figure 3:** A chat excerpt from the Rori WhatsApp interface and a simplified view of the conversation phases.

The statistical model we used was OpenAI’s moderation API, which predicts the presence of five high-level content categories and six sub-categories [35]. Each message is given a score between 0 and 1 reflecting how likely that message is to contain content in that category. We set the per-category thresholds for which we would take system action based on the red-teaming exercise.

**System moderation actions.** Based on the assessed risk of the message, we took one of two moderation actions in response to student messages. We classified self-harm, sexual/minors, and the two /threatening sub-categories as high risk messages and the rest as low risk. In response to low risk messages, we drop the students most recent message from the prompted context and ask them to continue the conversation with a more appropriate message. In response to high risk messages, we end the conversation immediately with the message: “That sounds like a serious topic, and a real person needs to look at this. They might try to contact you to check on you. Until someone has reviewed this, Rori will not reply.” We make an open source reference implementation of our moderation system available on GitHub.[1]

**Educator red-teaming.** To evaluate the acceptability of the conversation design and the safety guardrails, we conducted an asynchronous red-teaming exercise. There is considerable variation in red-teaming exercises [36]; the purpose of our exercise was to qualitatively assess the effectiveness of the safety guardrails and to quantitatively set initial per-category moderation thresholds. We recruited 17 Rising Academies educators and system designers to adversarially probe the conversation design. Across 57 conversations, we received negative feedback on 39 messages that should have been flagged, setting the thresholds appropriately. After small tweaks to the prompts, we observed no obviously negative conversational experiences. We return to the topic of subtly negative experiences in the discussion, but we determined there to be minimal risk in proceeding with a full usability assessment with students.

**Monitoring.** To ensure the safety of Rori student users, we designed a continual monitoring procedure. We imple-

[1] https://github.com/DigitalHarborFoundation/chatbot-safety

**Table 1**
Counts of students, conversations, and messages across two studies.

| | Students | Conversations | Messages |
|---|---|---|---|
| Usability Test | 109 | 252 | 3,722 |
| Field Deployment | 8,168 | 8,755 | 126,278 |

mented data dashboards to review the most recent and the riskiest conversations. Messages flagged as high-risk generate an email alert to an internal team. We designed a basic reporting protocol for use with student users in the event of particular sensitive disclosures e.g. sexual abuse or suicidal thoughts.

# 3. Study 1: Student Usability Test

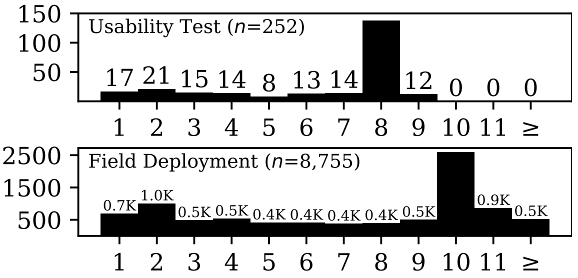


**Figure 4:** Conversation length (as number of student messages) for all conversations. Completion rate was higher during the usability test (59.5%) than the field deployment (38.9%).

**Table 2**
Highest and 99th percentile of the OpenAI moderation scores observed during the two studies. The highest possible value is 1.

| | Usability Test | | Field Deployment | |
|---|---|---|---|---|
| Source | Q99 | Max | Q99 | Max |
| GPT-3.5 | 0.000 | 0.010 | 0.003 | 0.044 |
| Student | 0.002 | 0.045 | 0.030 | 0.989 |

In December 2023, 109 in-school students across 6 total classrooms were instructed to use the growth mindset generative chat during a regularly-scheduled study hall using Rori for math skills practice [29]. 252 conversations occurred between December 13th and 15th. 60% of the conversations were completed; the distribution of conversation lengths is shown in Figure 4.[2]

At the end of the conversation, we asked students to rate the conversation from one to five stars.[3] The distribution of ratings is shown in Table 3. Of the rated conversations, 16 conversations (6.3%) were rated less than five stars. Qualitative investigation of those 16 low-rated conversations reveals no clear difference between those and 5-star conversation; student messages in low-rated conversations were non-significantly more likely to be single-word responses (75.4% low-rated vs 65.4% five-star, $\chi^2$=0.82, d.f.=1, $p$=0.36).

No student or GPT-3.5 student messages were flagged by the safety filter. In particular, most GPT-3.5 and student messages received low moderation scores across all categories. Table 2 shows summary statistics for the highest score received across all categories: the highest-scoring GPT-3.5 message received a score of 0.01 ("Oh, it seems like you might not understand the question. Let me rephrase it. Do you think that being smart is something that you can choose to be, rather than something that you are born with?"), while the highest-scoring student message received a score of 0.05 (a typo).

