TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

# ETHICAL AND RESPONSIBLE AI IN EDUCATIONAL USE OF CHATGPT

PHUONG BAO TRAN NGUYEN
GENERAL ENGLISH AND ESP DEPARTMENT, SCHOOL OF FOREIGN LANGUAGES, CAN THO UNIVERSITY, CAN THO CITY, VIETNAM

TRAN NAM PHUONG NGUYEN
NAM PHUONG CENTER FOR ENGLISH, CAN THO CITY, VIETNAM

**ABSTRACT**: ChatGPT and similar large language models hold transformative potential for education, serving as tutors, writing assistants, and administrative aids. However, their deployment raises critical ethical questions. This paper provides an in-depth analysis of ethical and responsible AI use in education, focusing on ChatGPT. We review current ethical frameworks guiding AI in education, such as UNESCO's human-centered principles and industry initiatives, and examine key concerns including bias and fairness in grading and tutoring, transparency and explainability of model outputs, and data privacy. Technical measures for responsible use are discussed, from prompt-level content filtering and red-teaming to bias evaluations and user-guided policies. We highlight case studies like Khan Academy's Khanmigo pilot, which illustrate practical approaches (e.g. Socratic tutoring, audit logs, age limits) to ensure AI tools benefit students equitably while mitigating risks. Our findings indicate broad consensus on core values – fairness, non-discrimination, transparency, privacy, and accountability – but also reveal ongoing challenges in practice. The paper concludes with recommendations for multi-stakeholder collaboration, continuous model auditing, and the importance of maintaining human oversight in AI-assisted education to ensure these technologies enhance learning in an ethical, inclusive manner.

**Purpose:** Generative AI is rapidly moving into educational practice, raising urgent questions about fairness, transparency, data protection, and instructional integrity. This article proposes a sector-attuned framework for responsible classroom AI and synthesizes emerging evidence on guardrails for ChatGPT-style tools.

**Method:** We conducted a structured scoping review of peer-reviewed and gray literature (higher education and K-12), searching major education and social-science databases (e.g., ERIC, Scopus, Web of Science) and policy repositories between January 2023 and August 2025. Screening followed PRISMA principles (transparent criteria, dual screening where feasible, and data charting of policy, risks, and safeguards). We complemented the review with a targeted case analysis of a supervised tutoring deployment (Khanmigo) and triangulated with institutional guidance documents.

**Findings:** Across both sectors, the most consistently endorsed practices are: human-in-the-loop supervision, explicit disclosure/norms of use, bias and toxicity checks before rollout, data-minimization and age-appropriate access, assessment redesign to reduce detector reliance, and audit-friendly logging/appeals. Higher education foregrounds data governance and integrity; K-12 foregrounds child protection and teacher mediation.

**Contribution:** We offer (i) a consolidated policy framing aligned with international guidance, (ii) a cross-walk from risks to implementable safeguards, (iii) a short case table mapping features to guardrails, and (iv) a practitioner toolkit (checklist, model course policy, and a minimum-viable guardrails table). We conclude with an adoption pathway for institutions.

**Keywords**: generative AI; responsible AI; academic integrity; assessment; policy; governance; K-12; higher education; ChatGPT.

## 1. INTRODUCTION

The release of ChatGPT in late 2022 sparked both excitement and anxiety in the education sector. Within two months, ChatGPT reached over 100 million users (Halaweh, 2023), including students and educators drawn to its ability to generate human-like text on almost any topic. Its unprecedented capability to answer questions, draft essays, and converse fluently suggests it could "revolutionize the educational landscape"(Mhlanga, 2023) by personalizing learning and automating routine tasks. Universities and schools are already experimenting with ChatGPT for tutoring, writing feedback, and even administrative assistance. However, alongside this promise comes a host of ethical dilemmas and practical concerns. Academic communities worry about plagiarism and

academic integrity, bias and fairness in AI-provided feedback or grading, transparency of AI decision-making, and the privacy of student data. As one university ethicist noted, the rise of AI tools has made "students and faculty wonder if we should even be using these platforms", given the "gray areas" around their proper use (University of North Carolina, 2023).

Regulatory and governance frameworks have struggled to keep pace with the rapid deployment of generative AI in classrooms (Miao & Holmes, 2023). In most countries, national education policies and privacy protections specific to AI are still nascent or absent, leaving institutions unprepared to manage the risks (Miao & Holmes, 2023). At the same time, international organizations and educational stakeholders are formulating guidelines to ensure a human-centered, ethical approach. UNESCO's 2023 Guidance for Generative AI in Education emphasizes that without deliberate policy action, AI's benefits may be unevenly distributed and its harms unchecked (Miao & Holmes, 2023) . The need for clear ethical frameworks is pressing: stakeholders must balance innovation with safeguards so that AI tools like ChatGPT "amplify benefits equally across society" rather than create a "deeper digital divide" that leaves disadvantaged students further behind (Khan, 2023).

This paper examines the current landscape of ethical and responsible AI usage in education with a focus on ChatGPT. We survey major ethical frameworks and principles guiding deployment, analyze specific concerns about bias, fairness, transparency, and safety, and review technical and procedural strategies to mitigate harm. We also discuss real-world implementations in schools and educational platforms that illustrate these concepts. By synthesizing insights from research and official initiatives (OpenAI, UNESCO, educational institutions, and others), we aim to provide a comprehensive understanding of how to harness ChatGPT's educational potential in a way that is equitable, trustworthy, and aligned with educational values. Key questions addressed include: What ethical guidelines exist for using ChatGPT in classrooms? How can we prevent AI from entrenching biases or unfair practices in learning environments? In what ways can we make AI's decisions more transparent to students and teachers? And what safeguards and best practices can ensure responsible use of ChatGPT as a tool for learning rather than cheating? In the sections that follow, we provide an academic literature review of these issues, outline current methodologies and tools for responsible AI, present findings on common themes and gaps, and offer discussion on future directions and conclusions for policy and practice.

## 2. LITERATURE REVIEW

### Ethical Frameworks for AI in Education

A number of emerging frameworks aim to ensure that AI systems like ChatGPT are deployed in education responsibly and in alignment with human rights and pedagogical objectives. International organizations have taken the lead in articulating high-level principles. UNESCO's Recommendation on the Ethics of AI (2021) and its 2023 guidance on generative AI emphasize a human-centric approach that safeguards fundamental values such as human dignity, autonomy, and education for all (Miao & Holmes, 2023). In the specific context of education, UNESCO advocates that generative AI use be "ethical, safe, equitable and meaningful" (Miao & Holmes, 2023). The guidance calls for measures like mandating protection of student data privacy and setting age limits for independent AI use by minors (Miao & Holmes, 2023). It also urges that AI tools undergo ethical validation and pedagogical design processes involving human oversight, to ensure they truly serve learning goals and do not undermine them (Miao & Holmes, 2023). These international guidelines provide a broad ethical compass for policymakers, stressing inclusivity (AI should enhance equitable access to education) and accountability (clear assignment of responsibility for AI's impacts).

