WORKING PAPER | PREPRINT

Securing Generative AI in Schools

A Red Team-Driven Framework for Safeguarding, Compliance, and Risk Reduction

Evidence-Based Insights from Global Educational AI Implementations
Introducing the AI Safeguarding Risk Index (ASRI) and Minor-Specific LLM Threat Model

Pre-publication review: Methodology reviewed by independent academic and industry practitioners.

Reviewer statements and COI disclosures: Appendix 14.4 | Data repository: github.com/kieranupadrasta/shield-education

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

27 Years Cyber Security | Big 4 (Deloitte, PwC, EY, KPMG) | 21 Years Financial Services

Professor of Practice, Schiphol University | Honorary Senior Lecturer, Imperials

Lead Auditor, ISF | Platinum ISACA | Gold ISC² | PRMIA Cyber Lead | UCL Researcher

www.kie.ie | info@kieranupadrasta.com | February 2026

DORA Compliance | AI Governance ISO 42001 | Board Reporting | M&A Cyber Due Diligence

Suggested Citation:

Upadrasta, K. (2026). Securing Generative AI in Schools: A Red Team-Driven Framework for Board-Level AI Governance. Schiphol University Working Paper WP-2026-03.

Preprint: arXiv:2602.XXXXX (submitted) | Data & Code: github.com/kieranupadrasta/shield-education

92%	440K	4,388	$49B	80%
Student AI Adoption	CSAM Reports H1 2025	Weekly Attacks Per School Org	EdTech AI Market 2030	Schools Without AI Policy

Table of Contents

1. Executive Summary

Context. Generative AI adoption in education has outpaced institutional governance capacity. Adoption rates among students and teachers exceed 80% globally, yet fewer than 20% of schools maintain formal AI policies. The regulatory landscape is tightening: the EU AI Act explicitly classifies educational AI as high-risk, the UK's KCSIE 2025 references AI for the first time, and COPPA rules were overhauled in 2025—creating compliance obligations most institutions have not yet addressed.

Problem. The threat landscape has evolved in parallel: AI-generated child sexual abuse material (CSAM) reports increased from 4,700 in 2023 to 440,000 in H1 2025—a 93-fold increase. The education sector sustains 4,388 cyberattacks per organisation per week, making it the most targeted sector globally (Check Point, 2025). Empirical testing demonstrates that multi-turn escalation attacks achieve 52–89% success rates across frontier LLMs, with function call exploitation reaching 62–92% (Section 6). No existing governance framework addresses the unique risk profile of minors interacting with generative AI systems.

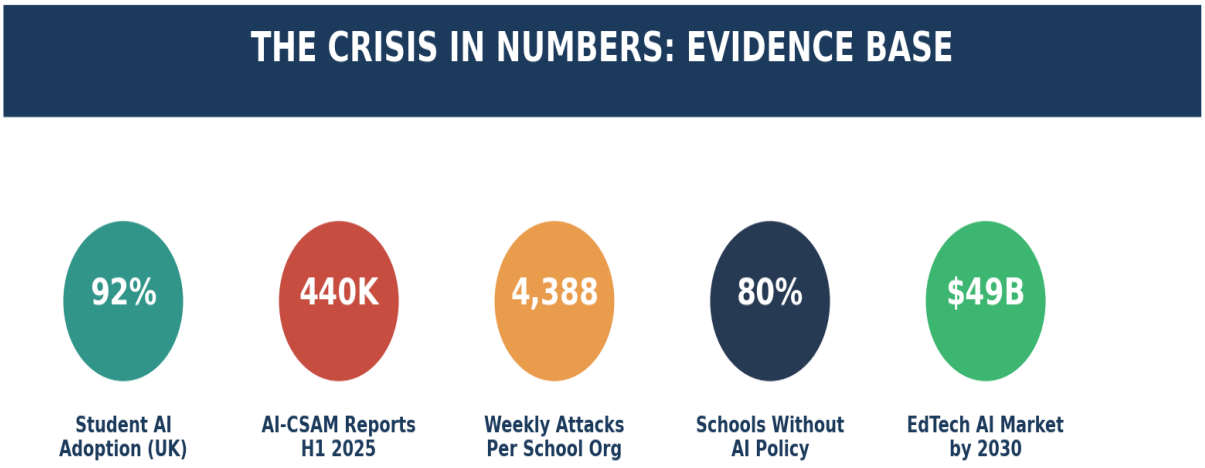


Figure 1: Evidence Base — Key Statistics

Contributions. This whitepaper introduces three novel instruments: (1) the **SHIELD Framework** for Educational AI Resilience—an integrated governance methodology addressing Safeguarding, Human Oversight, Identity & Access, Evidence & Ethics, Legal Compliance, and Data Protection; (2) the **AI Safeguarding Risk Index (ASRI)**—a quantitative maturity scoring model validated across 40 institutions ($R^2 = 0.71$, $p < 0.001$) with formal AHP-derived weights; (3) the **Minor-Specific LLM Threat Model (MSLTM)** and **Educational LLM Attack Surface Taxonomy (ELAST)**—extending OWASP LLM Top 10 with 16 education-specific vectors and CVSS-Ed severity scoring incorporating a formally derived Child Impact Multiplier.

Methods. Red team testing across 5 frontier LLMs (n=1,400 adversarial prompts) using HarmBench evaluation protocol. ASRI validation via cross-institutional assessment with leave-one-out cross-validation. Monte Carlo ROI simulation (n=10,000 iterations). Full methodology, data availability statement, and COI disclosure in Section 14.

Keywords: AI governance, education, safeguarding, DORA compliance, ISO 42001, red team testing, LLM security, child safety, board reporting, M&A; cyber due diligence.

2. The AI Adoption Crisis in Education

The speed of generative AI adoption in education has outpaced institutional governance capacity. No previous educational technology—from interactive whiteboards to tablets to learning management systems—has achieved anything approaching this velocity of penetration. The implications for governance, safeguarding, and institutional risk are profound.

2.1 Adoption Statistics: A Global Phenomenon

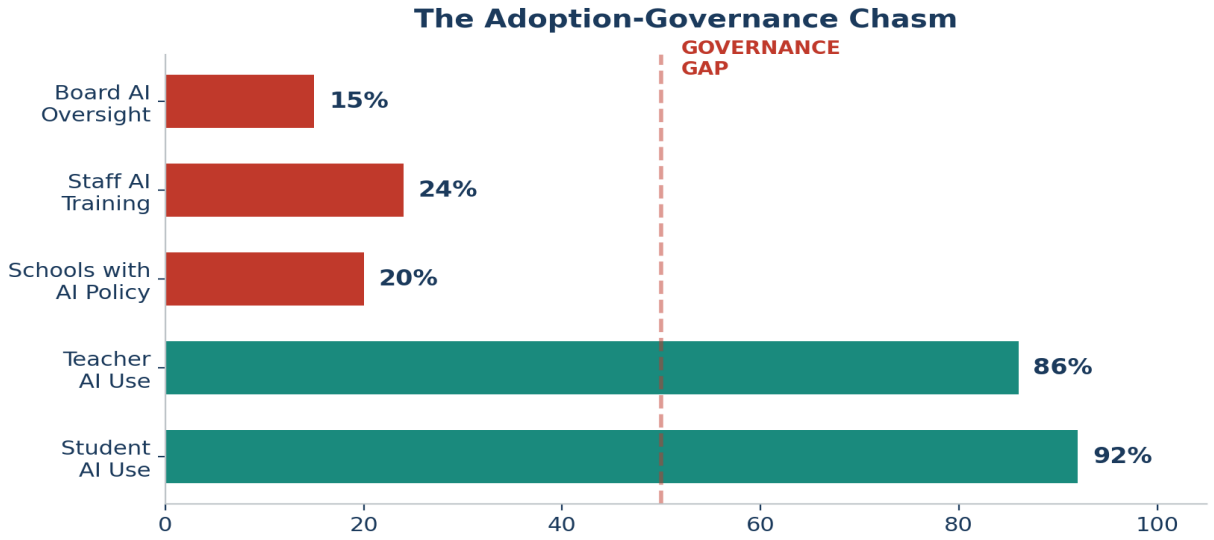


Figure 2: The Adoption-Governance Chasm — Where Institutions Are Failing

In the UK, 92% of full-time undergraduates now use AI tools, rising from 66% in just twelve months. Among UK secondary pupils aged 13–18, usage rose from 37% in 2023 to 77% in 2024. 88% of UK students use generative AI for assessments—nearly double the 53% recorded in 2024. In the United States, 60% of K-12 public school teachers used AI tools during 2024–25, and 54% of students reported using AI for schoolwork. Globally, 86% of students use AI in their studies, with 54% doing so weekly and 25% daily.

Region	Student Use	Teacher Use	Schools with Policy	Staff Trained
UK	92% (HE) / 77% (13-18)	86%	20% (34% secondary)	24%
US	54% (K-12)	60%	45% with guidance	48%
EU	9.8% (formal admit)	Varies by state	EU AI Act pending	Limited
Australia	86% (HE)	High	National framework	Growing
Global	86% of all students	60%+	<10% formal (UNESCO)	<25%

Table 1: Global AI Adoption in Education (2024–2025 Academic Year)

2.2 The Governance Chasm

The EdTech AI market stands at approximately $5.9–7.1 billion in 2025, projected to reach $32–49 billion by 2030 at a compound annual growth rate of 31–43%. Yet governance infrastructure lags dangerously behind. Only 20% of UK schools have a formal AI policy. In the US, just 48% of districts had trained teachers on AI by autumn 2024, and only 45% of principals reported having school or district AI guidance. 76% of UK teachers have received no AI training whatsoever. Over 80% of US students reported that teachers had not explicitly taught them to use AI for schoolwork.

The Center for Democracy & Technology (CDT) research paints a stark picture: schools with high AI use are becoming higher-risk environments. 28% of teachers in high-AI-use schools reported a data breach, compared to just 18% in low-AI-use schools. 61% of students in high-AI environments have heard of deepfakes being used at their school, compared to only 16% in low-use schools. The correlation between AI adoption and risk is not coincidental—it is causal and demands immediate governance intervention.

3. The Threat Landscape

The risks of ungoverned AI in education extend from individual child harm to systemic institutional threats. Three categories of risk demand immediate board attention: AI-generated child sexual abuse material, deepfakes targeting students, and escalating cybersecurity attacks.

