Proceedings of Machine Learning Research 339:123–128, 2026 IRAISE Workshop at the Festival of Learning 2026

When Learning Signals Become Safety Signals: A Bounded-Confidentiality Framework for Educational AI Agents

Hanjing Shi HASA23@LEHIGH.EDU and **Dominic DiFranzo** DJD219@LEHIGH.EDU
Lehigh University

Abstract

Educational AI agents are usually introduced as tutors, homework helpers, or study companions. As children use them regularly, learning help can become a disclosure site: a student may begin with a math problem and then describe bullying, fear of adult reaction, or school avoidance. This position paper asks how child-facing educational agents should preserve student trust while deciding what stays private, what can be shared under student control, and what must be routed to accountable adults. Full parental visibility would turn help-seeking into surveillance; full secrecy would leave serious harm unsupported. We argue for bounded confidentiality as a communication-governance framework: ordinary tutoring remains private unless student choice or credible danger creates a reason to involve adults. The paper specifies a response ladder for minimal records, student-reviewed communication, care-oriented adult prompts, and escalation to locally accountable humans.

Keywords: educational AI, child safety, bounded confidentiality, privacy, safeguarding, AI agents

1. Introduction

AI learning assistants are moving from occasional tools to persistent companions in children's learning lives. A student might ask for help with a math worksheet or ask an agent to explain why an answer is wrong. If the interaction stays academic, the design problem is already hard enough: the agent must support learning without doing the work for the student. But children do not always keep their lives inside school-subject boundaries. Academic struggle, school avoidance, and social harm can surface in the same learning interaction, making disclosure part of help-seeking rather than a separate mental-health feature.

When a homework exchange turns into a disclosure about school avoidance, peer harm, or fear of adult reaction, the agent is no longer handling only a learning signal. It is handling a communication boundary: the student may need adult support, yet may also fear that disclosure will be punished, exposed, or made worse. The agent's role is limited to learning support, communication scaffolding, and safe routing when danger is credible.

This paper addresses a narrow gap: the agent needs a bounded way to mediate between student trust, adult care, and safeguarding duties. The problem is not that schools lack adults who can respond. It is that a child's disclosure may pass through a system designed as a tutor before it reaches any adult at all. Bounded confidentiality names the design layer between private help seeking and accountable human response. Teachers, counselors, parents, and district procedures remain the accountable human system around the agent.

The contribution is a design construct rather than a clinical model. We treat child-facing educational AI as a learning-safety disclosure site and use bounded confidentiality to connect disclosure levels with AI actions, record retention, sharing rules, and human roles.

2. Adjacent work and gap

Adjacent work leaves this boundary unresolved. AI tutors focus on academic help. Academic help seeking relates to broader learning strategies and instrumental achievement behavior (Karabenick and Knapp, 1991). Avoiding help can be shaped by perceived threat, competence, and classroom norms (Ryan et al., 2001). Interactive learning-environment research adds that help design affects whether students seek support productively (Aleven et al., 2003). FamLearn shows that large language models can mediate caregiver coordination in everyday learning (Li et al., 2025). Family-school partnership work treats parent engagement as capacity-building rather than one-way information delivery (Mapp and Kuttner, 2013). These lines of work support parent-facing care prompts, but not the case in which a child wants help without giving parents the full transcript. Child-threat chatbot work shows why children may want conversational support when revealing an online threat to an adult feels difficult (Piccolo et al., 2021). AI companion safety work further cautions against simple keyword escalation because parents and experts judge youth risk differently (Yu et al., 2025). Policy reports identify mandated reporting, student privacy, pedagogy, and benchmarking as research areas for AI companions in education (EDSAFE AI Alliance, 2026), but the educational-agent ladder from private help to supported disclosure and escalation remains under-specified.

School surveillance systems mark the line this paper avoids. AI-powered monitoring tools scan student accounts and devices; reporting by the Associated Press documented privacy failures, sensitive unredacted student documents, and false alarms that sometimes led to disciplinary or law-enforcement responses (Lurye, 2025). A bounded-confidentiality agent works from ordinary learning interactions, student disclosures, and school support processes rather than broad device, file, or social-media monitoring. MTSS is a plausible institutional home because it already connects academic, behavioral, and social-emotional support around school teams (California Department of Education, n.d.). Educational AI adds student self-disclosure and learning interaction traces only as limited support signals and communication scaffolds.

3. Policy and practice constraints

In U.S. school contexts, law and policy shape these design constraints. FERPA gives parents, and later eligible students, rights to inspect education records (U.S. Department of Education, n.d.a). A Department of Education FAQ also says schools need not create or maintain records unless an inspection request is already pending (U.S. Department of Education, n.d.c). Raw AI-child transcripts should therefore not become the default record of sensitive disclosures. FTC guidance has treated schools as able to act as intermediaries for COPPA notice and consent when educational uses are for the school's benefit and not other commercial purposes, while also stressing minimization and retention limits (Federal Trade Commission, n.d.). Sensitive disclosures should stay out of advertising, engagement optimization, and broad student profiling.