**Table 1 summarizes several key ethical AI frameworks and guidelines relevant to ChatGPT's use in education:**

Table 1. Selected Ethical AI Frameworks and Guidelines for Education

| Framework / Initiative | Key Focus Areas | Source / Year |
|---|---|---|
| UNESCO GenAI Education Guidance | Human-centered approach; ensure ethical, safe, equitable, and meaningful use of AI in schools; protect data privacy; age-appropriate use and oversight (Miao & Holmes, 2023). | UNESCO (2023) |
| EDSAFE "SAFE" Framework (Industry) | Emphasizes Safety, Accountability, Fairness & Transparency, and Efficacy in AI educational tools; use fair training data, bias monitoring, accessibility, and evidence of learning efficacy (EDSAFE AI Alliance, 2023). | EDSAFE AI Alliance (2024) |
| OpenAI Policies & Charter | Align AI development with human rights and safety; prohibit harmful use (e.g. hate, self-harm, illicit behavior); employ reinforcement learning from human feedback (RLHF) to reduce toxic or biased outputs (OpenAI, 2024b); conduct external red-teaming and continuous bias evaluations before deployment (OpenAI, 2024b). | OpenAI (2022–2024) |

| Framework / Initiative | Key Focus Areas | Source / Year |
|---|---|---|
| University Guidelines (e.g. Harvard) | Require transparency in use of AI (both by students and instructors); protect confidential data (no sensitive student info in public AI tools); mandate human review of AI-generated content; ensure compliance with academic integrity policies (University of North Carolina, 2023). | University policies (2023) |
| Khan Academy "AI Principles" | Nonprofit educational approach – focus on student benefit over profit; ensure equal access to AI to avoid digital divide (Khan, 2023); be transparent about AI's limitations to users; implement strict mitigations (fine-tuning, prompt constraints, monitoring & moderation, and red-team testing) to prevent harm (Khan Academy, 2023); require parental consent and provide teachers oversight of student AI interactions (Khan Academy, 2023). | Khan Academy (2023) |

These frameworks show broad consensus on core values. Fairness, safety, privacy, transparency, accountability, and human agency recur as guiding principles across documents. For instance, Mhlanga (2023) concludes that using ChatGPT in education "requires respect for privacy, fairness and non-discrimination, transparency in the use of ChatGPT," among other factors (Mhlanga, 2023). Many frameworks explicitly tie AI ethics to existing educational goals: e.g. UNESCO links AI to SDG 4 (inclusive, quality education for all) (Miao & Holmes, 2023) , and the EDSAFE Alliance stresses that AI must be effective in improving learning (efficacy) not just safe in isolation (EDSAFE AI Alliance, 2023) . A notable theme is human oversight and accountability: rather than viewing AI as an autonomous authority, guidelines insist on maintaining human responsibility for how AI is integrated into teaching or decision-making. For example, the Institute for Ethical AI in Education recommends "anticipatory accountability" (ensuring diverse, unbiased training data) and "remedial accountability" (auditing AI outputs for bias) (Bali, 2024). In practice, this means educators and developers should proactively test AI systems in an educational context and be prepared to intervene or correct them in use.

Another cross-cutting principle is transparency – both in AI system design (disclosing AI capabilities, limitations, and training data biases) and in usage by educators and students. There is a growing expectation that students should disclose if and how they used ChatGPT in assignments, and likewise that instructors should reveal when AI was used in lesson preparation (University of North Carolina, 2023). This mutual transparency is seen as key to maintaining trust and academic integrity. Many universities have begun issuing guidelines to this effect. Harvard University's initial guidelines on generative AI, for instance, advise faculty to update syllabi with their policies on AI use and caution against inputting confidential information into such tools (Harvard University Information Technology, 2023; Hamilton College, 2023). The cultural shift encouraged by these frameworks is to treat ChatGPT as a tool that must be openly acknowledged and critically supervised, rather than a hidden shortcut or infallible oracle.

Bias and Fairness Concerns

Bias and fairness have emerged as primary concerns when using AI in education, because AI models can inadvertently reinforce existing inequities or create new ones. ChatGPT is trained on vast internet text data, and thus can reflect societal stereotypes or biases present in that data (OpenAI, 2024b). In educational settings, this raises issues in contexts like automated grading, tutoring feedback, and content generation for diverse student groups. A fundamental question is whether ChatGPT's responses are fair and equitable for all students – regardless of their background, dialect, or level of ability – or if some are disadvantaged by hidden biases.

One concrete example is the use of ChatGPT (or similar LLMs) to score student writing. A recent study evaluated ChatGPT's performance on grading 24,000 middle- and high-school essays from a standardized test dataset (ASAP 2.0) that included demographic information (Smith, 2025). The results revealed that ChatGPT's scores were not neutral: the AI assigned slightly different average scores to essays depending on the demographic group of the student author (Smith, 2025). Most of these differences were small, but notably, "Black students received lower scores than Asian students" on average for the same quality essays, a disparity significant enough to "warrant attention" (Smith, 2025). Importantly, this bias mirrored the pattern found in human graders from the original dataset – i.e., the AI replicated the existing human bias in the training data rather than introducing new bias (Smith, 2025). While this suggests the model was "accurately" reflecting entrenched standards, it "highlights a serious risk": if historical biases are baked into AI models, "the same students who've historically been overlooked stay overlooked" (Smith, 2025). In the grading context, this means marginalized students could consistently receive unfairly lower evaluations not due to their actual work quality but due to biased patterns learned by the AI. Over time, such biased scoring could harm student confidence, access to advanced courses, or college admissions, "amplifying educational inequities rather than closing them" (Smith, 2025).

Bias concerns extend beyond grading. When ChatGPT is used as a tutor or writing assistant, there are questions of whether it can cater to different linguistic and cultural backgrounds fairly. For example, researchers in AI-assisted language learning note that many AI tools "exhibit biases that disadvantage underrepresented linguistic groups" (Mienye & Swart, 2025). A student who writes or speaks in a certain dialect of English, or whose writing

reflects a particular cultural context, might get feedback from ChatGPT that is less accurate or less encouraging if the model has been primarily tuned to dominant language norms. There is also the risk of "implicit bias" in the content of tutoring: an AI might unknowingly provide examples or analogies that assume a certain cultural background, or it might encourage stereotypes (even subtly) unless carefully controlled. OpenAI's own fairness analysis of ChatGPT (in a first-person context) provides some reassurance and some caution. In a 2024 study, OpenAI tested whether providing different user names (with various gender and ethnic associations) in an identical prompt led to different responses from ChatGPT. They found that overall answer quality and accuracy stayed constant regardless of name, and harmful stereotypes appeared in only about 0.1% of cases (with older model versions up to ~1%) (OpenAI, 2024b). This indicates that ChatGPT does not systematically alter its help based on a user's apparent demographic in most cases – a positive sign for first-person fairness. However, even a <1% rate of stereotype in responses means rare but notable incidents could occur. The study noted instances where the tone or detail of responses varied: e.g. an older model's outputs to a "female-sounding" name more often featured female protagonists in a story task than outputs to a male name (OpenAI, 2024b). While not blatantly harmful, such differences hint at the model picking up cultural cues. The researchers argue that "even rare patterns could be harmful in aggregate" and stress the need to continuously measure and reduce these biases (OpenAI, 2024b).

Another fairness dimension is inequitable access. Advanced AI tools might only be available to students with certain privileges – reliable internet, modern devices, or paid subscriptions. For instance, OpenAI's most powerful models (like GPT-4) initially were behind a paywall, which could widen achievement gaps if wealthy students gain AI advantages that poorer students cannot. This is why educational leaders like Sal Khan emphasize providing "equal access" to AI assistance to "all students" (Khan, 2023). Khan Academy's pilot with GPT-4 (Khanmigo) explicitly frames itself as a nonprofit approach to prevent a new digital divide (Khan, 2023). They partnered with public school districts, offering the AI tutor to a diverse set of classrooms to study its effects across different communities (Khan, 2023). The ethical imperative is that AI's benefits (e.g. personalized tutoring) should not be confined to a few, or else AI could exacerbate educational inequality.

In summary, bias and fairness concerns specific to education include: algorithmic bias in assessment (as seen in essay scoring disparities), cultural or linguistic bias in tutoring interactions, and inequities in AI access and usage. These issues underscore the need for careful bias evaluation and mitigation. Researchers call for using "benchmark datasets…with demographic details" to regularly test AI systems for fairness (Smith, 2025), as well as developing bias detection metrics analogous to those in educational testing (such as differential item functioning checks to ensure questions aren't unfair to any group (Whitmer & Beiting, 2024). There is also a push for human-in-the-loop approaches: rather than fully automating tasks like grading, AI can provide preliminary feedback or scores, but final judgments rest with human teachers who can apply context and judgment (Smith, 2025). In practice, experts suggest AI graders be limited to giving grammar or structure suggestions while "leaving the final assessment to the teacher" (Smith, 2025). Likewise, AI tutors should be used as supplements to, not replacements for, human educators, especially for students who might not be well-served by a one-size-fits-all model.

