Microsoft

Microsoft Trustworthy AI

The Future of Learning Starts with Trust

A practical guide for K-12 education leaders driving responsible AI adoption

Table of contents

Executive summary

This brief provides K-12 education leaders with a practical guide for adopting Trustworthy AI, with a focus on security, safety, and privacy. As AI becomes more common in classrooms, trust remains essential, especially in how systems handle sensitive data, address digital safety, and align with school values. Microsoft supports schools with secure, responsible AI solutions grounded in privacy, transparency, and inclusive design. Learn how Microsoft empowers schools to achieve more by fostering trust in its AI systems while protecting data and defending against threats.

Key takeaways

- Trustworthy AI in education requires secure, safe, and private systems built on strong commitments and proven capabilities.
- Microsoft's privacy-by-design approach supports compliance and builds trust with students, families, and communities.
- Microsoft 365 Copilot inherits permissions, sensitivity labels, and retention policies to protect sensitive data and enable responsible governance.
- AI content filters and oversight tools can help schools maintain safer, more productive learning environments.

Introduction

The trust gap in edtech

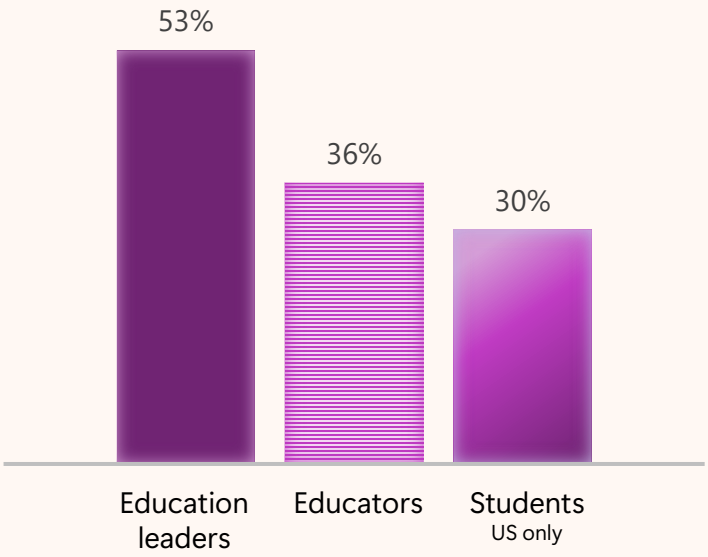
Trust is central to every technology decision in education. Leaders, educators, and IT professionals need to be confident that their tools are safe, secure, and private. With AI use increasing in classrooms, concerns around how systems use data are rising.

According to the 2025 AI in Education: A Microsoft Special Report, approximately 20% of education professionals are concerned about data privacy and security. Education was also the third most-targeted industry for cyberattacks in Q2 2024.[1]

Microsoft empowers schools to achieve more by developing trust in its AI systems and the security products that protect data and defend against threats. Built on responsible AI principles and a commitment to security and privacy, Microsoft offers AI solutions that are as trustworthy as they are impactful for teaching, learning, and school operations. This commitment enables schools to innovate while protecting sensitive student data and promoting equitable learning.

Education leaders lead in AI use

What percentage of school community members use AI daily for school-related purposes?[2]



"With the right guardrails, cutting-edge technology can be safely introduced to the world to help people be more productive and go on to solve some of our most pressing societal problems."

—Natasha Crampton, Chief Responsible AI Officer, Microsoft

[1] Cyber Signals Issue 8 | Education under siege: How cybercriminals target our schools

[2] 2025 AI in education: A Microsoft Special Report

Keeping AI secure in education

At Microsoft, our approach to AI in education is grounded in our Trustworthy AI principles, privacy, safety, security. These principles guide how we build technology and how we help schools adopt AI with confidence.

Generative AI creates new opportunities for education to achieve more, and it introduces new considerations for leaders around data privacy, content safety, and cybersecurity. With Microsoft's built-in protections, governance controls, and responsible AI guardrails integrated into the tools schools already use, institutions can adopt AI confidently while focusing on innovation and learning.

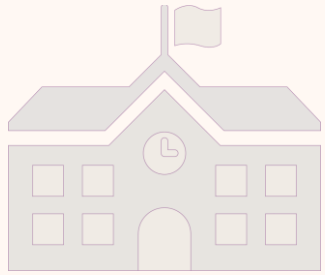
Microsoft's Secure Future Initiative (SFI) and responsible AI principles help us address AI vulnerabilities such as prompt injection, inversion attacks, and harmful content generation, enabling institutions to protect sensitive data while safely exploring AI's potential.