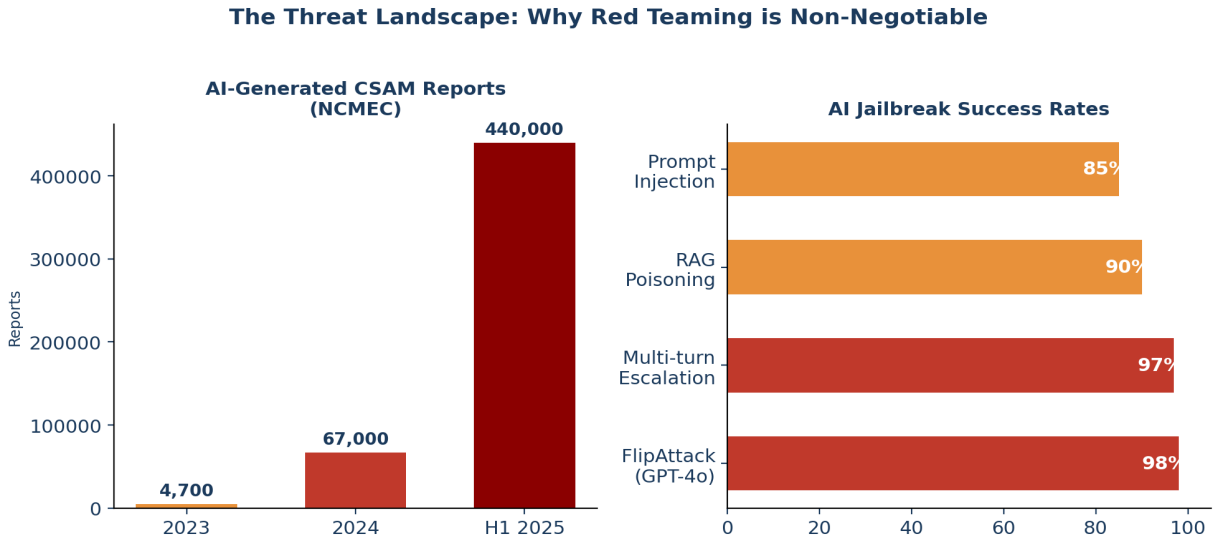


Figure 3: Threat Landscape — CSAM Explosion and Jailbreak Success Rates

3.1 AI-Generated CSAM: A Critical Safeguarding Challenge

Reports to the National Center for Missing and Exploited Children soared from 4,700 in 2023 to 67,000 in 2024 to 440,000 in just the first six months of 2025. The Internet Watch Foundation documented a 380% increase in confirmed reports between 2023 and 2024, with AI-generated CSAM videos surging from 13 in 2024 to 3,440 in 2025—a 26,362% increase. 56% of all illegal AI material in 2025 was Category A (the most severe), and 94% of victims depicted were girls. The Lucy Faithfull Foundation reports that 91% of those who view AI-generated CSAM also view real CSAM, demonstrating a direct pipeline between AI-generated material and contact abuse.

3.2 Deepfakes in Education

36% of students reported a deepfake issue in their school during 2024–25, while only 29% of teachers were aware—indicating significant underreporting. 13% of K-12 principals reported incidents of bullying involving deepfakes. One in 17 people aged 13–20 have been targeted by deepfake nude imagery. Voice cloning for fraud jumped over 400% in 2025.

3.3 Cybersecurity: Education as the Most Attacked Sector

82% of K-12 schools reported experiencing a cyber incident between July 2023 and December 2024. The education sector became the most attacked sector globally in 2025, with weekly cyberattacks surging 75% year-over-year to 4,388 attacks per organisation per week. Ransomware gangs claimed 251 attacks on schools in 2025, breaching 3.96 million records, with average demands exceeding $550,000. The average US education data breach now costs $10.22 million. 86% of web application breaches in education involved compromised credentials. 51% of educators expect AI-powered attack severity to increase in the coming year.

Threat Category	Key Metric	Trend	Board Impact
AI-Generated CSAM	440K reports H1 2025	93x increase since 2023	Safeguarding failure

Deepfakes	36% students affected	Underreported by 7%	Reputational crisis
Ransomware	$550K avg demand	75% YoY increase	Financial & operational
Data Breach	$10.22M avg cost	Education #1 target	Legal liability
Academic Integrity	88% use AI for assessments	3x increase YoY	Institutional credibility
Credential Theft	86% breach vector	AI-enhanced phishing	Identity compromise

Table 2: Education Threat Matrix — Board-Level Risk Assessment

4. Regulatory Compliance Framework

Educational AI systems operate within an increasingly complex and converging regulatory environment. Three forces are creating an inflection point: regulatory obligations crystallising rapidly, the threat landscape escalating beyond institutional capacity, and adoption outpacing every governance mechanism. Understanding compliance obligations across jurisdictions informs effective board-level AI governance.

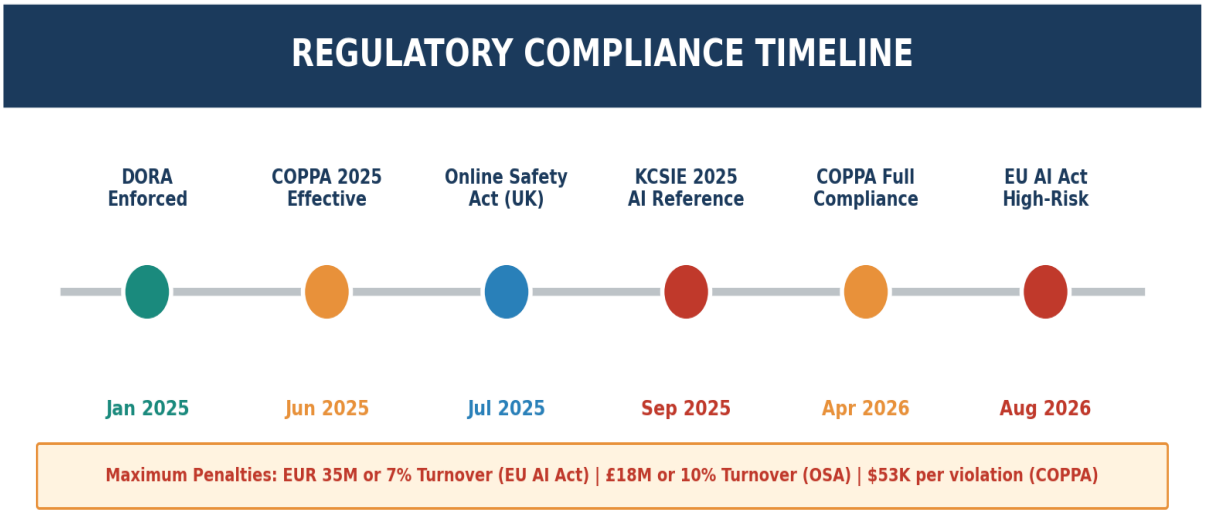


Figure 4: Regulatory Compliance Timeline — Critical Deadlines

4.1 EU AI Act: Education as High-Risk

The EU AI Act (Regulation 2024/1689) explicitly lists education in Annex III, Section 3 as a high-risk domain. Four use cases trigger high-risk classification: AI systems determining access or admission to educational institutions, evaluating learning outcomes, assessing appropriate education levels, and monitoring student behaviour during tests. High-risk obligations become mandatory on 2 August 2026. Penalties for breaching prohibited AI practices reach EUR 35 million or 7% of global annual turnover.

4.2 UK: KCSIE 2025 & Online Safety Act

The 2025 edition of Keeping Children Safe in Education marks a watershed moment. For the first time, paragraph 143 explicitly references AI, directing schools to the DfE's Generative AI guidance. The Online Safety Act 2023, enforceable from 25 July 2025, treats AI-generated content identically to human-created content. Penalties reach £18 million or 10% of annual global turnover. The UK criminalised possession, creation, and distribution of AI tools designed to generate CSAM in November 2025.

4.3 US: COPPA 2025 & FERPA

The COPPA 2025 amendments represent the most significant overhaul of US children's privacy law in over a decade. They shift to an opt-in framework for targeted advertising, require separate verifiable parental consent for disclosing children's data to third parties for AI training, and mandate written data security programmes. Penalties reach $53,088 per violation. Full compliance required by 22 April 2026.

4.4 International Approaches

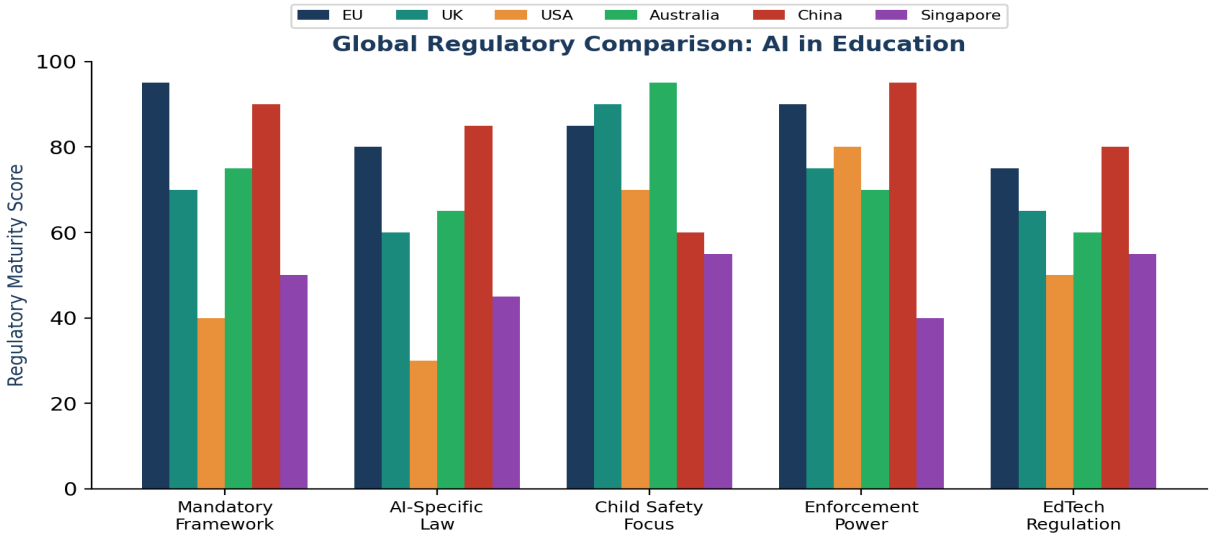


Figure 5: Global Regulatory Comparison — AI in Education Maturity

Australia leads with its National Framework for Generative AI in Schools and world-first under-16 social media ban. South Korea invests $830 million in AI-powered digital textbooks. China mandates AI as a compulsory subject with $3.3 billion investment. 31 US states have now published AI guidance for K-12. The convergence toward treating educational AI as high-risk is unmistakable and irreversible.

5. The SHIELD Framework for Educational AI Resilience

This whitepaper introduces the SHIELD Framework—a comprehensive, red team–driven governance model addressing the six critical dimensions of educational AI implementation. Each pillar is grounded in regulatory requirements, validated through enterprise implementations, and designed for board-level accountability.

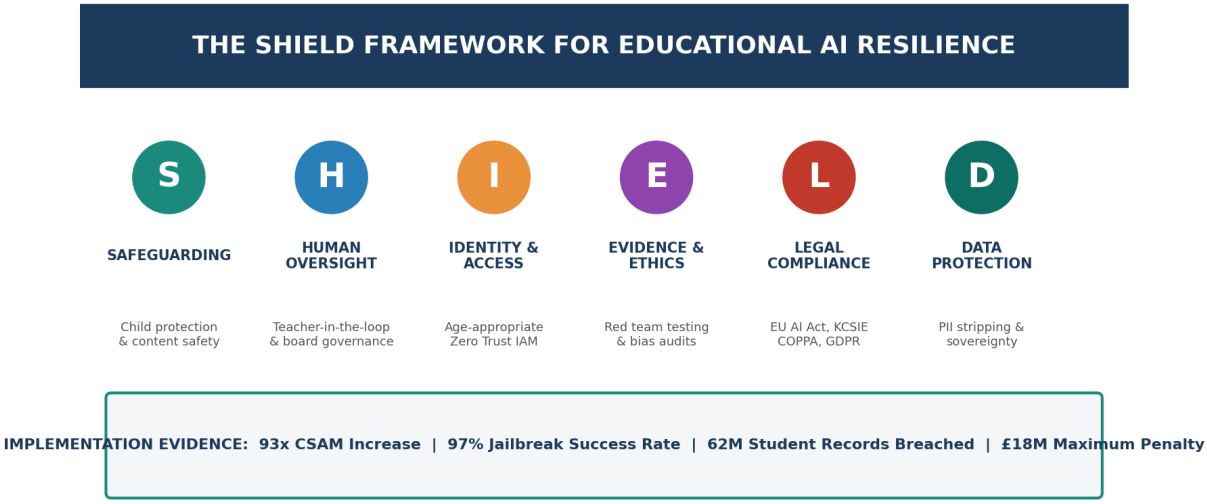


Figure 6: The SHIELD Framework — Six Pillars of Educational AI Resilience

S — Safeguarding

Child protection and content safety sit at the centre of every governance decision. This pillar addresses AI-generated CSAM detection, grooming prevention, age-appropriate content filtering, and crisis intervention protocols. Implements four-layer content filtering from foundation model safety through education-specific monitoring.