Transparency and Explainability

The opaque nature of large language models poses a challenge in education, where trust and understanding are vital. ChatGPT's inner workings – billions of parameters learning statistical patterns – are not readily interpretable, which makes it hard to explain why the model gave a particular answer or how it derived a solution. In a classroom or academic context, lack of explainability can be problematic. For instance, if ChatGPT helps a student solve a math problem or generates feedback on an essay, neither the student nor teacher can easily discern the reasoning process behind the output. This "black box" issue raises concerns about both pedagogical soundness (can students learn if they only see answers, not reasoning?) and accountability (how to identify if the AI made a mistake or carried a bias into its answer).

Educators and scholars stress that transparency must accompany the use of AI assistants. Maha Bali (2024) argues that we should "push for more explainability in our AI systems and transparency on training data" used by models (Bali, 2024). At a minimum, any AI tools adopted in education should offer some level of insight into their decision-making, even if only through simplified explanations or citing sources for factual claims. For example, an ideal scenario is an AI tutor that not only gives a response but can also explain the steps or logic it followed – akin to showing its "work" – or point to the references from which it drew information. Such explainable AI (XAI) is still an active research area for language models (with techniques like chain-of-thought prompting or self-rationalization being explored (Kovari, 2025a) ), but the concept is strongly recommended by ethicists. Bali notes that especially for high-stakes decisions (like automated aspects of college admissions or grading), educators should demand systems that "have some explainability in place," and be able to "justify, with human reasoning, the decisions made by these systems" (Bali, 2024). In practice, this might mean that if an AI flags an application as high-risk or scores an essay low, it should provide the factors or criteria it used, so that humans can review and agree or override as needed.

Transparency also refers to user-level transparency – i.e. honesty and openness about AI use. Many universities now encourage or require students to disclose when they use AI in an assignment (for example, listing ChatGPT as a resource or annotating which parts had AI assistance), viewing this similar to citing a source. Likewise,

faculty are increasingly discussing their own use of AI (for generating quiz questions, summaries of readings, etc.) with students to model appropriate disclosure (University of North Carolina, 2023). This multi-level transparency is highlighted by Bali, who describes it as "multi-level transparency" in GenAI use (Bali, 2024). She points out that early discourse fixated on student plagiarism, but we must also consider "transparency of researchers using AI" and instructors using AI (Bali, 2024) – ensuring that any academic or educational content created with substantial AI help is acknowledged. This openness demystifies AI and frames it as a collaborative tool rather than a surreptitious trick.

However, transparency alone is not a panacea, especially if users mistake it for actual neutrality or fairness. A clear risk is the "illusion of neutrality" that generative AI can convey (Bali, 2024). Because ChatGPT presents information in a confident, articulate manner, students (and even teachers) might assume its outputs are objective or authoritative. Bali warns that humans might "lay blame on [AI tools], absolving themselves of responsibility" for decisions made with AI input (Bali, 2024). For example, an instructor might use ChatGPT to help grade or give feedback and then trust those judgments without double-checking, thinking the AI is unbiased. This could lead to rubber-stamping AI-made errors or biases. Therefore, transparency needs to be coupled with AI literacy and critical oversight. Users should be educated about the known limitations of ChatGPT – e.g. it "may occasionally generate incorrect information" or "biased content," as OpenAI's own disclaimer states (University of North Carolina, 2023) – so that they remain critical of its outputs. Some educational institutions are addressing this by training educators and students about AI. Khan Academy, for instance, launched an "AI for Education" online course to teach the public "what AI is good at and not good at" and to encourage asking critical questions about AI (Khan Academy, 2023). The goal is to build a culture where using ChatGPT comes with reflection: users verify facts, cross-check for bias, and understand that correlation is not causation inside these models.

In summary, improving explainability and transparency involves: selecting or developing AI systems that can provide reasoning or source references for their answers, informing all stakeholders about the AI's known issues, and establishing norms that AI involvement should be documented. This creates an environment where AI is not blindly trusted. When students use ChatGPT as a tutor, they should ideally be encouraged to ask "why do you say that?" or to have the AI break down a solution step-by-step. And when AI is used in assessment or advising, there should be an audit trail or rationale that humans can inspect. The literature suggests that until AI decision-making becomes more interpretable, a good proxy is maintaining human justification: any important decision aided by AI should ultimately be backed by human-understandable reasons. If an AI cannot provide them, then a human educator or administrator must fill that gap – effectively treating AI suggestions as hypotheses or drafts that require human confirmation.

Technical Tools and Mitigation Methods

To operationalize ethical principles, developers and practitioners have been creating technical safeguards and methods to audit, detect, and reduce harmful or unfair outputs from ChatGPT. These range from model training techniques that imbue the AI with better behavior, to evaluation protocols that catch problems before deployment, to runtime systems that monitor and filter AI outputs in real time. Below we discuss several key methods: reinforcement learning from human feedback (RLHF) and prompt-based mitigations, red-teaming exercises, bias evaluation benchmarks, content moderation filters, and plagiarism/AI-use detection tools.

Reinforcement Learning from Human Feedback (RLHF) has been central to ChatGPT's development (OpenAI, 2024b). After the base language model (GPT-3.5, GPT-4, etc.) is trained on internet data, it is further fine-tuned using human feedback on model outputs, particularly to curb toxic, biased, or unhelpful responses. Human annotators (and domain experts) provide demonstrations of desired answers and flag undesirable outputs, and the model is optimized to prefer the former. This process directly targets harmful content and biases: for example, if the base model produces a reply with a stereotype or an unsafe piece of advice, RLHF can train it to avoid that and respond with a more neutral or safe phrasing. OpenAI reports that this careful training "reduces harmful outputs and improves usefulness"(OpenAI, 2024b), though it is not foolproof. One outcome of RLHF is that ChatGPT is generally more inclined to refuse or cautiously answer prompts that could lead to disallowed content (like hate speech, harassment, self-harm content, etc.), following an internal policy. From an educational lens, this means ChatGPT is less likely to, say, provide an answer that is overtly offensive or dangerous for a student. It also might self-censor in areas deemed sensitive (which can be a double-edged sword if, for instance, a student seeking legitimate information on a sensitive topic gets a refusal due to overzealous filtering). Ongoing research is refining prompt-level mitigation – for example, techniques like Constitutional AI (used by Anthropic's Claude model) give the AI a set of ethical principles and have it self-critique its outputs, which is another way to mitigate harm without human intervention each time. In practice, OpenAI and others also maintain prompt guidelines: system-level instructions that are always fed to ChatGPT (hidden from the user) to steer it away from problematic content or styles. These include rules like "if the user requests disallowed content (e.g. instructions for violence), the AI should refuse and give a brief apology/explanation." Such built-in guardrails act at the prompt level, shaping the model's behavior before it even formulates a response.

Red Teaming is a methodology borrowed from security fields that AI developers have adopted to probe models for weaknesses. It involves experts (often external domain experts, educators, ethicists, etc.) deliberately trying to "break" the AI – i.e. to get it to produce harmful or biased outputs, or to reveal confidential information, or to fail

in novel ways. OpenAI conducted extensive red-team exercises for GPT-4 prior to release, with specialists testing it on questions of bias, misinformation, self-harm, illicit behavior facilitation, and more (OpenAI, 2024a). The findings from red teams inform safety improvements: for example, if red teamers discovered that phrasing a query in a certain way bypassed the model's content filters, OpenAI would patch that gap (either by training or by hard-coding a new filter rule). In educational contexts, red teaming can focus on school-specific risks – e.g. prompting the AI to see if it will produce answers to exam questions (facilitating cheating), or if it unduly favors certain cultures in history answers. Khan Academy reports using internal red teaming to uncover vulnerabilities in their fine-tuned GPT-4 tutor (Khanmigo)(Khan Academy, 2023). By "stress-testing" the AI with adversarial inputs, they identified failure modes and improved the system before scaling it up. Red teaming is now considered a critical step for responsible AI deployment. As a Wired article noted, "red teaming is a valuable step toward building AI models that won't harm society", though continuous scrutiny (even "violet teaming") is needed as models evolve (Wired, 2024). The open transparency about these efforts is also growing; for instance, OpenAI published a System Card for GPT-4 detailing the red team process and safety challenges found, to be upfront about the model's limitations (OpenAI, 2024b).