It may be that the moderation API's implicit values diverge from our own, such that false negatives occur and harmful student messages are not flagged. To check, we randomly sampled 100 student conversations, finding no false negatives. Qualitatively, while some student messages were playful or inappropriate in ways that would likely trigger a response from a human tutor, we found our prompt for GPT-3.5 effective at producing appropriate redirections back to the current topic.

Taken together, these results suggest that the semi-structured growth mindset conversation is acceptable for broader use. Critically, the conversation design was effective at preventing messages that would trigger the safety filter: we identified no obviously unacceptable student messages. We made minor adjustments to the prompts and proceeded to a field deployment.

# 4. Study 2: Field Deployment

The growth mindset conversation was deployed publicly on Feburary 13, 2024 for non-school users of Rori and incorporated as a component of the on-boarding process before math skills practice begins. We analyzed the 126,278 messages between the feature launch and May 1, 2024.

## 4.1. Did GPT-3.5 generate objectionable outputs?

No. Quantitatively, the highest-scoring system message produced received a score of 0.044. During continual monitoring, the researchers annotated GPT-3.5 messages and determined none of them to be objectionable. The most controversial messages were those generated in response to student's objectionable messages, which we discuss in the next sections.

## 4.2. Did students write objectionable messages?

Yes, but not very much. 0.31% of student messages received a score in any moderation category of at least 0.1. Fewer than 8 in 10000 messages were flagged. Table 4 summarizes the moderation scores per-category. The most common negative messages were harassing or sexual. Only one message was flagged as high risk. After investigation by the team, it was determined to be a false positive by the OpenAI moderation model—the message should have been classified as low risk, as it contained violent language that merited corrective action but did not evidence self-harm. From an investigation of the 27 conversations with flagged messages, all flagged messages were determined to merit corrective action.

[2] Due to a bug that under-counted student messages, some conversations continued an extra turn.

[3] Feedback request message: "Thank you for your time! How much did you like the conversation?" A response modal labeled "Give us some ⭐s!" has quick-reply buttons.

**Table 3**
Student conversation ratings during Study 1.

| Rating | none | ★★★★★ | ★★★★ | ★★★ | ★★ | ★ |
|---|---|---|---|---|---|---|
| # conversations | 125 | 126 | 4 | 5 | 2 | 5 |

**Table 4**
OpenAI moderation scores by category for the 54,384 student messages sent during the field deployment. In addition to the 99th percentile and maximum observed score over all student messages, we show the number of messages with a score greater than 0.1 and greater than 0.5.

| Category | Q99 | Max | $n \geq 0.1$ | $n \geq 0.5$ |
|---|---|---|---|---|
| Harassment | 0.011 | 0.989 | 141 | 36 |
| Sexual | 0.012 | 0.914 | 28 | 5 |
| Hate | 0.002 | 0.524 | 3 | 1 |
| Violence | 0.001 | 0.959 | 2 | 1 |
| Self-harm/intent | 0.001 | 0.743 | 1 | 1 |
| Self-harm | 0.001 | 0.531 | 1 | 1 |
| Harassment/threatening | 0.000 | 0.451 | 1 | 0 |
| Hate/threatening | 0.000 | 0.087 | 0 | 0 |
| Violence/graphic | 0.000 | 0.081 | 0 | 0 |
| Self-harm/instructions | 0.000 | 0.072 | 0 | 0 |
| Sexual/minors | 0.007 | 0.024 | 0 | 0 |

### 4.3. Did GPT-3.5 respond appropriately?

We investigated the messages generated in response to student messages that were near the safety filter thresholds but remained unflagged. 48 unflagged conversations contained a message with a moderation score of at least 0.1. 40 of these conversations included at least one student message that warranted caution or a corrective statement from the system response, and we deemed the GPT-3.5-generated responses to be appropriately corrective in 37 of those cases. In 3 cases, the generated response ignored or equivocated when a corrective message would have been warranted. This is a subtler form of bad response: the yea-sayer effect [30].

## 5. Discussion

### 5.1. Key Findings

In this workshop paper, we described a system for conducting safe generative chats inside of an existing ITS. We found that the semi-structured conversation design we used eliminated imposter effects, while safety filters for students' inputs eliminated instigator effects. We found that it was surprisingly straightforward to develop a prompt for GPT-3.5 to respond appropriately to the vast majority of student messages [37].