The table on the next page outlines the top threats our solutions help mitigate, alongside the specific tools Microsoft offers to address them, so that schools can pursue AI innovation while maintaining compliance, safety, and trust.

"We urgently need to start talking about the guardrails we put in place to protect people and ensure this amazing technology can do its job of delivering immense value to the world."

—Mustafa Suleyman, Chief Executive Officer, Microsoft AI

Generative AI security threats



Threat	Impact on Education	Microsoft Solution
Poisoning Introduces harmful or misleading data into the AI training process to distort its outputs our behavior.	Can result in AI generating inaccurate or unethical responses that erode trust in school systems.	Microsoft Purview (Data governance, Labeling, Provenance checks, Telemetry)
Evasion Alters input data in ways that bypass security controls—such as. embedding threats in hidden formats.	Can spread harmful content across a district before it's detected, making cleanup difficult.	Microsoft Defender (Policy enforcement) Azure AI Content Safety (Multi-layer filtering
Inversion Repeated queries are used to reverse-engineer the AI model or extract training data.	Risks unauthorized access to student records or sensitive content.	Microsoft Purview (Access control, Encryption) Azure AI Content Safety (Data minimization)
Prompt injection Inserts malicious or manipulative code into prompts to override system behavior.	Can lead to AI models generating unsafe responses or accessing restricted data.	Azure AI Content Safety (App guardrails)
Functional extraction Uses repeated AI queries to infer how a model works or replicate it for misuse.	Attackers may mimic school communications or generate misinformation that appears legitimate.	Microsoft Defender for Cloud Apps (Rate-limiting, Telemetry) Azure AI Content Safety (Abuse detection)

Trustworthy AI in education

Every AI innovation at Microsoft is grounded in a comprehensive set of AI principles, policies and standards. This includes foundational commitments, such as our Secure Future Initiative, Responsible AI principles, and privacy principles. These commitments give you confidence that you control your data, and your data is secure in any state, whether at rest or in transit. We're transparent about where data is located and how it's used, and we're committed to making sure AI systems are developed responsibly.

These commitments also ensure that the AI systems we build have the right privacy, safety, and security in mind from the start. We use our own best practices and learnings to provide you with capabilities and tools to help you build your own AI applications that share the same high standards that we strive for.

Whether you are an enterprise leader, an AI developer, or a copilot enthusiast, **Microsoft provides the foundation you need to build and use generative AI that you can trust.**

"Trust in the technology, ultimately is going to be core to all the diffusion. If you don't trust it, you're not going to use it, and that's not going to be great for anyone."

—Satya Nadella, Chief Executive Officer, Microsoft

Security

Microsoft is committed to providing the secure foundation schools need to confidently adopt AI. The Secure Future Initiative (SFI) is our comprehensive approach to safeguarding data and maintaining operational continuity as threats evolve. Built through significant, ongoing investments, SFI focuses on developing cutting-edge security technologies, fostering strong partnerships, and empowering schools through education and support.

Microsoft AI solutions like Microsoft 365 Copilot Chat, are built on these principles with enforcement protocols and continuous threat updates. Microsoft 365 Copilot uses secure, encrypted tenants and Azure AI services to maintain comprehensive data privacy protections.

Three core principles of Secure Future Initiative

Secure by design
Security comes first when designing any product or service.

Secure by default
Security protections are turned on, require no extra effort, and aren't optional

Secure operations
Security controls and monitoring continuously improve to meet evolving cyberthreats

"Microsoft runs on trust, and trust must be earned and maintained. Our pledge to our customers and our community is to prioritize your cybersafety above all else."

—Charlie Bell, Executive Vice President of Security, Microsoft

Secure Future Initiative

The six pillars include goals and actions that define our approach

Protect identities and secrets

Reduce the risk of unauthorized access by implementing and enforcing best-in-class standards across all identity and secrets infrastructure, plus user and application authentication and authorization.

Protect tenants and isolate production systems

Protect all Microsoft tenants and production environments using consistent, best-in-class security practices and strict isolation to minimize breadth of impact.

Protect networks

Protect Microsoft production networks and implement network isolation of Microsoft and customer resources.

Protect engineering systems

Protect software assets and continuously improve code security through governance of the software supply chain and engineering systems infrastructure.