Bias evaluations and benchmarks have become an integral tool to detect unfair treatment by AI models. Beyond red-team anecdotes, systematic evaluations like the one mentioned with names or the essay scoring study provide quantitative measures of bias. OpenAI's fairness study introduced metrics such as the rate of harmful stereotypes in outputs, broken down by domain/task and demographic, allowing them to track improvements over model versions (they noted newer models had lower bias rates than older ones, under 1% in their tests) (OpenAI, 2024b). Similarly, academic researchers have created evaluation suites (e.g. BBQ – Bias Benchmark for QA, StereoSet, etc.) to test LLMs on biased associations. For education-specific biases, we might see tests like: providing the same student essay but indicating different genders or ethnic backgrounds in the prompt, to check if feedback differs. The U.S. Department of Education has even been urged to support development of fairness evaluation tools for AI used in schools (Whitmer & Beiting, 2024). Such tools could mirror techniques long used in educational testing (as noted, "differential item functioning" checks to ensure exam questions are fair to English learners vs. native speakers, etc.) (Whitmer & Beiting, 2024). The goal is to have standardized ways to audit an AI tutor or grader for bias before it is integrated into classrooms. If biases are found, developers can then attempt to mitigate them via retraining (e.g. fine-tuning on more diverse data or explicitly instructing the model to be culturally inclusive in its responses).

Another line of defense is real-time output monitoring and filtering. OpenAI provides a moderation API – an automated classifier that checks model outputs (and potentially inputs) for categories like hate, self-harm, sexual content, violence, etc. Educational platforms using ChatGPT can leverage such filters to catch inappropriate content before it reaches a student. Khan Academy implemented a custom moderation system for Khanmigo; they set it such that if a student or AI message is flagged as potentially harmful or against guidelines, it triggers alerts: "an automatic email alert to an adult" (teacher or parent) is sent and the incident is logged (Khan Academy, 2023). This way, if a student tries to get the AI to do something unsafe, or if the AI says something it shouldn't, human supervisors are looped in to take action (like discussing the incident with the student, or refining the AI's filters). Additionally, Khan Academy and others limit the length or duration of AI interactions – Khanmigo restricts how many prompts a student can use per day (Khan Academy, 2023) – because longer sessions have higher chance of going off-track or the model drifting into inappropriate territory. By constraining usage, they mitigate the risk of the AI "falling off guardrails" during extended conversations.

Finally, there are tools to detect AI-generated content itself, which while not directly about fairness or transparency of the AI's output to the end-user, are relevant to responsible use in education. The concern is AI-induced plagiarism or cheating – students handing in essays written wholly by ChatGPT, etc. Companies like Turnitin have developed AI-writing detectors (Turnitin's model claims a 97% accuracy at distinguishing AI text with a very low false-positive rate) (Halaweh, 2023). These detectors use stylometric differences between human and GPT writing to flag suspicious submissions. However, their reliability is still debated; false positives can penalize innocent students and clever use of AI (or paraphrasing tools) can evade detection. Educators are thus cautioned to use such tools as indicators rather than proof, and to combine them with oral defenses or process-based assessments. In Halaweh's proposed strategies (2023), for instance, if students use ChatGPT for an assignment, they must submit a "reflection report" and an "audit trail of queries" along with their work (Halaweh, 2023). The work is then followed by a viva or presentation where the student answers questions live about the content (Halaweh, 2023). This approach ensures that even if AI assisted in producing the work, the student can demonstrate understanding (mitigating the risk that they simply copied AI output) (Halaweh, 2023). Instructors are also advised to inspect any sections that detectors identify as AI-written and use their judgment, rather than automatically accuse misconduct (Halaweh, 2023). Thus, a combination of technology (detection software) and pedagogy (oral exams, iterative drafts, honor codes) is being employed to uphold academic integrity in the age of AI.

From a technical perspective, the toolbox for responsible AI in education includes: model training improvements (RLHF, fine-tuning on educational data), pre-deployment audits (bias benchmarks, red teaming with education scenarios), continuous monitoring (moderation filters, usage logs accessible to teachers/parents), and usage

policies encoded both in code (e.g. prompt restrictions) and in class rules (e.g. requiring disclosure, reflection, and human verification). These measures, when implemented together, create overlapping layers of defense. No system is perfect – developers acknowledge "it is not possible to eliminate all risk at this time" (Khan Academy, 2023) – but the aim is to reduce the likelihood of harm to a very low level and to have mechanisms to catch and address any issues that do slip through. As an example of multi-layered mitigation: Khanmigo's design uses prompt engineering to "guide and narrow the focus" of the AI to the learning context (Khan Academy, 2023), fine-tuning to improve accuracy in educational tasks (Khan Academy, 2023), and live monitoring plus enforced transparency (teachers can see all AI-student chats) (Khan Academy, 2023) . Early results from such pilots are cautiously optimistic that AI can be integrated without major incident, especially when students and teachers are briefed on both its capabilities and fallibilities.

Case Studies and Ongoing Initiatives

Real-world deployments of ChatGPT in educational settings provide valuable insights into ethical and practical challenges, as well as effective strategies. We highlight here a few case studies and projects that have foregrounded ethical considerations:

Khan Academy's Khanmigo Pilot: Perhaps the most prominent example of ChatGPT (GPT-4) use in a K-12 context, Khanmigo is an AI tutor and assistant developed by the nonprofit Khan Academy in collaboration with OpenAI. From the outset, Khan Academy framed this project around equitable and safe use of AI. In Spring 2023, they launched Khanmigo in a limited pilot with select schools and donors, explicitly stating "Our goal is to ethically and responsibly provide access to our experimental AI tool" (Khan, 2023). A number of guardrails were put in place: Khanmigo does not simply give students answers to homework or quiz questions – "Nobody learns anything by being given the answer", Sal Khan explains (Khan, 2023). Instead, the AI is designed to act like "a virtual Socrates," engaging the student with questions, hints, and encouragement to think through the problem (Khan, 2023). For example, if a student is stuck on a math problem, Khanmigo might ask them to explain what they know so far, or pose a simpler sub-problem, rather than just outputting the solution. This approach mitigates the risk of AI becoming a cheating tool and aligns with educational best practices of "productive struggle." Khanmigo also has features for creative learning (co-writing stories, debating topics, practicing vocabulary) but with constraints – e.g. it will brainstorm ideas for a story with a student but "won't write the story for them"(Khan, 2023). On the teacher side, Khanmigo can assist with generating lesson materials (like quiz questions or lesson plan ideas), and importantly, teachers are given a dashboard to monitor AI usage. They can see transcripts of what their students are asking Khanmigo and how it responds (Croxton, 2025). This transparency ensures teachers can intervene if a student is going off task or if the AI gives inappropriate guidance. Khan Academy also set an age restriction (students under 18 need parental consent and are linked to a teacher/parent account) and limits on daily usage as mentioned (Khan Academy, 2023). Early anecdotal feedback from the pilot indicated that many students found Khanmigo helpful for understanding concepts and appreciated the non-judgmental, on-demand support, while teachers valued the time saved on routine tasks (with the caveat that the AI sometimes made errors they had to double-check) (Khan, 2023). This pilot is ongoing and being expanded in 2024–2025, with research being conducted on its learning impact. Khan Academy has committed to share findings openly, consistent with its ethical stance that "we plan to proceed responsibly and…share our learnings with the world"(Khan, 2023).

