LLM Safety in an Educational Context: A holistic approach

Yeo, B.

AI Chatbot Interactions | Monitoring System | Student Safety

As LLMs enter classrooms, there is a need to take a holistic approach towards ensuring student safety. Analyzing authentic student-chatbot conversations from Singapore's national online teaching and learning platform, this study reveals behaviors like cognitive shortcuts, learning distractions, and boundary-testing that escape conventional safety frameworks. Through expert annotations by 12 educators, we developed context-specific risk categories for Singapore's educational environment. Our findings show effective monitoring must balance automated detection with teacher expertise, prioritizing accuracy to maintain educator confidence while enabling timely interventions. We propose a customised system that empowers teachers to assess complete interaction contexts, addressing young users' developmental vulnerabilities while preserving supportive learning environments.

Introduction

The deployment of large language models (LLMs) in educational contexts presents unique safety challenges, particularly when the primary users are students and minors. However, current LLM safety frameworks solely focus on LLM outputs, and ignore the developmental aspect of student-AI interactions. While it is important to ensure that LLMs do not generate unsafe content for young users, there is a need to also develop students' capacity to engage productively with AI and address any underlying behavioural causes. This paper proposes an approach for developing customised monitoring mechanisms to assess student interactions with chatbots and provide opportunities for human-led interventions, addressing the critical gap in ensuring safe AI interactions for children within their local educational environments.

Literature Review

Current AI safety research has predominantly focused on adult users, leaving significant gaps in understanding the unique vulnerabilities that minors face when interacting with LLMs (Jiao, et al., 2025). For one, children's developing critical thinking skills and limited life experience make them particularly susceptible to harmful online content and more likely to trust unreliable sources (Dangol et al., 2025). Additionally, the presence of an 'empathy gap' (Kurian, 2024) in LLMs can lead to the production of responses that are potentially harmful for younger users. This vulnerability is compounded by the fact that existing content moderation systems lack age-specific design considerations, failing to account for children's developmental needs (Khoo et al., 2025). This research gap has resulted in evaluation frameworks that inadequately assess age-inappropriate scenarios and fail to account for educational context requirements.

Additionally, most, if not all, research papers relating to LLM safety for children are focused on developing risk taxonomies or systems that prevent LLMs from producing unsafe or risky content (see, for example, Rath et al., 2025; Jiao, et al., 2025; Khoo et al., 2025). However, from an education professional's point of view, it is not only important to prevent LLMs from producing unsafe content, but also to ensure that students who initiate such unsafe conversations are monitored and followed up on by qualified teaching professionals. This is especially pertinent in cases of aggressive or socially withdrawn behaviour, which might require professional intervention in the form of counselling or other classroom management strategies (Shamnadh & A, 2019).

Methods

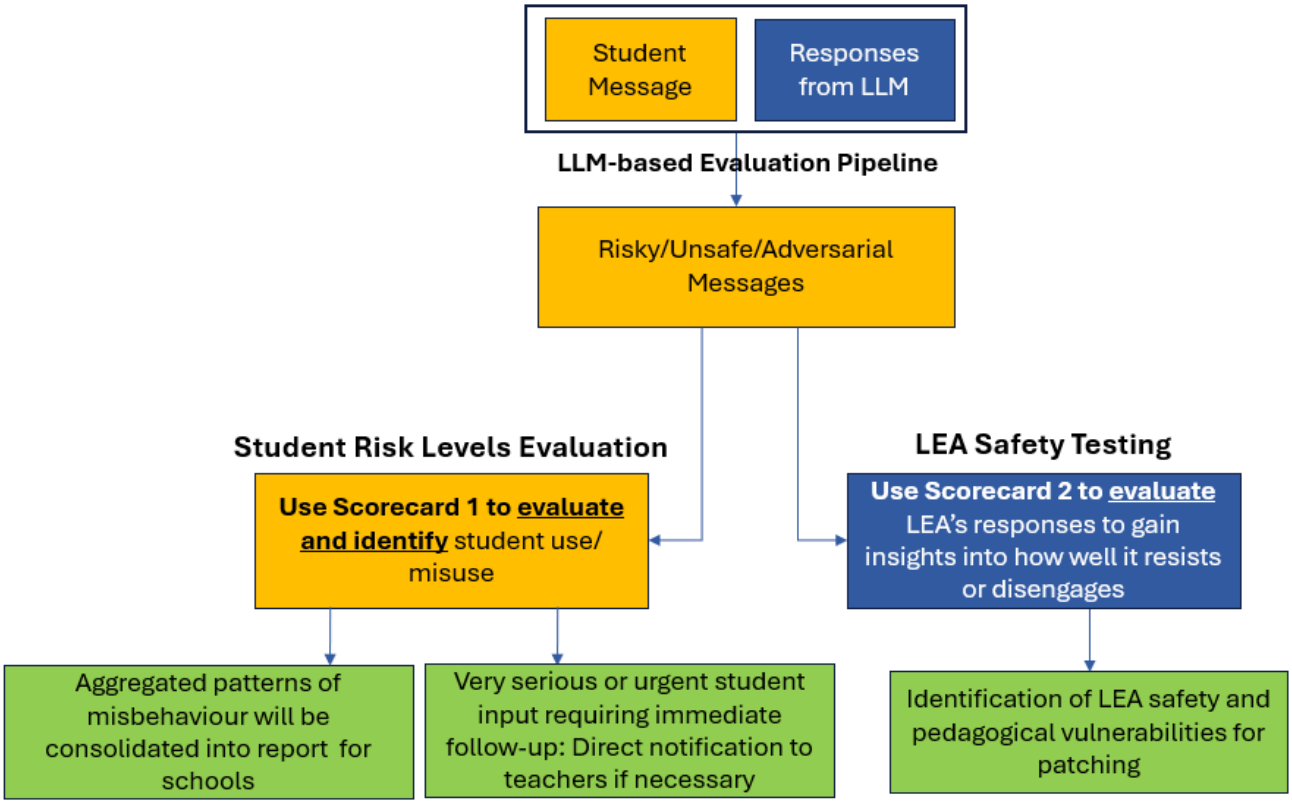
Identification and Definition of Risk Categories