Monitor and detect threats

Provide comprehensive coverage and automatic detection of cyberthreats for Microsoft production infrastructure and services.

Accelerate response and remediation

Prevent exploitation of vulnerabilities discovered by external and internal entities through comprehensive and timely remediation.

Build trust through secure AI adoption

With over 80% of education leaders concerned about data when interacting with AI,[3] Microsoft 365 Education plans provide targeted data protection when school deploy:

Microsoft Purview safeguards privileged information through data classification, tagging, and labeling to control access to data within AI systems.

Microsoft Defender works alongside Purview to provide threat detection across identities, endpoints, apps, and AI systems.

Along with data loss and insider risk protection found in the Copilot Control System, these products provide end-to-end protection, data governance, and compliance for AI workloads.

Compliance with emerging AI standards

Microsoft's approach to trustworthy AI is aligned with key regulatory frameworks, helping schools meet evolving legal and policy requirements. This includes the EU AI Act, where Microsoft 365 solutions support compliance for education institutions operating in Europe.

Additionally, Microsoft follows the White House's guidance on AI security and the NIST AI Risk Management Framework in the United States, reinforcing our commitment to secure, transparent, and responsible AI adoption in schools.

While regulations continue to evolve, Microsoft builds AI solutions with compliance in mind, so institutions can focus on innovation without worrying about legal complexity.

Explore the path to securely adopting AI in Accelerate AI transformation with strong security

[3] First annual generative AI study: Business rewards vs. Security risks

Customer success stories

Schools and institutions around the world are using Microsoft Security solutions to protect digital learning environments while encouraging innovation. Prince William County Public Schools in Virginia and Fulton County Schools in Georgia are a few schools making AI adoption safer and more effective with Microsoft.

Prince William County Public Schools

Challenge

Prince William County Schools, one of Virginia's largest school districts, faced increasing cybersecurity threats that put sensitive student and staff data at risk. With limited IT staff and growing digital learning needs, PWCS needed a way to strengthen protection without overburdening resources.

Solution

PWCS adopted Microsoft Security solutions, including Microsoft Defender for Endpoint and Microsoft Sentinel, within its existing Microsoft 365 A5 licensing. This integration enabled the district to streamline threat detection, automate response, and improve visibility across all devices, all while staying compliant with data privacy regulations.

Impact

With Microsoft's unified security platform, PWCS reduced response times from days to minutes, freeing up IT staff to focus on strategic goals. The district now proactively safeguards its learning environment, helping ensure a safer, more resilient experience for students and educators.

"The biggest reward for us is when a counselor or principal thanks us for passing on a compliance alert from Microsoft Purview that showed that a student needed support."

– AJ Phillips, Director of Information & Instructional Technology, Prince William County Public Schools

Prince William County Public Schools creates a more cybersafe classroom with Microsoft Purview

Customer success stories

Fulton County Schools

Challenge

Fulton County Schools, one of Georgia's largest K–12 districts, sought to prepare students for a future shaped by AI, while reducing administrative burden for educators. The district faced challenges in responsibly integrating new technologies at scale, safeguarding student data, and ensuring equitable access to digital learning. Leadership needed a secure, inclusive solution that would support innovation without compromising privacy, trust, or educational equity.

Solution

Fulton County Schools adopted Microsoft 365 Copilot and Copilot Chat to help transform instruction and school operations. The district established an AI task force, identified high-impact scenarios, and provided professional development to support responsible AI integration into teaching and learning. Students used Copilot Chat to brainstorm ideas, write content, and explore concepts interactively. Microsoft 365's built-in, enterprise-grade security helped the district protect student data and maintain compliance. Administrators used Copilot to simplify planning, reporting, and communications, reducing manual workload across schools.

Impact

With Microsoft 365 Copilot, the district enhanced student engagement and improved operational efficiency. Students gained confidence through personalized learning experiences. Educators reclaimed valuable time to focus on instruction. Tasks that once took weeks or months, such as compiling reports, were completed in days. Fulton County Schools now delivers safer, more inclusive, and more engaging learning environments. The adoption of Microsoft 365 Copilot reflects the district's commitment to using AI responsibly to support student success and educator well-being.

"Copilot Chat allowed us to explore AI within a safe environment. It's not about automation; it's about equity—giving every student the opportunity to learn in a way that works best for them."