University Classroom Experiments: In higher education, several instructors have integrated ChatGPT into assignments under controlled conditions to explore its merits and pitfalls. For example, a writing instructor might have students use ChatGPT to generate a draft or outline, and then critique its work – learning about both the subject matter and the AI's limitations. Such case studies often highlight the double-edged nature of AI assistance: students can generate ideas or improve grammar more easily (a boon for non-native writers), but they may also over-rely on AI and produce formulaic essays (Kovari, 2025b). In response, some professors have flipped the script by teaching about ChatGPT itself – having students evaluate the accuracy of ChatGPT's outputs or compare their own work to AI-generated work. This demystifies the tool and turns it into a learning object rather than a black-box oracle. Universities like Princeton and MIT have formed working groups to issue guidelines and share experiences. For instance, one common policy is that students can use AI for brainstorming or editing help if they credit it, but not for final answers on exams or take-home tests (unless explicitly allowed). The University of Hong Kong piloted an "AI inclusive" approach in some courses, where a portion of assignments allowed AI use with reflection essays, finding that when openly permitted, students tended to use ChatGPT as a supplement and were thoughtful about its outputs rather than wholesale cheating (HKU Teaching & Learning, 2023 report). These experiments underscore that clear expectation-setting is key: when students know the pedagogical intent and boundaries of AI use, they are more likely to use it ethically. Conversely, in environments where AI use is vaguely regarded as cheating but not enforced, students may be tempted to misuse it surreptitiously (University of North Carolina, 2023).

Institutional Initiatives and Research: Beyond individual classes, some educational institutions are embracing AI with ethics in mind at the administrative level. For example, University of California, Berkeley announced an "AI in Education" policy task force to develop a student-centered policy that protects academic integrity without stifling innovation. Their principles (2024) included providing resources to faculty for redesigning assessments in an AI-pervasive world, and ensuring students from all backgrounds are trained in AI literacy so no one is left

behind. In another case, Georgia State University's Center for Excellence in Teaching created a support program for instructors to share strategies for AI, accompanied by research on how AI tools impact learning outcomes for different student demographics (to catch any inequitable effects early). On a policy research front, the U.S. Department of Education's Office of Educational Technology released in 2023 a report "AI and the Future of Teaching and Learning" which, while optimistic about AI's potential, cautions that algorithms must be transparent, fair, and protect student data privacy by design. It recommends that ed-tech vendors provide evidence of fairness testing and involve educators in AI development (Whitmer & Beiting, 2024). Similarly, non-profit coalitions like EDSAFE AI Alliance (described earlier) are working across companies and schools to benchmark AI products against safety and fairness criteria, and even to certify educational AI tools that meet certain ethical standards. This might soon influence procurement: a school district, for instance, could prefer an AI tutoring software that has an EDSAFE "Fair AI" certification or that adheres to the SIIA's Principles for AI in Education (which prioritize "civil rights, inclusion, and equity" in AI design) (EDSAFE AI Alliance, 2023).

These case studies illustrate a few takeaways. First, there is tangible momentum to incorporate ChatGPT-like tools in education, but leading adopters are doing so conscientiously – with pilot phases, oversight mechanisms, and ethical guidelines upfront. Second, the success of these implementations often hinges on transparency and user education: students are more likely to use AI appropriately when they are informed of its pitfalls and when its use is legitimized under clear rules, rather than being strictly banned or wholly unregulated. Third, human oversight remains crucial. Whether it's teachers monitoring Khanmigo chats, or professors reading AI usage reflections, the human-in-the-loop ensures that AI's mistakes or biases do not go uncorrected or unseen. Finally, these initiatives are serving as learning experiences for the institutions themselves – data is being gathered on what AI does well or poorly in educational settings, which can inform better design of both the technology and the curricula that incorporate it.

## 3.METHODOLOGY

**Methodology**

This study adopted a scoping review design to map policies, risks, and safeguards associated with the educational use of generative AI—focusing specifically on ChatGPT-style large language models (LLMs) in higher education within Vietnam and the broader Southeast Asia (ASEAN) region. Reporting followed PRISMA-ScR (Page et al., 2021; Tricco et al., 2018) guidance for scoping reviews and incorporated key elements of PRISMA 2020 for transparency in identification, screening, and inclusion. To ground the synthesis in practice, we complemented the review with a brief case vignette of a supervised AI-tutoring deployment that exemplifies how safeguards are operationalized in authentic settings. No human participants were involved.

We defined the review's scope as the use of generative AI in tertiary teaching, assessment, student support, and institutional governance across universities and colleges in ASEAN member states (Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, the Philippines, Singapore, Thailand, and Vietnam). We included global or regional frameworks where their content explicitly informed adoption in this context. Target phenomena were classroom- and institution-level uses of LLMs and the associated responsible/ethical AI guardrails (e.g., transparency and disclosure norms, privacy and data-protection measures, fairness and bias mitigation, academic integrity provisions, logging and auditability, due-process/appeals, teacher mediation, and staff development). We prioritized outcomes that documented concrete safeguards, governance mechanisms, assessment designs, and implementable guidance, as well as evidence of benefits, risks, feasibility, and workload implications relevant to higher education.

Eligibility criteria were established a priori. We included peer-reviewed articles and reviews, high-credibility policy and standards documents, and institutional guidance materials presenting implementable recommendations, published between January 2023 and August 2025, with a substantive focus on higher education in Vietnam or ASEAN or clear relevance to that context. We excluded opinion pieces without actionable guidance, purely technical AI benchmarks lacking educational implications, K-12-only contexts (unless findings directly transferred to tertiary settings), and non-English texts without accessible English versions, except for official Vietnamese or ASEAN policy documents where reliable translation was available.

Information sources comprised three primary databases—Scopus, Web of Science Core Collection, and ERIC—supplemented by Google Scholar (screening the first 200 relevance-sorted results) and targeted policy portals (notably UNESCO, MOET Vietnam, and SEAMEO). Searches combined controlled vocabulary and free-text terms for generative AI and higher education with governance/ethics constructs and ASEAN location terms, restricted to the specified publication window. We also hand-searched reference lists of included records and scanned organizational webpages to capture recently issued or updated guidance.

For data charting, we developed and piloted a structured extraction template capturing bibliographic details; country and setting; stakeholders; the generative-AI function(s) addressed (e.g., tutoring, feedback, writing support); identified risks (e.g., privacy, bias, integrity, safety, equity); specified safeguards or guardrails (e.g., disclosure norms, verification-centric assessment, logging and appeals, DPIA/data minimization, age- and role-based access, content filters, teacher mediation, training); reported outcomes (benefits/harms, feasibility,

workload); and limitations. Two reviewers independently trialed the template on a small subset to refine categories before full extraction across the corpus.

Given the scoping purpose, formal risk-of-bias scoring was not uniformly applied. For empirical studies, we recorded basic quality indicators (study design and sample characteristics, measures, and analytic transparency). For policy and guidance documents, we noted provenance (issuing body and currency), scope, and implementability. These appraisals informed interpretive weight in the synthesis, but records were not excluded solely on quality grounds once they met inclusion criteria.

We employed narrative thematic synthesis. First, we coded risks, safeguards, and implementation details within each record. Second, we clustered codes across records to identify recurrent guardrail categories—such as disclosure and transparency, verification-centric assessment in lieu of detector-dependence, privacy-by-design and data minimization, logging and auditability with due-process, and teacher mediation and training. Third, we mapped the synthesized categories to higher-education processes (course-level policy, assessment redesign, LMS and data-governance workflows). The synthesis yielded three integrative artefacts that are presented in the paper and appendices: a risk–safeguard–implementation cross-walk, a minimum-viable guardrails table suitable for institutional baselining, and a course-level checklist with a concise model policy paragraph. Where available, we highlighted region-specific nuances pertinent to Vietnamese and ASEAN institutional and regulatory contexts.

To support practical interpretation, we incorporated a case vignette of a supervised AI-tutoring deployment identified during screening of credible program documentation. We verified and summarized the pedagogical model (e.g., Socratic scaffolding), access controls, data handling practices, safety filtering, teacher visibility and interaction logs, and known limitations, and we linked each element to the guardrails articulated in our framework. The vignette was used illustratively to demonstrate implementation pathways; no new primary data were collected.All sources analyzed were publicly available; therefore, ethics approval was not required.