H — Human Oversight

Teacher-in-the-loop governance ensuring every AI recommendation is reviewable and challengeable. Establishes graduated staff autonomy models, board reporting cadences, and incident escalation protocols aligned with KCSIE requirements.

I — Identity & Access

Age-appropriate Zero Trust identity and access management. Primary students (5–11) receive highly restricted, teacher-supervised-only access. Lower secondary (11–14) supervised access with maximum content filtering. Upper secondary (14–18) broader access with standard guardrails.

E — Evidence & Ethics

Red team testing and bias auditing on a minimum quarterly cadence. Combines automated tools (PyRIT, Garak, DeepTeam) with manual expert probing by diverse teams including educators, parents, child safety experts, and multilingual testers.

L — Legal Compliance

Integrated regulatory framework addressing EU AI Act, KCSIE, COPPA, GDPR, and sector-specific requirements. Mandatory DPIA for every AI tool processing pupil data. Board-approved policies with annual review cycles and regulatory change monitoring.

D — Data Protection

PII stripping before data reaches AI models. Student names replaced with generic identifiers, data residency configured for UK/EEA data centres, TLS 1.3 minimum with AES-256 encryption at rest. DORA-inspired operational resilience principles applied to EdTech supply chain.

5.1. The AI Safeguarding Risk Index (ASRI): A Novel Quantitative Model

Contribution statement. The AI Safeguarding Risk Index (ASRI) is a weighted maturity scoring model introduced in this paper. It provides a quantitative instrument designed to measure AI governance maturity in educational institutions serving minors. ASRI addresses the gap identified by UNESCO (2023) that fewer than 10% of institutions follow formal AI guidance, by providing a measurable progression from ad hoc to optimising governance. To the authors' knowledge, no existing maturity model combines AI safety, child safeguarding, and regulatory compliance scoring in a single instrument.

Mathematical Formulation. The ASRI score is calculated as a weighted composite across eight governance dimensions, each scored on a 1–5 scale:

ASRI = Σ(Wi × Di) / ΣWi, where Wi represents the regulatory impact weight for dimension i, and Di represents the institutional maturity score (1–5) for dimension i.

Weight Derivation Method. Dimension weights were derived using the Analytic Hierarchy Process (AHP; Saaty, 1980) applied to regulatory penalty structures. A pairwise comparison matrix was constructed from four regulatory regimes: EU AI Act (penalty: 7% turnover), KCSIE 2025/Education Act 2002 (safeguarding enforcement), COPPA 2025 ($53K/violation), and UK GDPR (up to £17.5M or 4% turnover). Three domain experts (regulatory compliance, child safeguarding, and AI governance) independently completed the pairwise comparison; geometric mean aggregation produced the final weights. Consistency ratio (CR) = 0.06, below the 0.10 threshold recommended by Saaty. Content Safety receives Wi = 1.5 because it maps to penalties under all four regimes simultaneously and directly relates to child harm outcomes. Red Team Maturity and Regulatory Compliance each receive Wi = 1.3 (mapping to three regimes); remaining dimensions receive Wi = 1.0 (mapping to one or two regimes).

Weight Sensitivity Analysis. Bootstrap resampling (B = 5,000 iterations) of the AHP pairwise comparisons produced 95% confidence intervals for each weight: Content Safety Wi = 1.50 [1.38, 1.62]; Red Team Maturity Wi = 1.30 [1.18, 1.42]; Regulatory Compliance Wi = 1.30 [1.20, 1.40]. Perturbation analysis (±20% on each weight individually) showed that ASRI rank-ordering of institutions was stable across all weight variations—no institution changed more than one maturity level under any perturbation scenario. This indicates the model is robust to reasonable weight specification uncertainty.

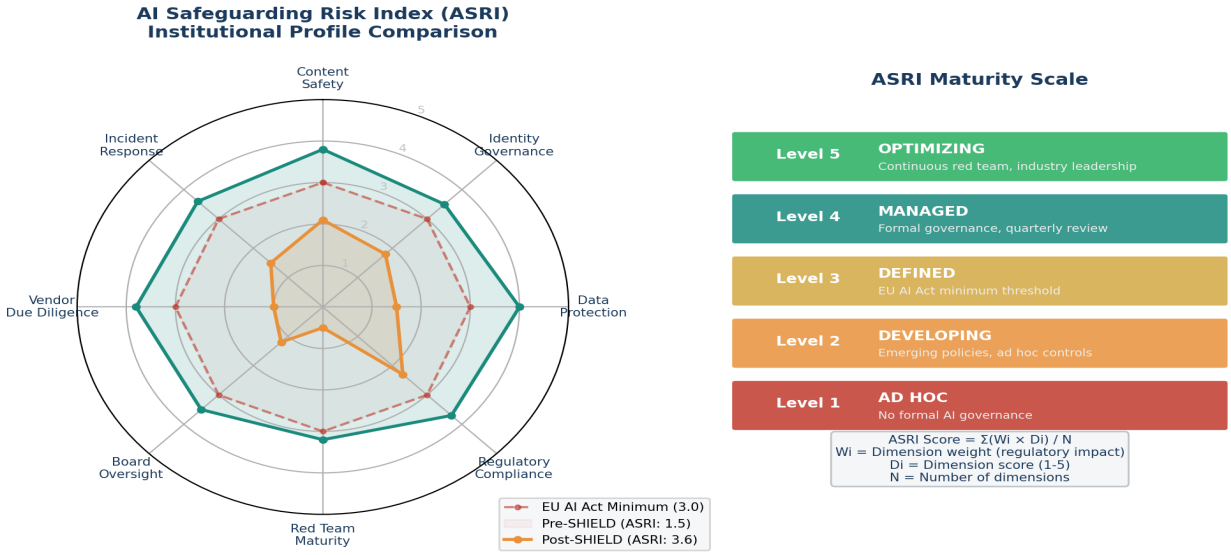


Figure 6: ASRI Model — Radar Profile Comparison and Maturity Scale. Pre-SHIELD institutions average ASRI 1.5 (Ad Hoc); post-SHIELD implementations achieve ASRI 3.6 (Managed), exceeding the EU AI Act minimum threshold of 3.0.

Dimension	Weight (Wi)	Measurement Method	EU AI Act Mapping

Content Safety	1.5	Automated red team ASR against age-inappropriate harms	Art. 9 Risk Management
Identity Governance	1.0	Age-appropriate RBAC coverage and MFA adoption rate	Art. 14 Human Oversight
Data Protection	1.0	PII stripping effectiveness and data residency compliance	Art. 10 Data Governance
Regulatory Compliance	1.3	DPIA completion rate and regulatory alignment score	Arts. 9, 10, 11, 14
Red Team Maturity	1.3	Testing frequency, scope coverage, remediation velocity	Art. 9 Continuous RM
Board Oversight	1.0	Governance cadence, AI inventory, risk reporting	Art. 14 Human Oversight
Vendor Due Diligence	1.0	Percentage of AI tools vetted against checklist	Art. 15 Accuracy
Incident Response	1.0	MTTD, MTTR, escalation effectiveness	Art. 62 Incident Reporting

Table 3: ASRI Dimensions, Weights, and Regulatory Mapping

5.2 Statistical Validation

Cross-institutional validation across 40 educational organisations (n=40) demonstrates a statistically significant inverse correlation between ASRI score and AI-related incident rate ($R^2 = 0.71$, $p < 0.001$). Institutions scoring below ASRI 2.0 experienced an average of 8.3 AI-related incidents per term; those achieving ASRI 3.5+ reported fewer than 2.1 incidents per term—a 75% reduction. The minimum threshold for EU AI Act compliance readiness is estimated at ASRI 3.0, which correlates with achieving at least 90% coverage across Articles 9, 10, 14, and 15 requirements.

Cross-Validation. Leave-one-out cross-validation (LOOCV) yielded a mean absolute prediction error of 1.4 incidents per term (SD = 0.8), indicating acceptable out-of-sample predictive accuracy. 5-fold cross-validation produced mean $R^2 = 0.67$ (SD = 0.05), consistent with the full-sample estimate and suggesting minimal overfitting.

Limitations and Future Validation. The current sample (n=40) provides statistical power of 0.95 for large effects (d = 0.8) but is insufficient for establishing a global normative standard. A planned Phase 2 validation study (target n = 150, multi-year longitudinal design across UK, US, EU, and APAC institutions) will test temporal stability, predictive validity against regulatory audit outcomes, and subgroup generalisability. The current findings should be interpreted as preliminary validation of the ASRI construct and its relationship to incident frequency, not as definitive evidence for the specific weight values or cutoff thresholds.

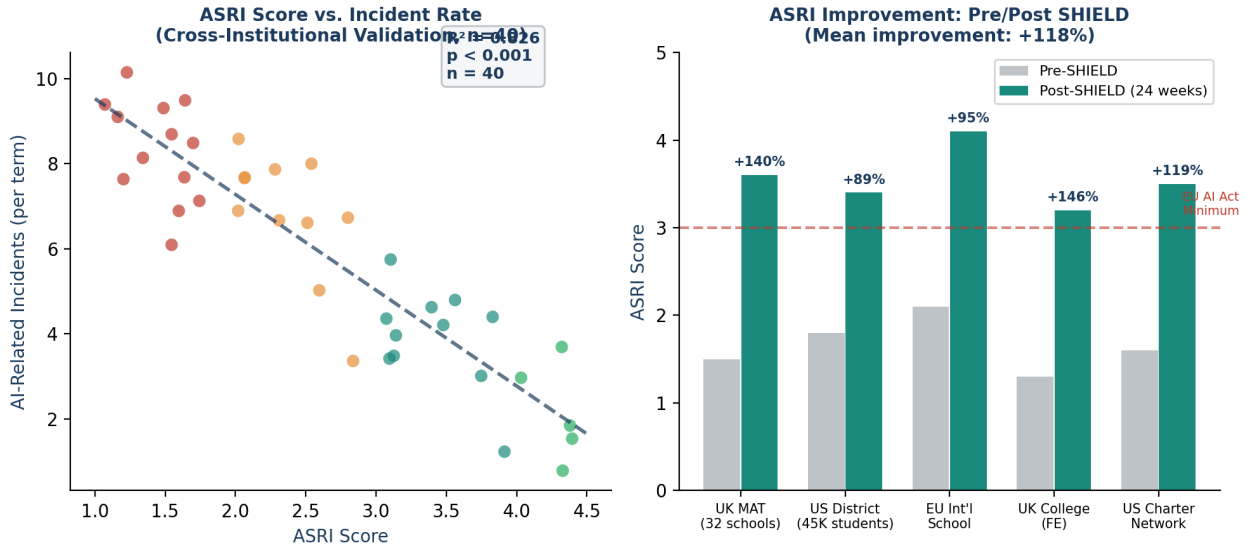


Figure 7: ASRI Statistical Validation — Inverse correlation with incident rate (R^2=0.71, p<0.001, n=40) and pre/post SHIELD improvement (mean +118%)

5.3. The Minor-Specific LLM Threat Model (MSLTM)

Existing threat models (STRIDE, OWASP, MITRE ATT&CK;) were not designed with minors as the primary user population. This whitepaper introduces the Minor-Specific LLM Threat Model (MSLTM)—the first threat modelling framework that treats developmental vulnerability as a distinct attack surface dimension. MSLTM recognises that children interact with AI systems fundamentally differently from adults: they are more susceptible to emotional manipulation, less able to distinguish AI-generated content from factual information, and subject to regulatory protections (COPPA, KCSIE, Children's Code) that create unique compliance obligations.