– Dr. Joe Phillips
Chief Information Officer, Fulton County Schools

Fulton County School's adoption of Microsoft 365 Copilot empowers educators and students to innovate securely and confidently with AI.

Safety

In K–12 schools, safety isn't just a priority when educators and students use AI—it's a necessity. Every interaction with technology in the classroom carries a responsibility to protect student well-being, privacy, and trust. As AI becomes a powerful tool for personalized learning and administrative efficiency, schools must ensure these innovations do not introduce new risks such as harmful content, bias, or data misuse.

Microsoft's six responsible AI principles, established in 2018, guide how we design, build, and deploy AI systems that are not only innovative but also safe, reliable, and aligned with educational values. These principles are the foundation for every AI solution we deliver—helping schools embrace the benefits of AI while safeguarding what matters most: students, educators, and their communities.

Responsible AI principles

Fairness
AI systems should treat all people fairly.

Reliability and safety
AI systems should perform reliably and safely.

Privacy and security
AI systems should be secure and respect privacy.

Inclusiveness
AI systems should empower everyone and engage all people, regardless of their backgrounds.

Transparency
AI systems should be understandable.

Accountability
People should be accountable for AI systems.

"We must protect youth safety and privacy online and ensure that technology – including emerging technologies such as AI – serves as a positive force for the next generation."

—Brad Smith, Vice Chair and President, Microsoft

Strengthen Trust and Safety

As institutions integrate AI into learning, research, and operations, ensuring responsible use is essential. Azure AI Content Safety help education leaders uphold safety, security, and compliance by providing built-in tools to monitor and manage AI-generated content.

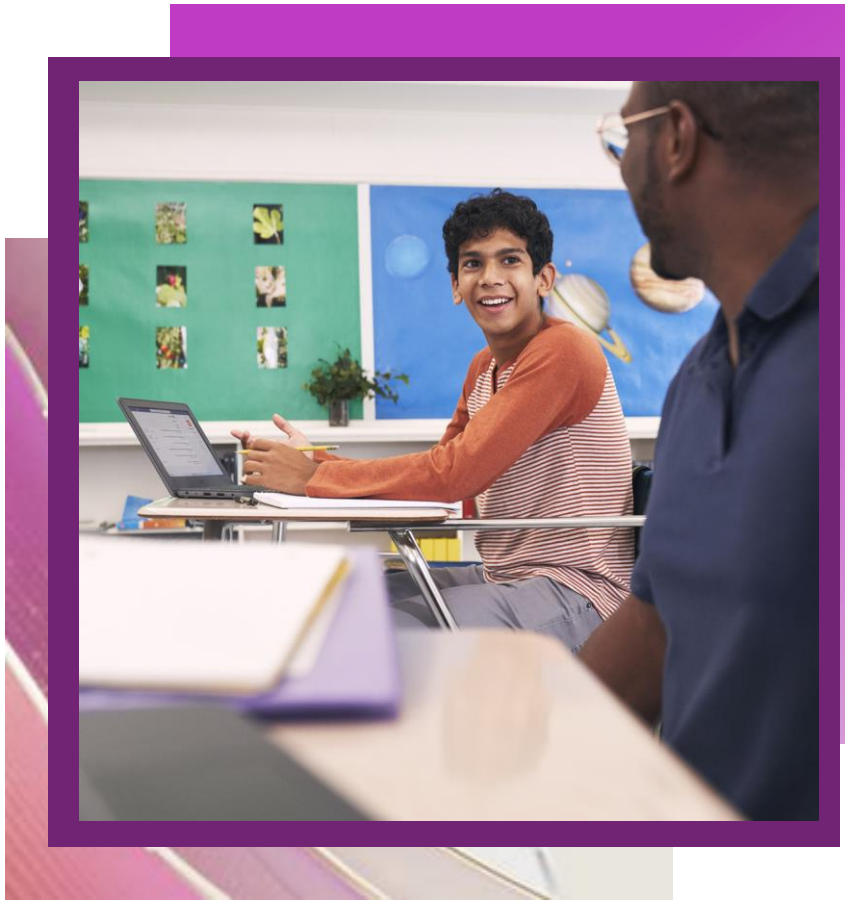
1. **Block harmful content**
   Detect and block violence, hate, sexual, and self-harm content. Configure severity thresholds for your specific use case and adhere to your responsible AI policies.

2. **Create custom AI filters**
   Create unique content filters tailored to your requirements using custom categories. Quickly train a new custom category by providing examples of content you need to block.