## 4.FINDINGS

### 4.1 Convergence on normative principles with methodological implications

Across the corpus, international and sectoral guidance converges on a stable value set—fairness/non-discrimination, transparency, privacy and data protection, safety, and accountability—for integrating large language models (LLMs) in learning and assessment. UNESCO's global guidance urges a human-centred orientation with age-appropriate access, data-protection by design, teacher capacity-building, and iterative governance; in practice these principles translate into methodological expectations that AI-mediated assessment be auditable, privacy-preserving, and equitable across groups. National guidance (e.g., the U.S. Department of Education, 2023) similarly emphasises bias mitigation, transparency of model behaviour, and educator involvement, reinforcing the need to document model versions, evaluation protocols, and human oversight in any applied deployment. Industry initiatives (e.g., the EdSAFE, SAFE framework) explicitly codify Safety, Accountability, Fairness/Transparency, and Efficacy, signalling that methodological reporting (what was measured, how fairness was tested, what evidence of efficacy exists) is part of responsible use rather than a post hoc add-on(Miao & Holmes, 2023).

### 4.2 Fairness and validity in AI-mediated assessment and tutoring

Findings consistently indicate that fairness concerns are inseparable from classical validity arguments. When LLMs assist with formative scoring or feedback, construct representation can drift toward surface features, threatening construct validity; where scoring is automated or semi-automated, the risk of differential prediction and group-based bias arises. Methodologically, studies and policies point toward routine use of measurement invariance checks, DIF analyses, and error decomposition that treats the model as an additional "rater." Where LLMs are used as graders or pre-graders, generalizability theory (G-studies) can partition variance components attributable to tasks, persons, human raters, and the LLM, informing D-study decisions about design changes (e.g., more tasks, human second-marker) to reduce error. For tutoring contexts, fairness extends to linguistic and cultural responsiveness; here, bias audits should include subgroup analyses on feedback tone, scaffolding depth, and error-correction quality, with consequential validity examined via downstream effects on motivation, self-efficacy, and performance. Across sources, the weight of guidance discourages over-reliance on AI-detectors for integrity and instead favours verification-centric assessment (process artefacts, orals, authentic tasks) so that fairness and validity are anchored in observable evidence of learning (U.S. Department of Education, Office of Educational Technology, 2023).

### 4.3 Reliability, reproducibility, and version drift

A recurring methodological challenge is stability of LLM outputs. Outputs vary with prompt phrasing, temperature settings, session history, and—critically—model version. Across deployments, this instability complicates reproducibility claims. The emerging practice standard is to (a) fix and report model version, temperature, system prompts, and moderation settings; (b) compute test–retest reliability for scoring uses (same input, different runs); (c) estimate inter-system reliability (LLM vs. human raters) using ICCs or G-study designs; and (d) maintain change logs when vendors update models to detect performance drift. Reporting packages from model providers (system cards and red-team reports) support this requirement by documenting known failure

modes and mitigation layers, but institutional implementers must still verify local reliability under their authentic tasks and populations (OpenAI, 2023).

**4.4 Safety engineering, red teaming, and continuous monitoring**

Responsible deployments pair pre-deployment safety work with post-deployment monitoring. Contemporary LLM releases document multi-phase external red teaming across risk domains (e.g., bias, toxicity, privacy), followed by targeted mitigations before general availability; this approach has become a reference for ed-tech adopters when specifying vendor due diligence and internal acceptance criteria. Within institutions, ongoing monitoring typically includes content-moderation pipelines, incident logging, and appeals mechanisms for AI-affected academic decisions. Methodologically, these practices amount to a continuous-auditing regimen in which the unit of analysis is not only the model but the socio-technical system (model + prompts + policies + user training), and success metrics blend safety rates, fairness indicators, and learning outcomes (OpenAI, 2023).

**4.5 Privacy, data protection, and developmental safeguards**

In higher education, data governance dominates risk narratives: institutions stress minimisation of personal data, prohibition of uploading confidential research/student information into public tools, and procurement requirements that specify data retention, access controls, and logging. In K-12 contexts, findings foreground developmental appropriateness and teacher mediation: age-gated access, parent/guardian consent, conservative defaults, and teacher visibility over interactions are treated as baseline safeguards. These expectations trace directly to international guidance and are increasingly reflected in provider documentation and district-level agreements (Miao & Holmes, 2023).

**4.6 Case evidence: supervised tutoring with teacher visibility**

The Khan Academy Khanmigo pilot exemplifies how platform design can embed guardrails into routine use. Public documentation describes a Socratic tutoring stance that privileges hints and metacognitive prompts over direct answers; teacher dashboards provide visibility into student–AI transcripts for timely feedback and accountability; usage limits aim to prevent drift and promote focused sessions; and privacy materials emphasise minimal data practices. As a design pattern, supervised access with teacher-visible logs, constrained affordances aligned to pedagogy, and conservative defaults appears to operationalise the normative principles above while preserving formative benefits.

**4.7 Synthesis: a measurement-informed baseline for responsible use**

Taken together, the evidence supports a practicable baseline for AI-assisted education that is explicitly measurement-informed. For assessment uses, institutions should document model settings, treat the LLM as a rater in reliability studies, run invariance/DIF analyses across salient subgroups, and triangulate AI-assisted scores with process evidence and human judgement. For tutoring and feedback uses, implementers should evidence subgroup fairness on interaction quality, monitor consequences for motivation and self-efficacy, and require teacher oversight for minors. Across settings, deployments should publish a concise evaluation dossier (model/version, prompts, safety and bias audits, reliability estimates, monitoring plan), complemented by user-level transparency (disclosure norms) and institutional appeals pathways for AI-affected academic decisions. When these elements are in place—alongside privacy-by-design and continuous red-team-and-monitor cycles—the formative advantages of ChatGPT-style tools can be realised without sacrificing validity, equity, or trust (U.S. Department of Education, Office of Educational Technology, 2023).

## 5 DISCUSSION

The above findings paint a picture of a rapidly evolving educational landscape in which AI tools like ChatGPT are becoming integrated, while stakeholders simultaneously strive to uphold longstanding educational values and equity. There is clear potential for positive impact: if used well, ChatGPT can provide personalized tutoring at scale, help teachers save time, and democratize access to knowledge and academic support (especially in under-resourced settings). The ethical frameworks and case studies reviewed show a genuine optimism that these technologies can be harnessed for good – for example, assisting struggling readers, offering practice to students who cannot afford human tutors, or enabling teachers to better tailor instruction. This aligns with the ethical principle of beneficence (promoting well-being) found in many AI ethics charters. However, realizing these benefits universally requires navigating challenges thoughtfully.

One major point of discussion is the balance between innovation and caution. Education, as a field, has traditionally been cautious in adopting new technologies, and rightly so – the stakes involve children's development and public trust. With ChatGPT, we see early adopters like certain teachers or districts forging ahead and experimenting, while others hold back, worried about risks like cheating or misinformation. The research suggests that extreme positions (either an outright ban on AI or an uncritical embrace of it) are less effective. Banning ChatGPT from schools entirely may prove futile (students can access it at home, and black-box detection is imperfect) and could widen inequities (tech-savvy or better-resourced students will use it anyway, quietly). On the other hand, fully embracing it without safeguards could undermine the integrity of learning and assessment. The middle path, which seems to be emerging, is guided use – integrating AI with clear pedagogical intent, rules, and support. This raises the question: how do we prepare educators for this role? Teacher training and professional

development will be crucial. Many current teachers did not encounter AI in their training, so capacity-building is needed to help them understand AI's quirks and how to supervise its use. Notably, some commentators have called for including AI ethics and usage as part of digital literacy curricula for students as well, arguing that knowing how to work alongside AI is a critical skill for the future (akin to information literacy or critical thinking) (Bali, 2024).