Instead, our attention was drawn to the more frequent and more challenging problem of how to deal with inappropriate or otherwise sensitive student messages. In some ways this challenge in analogous to the challenges of content moderation on online platforms, where the context in which a comment exists is important, and policies that are reasonable in many cases might be ineffective in edge cases. As an example: how to handle questions regarding contentious political or historical topics? In many cases acknowledging that there are different valid opinions is a good pedagogical approach, but in particularly sensitive or egregious examples this "both-sideism" can be inappropriate [38]. However, these are the types of challenges teachers deal with constantly, and we believe that there is a research opportunity here at the intersection of content moderation and classroom management to develop appropriate system actions in response to objectionable student messages.

Another important finding was that the process of red-teaming was effective in its primary goal of identifying potential risk. It had other benefits we did not expect: building organizational confidence. We found that being transparent about the shortcoming of our V1 approach and including designers, educators, and researchers in the evaluation process had the dual benefit of improving trust and soliciting higher-quality feedback to improve the design.

### 5.2. Limitations & Future Research

The specific moderation actions we implemented are reasonable starting points, and by classifying messages at two risk levels we are able to positively redirect conversations with pre-vetted messages [39]. While these corrective messages were written by educators, in the future we hope that approaches from culturally-responsive classroom management might be combined with soliciting cultural background information from students so that behavioral expectations can be communicated more clearly and correctives can be applied more appropriately [40, 41].

In the event of more serious disclosures, as with the messages we classify as high risk, we argue that our choice to automatically end the conversation and move to human review rather than attempting to generate an appropriate LLM response in the moment is the more ethical one [42]. However, the specific approach we used of ending the conversation is not ideal; we might consider technical infrastructure that starts an in-chat support session with a human or otherwise connects explicitly to contacts at the student's school.

We did observe evidence of the yea-sayer effect in response to some objectionable student messages; future work should explore opportunities for mitigating this effect. In the mean time, designers should monitor for the prevalence of yea-saying and consider technical approaches that explicitly model the appropriate corrective behavior.

## Acknowledgments

We would like to thank Millie-Ellen Postle, Hannah Horne-Robinson, and the staff of Rising Academies for their contributions. This work was supported by the Learning Engineering Virtual Institute (LEVI) and the Digital Harbor Foundation.

## References

[1] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer,

A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195. doi:10.1016/j.lindif.2023.102274.

[2] A. Caines, L. Benedetto, S. Taslimipoor, C. Davis, Y. Gao, O. Andersen, Z. Yuan, M. Elliott, R. Moore, C. Bryant, M. Rei, H. Yannakoudakis, A. Mullooly, D. Nicholls, P. Buttery, On the application of Large Language Models for language teaching and assessment technology, 2023. URL: http://arxiv.org/abs/2307.08393, arXiv:2307.08393 [cs].

[3] Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.-E. Postle, W. Xing, Retrieval-augmented Generation to Improve Math Question-Answering: Tradeoffs Between Groundedness and Human Preference, in: NeurIPS'23 Workshop on Generative AI for Education (GAIED), arXiv, New Orleans, 2023. URL: http://arxiv.org/abs/2310.03184. doi:10.48550/arXiv.2310.03184, arXiv:2310.03184 [cs].

[4] S. Upadhyay, E. Ginsberg, C. Callison-Burch, Improving Mathematics Tutoring With A Code Scratchpad, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 20–28. URL: https://aclanthology.org/2023.bea-1.2. doi:10.18653/v1/2023.bea-1.2.

[5] S. Hobert, R. M. v. Wolff, Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents, Wirtschaftsinformatik 2019 Proceedings (2019). URL: https://aisel.aisnet.org/wi2019/track04/papers/2.

[6] B. Khosrawi-Rad, H. Rinn, R. Schlimbach, P. Gebbing, X. Yang, C. Lattemann, D. Markgraf, S. Robra-Bissantz, Conversational Agents in Education – A Systematic Literature Review, ECIS 2022 Research Papers (2022). URL: https://aisel.aisnet.org/ecis2022_rp/18.

[7] Z. A. Pardos, S. Bhandari, Learning gain differences between ChatGPT and human tutor generated algebra hints (2023). URL: https://arxiv.org/abs/2302.06871. doi:10.48550/ARXIV.2302.06871.

[8] S. Sonkar, L. Liu, D. B. Mallick, R. G. Baraniuk, CLASS Meet SPOCK: An Education Tutoring Chatbot based on Learning Science Principles, 2023. URL: http://arxiv.org/abs/2305.13272, arXiv:2305.13272 [cs].

[9] P. P. Liang, C. Wu, L.-P. Morency, R. Salakhutdinov, Towards Understanding and Mitigating Social Biases in Language Models, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 6565–6576. URL: https://proceedings.mlr.press/v139/liang21a.html, iSSN: 2640-3498.