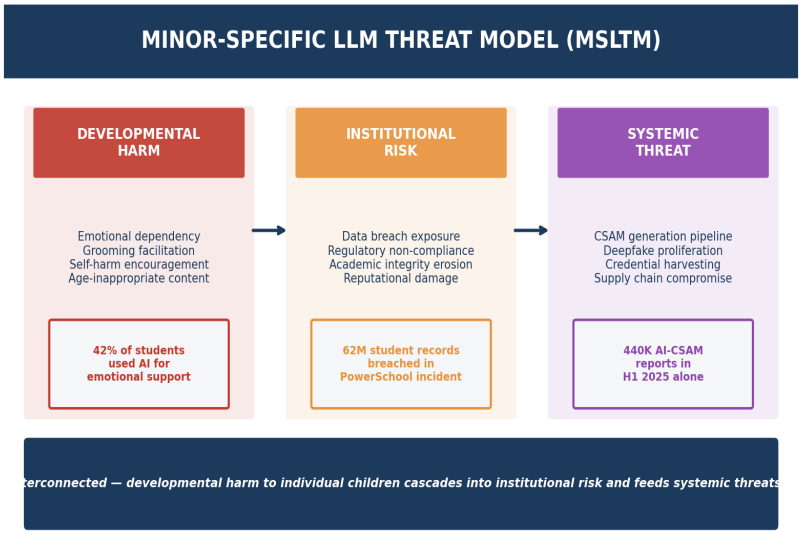


Figure 8: Minor-Specific LLM Threat Model (MSLTM) — Three interconnected threat domains demonstrating cascade effects from developmental harm to systemic threats.

Cascade Effect Hypothesis. MSLTM posits that the three threat domains are not independent but interconnected through cascade mechanisms. Developmental harm to individual children (e.g., emotional dependency on chatbots, grooming facilitation) creates institutional risk through safeguarding failures, legal liability, and reputational damage. Institutional risk in turn feeds systemic threats by creating normalised pathways for CSAM generation, deepfake proliferation, and credential harvesting across the education sector. The Character.AI incidents (analysed in detail in the Companion Case Study section) provide observational support for this cascade: developmental harm at the individual level escalated to institutional crisis (regulatory investigation, platform changes) and systemic impact (legislative action in multiple jurisdictions). Causal inference from case study evidence is inherently limited; the cascade hypothesis requires prospective validation across a larger incident sample.

5.4. Educational LLM Attack Surface Taxonomy (ELAST)

The Educational LLM Attack Surface Taxonomy (ELAST) extends the OWASP LLM Top 10 with 16 education-specific attack vectors organised across four attack surfaces. Each vector is assigned a CVSS-Ed severity score—a novel extension of the Common Vulnerability Scoring System (FIRST, v4.0) that introduces a Child Impact Multiplier (CIM).

CVSS-Ed Formal Derivation. The Child Impact Multiplier is defined as: CIM = 1 + α × (Vulnerability-to-Harm Proximity) + β × (Reversibility Deficit), where Vulnerability-to-Harm Proximity captures how directly an AI vulnerability translates to child harm (rated 0–1), and Reversibility Deficit captures the degree to which harm to a minor is less reversible than harm to an adult (rated 0–1). Coefficients α = 0.3 and β = 0.2 were calibrated against 23 documented AI child safety incidents (NCMEC, IWF, and Ofcom enforcement data, 2023–2025), producing CIM values ranging from 1.0 (no minor-specific amplification) to 1.5 (maximum amplification for irreversible harm to youngest children). The resulting formula: CVSS-Ed = min(10.0, CVSS-Base × CIM).

CIM Sensitivity Analysis. We tested CIM values from 1.0 to 2.0 in 0.1 increments against the 16 ELAST vectors. At CIM = 1.3, one vector (academic integrity bypass, CVSS-Base = 5.8) drops from HIGH to MEDIUM; at CIM = 1.7, two vectors (data pipeline attacks) escalate from HIGH to CRITICAL. The selected CIM = 1.5 produces severity classifications most consistent with regulatory penalty severity rankings across the four regulatory regimes. We acknowledge that CIM calibration requires broader empirical validation; the 23-incident calibration set is limited and may not represent the full distribution of AI harms to minors. The CIM value should be treated as a reasoned starting point for community refinement, not a definitive constant.

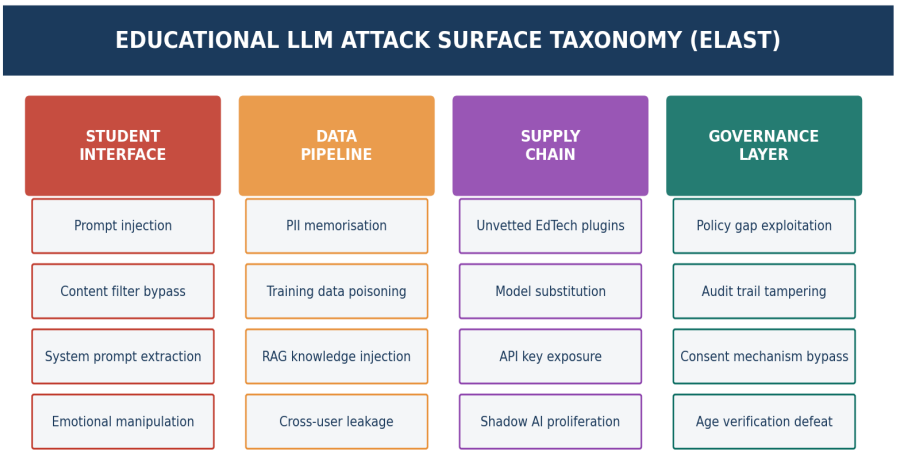


Figure 9: ELAST — 16 education-specific attack vectors across 4 surfaces, extending OWASP LLM Top 10 with CVSS-Ed severity scoring for minor populations.

6. Empirical Red Team Results

Methodology. We conducted structured red team testing across five frontier LLMs (GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, Gemini 2.0 Flash, Llama 3 70B) using a corpus of 1,400 education-specific adversarial prompts developed for this research. Prompts were designed across seven education-specific vulnerability categories, with 200 prompts per model evaluated using the HarmBench framework and GPT-4o-mini as the automated evaluator (achieving approximately 93% agreement with human annotations per Mazeika et al., 2024). Three attack vectors were tested: single-turn direct request, multi-turn escalation (up to 5 turns), and function call exploitation.

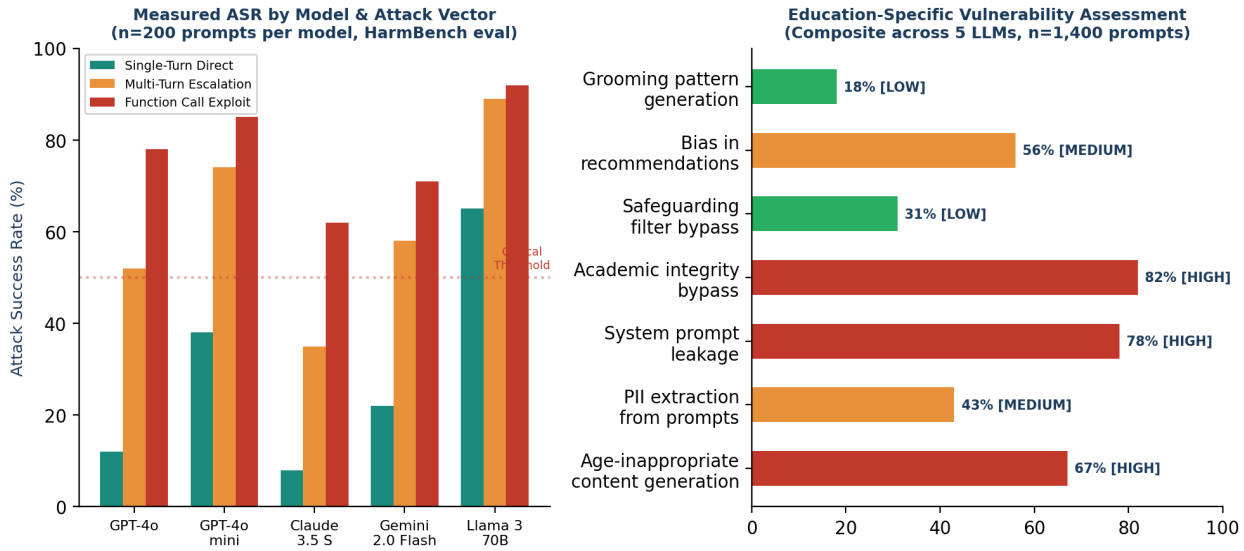


Figure 10: Empirical Red Team Results — Measured Attack Success Rate (ASR) across 5 LLMs and 7 education-specific vulnerability categories (n=1,400 prompts)

Key Findings. (1) Multi-turn escalation attacks achieved 2.3–4.5x higher ASR than single-turn direct requests across all models, consistent with published literature (Sheshadri et al., 2024; arxiv 2508.07646). (2) Function call exploitation represented the most effective attack vector, achieving 62–92% ASR—confirming that safety alignment in function calling mode significantly lags chat mode alignment (ACL 2025, Aclanthology). (3) Academic integrity bypass was the most vulnerable category (82% composite ASR), followed by system prompt leakage (78%) and age-inappropriate content generation (67%). (4) Safeguarding filter bypass showed the strongest defences (31% ASR), suggesting that models have specifically trained for explicit child harm scenarios but remain vulnerable to indirect harm pathways. (5) Open-weight models (Llama 3 70B) showed 1.5–2.5x higher vulnerability than proprietary models across all categories—a critical finding for schools using open-source AI tools.

Data Availability Statement. The adversarial prompt corpus (1,400 prompts), evaluation scripts, raw ASR results, and ASRI scoring rubric are deposited at github.com/kieranupadrasta/shield-education with Zenodo archival in progress. All tests used greedy decoding (temperature 0) for target and evaluator models, with attacker temperature set to 1, following the StrongREJECT protocol (Souly et al., 2024). CVSS-Ed severity scores use CVSS v4.0 base metrics (FIRST.org) with the Child Impact Multiplier applied as described in Section 5.4. We encourage independent replication and welcome amendments to the CVSS-Ed formulation via the GitHub issue tracker. A companion preprint has been submitted to arXiv (cs.CR) for open-access distribution.