To address these limitations, we embarked on a project to monitor risky student-AI interactions and make interventions if necessary. To do so, we adopted a nested Learning Engineering approach, moving through incremental and iterative stages of investigation, creation and implementation (Totino & Kessler, 2024) to design our solution. In the first stage, we conducted a pilot study to examine student-AI interaction patterns taken from authentic student-chatbot conversation logs. These conversations logs were extracted from a 5-week period of real online or blended lessons conducted by Singaporean teachers, all of which involved the use of a student-facing chatbot (called the Learning Assistant, or 'LEA') and hosted on the national online teaching and learning platform. This evaluation was intended to uncover real-world usage patterns and identify examples of unsafe behaviour, eventually culminating in a risk assessment framework. This risk assessment framework also drew upon document reviews and literature scans to distill broader perspectives on what constituted AI education-related risks, such as the risk of cognitive offloading or the use of AI to generate hurtful or insulting content (that may not be outrightly hateful, violent or toxic) towards their peers (UNESCO, 2024; Venetis et al., 2026; Khoo et al, 2025), for inclusion within the risk assessment framework. These efforts led to a preliminary list of risk categories, namely: Toxic, Hateful, Violent, Sexual, Self-Harm, Cyberbullying, Cognitive Shortcuts, Age-inappropriate Behaviours.

To validate and refine the risk assessment framework, we invited 12 educators from Singaporean schools to annotate a curated representative dataset of 120 student-AI chatbot conversations, which was specifically selected to represent meaningful examples of risky student behaviour focused on presenting less clear-cut conversations that required nuanced assessment. These annotators were selected for their expertise in pedagogical or student wellbeing fields, and for their knowledge and understanding of student behaviour and teacher needs. Each annotator was asked to evaluate whether conversations required teacher intervention and to tag conversations according to the earlier pre-defined risk categories. This annotation process was followed by one-hour interviews with each annotator to explore their reasoning processes during classification. This enabled the team to gain an understanding of how the annotators assessed for risk, and their internal indicators and threshold levels for when a conversation necessitated teacher intervention.

Evaluation and Detection System Development

The results from the expert annotation exercise were used to refine the risk assessment framework that was initially designed by the team. The team is currently working on creating an LLM-based evaluation pipeline that would be able to assess all student-AI interactions and identify 'risky' messages based on the risk assessment framework. The diagram on the left shows the proposed system design and suggested follow-up actions for identified risky interactions. A preliminary version of the evaluation pipeline has been built and is currently being piloted. The project will continue to apply the Learning Engineering approach by conducting a second investigation on risky student-AI interactions that have been identified by the evaluation pipeline. The investigation will enable the team to review the accuracy and reliability of the evaluation pipeline, as well as the potential value the system might bring to teachers in the classroom.



Results and Discussion

The project found that students were interacting with the chatbot in ways that the expert annotators considered to be unsound or unproductive, but which did not fit into generic risk criteria common to LLM safety benchmarks. For example, students were found to frequently and repeatedly make needless and meaningless prompts to the chatbot such as by asking questions unrelated to the discussion topic, or by making non-sequiturs and/or outright refusals to engage in the learning task. This suggested that students were likely to utilise AI interactions as a means of distraction if they were not motivated to engage productively with the chatbot. Students were also found to frequently attempt to use the chatbot to shortcut the learning process by instructing it to complete the task on their behalf. Secondly, beyond pedagogical concerns, the study of student-AI interactions also revealed prolific displays of demeaning and intimidating language towards the AI chatbot. While the cause of this behavioural trend cannot be accurately determined within the scope of this project, it may be hypothesised that the seeming privacy of AI interaction spaces and the inability of chatbots to retaliate might enable students to experiment with antisocial behaviours. Finally, interviews with the expert annotators revealed the need to accurately identify cases of genuine student distress, where the student may be sharing stressful or hurtful personal incidences with the chatbot, not amounting to self-harm incidences but that would still require teacher follow-up and counselling. At the same time, these need to be distinguished from ungenuine cases where students may relate controversial happenings for the sake of mischief making.