[10] R. Navigli, S. Conia, B. Ross, Biases in Large Language Models: Origins, Inventory, and Discussion, Journal of Data and Information Quality 15 (2023) 10:1–10:21. URL: https://dl.acm.org/doi/10.1145/3597307. doi:10.1145/3597307.

[11] R. S. Baker, A. Hawn, Algorithmic Bias in Education, International Journal of Artificial Intelligence in Education 32 (2022) 1052–1092. URL: https://doi.org/10.1007/s40593-021-00285-9. doi:10.1007/s40593-021-00285-9.

[12] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, D. Gašević, Practical and ethical challenges of large language models in education: A systematic scoping review, British Journal of Educational Technology 55 (2024) 90–112. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13370. doi:10.1111/bjet.13370, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370.

[13] OpenAI, GPT-4 Technical Report, 2023. URL: http://arxiv.org/abs/2303.08774, arXiv:2303.08774 [cs].

[14] Y. Tao, O. Viberg, R. S. Baker, R. F. Kizilcec, Auditing and Mitigating Cultural Bias in LLMs, 2023. URL: http://arxiv.org/abs/2311.14096. doi:10.48550/arXiv.2311.14096, arXiv:2311.14096 [cs].

[15] P. K. Murphy, P. A. Alexander, Situating Text, Talk, and Transfer in Conceptual Change: Concluding Thoughts, in: International Handbook of Research on Conceptual Change, 2 ed., Routledge, 2013. Num Pages: 19.

[16] S. Girard, H. Johnson, What Do Children Favor as Embodied Pedagogical Agents?, in: V. Aleven, J. Kay, J. Mostow (Eds.), Intelligent Tutoring Systems, Springer, Berlin, Heidelberg, 2010, pp. 307–316. doi:10.1007/978-3-642-13388-6_35.

[17] E. J. Goldman, A.-E. Baumann, D. Poulin-Dubois, Preschoolers' anthropomorphizing of robots: Do human-like properties matter?, Frontiers in Psychology 13 (2023). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1102370/full. doi:10.3389/fpsyg.2022.1102370, publisher: Frontiers.

[18] Y. Xu, L. Hu, J. Zhao, Z. Qiu, Y. Ye, H. Gu, A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias, 2024. URL: http://arxiv.org/abs/2404.00929, arXiv:2404.00929 [cs].

[19] A. Agiza, M. Mostagir, S. Reda, Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in LLMs, 2024. URL: http://arxiv.org/abs/2404.08699, arXiv:2404.08699 [cs].

[20] M. Falkiner, D. Thomson, A. Day, Teachers' Understanding and Practice of Mandatory Reporting of Child Maltreatment, Children Australia 42 (2017) 38–48. doi:10.1017/cha.2016.53.

[21] K. A. Beck, J. R. P. Ogloff, A. Corbishley, Knowledge, Compliance, and Attitudes of Teachers toward Mandatory Child Abuse Reporting in British Columbia, Canadian Journal of Education / Revue canadienne de l'éducation 19 (1994) 15–29. URL: https://www.jstor.org/stable/1495304. doi:10.2307/1495304, publisher: Canadian Society for the Study of Education.

[22] M. R. Haney, Ethical Dilemmas Associated With Self-Disclosure in Student Writing, Teaching of Psychology (2004). URL: https://www.tandfonline.com/doi/abs/10.1207/s15328023top3103_2. doi:10.1207/s15328023top3103_2, publisher: Lawrence Erlbaum Associates, Inc.

[23] E. Berger, N. Chionh, A. Miko, School Leaders' Experiences on Dealing with Students Exposed to Domestic Violence, Journal of Family Violence 37 (2022) 1089–1100. URL: https://doi.org/10.1007/s10896-021-00310-4. doi:10.1007/s10896-021-00310-4.

[24] E. J. Sabornie, D. L. Espelage (Eds.), Handbook of Classroom Management, 3rd edition ed., Routledge, 2022.

[25] R. J. Marzano, A Handbook for Classroom Man-

agement that Works, ASCD, 2005. Google-Books-ID: BMOQFLa0fcEC.

[26] J. Levin, J. F. Nolan, What Every Teacher Should Know About Classroom Management, 1st edition ed., Pearson, 2002.

[27] M. Falkiner, D. Thomson, B. Guadagno, A. Day, Heads you win, tails I lose: The dilemma mandatory reporting poses for teachers, Australian Journal of Teacher Education (Online) 42 (2020) 93–110. URL: https://search.informit.org/doi/abs/10.3316/INFORMIT.088816132023302. doi:10.3316/informit.088816132023302, publisher: Edith Cowan University.