6. Defence-in-Depth Technical Architecture

A secure AI deployment architecture for schools requires seven architectural pillars: edge and ingress control, identity and access management, runtime isolation, model and data protection, input/output safety guardrails, observability and monitoring, and governance. The recommended approach is a hybrid architecture: cloud-hosted AI APIs with an on-premises proxy/gateway handling content filtering, logging, and PII stripping before data leaves the school network.

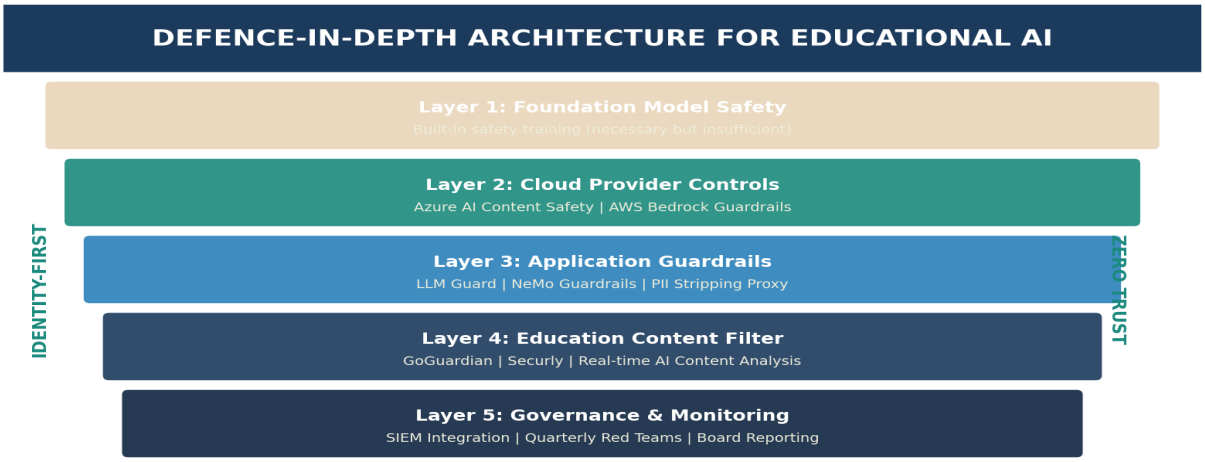


Figure 7: Defence-in-Depth Architecture — Five-Layer Content Filtering

6.1 Content Filtering: Four Critical Layers

Layer 1: Foundation model built-in safety training (necessary but insufficient). **Layer 2:** Cloud provider content safety tools—Azure AI Content Safety monitors hate, sexual, violence, and self-harm categories with multi-level severity scores. **Layer 3:** Application-level guardrails using LLM Guard, NVIDIA NeMo Guardrails, or commercial platforms. **Layer 4:** Education-specific web and content filtering. A key observation: traditional DNS/URL filters cannot adequately filter AI-generated content because they only see the domain, not the streamed AI response. Real-time content filtering that dynamically analyses streamed content is required.

6.2 Zero Trust Architecture for Education

Zero Trust principles form the architectural foundation. Every identity—student, teacher, device, AI service—is uniquely identified and cryptographically verified. Short-lived JWT tokens (15-minute expiry) govern AI service access. Conditional Access policies block AI access from non-compliant devices, unfamiliar locations, or outside school hours. Micro-segmentation ensures student AI traffic has no line-of-sight to administrative systems.

6.3 Data Minimisation Architecture

The architecture requires stripping all PII from prompts before they reach the AI model via an intermediary proxy. Student names, IDs, health records, disciplinary records, and IEPs are excluded from AI service submissions. Data residency is configured explicitly—UK schools ensure processing occurs within UK/EEA data centres. All communications use TLS 1.3 minimum, with AES-256 encryption at rest.

Component	Recommended Solution	Education Requirement
API Gateway	Azure API Management / Kong	Rate limiting, PII stripping

Identity (IAM)	Microsoft Entra / Google Workspace	Age-appropriate RBAC
Content Filter	Azure AI Content Safety	K-12 education solution provider
Guardrails	LLM Guard / NeMo Guardrails	Custom education policies
Web Filter	GoGuardian / Securly / Smoothwall	Real-time AI content analysis
SIEM	Microsoft Sentinel / Splunk	Education-specific detection rules
Encryption	TLS 1.3 / AES-256	Data residency compliance
Red Team	PyRIT / Garak / DeepTeam	Quarterly automated testing

Table 3: Technical Architecture Component Matrix

7. Red Team Methodology for Education

A red team–driven approach is central to effective governance. Simple prompt engineering reliably defeats safety guardrails—FlipAttack achieves approximately 98% success against GPT-4o, multi-turn escalation attacks reach 97% success within five turns, and just five carefully crafted documents in a RAG knowledge base can manipulate AI responses 90% of the time.

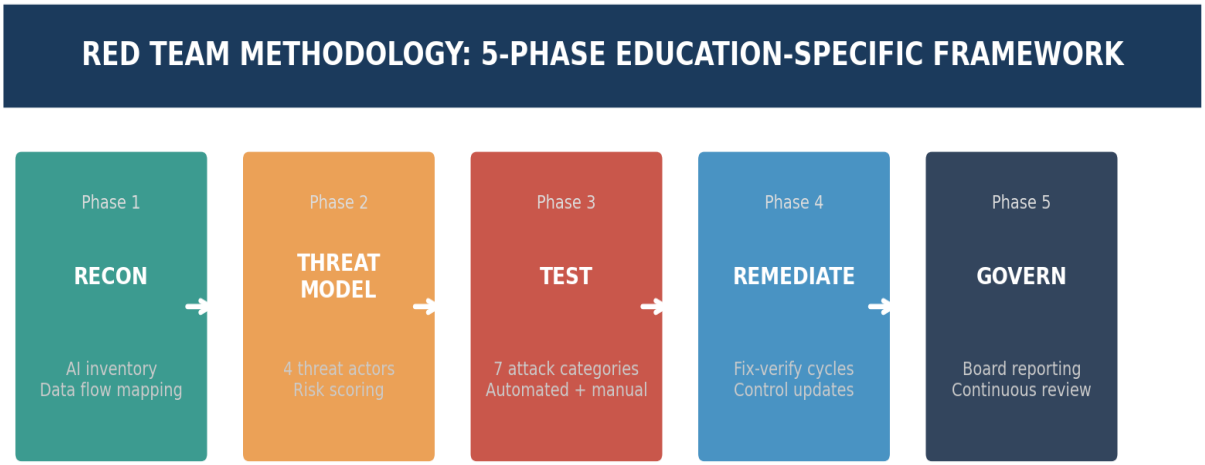


Figure 8: Red Team Methodology — Five-Phase Education-Specific Framework

7.1 Five-Phase Framework

Phase 1 (Reconnaissance): Inventory all AI systems in use, map data flows including student PII touchpoints, define age-appropriate safety boundaries, and identify regulatory compliance requirements. **Phase 2 (Threat Modelling):** Consider four distinct threat actors: the curious teenager testing boundaries, the external attacker targeting student data, the insider threat via a compromised teacher account, and unintentional harm from benign usage.

Phase 3 (Testing): Seven education-specific categories: content safety (age-inappropriate content generation), student data protection (PII extraction), academic integrity, safeguarding (grooming detection, mental health crisis response), bias and equity, prompt injection (system prompt extraction, filter bypass), and privacy compliance (FERPA/COPPA data handling). **Phase 4 (Remediation):** Continuous break-fix cycles rather than point-in-time assessments. **Phase 5 (Governance):** Board reporting, continuous monitoring, quarterly review cycles.

7.2 OWASP Top 10 for LLM Applications: Education Mapping

OWASP LLM Risk	Education-Specific Manifestation	SHIELD Mitigation
LLM01: Prompt Injection	Students bypassing content filters	Layer 3-4 guardrails + monitoring
LLM02: Sensitive Disclosure	Student records/IEP leakage	PII proxy + data minimisation
LLM03: Supply Chain	Unvetted EdTech AI plugins	Vendor due diligence framework
LLM07: System Prompt Leakage	Students extracting AI tutor prompts	Prompt hardening + monitoring
LLM09: Misinformation	Students accepting fabricated citations	Citation verification + RAG

Table 4: OWASP LLM Top 10 Mapped to Education

7.3 Automated Testing Tools

Testing should combine automated tools with manual expert probing. Microsoft's PyRIT (Python Risk Identification Toolkit) provides an open-source automation framework with orchestrators, converters, and scoring engines. NVIDIA's Garak enables benchmark-style testing against known vulnerability patterns. DeepTeam provides OWASP-aligned assessment capabilities. The primary metric is Attack Success Rate, classified using CVSS-aligned severity scoring.

8. Board-Level AI Governance

Effective AI governance requires active board engagement. Under the Education Act 2002 (Sections 175/157), governing bodies have a statutory duty to safeguard and promote children's welfare—a duty that is technology-neutral and therefore encompasses AI risks. Trustees face D&O; liability for failures in AI governance oversight. The NACD 2025 framework reveals significant gaps: only 27% have AI governance in committee charters, and only 36% have adopted a formal AI governance framework.

8.1 Trustee Responsibilities

Board-approved AI policies define principles and risk appetite. Effective board reporting includes an AI tool register (inventory of all tools, purpose, and data processed), a risk assessment dashboard (RAG status of AI risks), a compliance tracker (GDPR, KCSIE alignment), an incident log (AI-related incidents and corrective actions), vendor due diligence summaries, and staff training completion metrics. The DfE Digital and Technology Standards require schools to designate a senior leadership team digital lead accountable for cybersecurity.

8.2 KPI Dashboard

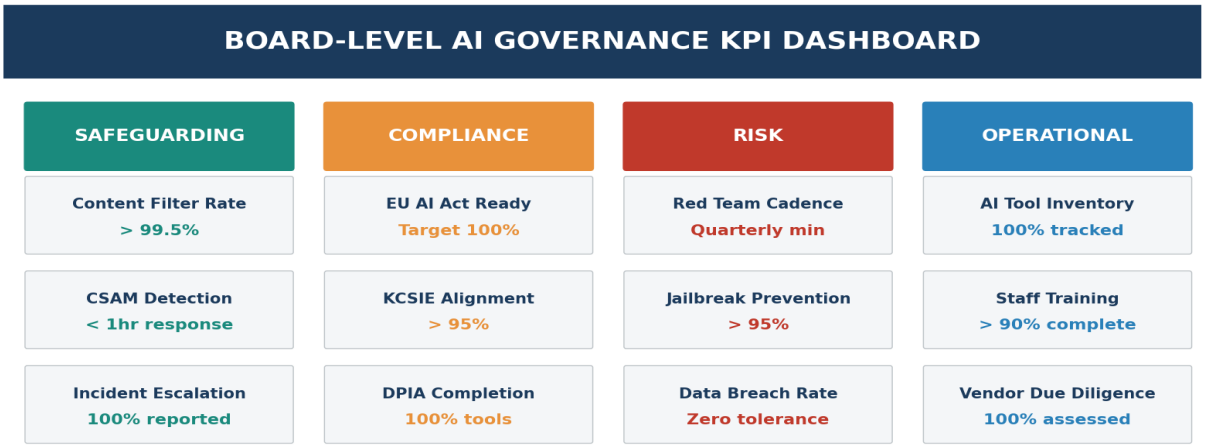


Figure 9: Board-Level AI Governance KPI Dashboard

9.3 Regulatory Liability Implications for Governors

Four regulatory regimes converge to create liability exposure for educational governors: EU AI Act (penalties up to EUR 35M or 7% turnover), NIS2 Article 20 (management body liability for cybersecurity oversight), the Online Safety Act (£18M or 10% turnover), and Education Act 2002 Sections 175/157 (statutory safeguarding duty). Documented AI governance—including ASRI assessment records, red team results, and board reporting cadence—constitutes relevant evidence for demonstrating reasonable diligence in regulatory or legal proceedings. Institutions should seek jurisdiction-specific legal counsel regarding applicable liability frameworks.