PPRA adds a boundary around surveys, analyses, or evaluations concerning protected areas such as mental or psychological problems, sex behavior or attitudes, self-incriminating behavior, and family relationships (U.S. Department of Education, n.d.b). An agent that asks broad mental-health or family questions may no longer be acting like a tutor. Safety-check questions should be minimal, non-leading, triggered by the student's own disclosure, and governed by district policy. School counseling ethics already use bounded confidentiality: the primary obligation is to the student, but law, school policy, parental rights, and serious foreseeable harm can limit confidentiality (American School Counselor Association, 2024). AI can borrow that shape without pretending to be a counselor.

4. Bounded learning-safety mediation

To make bounded confidentiality operational, the system has to be narrow about what it treats as evidence. It should rely on ordinary educational use, student-initiated disclosure, existing school support data, or deliberately entered human observations. It should leave out social media scraping, device-wide monitoring, private-file scanning, and always-on sentiment analysis. Student disclosures, learning interaction signals, school operational signals, and lightweight human observations can all matter, but they do not mean the same thing. A child saying "please do not tell my parents" is not the same as a missing assignment. A sudden shift from active participation to avoidance may matter, but it remains a weak support signal, not proof of bullying or anxiety. Attendance and course performance can help a support team notice patterns across classes, while bullying or family harm remain matters for human follow-up. Human observations should be observable facts, such as "missed two assignments," not character judgments.

The agent should produce a support hypothesis with uncertainty and a next safe communication step: "The student disclosed school avoidance and has recent missing assignments. Cause unknown. A non-disciplinary check-in may be appropriate." MTSS placement, risk diagnosis, and legal reporting thresholds remain human and institutional decisions. Before deployment, schools should name level recipients, false-alert reviewers, and parent-involvement rules for student-chosen, required, or potentially harmful cases; these are deployment policy, not runtime improvisation. When parent involvement is contested because the student refuses it, a designated school official believes policy requires parent contact, or a parent may be part of the risk, the agent should route a minimal summary to that human reviewer rather than send a parent prompt or promise silence.

Initial level assignment should be conservative and reviewable. Table 1 sketches a four-level ladder, from private homework confusion to immediate safety escalation, but the table is not meant to automate judgment. The hard cases sit between levels, where a student wants help but fears exposure, or where repeated avoidance suggests that ordinary learning support may no longer be enough. The main error modes are different: under-escalation can delay protection, while over-escalation can break trust, expose private information, or punish help seeking. A deployable system therefore needs ambiguity handling: brief non-leading safety questions, visible uncertainty, named human review for borderline cases, and audits of false alerts and missed alerts after use.

By slowing the path from a child's words to an adult-facing record, the ladder keeps support from becoming either surveillance or unsafe secrecy. Routine help should not create a transcript for inspection. When sharing is appropriate, the system should favor student-reviewed summaries and care prompts over evidence packets. Escalation still belongs to locally named humans, and the agent should reflect without labeling, route without adjudicating, and retain only what the triage task requires.

Table 1: Bounded learning-safety mediation response ladder.

Level	Signal	AI response	Record retained	Sharing rule	Human role
1. Private	Homework confusion.	Tutor and scaffold.	Learning tag; no transcript.	No sharing by default.	None.
2. Student choice	Fear of parent reaction.	Prepare message for student preview.	Local note unless sent.	Student chooses.	Chosen adult.
3. Supported	Peer harm and school avoidance.	Non-leading safety check; offer summary.	Minimal summary; no transcript.	Student-involved when safe.	Counselor or support team.
4. Escalation	Self-harm, sexual harm, adult abuse, or immediate danger.	Explain limit; escalate.	Minimum safety record.	Designated process.	Safeguarding role.

5. Scenario

Consider a homework-support agent in a middle school math class. A student who usually finishes practice begins leaving multi-step problems blank, asks for repeated hints, abandons the task, and writes: "I do not want to go to school. They keep making comments about me. Please do not tell my parents." The episode begins as ordinary academic support: unfinished work, repeated hints, and a task left behind. The final sentence changes the situation without making it a complete safety report. It points to school avoidance and possible peer harassment, but not whether there is immediate danger or what kind of adult help the student would accept.

Two simple responses both fail the student. An immediate transcript alert treats a hesitant disclosure as an evidence packet and may teach the student that asking for help means losing control over the story. Saying nothing to anyone is not enough either, because the student may still need adult support before the situation gets worse.

A bounded-confidentiality design first responds to the student, not the dashboard: "I can keep ordinary homework help private, but if someone is in immediate danger I need to involve a trusted adult. You do not have to share details. Are you safe right now? Has anyone threatened you today?" The agent does not ask for names, screenshots, or a full account before the safety check; those details belong in human follow-up if the student chooses or policy requires it.