[28] J. D. G. Goldman, Primary school student-teachers' knowledge and understandings of child sexual abuse and its mandatory reporting, International Journal of Educational Research 46 (2007) 368–381. URL: https://www.sciencedirect.com/science/article/pii/S0883035507000675. doi:10.1016/j.ijer.2007.09.002.

[29] O. Henkel, H. Horne-Robinson, N. Kozhakhmetova, A. Lee, Effective and Scalable Math Support: Evidence on the Impact of an AI- Tutor on Math Achievement in Ghana, 2024. URL: http://arxiv.org/abs/2402.09809. doi:10.48550/arXiv.2402.09809, arXiv:2402.09809 [cs].

[30] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, V. Rieser, Anticipating Safety Issues in E2E Conversational AI: Framework and Tooling, 2021. URL: http://arxiv.org/abs/2107.03451. doi:10.48550/arXiv.2107.03451, arXiv:2107.03451 [cs].

[31] S. Karumbaiah, R. Lizarralde, D. Allessio, B. Woolf, I. Arroyo, N. Wixon, Addressing Student Behavior and Affect with Empathy and Growth Mindset, International Educational Data Mining Society, 2017. URL: https://eric.ed.gov/?id=ED596572, publication Title: International Educational Data Mining Society ERIC Number: ED596572.

[32] D. S. Yeager, C. Romero, D. Paunesku, C. S. Hulleman, B. Schneider, C. Hinojosa, H. Y. Lee, J. O'Brien, K. Flint, A. Roberts, J. Trott, D. Greene, G. M. Walton, C. S. Dweck, Using Design Thinking to Improve Psychological Interventions: The Case of the Growth Mindset During the Transition to High School, Journal of educational psychology 108 (2016) 374–391. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981081/. doi:10.1037/edu0000098.

[33] R. F. Kizilcec, D. Goldfarb, Growth Mindset Predicts Student Achievement and Behavior in Mobile Learning, in: Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, L@S '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–10. URL: https://dl.acm.org/doi/10.1145/3330430.3333632. doi:10.1145/3330430.3333632.

[34] S. Jhaver, A. Q. Zhang, Q. Z. Chen, N. Natarajan, R. Wang, A. X. Zhang, Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 289:1–289:33. URL: https://dl.acm.org/doi/10.1145/3610080. doi:10.1145/3610080.

[35] Moderation - OpenAI API, 2024. URL: https://platform.openai.com/docs/guides/moderation/overview.

[36] M. Feffer, A. Sinha, Z. C. Lipton, H. Heidari, Red-Teaming for Generative AI: Silver Bullet or Security Theater?, 2024. URL: http://arxiv.org/abs/2401.15897. doi:10.48550/arXiv.2401.15897, arXiv:2401.15897 [cs].

[37] A. Narayanan, S. Kapoor, S. Lazar, Model alignment protects against accidental harms, not intentional ones, 2023. URL: https://www.aisnakeoil.com/p/model-alignment-protects-against.

[38] R. Smith, How "both-sideism" harms health, BMJ 378 (2022) o2136. URL: https://www.bmj.com/content/378/bmj.o2136. doi:10.1136/bmj.o2136, publisher: British Medical Journal Publishing Group Section: Opinion.

[39] D. Leach, S. Helf, Using a Hierarchy of Supportive Consequences to Address Problem Behaviors in the Classroom, Intervention in School and Clinic 52 (2016) 29–33. URL: https://doi.org/10.1177/1053451216630288. doi:10.1177/1053451216630288, publisher: SAGE Publications Inc.

[40] C. S. Weinstein, S. Tomlinson-Clarke, M. Curran, Toward a Conception of Culturally Responsive Classroom Management, Journal of Teacher Education 55 (2004) 25–38. URL: https://doi.org/10.1177/0022487103259812. doi:10.1177/0022487103259812, publisher: SAGE Publications Inc.

[41] R. Skiba, H. Ormiston, S. Martinez, J. Cummings, Teaching the Social Curriculum: Classroom Management as Behavioral Instruction, Theory Into Practice 55 (2016) 120–128. URL: https://doi.org/10.1080/00405841.2016.1148990. doi:10.1080/00405841.2016.1148990, publisher: Routledge _eprint: https://-doi.org/10.1080/00405841.2016.1148990.

[42] L. Stapleton, S. Liu, C. Liu, I. Hong, S. Chancellor, R. E. Kraut, H. Zhu, "If This Person is Suicidal, What Do I Do?": Designing Computational Approaches to Help Online Volunteers Respond to Suicidality, 2024.