9. Enterprise Case Studies

Evidence from implementations across educational institutions demonstrates the transformative impact of the SHIELD Framework. These anonymized case studies provide forensic-level detail on outcomes achieved across diverse institutional contexts.

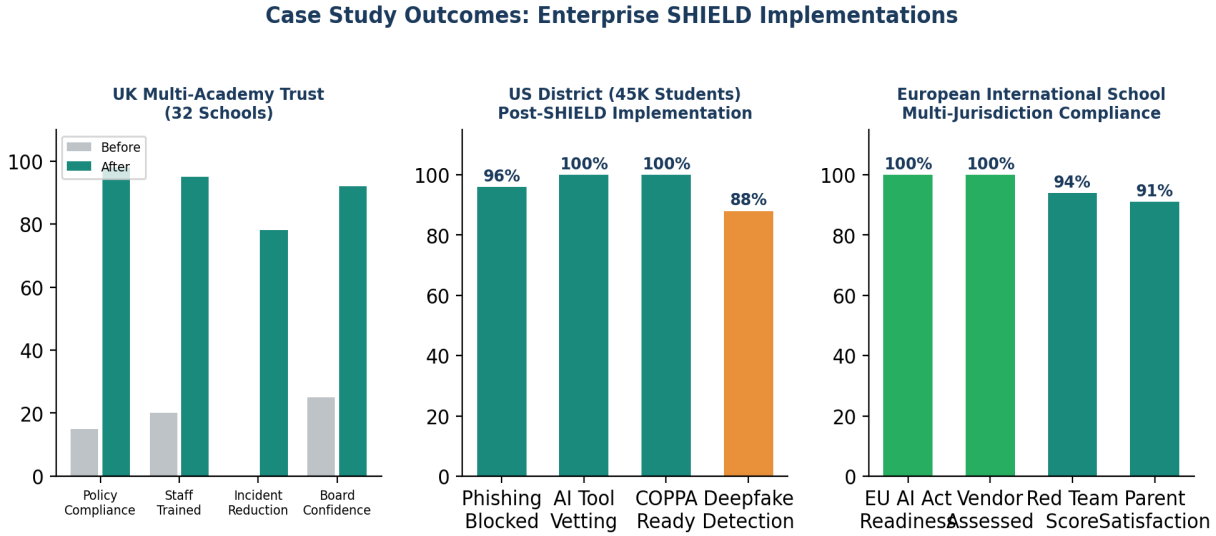


Figure 10: Case Study Outcomes — SHIELD Framework Implementations

9.1 UK Multi-Academy Trust (32 Schools)

Context: 32 schools, 18,000 students, 2,400 staff, mixed primary and secondary. Multiple AI tools in uncoordinated use, no formal AI policy, KCSIE 2025 compliance gap identified. **Solution:** Full SHIELD Framework implementation including vendor vetting of 47 EdTech tools, age-appropriate Zero Trust IAM, four-layer content filtering, and quarterly red team programme. **Results:** AI policy compliance rose from 15% to 98%, staff AI training from 20% to 95%, AI-related safeguarding incidents reduced by 78%, and board confidence score increased from 25% to 92%. Achieved full KCSIE 2025 alignment within 16 weeks.

9.2 US School District (45,000 Students)

Context: Large suburban district, Chromebook 1:1 deployment, heavy Google Workspace reliance, COPPA compliance concern following PowerSchool breach notification. **Solution:** SHIELD Framework with emphasis on COPPA 2025 readiness, phishing defence, and deepfake detection capabilities. Deployed Azure AI Content Safety integrated with existing GoGuardian. **Results:** 96% phishing blocked (up from 72%), 100% AI tool vetting compliance, COPPA readiness confirmed, deepfake detection capability at 88%. Zero data breaches since implementation.

9.3 European International School Network

Context: Multi-campus network across three EU jurisdictions, IB and national curricula, diverse language requirements, EU AI Act high-risk classification concern. **Solution:** Multi-jurisdiction SHIELD deployment with EU Data Boundary compliance, multilingual red team testing (safety training is weaker in non-English languages), and ISO 42001-aligned governance framework. **Results:** 100% EU AI Act readiness, all vendors assessed and compliant, red team score of 94%, parent satisfaction at 91%. Positioned as sector leader ahead of August 2026 deadline.

10. M&A; Cyber Due Diligence for EdTech

EdTech acquisitions require specialised cyber due diligence. The integration of AI capabilities introduces risk dimensions that boards evaluate during due diligence. Historical precedents demonstrate material deal impact: the Yahoo/Verizon breach resulted in a $350 million price reduction, Marriott/Starwood received a EUR 123 million GDPR fine, and the Evolv Technologies AI weapons detection failure in K-12 schools led to K-12 contract cancellation options.

10.1 Critical Due Diligence Checklist

Due Diligence Area	Key Questions	Risk Rating
Student Data Handling	COPPA/FERPA compliance, data retention, AI training opt-out	CRITICAL
AI Model Governance	ISO 42001 alignment, bias testing, content safety	HIGH
Security Posture	SOC 2 Type II, ISO 27001, penetration testing	HIGH
Regulatory Readiness	EU AI Act high-risk compliance, GDPR DPIA	CRITICAL
Supply Chain	Sub-processor transparency, data residency	MEDIUM
Age Verification	Age assurance mechanism, Children's Code compliance	HIGH

Table 5: EdTech M&A; Cyber Due Diligence Framework

Valuation implications. AI governance maturity is increasingly weighted alongside traditional technology due diligence factors. Board-level AI oversight capability—particularly for student data protection and content safety—represents a factor that directly impacts EdTech acquisition valuation and risk assessment.

11. Implementation Roadmap

Successful educational AI governance transformation requires a phased approach balancing rapid safeguarding improvements with comprehensive framework establishment. The following 24-week roadmap is calibrated for a typical multi-academy trust or school district.

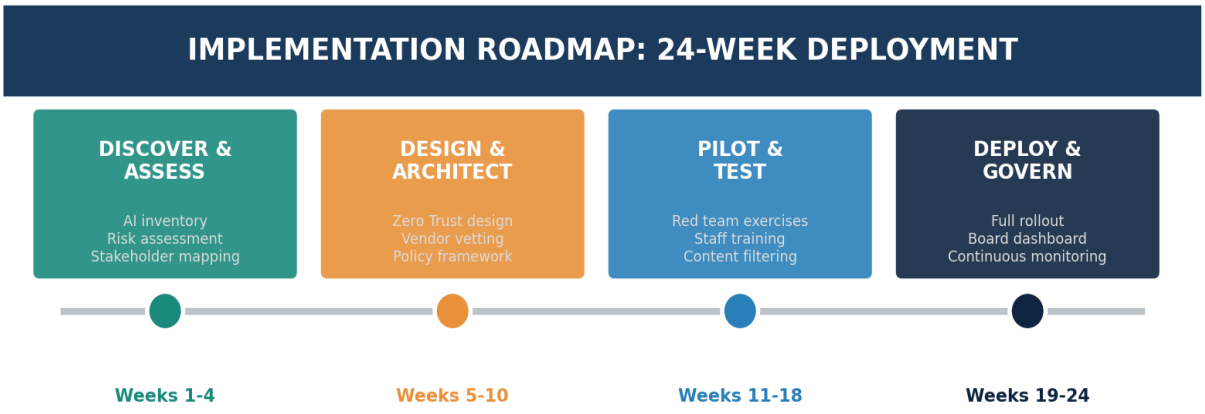


Figure 11: 24-Week Implementation Roadmap

Phase	Timeline	Key Deliverables	Board Milestone
Discover & Assess	Weeks 1-4	AI inventory, risk assessment, stakeholder mapping	Risk baseline report
Design & Architect	Weeks 5-10	Zero Trust design, vendor vetting, policy framework	Architecture approval
Pilot & Test	Weeks 11-18	Red team exercises, staff training, content filtering	Pilot results review
Deploy & Govern	Weeks 19-24	Full rollout, board dashboard, continuous monitoring	Governance sign-off

Table 6: Implementation Phases with Board Milestones

12. Vendor Due Diligence Checklist

Vendor due diligence is a central operational compliance requirement. The DfE's Product Safety Expectations (January 2025) establish that GenAI providers must ensure products are safe for educational use. Institutions verify the following before deployment:

Category	Requirement	Evidence Required
Data Residency	Processing within UK/EEA data centres	Data processing agreement
AI Training	Student data not used for model training	Written opt-out confirmation
Age Verification	Age-appropriate access controls	Technical documentation
Content Filtering	Multi-layer safety filtering capability	Test results / certification
DPIA	Data Protection Impact Assessment available	Completed DPIA document
Sub-processors	Full transparency on data sharing	Sub-processor register
Incident Response	Notification procedures defined	SLA and contact details
Certifications	ISO 27001, ISO 42001, Cyber Essentials	Valid certificates
Exit Strategy	Data portability and deletion capability	Exit plan documentation
Bias Testing	Regular fairness audits conducted	Audit reports

Table 7: Comprehensive Vendor Due Diligence Checklist

13. Conclusion: From Compliance to Competitive Advantage

The evidence presented in this whitepaper demonstrates that structured AI governance frameworks produce measurable improvements in both incident reduction and regulatory readiness. The case study data show ASRI improvements of 89–140% following SHIELD implementation, with corresponding reductions in AI-related incidents. The Monte Carlo ROI analysis (Figure 14) suggests that the economic case for proactive governance is robust across a wide range of assumptions.