3. **Identify and mitigate security threats**
   Safeguard your AI applications against prompt injection attacks and jailbreak attempts. Identify and mitigate both direct and indirect threats with prompt shields.

4. **Detect and correct Generative AI hallucinations**
   Identify and correct generative AI hallucinations and ensure outputs are reliable, accurate, and grounded in data with groundedness detection.

5. **Identify protected materials**
   Pinpoint copyrighted content and provide sources for preexisting text and code with protected material detection.



Discover resources and strategies to help you implement AI: Explore the business case for responsible AI in new IDC whitepaper

Customer success stories

Department for Education, South Australia

Challenge

The South Australia Department for Education wanted to incorporate generative AI in classrooms to enhance learning but needed strong safeguards to prevent harmful or inappropriate content exposure. The core question: how to empower students with AI responsibly.

Solution

In partnership with Microsoft, the Department launched EdChat, a generative AI–based educational chatbot backed by Azure AI Content Safety. The system includes built-in safeguards that detect and block risky or harmful content, while giving the Department control over moderation and filtering. They piloted EdChat with ~1,500 students and 150 teachers across eight schools.

Impact

The pilot confirmed that EdChat can be used safely in schools, giving students and teachers secure access to AI assistants. The built-in content safety features allowed educators to focus on learning instead of moderating responses. The initiative now positions the Department to scale AI-powered support across schools—fostering curiosity, improving access to information, and enabling a more dynamic, safe learning environment.

"We wouldn't have been able to proceed at this pace without having the content safety service in there from Day 1. It's a must-have."

–Simon Chapman, Director of Digital Architecture, South Australia Department for Education

EdChat: Pioneering safe generative AI in South Australian classrooms with Azure AI Content Safety

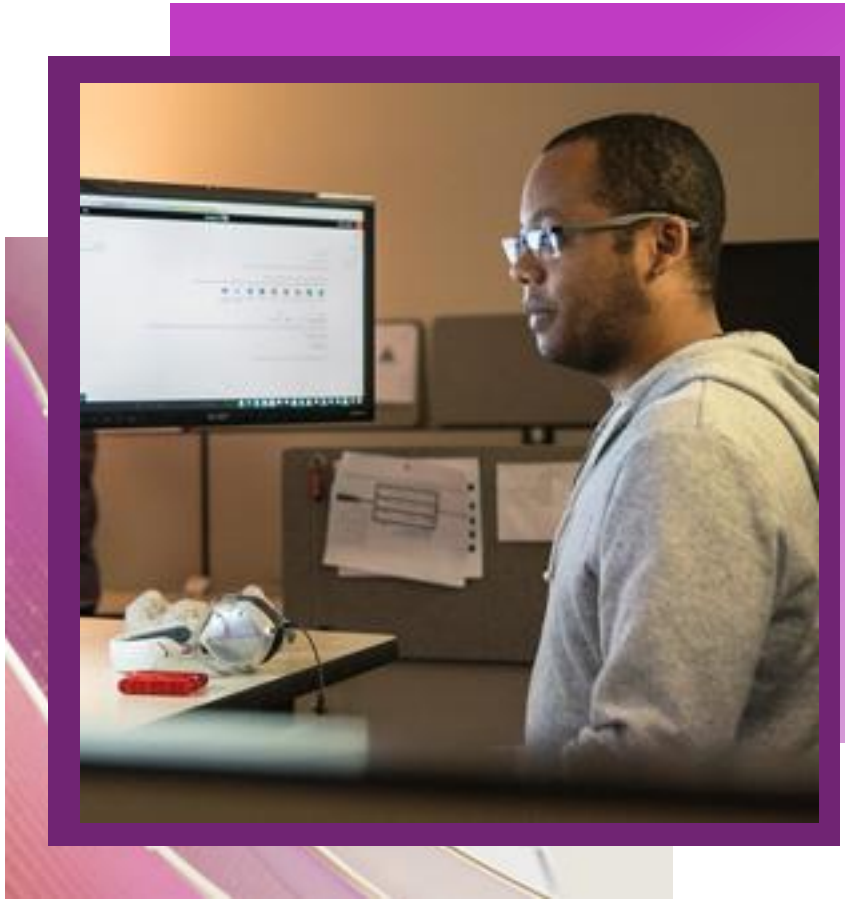
Privacy

At Microsoft, we treat privacy as a fundamental human right and embed protections into every product we build. For schools, this means the same enterprise-grade privacy commitments trusted in Microsoft 365 also extend to Copilot and AI solutions.