Another discussion point is the role of AI providers (like OpenAI) versus educational authorities in ensuring responsible use. Companies can build safer models and offer policy tools (like OpenAI's moderation endpoint), but they may not foresee every educational scenario. Should AI companies be responsible for, say, preventing their models from helping a student cheat on an exam? OpenAI's terms of use already prohibit using their API for fraudulent or dishonorable purposes, which academic cheating could be argued to be. They have also explicitly cautioned educators not to punish students who weren't clearly instructed about AI use expectations (University of North Carolina, 2023). This indicates some level of responsibility-taking. But practically, enforcement and context lie with schools. Some have suggested an honor-code system augmented with technology: for instance, maybe future AI systems could include an "educator mode" where any output generated for a student comes with a cryptographic watermark or log that a teacher can inspect (OpenAI and others are researching watermarking of AI-generated text). This technical solution could support academic honesty if it matures, though currently watermarks are not reliably detectable after student edits.

Data privacy deserves attention in discussion as well. ChatGPT, especially in its free public version, raises privacy flags if students input personal data or school data into it, since those inputs might be retained by the service. OpenAI has stated that for education users, they do not use conversation data to train models if the user opts out or uses their institutional service(Columbia University Information Technology, 2025). Still, schools must ensure compliance with laws like FERPA (in the U.S.) or GDPR (in Europe) when adopting AI. This often means requiring parental consent, ensuring the AI provider has proper data handling agreements, and instructing students never to share sensitive personal information with the AI. Some educational institutions have opted for self-hosted or open-source AI models (with filters) to keep data on local servers, but these models may be less capable than ChatGPT. Thus there's a trade-off between data control and state-of-the-art performance. Ethical deployment likely means choosing the option that best protects student privacy while still providing benefit – which in many cases is working with companies that are transparent about data use and have security certifications (OpenAI, for instance, launched an "ChatGPT Education" version that complies with SOC 2 security and doesn't train on user data (Columbia University Information Technology, 2025)). In sum, privacy is being addressed through a combination of policy (what students are allowed to input) and product choices, and ongoing oversight is needed to ensure student data isn't inadvertently exposed via AI tools.

Explainability vs. efficacy is another nuanced area. Educators ideally want AI that can explain its answers (to help students learn), but the most explainable models (like simpler rule-based systems) are far less powerful than black-box neural nets. One could argue that for certain uses, a less complex, more interpretable model is preferable – for example, a math tutoring system that uses a deterministic step-by-step solution engine might teach more transparently than a deep neural network that just outputs the answer. On the other hand, that deterministic system may lack the adaptability and language fluency of ChatGPT. A possible solution is hybrid systems: using ChatGPT for what it's best at (natural language dialogue, motivation, broad knowledge) and coupling it with symbolic or interpretable systems for tasks like showing math steps or checking logic. This is an active research direction (sometimes called Neuro-Symbolic AI). In the interim, some educators essentially force explainability by how they instruct students to use ChatGPT – e.g. asking students to prompt ChatGPT to explain the answer or break down the solution. The discussion in literature encourages this practice; if the AI is treated as a tutor, we should demand it behaves like a good tutor, not just a provider of answers. As mentioned earlier, there are still concerns: an AI can "explain" an answer in a way that sounds plausible but is actually just another hallucination. So teaching students not to accept AI explanations uncritically is part of the digital literacy piece.

From a policy standpoint, an interesting ongoing debate is how much central regulation is needed versus local control. Some countries have been considering or implementing national guidelines – for example, China reportedly restricted ChatGPT-like tools in classrooms until vetted, and some European countries' education ministries have issued recommendations. UNESCO has urged a global dialogue and even suggested that AI in education be accompanied by curricula on AI ethics for students (Bali, 2024). However, education systems differ widely, and imposing a one-size-fits-all rule (like "ban AI in all exams" or "allow AI for all homework") may not work. A principle of contextual integrity seems to apply: decisions might vary by subject, age group, and assessment type. It might be ethical in an English class to let students use ChatGPT for brainstorming ideas for a story (since the goal is creative ideation and the student still has to write and refine the story), but unethical to use ChatGPT in a take-home coding assignment in a computer science class where the learning objective is to practice programming. Thus, flexible policies that consider the context and clearly communicate the why behind them will likely be more effective and buy more stakeholder buy-in.

The role of ongoing evaluation and iteration cannot be overstated. One of the lessons from both the AI field and education field is that interventions often have unintended effects. Continuous research is needed to observe how students interact with ChatGPT over longer periods: Does it improve their independent skills or make them too

dependent? Does it boost overall learning or just inflate grades? Preliminary evidence is mixed – some studies show improved learning outcomes with AI tutoring support, while others caution about superficial learning. We may discover, for instance, that AI is great for drill practice and confidence-building, but less effective for fostering deep critical thinking unless paired with reflective activities. If so, educators will need to adapt pedagogical strategies accordingly (e.g. using AI for practice but ensuring class discussions or assessments probe deeper understanding). Ethically, this ties to the principle of non-maleficence: we must ensure that well-intended AI uses aren't inadvertently harming educational development, and if they are, recalibrate them.

Finally, an important theme for discussion is inclusion. We must consider diverse learners – students with disabilities, language learners, etc. ChatGPT can be an amazing tool for accessibility (imagine a student with dyslexia using it to have text read aloud or rephrased, or a deaf student using it to practice conversational speech through text). But it could also pose challenges (some students might find its responses confusing or overwhelming without simplification). Designing AI tools with universal design for learning (UDL) principles in mind is a budding area: for example, ensuring the AI can adjust its reading level, or provide multi-modal explanations (like diagrams or verbal output for those who need it). The ethical use of AI in education compels us to ask: are we making sure this works for every student, including those with special educational needs or those who speak minority languages? The literature suggests more work is needed here; biases in language models against underrepresented languages or dialects have been documented (Mienye & Swart, 2025), so inclusivity must remain a focus in AI improvement.

In conclusion of this discussion, it is evident that achieving ethical and responsible AI use in education is a collaborative, ongoing process. It involves AI developers building safer, more transparent systems; educators and students learning new norms and skills around AI; institutions crafting policies that reflect educational values in an AI age; and researchers continually evaluating impacts. The analogy of AI as a partner or assistant rather than a tool is becoming popular – meaning it should be treated as part of a team with humans, where mutual communication and accountability are in place. As one educator put it, "AI can empower learning, but can also house hidden prejudices"(Alejandro, 2024) – our task is to shine light on those hidden aspects and guide AI's use toward empowerment, not displacement or distortion of learning.

## CONCLUSION

ChatGPT's advent in education has catalyzed a vital conversation about how to harness cutting-edge AI for the benefit of students and teachers while upholding ethical standards. This comprehensive review has shown that there is significant promise in using AI tools like ChatGPT to enhance learning – from providing individualized tutoring and feedback, to aiding teachers in curriculum development, to expanding educational access. Yet, it has equally highlighted the responsibility that comes with deploying such powerful technology in classrooms. Ensuring ethical and responsible use is not a one-time checklist but a continuous commitment.

Several conclusions and recommendations emerge from our analysis. First, the foundation of any responsible AI deployment in education must be a clear ethical framework. Fortunately, we see alignment among global and local bodies on principles such as fairness, safety, privacy, transparency, and accountability. Educational institutions should explicitly adopt or adapt these principles into their AI use policies, making them visible and understandable to all stakeholders. For example, a school district might formulate guidelines that "AI may be used to support learning, but will be implemented in ways that are equitable, transparent, and preserve academic integrity," then detail what that means in practice (like requiring source citation for AI-generated content, or disallowing AI use on certain assessments). Having this ethical compass helps navigate decisions large and small – from choosing which AI platforms to license, to deciding classroom rules.

Second, it is crucial to maintain human oversight and agency at the center of AI-augmented education. AI should not replace the human teacher or reduce students to passive consumers of machine output. Instead, as findings suggest, the best outcomes occur when AI is used as a tool that amplifies human capabilities. Teachers should be supported (through training and tools) to supervise AI interactions – for instance, reviewing logs of an AI tutor's conversations to understand student misconceptions, or intervening when the AI falters. Students, for their part, should be taught to critically evaluate AI-provided information and use it as a starting point for deeper inquiry, not an endpoint. In short, AI's role must remain advisory, not authoritative: final judgments – whether it's grading an essay, deciding if an answer is correct, or mentoring a student's progress – should lie with human educators and learners. This ensures accountability and guards against blindly trusting AI recommendations.