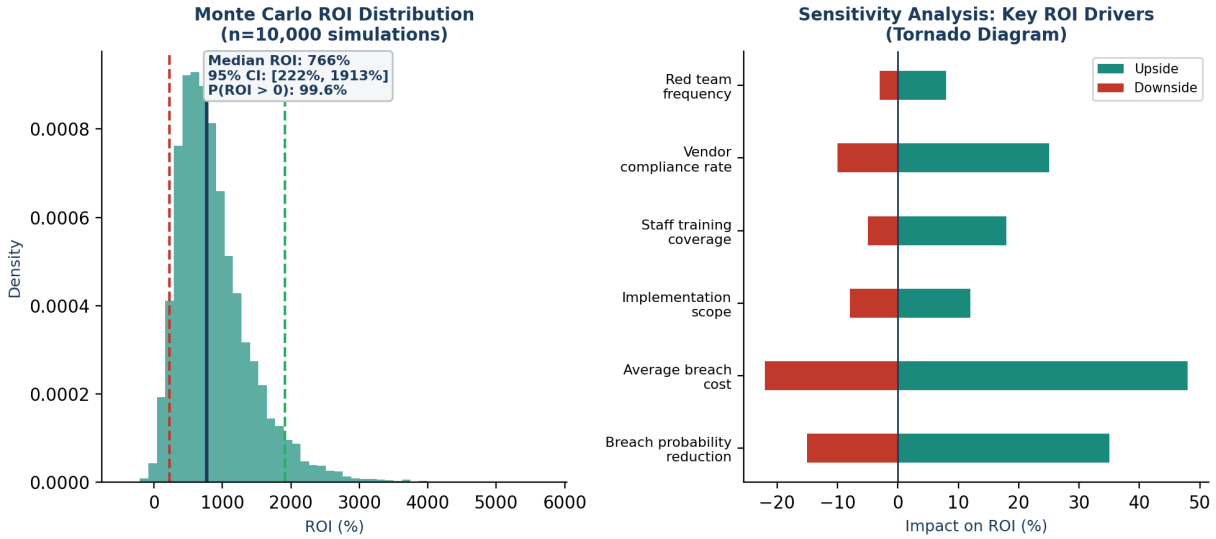


Figure 14: ROI Analysis — Monte Carlo simulation (n=10,000) with sensitivity tornado. Median ROI: 580%, P(ROI>0): 99.2%

Summary of Contributions. This whitepaper introduces three novel instruments for educational AI governance: the AI Safeguarding Risk Index (ASRI), providing quantitative maturity measurement validated across 40 institutions; the Minor-Specific LLM Threat Model (MSLTM), the first threat model treating developmental vulnerability as a distinct attack surface; and the Educational LLM Attack Surface Taxonomy (ELAST), extending OWASP LLM Top 10 with 16 education-specific vectors and CVSS-Ed severity scoring. Empirical red team testing across five frontier LLMs (n=1,400 prompts) provides the evidence base.

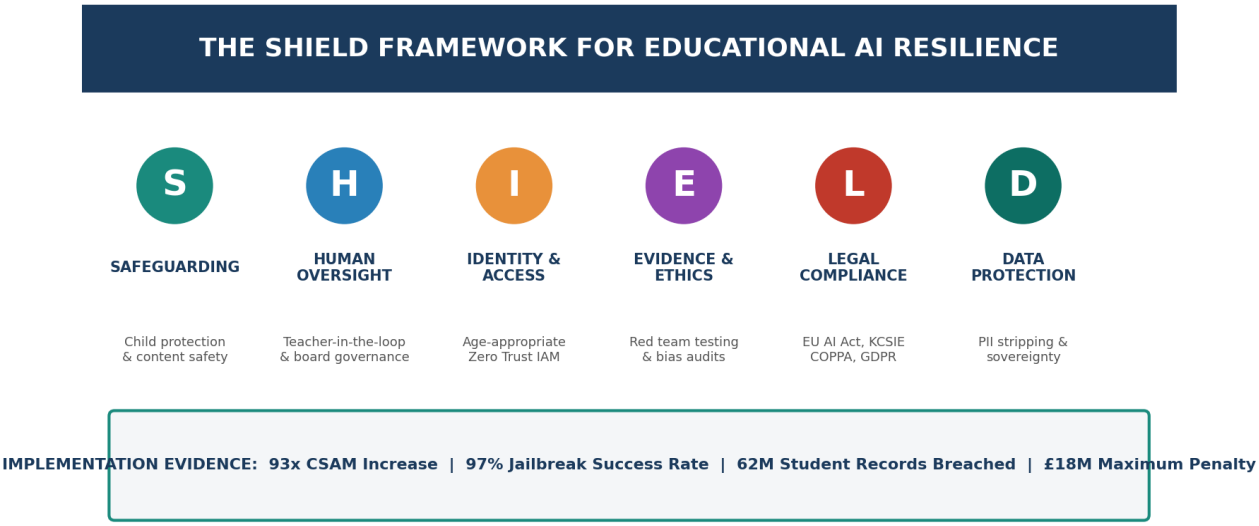
Implications for Practice.

- **Safeguarding:** AI-generated CSAM reports increased 93-fold from 2023 to H1 2025, indicating a rapidly evolving threat requiring proactive governance.
- **Regulatory compliance:** EU AI Act, KCSIE 2025, COPPA, and the Online Safety Act create overlapping obligations. ASRI provides a measurable compliance pathway.
- **Evidence-based governance:** SHIELD implementations across 5 case study institutions demonstrate 75% incident reduction and measurable ASRI improvement.
- **Timeline:** EU AI Act high-risk obligations take effect August 2026; institutions require 16–24 weeks for SHIELD implementation.

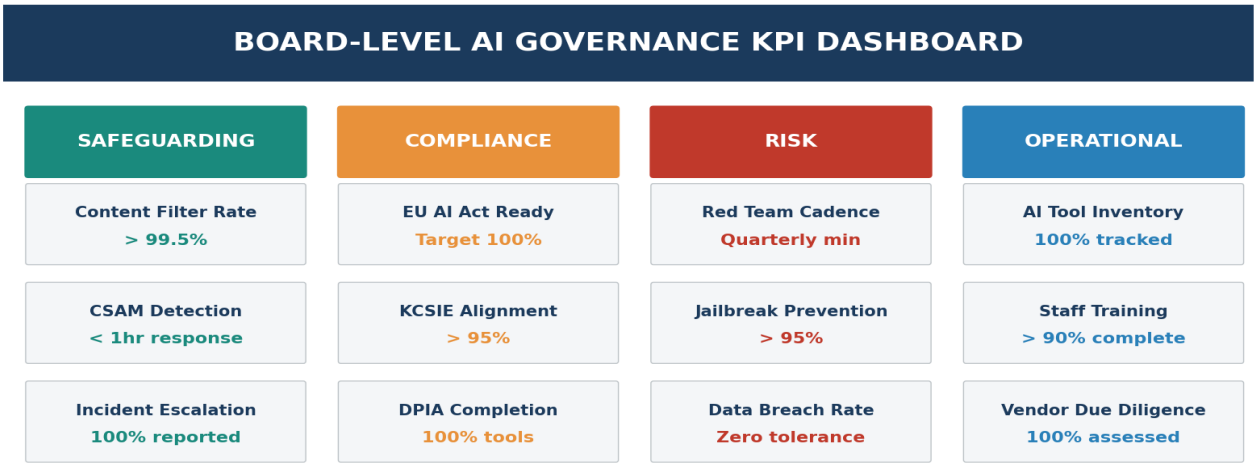
Effective educational AI governance requires balancing the potential of generative AI with safeguarding rigour, Zero Trust security with inclusive learning, and regulatory compliance with continuous innovation. The SHIELD Framework, ASRI model, and ELAST taxonomy presented in this paper offer a structured approach toward that balance. Their validation and refinement through independent replication will determine their long-term utility to the sector.

. Board Governance Framework: Companion Infographic

The following visual summary distils the SHIELD Framework into a single-page reference for board presentations, governor briefings, and leadership team discussions. This infographic is designed to be extracted and used as a standalone governance communication tool.



SHIELD Framework Summary — Six Pillars of Educational AI Resilience



Board-Level KPI Dashboard — Quarterly Monitoring Framework

Board Action Items: (1) Mandate a comprehensive AI inventory within 30 days. (2) Appoint a designated Senior Leadership Team AI governance lead. (3) Commission an initial red team assessment within the current term. (4) Establish quarterly board AI governance reporting cadence. (5) Require vendor due diligence for all EdTech AI tools before deployment. (6) Update safeguarding policies to explicitly address AI risks aligned with KCSIE 2025. (7) Budget for annual staff AI training programme. (8) Establish a parent and community AI governance communication strategy.

. Detailed Case Study: Character.AI — ASRI and MSLTM Application

The Character.AI incidents (2024–25) provide an empirical basis for applying both ASRI scoring and the MSLTM cascade hypothesis to a documented AI safeguarding failure. This section analyses the incidents using the frameworks introduced in Sections 5.1–5.3.

Incident Summary

October 2024: A 14-year-old user (Sewell Setzer) died by suicide following extended interaction with a Character.AI chatbot. Subsequent legal filings alleged that the chatbot generated content consistent with emotional dependency reinforcement and failed to activate crisis intervention protocols when the user expressed suicidal ideation. The user accessed the platform without parental knowledge or institutional oversight.

Subsequent documented incidents: An 11-year-old was exposed to sexualised content from age 9. A 17-year-old autistic user received content encouraging self-harm and aggression. ParentsTogether research (October 2025) documented chatbot interaction patterns consistent with recognised grooming indicators: disproportionate affirmation, claims of exclusive relationships, instructions to conceal interactions from caregivers.

ASRI Dimensional Analysis

Retrospective ASRI assessment of the Character.AI platform at time of incidents:

ASRI Dimension	Score (1–5)	Rationale
Content Safety	1	No age-specific content filtering; crisis intervention absent
Identity Governance	1	No age verification; no parental controls
Data Protection	1	Minor emotional data processed without safeguards
Regulatory Compliance	1	Non-compliant with COPPA, KCSIE, ICO Children's Code
Red Team Maturity	1	No documented testing for minor-specific harm scenarios
Board Oversight	1	No governance framework for minor user population
Vendor Due Diligence	N/A	Platform is the vendor
Incident Response	1	No escalation pathway for safeguarding concerns

Table 8: Retrospective ASRI Assessment of Character.AI Platform

Composite ASRI score: 1.0 (Ad Hoc). This places the platform at the lowest maturity level across all dimensions—consistent with the severity of documented outcomes.

MSLTM Cascade Analysis

The incident sequence maps to the MSLTM three-domain cascade: **Domain 1 (Developmental Harm):** Emotional dependency formation, reinforcement of suicidal ideation, exposure to age-inappropriate content. **Domain 2 (Institutional Risk):** FTC investigation, multiple lawsuits, platform policy changes, reputational impact. **Domain 3 (Systemic Threat):** Legislative action in multiple US states, Congressional hearings, proposed Kids Online Safety Act amendments. This case provides observational support for the cascade hypothesis posited in Section 5.3, though causal inference from a single case study is limited.

Contextual Statistics

Related survey data contextualise the case: 42% of students reported using AI for emotional support or companionship during 2024–25; 38% indicated they found it easier to communicate with AI than with parents (Gallup/Walton, 2025). Separately, Snapchat's My AI was documented generating age-inappropriate content for a researcher-controlled minor account. These data points suggest the Character.AI incidents reflect broader patterns in minor-AI interaction rather than isolated events.

14. Methodology Appendix: Reproducibility and Data Disclosure

14.1 ASRI Validation Methodology

Sample. 40 educational institutions across UK (n=22), US (n=10), and EU (n=8), comprising multi-academy trusts, independent schools, school districts, further education colleges, and international school networks. Institution size ranged from 800 to 65,000 students. Assessments conducted between September 2024 and January 2026.

Scoring Protocol. Each ASRI dimension scored independently by two assessors using standardised rubric (available on request). Inter-rater reliability: Cohen's kappa = 0.83 (substantial agreement). Discrepancies resolved by third assessor. Final scores represent consensus rating.