If no immediate danger is reported, the agent asks the student to choose a trusted adult, previews a counselor message, and logs only the disclosure category unless the student sends it. A counselor-facing summary might say: "Student reports repeated peer comments and school avoidance; no immediate threat reported; student wants help choosing an adult." Parent contact follows a separate path: by student choice or designated reviewer decision, the agent gives bounded support context without transcript access: listen first, avoid punishment, ask what would make school feel safer, and contact the counselor together when the student agrees.

If the student discloses self-harm, sexual harm, adult abuse, or immediate danger, the workflow shifts to the local safety process: minimum necessary information goes to locally designated humans, with the student told what is happening when possible.

6. Contribution and next steps

The ladder is a design and evaluation starting point. Before deployment, students should be able to explain what the agent can keep private, when sharing is optional, and when safety limits apply. Counselors and support teams need to judge whether minimal summaries are specific enough for a non-disciplinary check-in without requiring raw transcripts. Parent groups should review care prompts for tone, usefulness, and risk of punitive response. Incident audits should track false positives, false negatives, transcript exposure, delayed intervention, and whether students avoid the system after a disclosure.

The framework also needs cross-context testing. GDPR-style regimes, jurisdictions with different mandated-reporting thresholds, and schools with little counseling infrastructure change who can receive a summary and what counts as a support pathway. In low-resource settings, the answer may be a smaller ladder with clearer referral roles, not more automation. Future work should therefore include students, parents, school counselors, health-care professionals, and education policy experts in deciding the level boundaries and the acceptable records for each boundary.

References

American School Counselor Association. The School Counselor and Confidentiality. https://schoolcounselor.org/Standards-Positions/Position-Statements/ASCA-Position-Statements/The-School-Counselor-and-Confidentiality, 2024. Accessed 2026-04-26.

California Department of Education. Multi-Tiered System of Supports. https://www.cde.ca.gov/CI/CR/RI/index.asp, n.d. Accessed 2026-04-26.

Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73(3):277–320, 2003. doi: 10.3102/00346543073003277.

EDSAFE AI Alliance. S.A.F.E. by Design: Policy, Research, and Practice Recommendations for AI Companions in Education. https://www.edsafeai.org/safeaichatbots, 2026. Accessed 2026-04-26.

Federal Trade Commission. Complying with COPPA: Frequently Asked Questions. https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions, n.d. Accessed 2026-04-26.

Stuart A. Karabenick and John R. Knapp. Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. *Journal of Educational Psychology*, 83(2):221–230, 1991. doi: 10.1037/0022-0663.83.2.221.

Yao Li, Jingyi Xie, Ya-Fang Ling, He Zhang, Ge Wang, Gaojian Huang, Rui Yu, and Si Chen. "Learning Together": AI-Mediated Support for Parental Involvement in Everyday Learning, 2025. URL https://arxiv.org/abs/2510.20123.

Sharon Lurye. Takeaways from Our Investigation on AI-Powered School Surveillance. https://apnews.com/article/381fa82978f27eb85f20d03236820711, 2025. Associated Press.

Karen L. Mapp and Paul J. Kuttner. Partners in Education: A Dual Capacity-Building Framework for Family-School Partnerships. Technical report, SEDL, 2013. https://eric.ed.gov/?id=ED593896. Accessed 2026-04-26.

Lara S. G. Piccolo, Pinelopi Troullinou, and Harith Alani. Chatbots to support children in coping with online threats: Socio-technical requirements. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1502–1521. ACM, 2021. doi: 10.1145/3461778.3462114.

Allison M. Ryan, Paul R. Pintrich, and Carol Midgley. Avoiding seeking help in the classroom: Who and why? *Educational Psychology Review*, 13:93–114, 2001. doi: 10.1023/A:1009013420053.

U.S. Department of Education. Family Educational Rights and Privacy Act (FERPA). https://studentprivacy.ed.gov/faq/what-ferpa, n.d.a. Accessed 2026-04-26.

U.S. Department of Education. What Is the Protection of Pupil Rights Amendment (PPRA)? https://studentprivacy.ed.gov/faq/what-protection-pupil-rights-amendment-ppra, n.d.b. Accessed 2026-04-26.

U.S. Department of Education. Does an Educational Agency or Institution Have Discretion Over What Education Records It Decides to Create and Keep? https://studentprivacy.ed.gov/faq/does-educational-agency-or-institution-have-discretion-over-what-education-records-it-decides, n.d.c. Accessed 2026-04-26.

Yaman Yu, Mohi, Aishi Debroy, Xin Cao, Karen Rudolph, and Yang Wang. Principles of Safe AI Companions for Youth: Parent and Expert Perspectives, 2025. URL https://arxiv.org/abs/2510.11185.