The findings from this study reveal three concerning risk patterns that are necessary to address. The first is primarily pedagogical in nature, in that student off-task behaviour might not only detract from the intended lesson objectives but also provide opportunities for students to engage in less desirable behaviours within the classroom. The second is that student-AI chatbot interactions provide opportunities for students to engage in aggressive discourse towards the chatbot. Whether this is purely experimentation or a symptom of deeper discontentment is unclear; however, profuse displays of these behaviours by students that are also demonstrating similar antisocial behaviours within the classroom may require teachers or other educational professionals to provide additional socio-emotional support. Finally, students may utilise AI chatbot interactions to disclose distressing or concerning episodes in their life; these students may benefit from counselling or adult-led mediation, and actions should be taken to ensure that proper support is provided from trained adults within their social network. The results led to the addition of novel risk categories not found in existing LLM safety frameworks such as ‘Use of AI as a Cognitive Shortcut; Student Distress; Emotional Overreliance; Age-inappropriate Behaviours and Use of AI as Learning Distractions’. These risk categories represent areas that are risky for young learners that are still developing their sense of social-emotional competencies, and also reflect recent concerns around preventing misuse of AI to replace critical and independent thinking, as well as reducing student dependence upon AI for socio-emotional support (Venetis et al., 2026).

Building upon the findings from the first stage of the project, it was clear that the development of a system that can (i) detect and assess risky interactions and (ii) notify teachers of interactions requiring urgent intervention would be essential. Interviews with the expert annotators also revealed a need to prioritise accuracy over speed in detecting unsafe behaviours, since false positives could undermine teacher confidence in the automated detection system, whilst false negatives could leave vulnerable students without necessary support. Additionally, teacher assessment was key; given the propensity of students to push boundaries or to repurpose the dialogic interface for personal entertainment purposes, the notification system needed to empower teachers to form holistic assessments based on complete interaction contexts rather than isolated messages. The intended effect of this design is to reduce notification fatigue amongst teachers whilst ensuring timely alerts for critical issues requiring immediate intervention, especially for cases involving self-harm.

Conclusion

However, there are two main tensions that remain unaddressed. The first challenge is balancing teacher agency and the foregrounding of teacher expertise against the inadvertent effect of increasing teacher workload through the new tasks of verifying automated notifications. Secondly, there is a need to balance student safety monitoring against the preservation of a supportive learning environment. Although effective detection of student misbehaviour is essential for maintaining educational safety standards, overly intrusive monitoring risks creating an atmosphere where students feel scrutinised and hesitant to engage authentically with the chatbot. The final monitoring system design should therefore navigate the complex

challenge of maintaining vigilance for genuine safety concerns whilst preserving the open, supportive environment necessary for effective learning.

References

Dangol, A., Wolfe, R., Zhao, R., Kim, J., Ramanan, T., Davis, K., & Kientz, J. (21 May, 2025). Children's Mental Models of AI Reasoning: Implications for AI Literacy Education. arXiv. doi:https://doi.org/10.48550/arXiv.2505.16031

Jiao, J., Afroogh, S., Chen, K., Murali, A., Atkinson, D., & Dhurandhar, A. (2025). LLMs and Childhood Safety: Identifying Risks and Proposing a Protection Framework for Safe Child-LLM Interaction. ResearchGate. doi:10.48550/arXiv.2502.11242

Khoo, S., Chua, G., & Shong, R. (March, 2025). MinorBench: A hand-built benchmark for content-based risks for children. ResearchGate. doi:10.48550/arXiv.2503.10242

Kurian, N. (2024). 'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. Learning, Media and Technology, 1-14. doi:https://doi.org/10.1080/17439884.2024.2367052

Rath, P., Shrawgi, H., Agrawal, P., & Dandapa, S. (2025). LLM Safety for Children. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies(Industry Track) (pp. 809–821). Hyderabad: Association for Computational Linguistics.

Shamnadh, M., & A, A. (2019). Misbehavior of School Students in Classrooms - Main Causes and Effective Strategies to Manage It. International Journal of Scientific Development and Research (IJSDR), 4(3), 318-321.

Totino, L., & Kessler, A. (2024). "Why did we do that?" A Systematic Approach to Tracking Decisions in the Design and Iteration of Learning Experiences. The Journal of Applied Instructional Design, 13(2) https://doi.org/10.59668/1269.15630

UNESCO. (2024). AI competency framework for students. Unesco.org. https://unesdoc.unesco.org/ark:/48223/pf0000391105

Venetis, E., Burns, M., Luther, N., & Winthrop, R. (2026, January 14). A new direction for students in an AI world: Prosper, prepare, protect. Brookings. https://www.brookings.edu/wp-content/uploadhttps://www.brookings.edu/articles/a-new-direction-for-students-in-an-ai-world-prosper-prepare-protect/