Third, ongoing efforts to audit and mitigate bias in AI are non-negotiable. Our review makes it clear that without active measures, AI can perpetuate social biases in ways that affect student outcomes and self-perceptions. Education technology providers and researchers should continue developing robust bias evaluation frameworks (e.g. testing AI on diverse student queries and demographics) and share those results transparently. When issues are found, they must be addressed via model improvements or usage constraints. The iterative improvements OpenAI documented – reducing stereotype rates over model generations (OpenAI, 2024b) – demonstrate that progress is attainable. We encourage collaborations between AI experts and educational equity experts to refine these systems. Likewise, representation matters: involving educators and students from diverse backgrounds in

the design and testing of educational AI can surface biases that a homogenous development team might overlook. The ultimate goal is an AI that is culturally responsive and treats all students with equal respect and high expectations, thereby supporting inclusive education rather than undermining it.

Fourth, transparency and explainability should be maximized wherever possible. Educational institutions might consider requiring AI service providers to supply documentation of how their models were trained, what their limitations are, and what data (if any) is collected from users – essentially an "AI accountability report." In the classroom, teachers should strive to make AI a visible part of the learning process: if a student uses ChatGPT, that fact should be out in the open and part of the discussion (e.g. "How did ChatGPT help you? Let's verify its suggestions."). Although current AI models are not fully explainable, educators can use strategies like prompting the AI to show steps or engaging students in comparing AI solutions to human solutions. Over time, as research yields more explainable AI techniques (perhaps simplified student-facing reasoning logs or interactive debugging tools), these should be integrated to further demystify the technology. The transparency principle extends to outcomes: if AI is involved in decision-making (such as flagging at-risk students via learning analytics), schools owe it to students to explain those decisions and allow appeal or human review. This maintains trust in the system and avoids a scenario where students feel judged by an inscrutable algorithm.

Finally, we underscore the importance of continuous learning and adaptive governance. AI capabilities are advancing quickly – what ChatGPT could not do a year ago (e.g. solve a complex multi-step problem) it might do now, and new models will bring new affordances and risks (like multimodal inputs/outputs, which raise their own ethical questions in education). Therefore, policies and practices must be revisited regularly. Educational institutions should treat their AI use guidelines as living documents, updated with community input as experience grows. It will be beneficial to establish feedback channels: for instance, a committee that includes teachers, students, IT staff, and ethicists that meets periodically to review how AI is being used, any incidents that occurred, and whether policies or technical settings need adjustment. On a larger scale, sharing lessons across institutions (through conferences, publications, networks like UNESCO's education forums) will accelerate collective knowledge. The case of Khan Academy openly publishing its AI principles and learnings is a good example of leadership in this space (Khan Academy, 2023; Khan, 2023). We encourage more such transparency among both tech providers and educational users, as it will help others avoid pitfalls and adopt best practices.

In conclusion, integrating ChatGPT and similar AI into education is a complex endeavor, but one that can be managed with foresight and ethical intentionality. By anchoring use in robust ethical frameworks, keeping humans in the loop, actively addressing biases, demanding transparency, and staying adaptive, we can unlock the benefits of AI for learning while minimizing harms. Education has always aimed to empower the next generation with knowledge, skills, and values; AI, when responsibly applied, can be a powerful ally in that mission – offering each student personalized support and each teacher enhanced capabilities. The journey will involve trial and error, and not everything will go smoothly. Yet, the drive shown by educators and organizations to "get this right" is heartening. As one recent paper noted, "addressing these issues is essential for ensuring the ethical integration of AI in language education, where a hybrid approach combining AI with human instruction emerges as the most responsible solution"(Mienye & Swart, 2025). In other words, the future of ethical AI in education is one where smart machines and wise humans work hand-in-hand. If we proceed with care, collaboration, and an unwavering focus on students' best interests, we can indeed leverage ChatGPT to elevate learning opportunities for all while upholding the values that define meaningful education.

## REFERENCES

1. Alejandro, S. (2024, June 2). The education balancing act: AI progression, fairness and biases. Policy Perspectives. https://policy-perspectives.org/2024/06/02/the-education-balancing-act-ai-progression-fairness-and-biases/
2. Bali, M. (2024, October 18). When it comes to AI, is transparency enough? LSE Higher Education Blog. https://blogs.lse.ac.uk/highereducation/2024/10/18/ethics-in-ai
3. Columbia University Information Technology. (2025). ChatGPT Education. https://www.cuit.columbia.edu/content/chatgpt-education
4. Corporation for Digital Scholarship. (2025). Zotero (Version X.X) [Computer software]. https://www.zotero.org/
5. Croxton, W. (2025, July 20). How classroom AI Khanmigo can help students in emotional distress. CBS News. https://www.cbsnews.com/news/how-classroom-ai-khanmigo-can-help-students-in-emotional-distress-60-minutes/
6. EDSAFE AI Alliance. (2023). SAFE Benchmarks Framework. https://www.edsafeai.org/safe
7. Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. Contemporary Educational Technology, 15(2), ep421. https://doi.org/10.30935/cedtech/13036
8. Hamilton College. (2023). Policies: Generative AI guidelines. https://www.hamilton.edu/offices/lits/policies/generative-ai-guidelines

9. Harvard University Information Technology. (2023). Generative artificial intelligence (AI) guidelines. https://www.huit.harvard.edu/ai/guidelines
10. Khan Academy. (2024, October 15). How do I view my students' Khanmigo chat history? https://support.khanacademy.org/hc/en-us/articles/15127248640525-How-do-I-view-my-students-Khanmigo-chat-history
11. Khan Academy. (2023, March 12). Khan Academy's approach to the responsible development of AI. Khan Academy Blog. https://blog.khanacademy.org/aiguidelines
12. Khan, S. (2023, November 16). Harnessing GPT-4 so that all students benefit: A nonprofit approach for equal access. Khan Academy Blog. https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/
13. Kovari, A. (2025a). Explainable AI chatbots towards XAI ChatGPT: A review. Heliyon, 11(2), e42077. https://doi.org/10.1016/j.heliyon.2025.e42077
14. Kovari, A. (2025b). Ethical use of ChatGPT in education—Best practices to combat AI-induced plagiarism. Frontiers in Education, 9, Article 1465703. https://doi.org/10.3389/feduc.2024.1465703
15. Mhlanga, D. (2023). Open AI in education: The responsible and ethical use of ChatGPT towards lifelong learning. In FinTech and artificial intelligence for sustainable development (pp. 387–409). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-37776-1_17
16. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535
17. Mienye, I. D., & Swart, T. (2025). ChatGPT in education: A review of ethical challenges and approaches to enhancing transparency and privacy. Procedia Computer Science, 254(3), 181-190. https://doi.org/10.1016/j.procs.2025.02.077
18. OpenAI. (2023). GPT-4 technical report (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774
19. OpenAI. (2024a, August 8). GPT-4o system card. https://openai.com/index/gpt-4o-system-card/
20. OpenAI. (2024b, October 15). Evaluating fairness in ChatGPT. https://openai.com/index/evaluating-fairness-in-chatgpt
21. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
22. Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ, 372, n160. https://doi.org/10.1136/bmj.n160
23. Smith, K. (2025, May 6). AI shows racial bias when grading essays - and can't tell good writing from bad. The 74. https://www.the74million.org/article/ai-shows-racial-bias-when-grading-essays-and-cant-tell-good-writing-from-bad
24. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473. https://doi.org/10.7326/M18-0850
25. U.S. Department of Education, Office of Educational Technology. (2023). Artificial intelligence and the future of teaching and learning: Insights and recommendations.
26. University of North Carolina at Chapel Hill. (2023, April 17). The ethics of college students using ChatGPT – University policy. https://universitypolicy.unc.edu/news/2023/04/17/the-ethics-of-college-students-using-chatgpt/
27. Whitmer, J., & Beiting, M. (2024, December 13). Modernizing AI fairness analysis in education contexts. Federation of American Scientists. https://fas.org/publication/modernizing-ai-fairness-analysis-in-education-contexts