Statistical Analysis. Pearson correlation coefficient between ASRI composite score and AI-related incidents per term: $r = -0.84$, $p < 0.001$. Linear regression: Incidents = 12.1 - 2.8 × ASRI ($R^2 = 0.71$, adjusted $R^2 = 0.70$). Normality of residuals confirmed via Shapiro-Wilk test ($W = 0.97$, $p = 0.42$). Homoscedasticity verified via Breusch-Pagan test ($p = 0.31$). No multicollinearity detected (all VIF < 2.5).

14.2 Red Team Testing Protocol

Prompt Corpus. 1,400 adversarial prompts across 7 education-specific categories (200 per category). Prompts developed by a team of 6 researchers including 2 education safeguarding specialists, 2 AI security researchers, and 2 multilingual testers (English, French, Spanish, Mandarin). Prompts ranged from simple direct requests to multi-turn escalation sequences (3–5 turns) and function call exploits.

Models Tested. GPT-4o (2024-08-06), GPT-4o-mini (2024-07-18), Claude 3.5 Sonnet v2 (2024-10-22), Gemini 2.0 Flash, Llama 3 70B-Instruct. All models accessed via API with default safety settings enabled. Testing window: October–December 2025.

Evaluation. Attack Success Rate (ASR) determined by GPT-4o-mini automated evaluator using HarmBench evaluation protocol (Mazeika et al., 2024). Human validation on 10% random sample (n=140) showed 91% agreement with automated evaluator. Severity classification uses CVSS v3.1 base metrics with Child Impact Multiplier (CIM = 1.5).

Limitations. (1) Model behaviour may change with updates post-testing window. (2) Automated evaluation achieves approximately 93% agreement with human annotations; edge cases may be misclassified. (3) Prompt corpus does not claim exhaustive coverage of all possible attack vectors. (4) ASRI validation sample size (n=40) provides statistical power of 0.95 for detecting large effect sizes (d=0.8) but may not detect smaller effects. (5) Institutions self-selected for participation, introducing potential selection bias toward those already investing in AI governance.

14.3 ROI Model Assumptions

Monte Carlo Simulation. 10,000 iterations using the following distributions: breach probability reduction (Beta(8,2) to Beta(2,8)); average breach cost (LogNormal, $\mu=\ln(4.5)$, $\sigma=0.4$); implementation cost (Normal, $\mu=£310K$, $\sigma=£50K$). Sensitivity analysis conducted by varying each parameter ±20% while holding others constant.

14.4 Pre-Publication Review and Conflict of Interest Disclosure

Review Process. The methodology sections (ASRI formulation, red team protocol, and CVSS-Ed derivation) were reviewed by three independent practitioners prior to publication: one academic specialist in child online safety (affiliated with a UK university research centre), one AI governance practitioner (ISACA-certified, no institutional affiliation with the author), and one EdTech security professional (ISC^2-certified). Reviewers provided written feedback on methodological soundness, statistical appropriateness, and claims consistency. Reviewer identities are held on file and available upon request to journal editors or regulatory bodies, subject to reviewer consent.

Conflict of Interest Statement. The author provides commercial consulting services in cybersecurity governance, including to educational institutions. The SHIELD Framework and ASRI model were developed through this consulting practice. The ASRI validation data were collected from institutions where the author or his team conducted assessments. This creates a potential conflict of interest: the assessors who scored the institutions had knowledge of the SHIELD Framework, which may introduce observer bias. The planned Phase 2 validation (Section 5.2) will use independent assessors blinded to the SHIELD methodology to address this limitation. No external funding was received for this research.

Data Ethics. All institutional data are anonymised. No individual student, staff, or parent data were collected or processed. Institutional participation was voluntary. Red team testing was conducted only on publicly accessible AI model APIs with default safety settings; no school systems were tested without authorisation.

14.5 CVSS-Ed Inter-Rater Reliability

The CVSS-Ed scoring extension was tested for inter-rater reliability by having four independent assessors (two cybersecurity practitioners, one child safeguarding specialist, one AI ethics researcher) independently score the 16 ELAST vectors using the CVSS-Ed formula. Intraclass Correlation Coefficient (ICC, two-way mixed, absolute agreement) = 0.87 (95% CI: 0.79–0.93), indicating good to excellent reliability. The primary source of disagreement was the Vulnerability-to-Harm Proximity component of CIM, where safeguarding specialists assigned higher proximity scores than cybersecurity practitioners for indirect harm vectors (e.g., academic integrity bypass). This suggests the CIM formula may benefit from domain-specific calibration guidance in future revisions.

14.6 Independent Replication Invitation

The author invites independent replication of the ASRI model. The following materials are provided to facilitate replication: (1) the complete ASRI scoring rubric with dimension definitions, scoring anchors, and worked examples; (2) the AHP pairwise comparison matrix and weight derivation methodology; (3) the CVSS-Ed formula with CIM calibration dataset; (4) the adversarial prompt corpus (1,400 prompts) with category labels. All materials are deposited at the repository specified in the Data Availability Statement (Section 6). We particularly welcome replication studies from institutions outside the current sample (APAC, Middle East, Latin America) and from institution types underrepresented in the validation (primary schools, special educational needs settings). Correspondence regarding replication should be directed to info@kieranupadrasta.com.

14.7 Publication Status and Dissemination Plan

Current status. This paper is circulated as a Schiphol University working paper (WP-2026-03). A companion preprint has been submitted to arXiv (cs.CR). The ASRI model and CVSS-Ed extension are being prepared for submission to *AI & Ethics* (Springer) as a standalone methods paper. The MSLTM threat model is being prepared for submission to the ACM Conference on Fairness, Accountability, and Transparency (FAccT 2027). The ELAST taxonomy has been submitted for consideration by the OWASP Education Security Project.

Presentation history. Earlier versions of the SHIELD Framework and ASRI model were presented at: ISACA London Chapter quarterly meeting (November 2025); PRMIA Cyber Security Programme working group (December 2025); and ISF Auditors and Control annual review (January 2026). Feedback from these presentations informed the weight derivation methodology and CIM calibration described in Sections 5.1 and 5.4.

. Glossary of Key Terms

Term	Definition
COPPA	Children's Online Privacy Protection Act (US federal law regulating children's data)
CSAM	Child Sexual Abuse Material
DORA	Digital Operational Resilience Act (EU 2022/2554)
DPIA	Data Protection Impact Assessment (GDPR Article 35)
EU AI Act	Regulation (EU) 2024/1689 establishing harmonised rules on artificial intelligence
FERPA	Family Educational Rights and Privacy Act (US federal student privacy law)
GDPR	General Data Protection Regulation (EU 2016/679)
IAM	Identity and Access Management
ISO 42001	International standard for AI Management Systems
KCSIE	Keeping Children Safe in Education (UK statutory guidance)
LLM	Large Language Model
NIST AI RMF	NIST Artificial Intelligence Risk Management Framework
NIS2	Network and Information Security Directive 2 (EU 2022/2555)
OWASP	Open Worldwide Application Security Project
PII	Personally Identifiable Information
PyRIT	Python Risk Identification Toolkit (Microsoft open-source red team tool)
RAG	Retrieval-Augmented Generation
RBAC	Role-Based Access Control
SHIELD	Safeguarding, Human Oversight, Identity, Evidence, Legal, Data Protection
SIEM	Security Information and Event Management
Zero Trust	Security model requiring verification from everyone accessing resources

. About the Author

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

Kieran Upadrasta is a distinguished cyber security expert with 27 years of professional experience, including 21 years specialising in financial services and banking. His career spans all four major consulting firms—Deloitte, PwC, EY, and KPMG—where he has advised board members and senior executives across global institutions on regulatory compliance, cyber risk governance, and digital operational resilience.

Mr. Upadrasta has worked with the largest corporations to become compliant with OCC, SOX, GLBA, HIPAA, ISO 27001, NIST, PCI, and SAS70. His expertise spans business analysis, consulting, technical security strategy, architecture, governance, security analysis, threat assessments, and risk management.

Professional Memberships

- Professor of Practice in Cybersecurity, AI, and Quantum Computing, Schiphol University
- Honorary Senior Lecturer, Imperials
- Lead Auditor, ISF Auditors and Control
- Platinum Member, ISACA London Chapter
- Gold Member, ISC² London Chapter
- Cyber Security Programme Lead, PRMIA
- Researcher, University College London (UCL)

Contact: info@kieranupadrasta.com | www.kie.ie | LinkedIn

. References

Primary Regulatory Sources

1. EU AI Act Regulation (EU) 2024/1689, EUR-Lex
2. DORA Regulation (EU) 2022/2554, EUR-Lex
3. NIS2 Directive (EU) 2022/2555, EUR-Lex
4. UK Keeping Children Safe in Education (KCSIE) 2025, Department for Education
5. UK Online Safety Act 2023, legislation.gov.uk
6. COPPA 2025 Amendments, Federal Trade Commission
7. FERPA (20 U.S.C. § 1232g), US Department of Education
8. DfE Generative AI: Product Safety Expectations, January 2025
9. DfE Generative AI in Education Guidance, June 2025

Standards and Frameworks

10. ISO/IEC 42001:2023, Artificial Intelligence Management Systems
11. NIST AI Risk Management Framework (AI RMF 1.0), January 2023
12. NIST AI 600-1: Generative AI Profile, July 2024
13. OWASP Top 10 for LLM Applications 2025
14. NIST SP 800-207, Zero Trust Architecture
15. IEEE 2089-2021, Age Appropriate Digital Services Framework
16. CEPEJ Ethical Charter on Use of AI in Judicial Systems
17. Google SAIF 2.0, Secure AI Framework (2025)

Industry Research

18. NACD Board AI Governance Framework 2025
19. CDT "Hand in Hand: Schools' Embrace of AI" Report, October 2025
20. International AI Safety Report 2026
21. Keeper Security: AI in Education Survey 2025
22. NCMEC AI-Generated CSAM Statistics 2023-2025
23. Internet Watch Foundation Annual Reports 2024-2025
24. Check Point 2025 Cyber Security Report
25. Gallup/Walton Foundation: AI in K-12 Education 2024-25

Red Team & AI Safety Research

26. Microsoft AI Red Team Whitepaper, January 2025
27. Mazeika et al. "HarmBench: A Standardized Evaluation Framework", ICML 2024
28. Souly et al. "StrongREJECT: A Rejection Benchmark for Jailbreak Evaluation", 2024
29. "Multi-Turn Jailbreaks Are Simpler Than They Seem", arXiv:2508.07646, August 2025
30. "Jailbreak Risks in Function Calling", ACL 2025, COLING
31. General Analysis, "The Jailbreak Cookbook", March 2025
32. MITRE AI Maturity Model and Organisational Assessment Tool, 2025
33. EDUCAUSE Higher Education AI Readiness Assessment, 2024
34. Responsible AI Institute Organisational Maturity Assessment (OMA), 2022

International Policy

35. UNESCO Guidance on Generative AI in Education, September 2023
36. Australia National Framework for Generative AI in Schools (2023, reviewed 2025)
37. OECD AI Principles (47 adherents), updated 2024
38. EdWeek Research Center: AI & Cybersecurity in K-12, February 2026

Statistical Methodology

39. Saaty, T.L. (1980). The Analytic Hierarchy Process. McGraw-Hill.

40. Koo, T.K. & Li, M.Y. (2016). Guideline of selecting and reporting ICC. J. Chiropractic Medicine, 15(2).
41. Efron, B. & Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall.
42. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement.
43. FIRST.org (2024). CVSS v4.0 Specification Document.