Our privacy commitments are defined by four principles that guide how we collect, store, and protect your data.

Our approach sets us apart:

- You control your data. Schools decide how their data is used, with clear settings and choices.
- You know where your data is located and how it's used. Data from Copilot interactions is managed under the same commitments schools already rely on with Microsoft 365, with options to meet local policies and regulations.
- Your data is secure. Microsoft uses multiple layers of protection to keep data safe when it's stored or shared.
- Microsoft defends your data. We follow strict rules for government requests and apply the same protections for AI that schools already trust for email and files.



"We must protect youth safety and privacy online and ensure that technology–including emerging technologies such as AI–serves as a positive force for the next generation."

—Brad Smith, Vice Chair and President, Microsoft

Beyond privacy, Microsoft also helps schools navigate compliance and innovation responsibly:

- Customer data is not used to train foundation models: Your school's data stays yours, a major differentiator compared to other providers.
- Global compliance leadership: Microsoft solutions are built to support regulatory requirements such as FERPA, COPPA, and the EU AI Act.
- Additional safeguards beyond privacy: Our AI Assurance and Copyright Commitment provide additional safeguards to help schools innovate responsibly.

Privacy-first compliance solutions

The privacy principles are embedded in Microsoft security and compliance solutions that help schools both reduce institutional risk and adopt AI responsibly:

- Microsoft 365 Advanced Data Residency add-on (ADR) expands coverage of Microsoft 365 services and customer data, provides committed data residency in local datacenter regions, and offers prioritized tenant migration services.
- Microsoft Purview Compliance Manager support alignment with FERPA, COPPA, and state regulations while also ensuring data residency and protecting data in use within AI systems.

Together, these tools help institutions strengthen compliance, transparency, and governance over AI data handling—reducing risk while enabling schools to explore new opportunities with AI in teaching and learning. With Microsoft, privacy isn't a barrier to AI, it's the foundation that allows schools to innovate responsibly.

Discover how Microsoft helps education leaders balance innovation with responsibility—so you can lead your institution confidently into the AI era. Microsoft Trust Center: Data protection and privacy

Next steps

Build the future of learning on a foundation of trust

AI presents a powerful opportunity for teaching and learning, but realizing its full potential starts with thoughtful planning and trust. Microsoft provides the tools, guidance, and partnership to help schools build a secure and inclusive AI strategy that supports students, staff, and community.

Begin your journey towards trustworthy AI:

1. **Explore** Microsoft AI in education to responsibly accelerate learning, prepare students for the future, improve efficiency and security, and give back energy for what matters most.
2. **Review the** Classroom toolkit: Unlocking generative AI safely and responsibly to create an immersive and effective learning experiences.
3. **Evaluate AI systems** using our Checklist for evaluating the trustworthiness of AI systems to assess adherence to responsible AI principles.

AI adoption in education is not just about what's possible, it's about what's responsible. Microsoft is here to help you build a future where innovation is grounded in privacy, safety, and trust.

Learn more about AI tools in the classroom with Microsoft Education to see how you can use AI to personalize learning.

Microsoft

Checklist for evaluating the trustworthiness of AI solutions

Evaluating AI systems before adoption–whether it's a Microsoft solution or a third-party application–can put you on a path towards trustworthy AI experiences. Consider following these steps:

- ❑ **Privacy and security:** Assess the risk of breaches and tools that align to school values and regulations about privacy and security. Implement data permissions, governance, and threat protection tools.
- ❑ **Reliability and safety:** Review safety of AI systems at purchase and through ongoing monitoring. Perform regular stress testing, maintenance, feedback, and evaluation so that tools perform as expected.
- ❑ **Accountability:** Keep people at the center of AI solutions. Establish oversight and accountability to mitigate adverse impacts, implement adequate controls, and assess if tools are fit for purpose.
- ❑ **Inclusiveness:** AI should be accessible to people of all abilities. Follow accessible design principles and comply with the Individuals with Disabilities Education Act (IDEA) when creating or procuring any AI tool.
- ❑ **Transparency:** Communicate openly about how, when, and why AI is used in schools and classrooms. Clear communication builds confidence and improves use of tools.
- ❑ **Fairness:** To promote fairness, assemble a diverse AI team, address stereotypes and statistical bias in datasets, and employ expert human review in decisions using AI to prevent biased outcomes.