Building Resilient Educational AI Through Multimodal Context Fusion

Donggil Song[1], **Anne Lippert**[2]

[1]Department of Engineering Technology & Industrial Distribution, College of Engineering, Texas A&M University, College Station, TX, USA

[2]Department of Psychology, College of Arts and Sciences, Prairie View A&M University, Prairie View, TX, USA

creative@tamu.edu, amlippert@pvamu.edu

Abstract

Educational systems face escalating challenges: diverse learner needs, resource limits, and rapid technological change. Resilient AI is needed to remain useful as learner behavior, interface conditions, and infrastructure vary. We present a position on resilient educational AI through multimodal context fusion, demonstrated with a virtual reality (VR) algebra learning system. We define resilience operationally through three measurable indicators: (1) action-grounding accuracy, (2) tolerance to latency variation and fallback conditions, and (3) stability under changing learner behaviors across mission phases. In a controlled ablation using 30 scripted traces, the context-grounded system achieved 92% action reference accuracy and reduced hallucination events by 86% relative to a mission-only baseline. Complementary pilot logs from live VR use show sustained learner-initiated interaction and stable response timing, supporting the human-AI collaboration claim while remaining bounded by the absence of direct trust/usability surveys. We position multimodal context fusion as a practical design pattern for robust, adaptable, and scalable educational AI, with implications for broader resilience domains.

The Resilience Imperative

Educational communities worldwide face conditions analogous to broader resilience challenges: widening achievement gaps, heterogeneous learner trajectories, uneven access to support, and infrastructure constraints (UNESCO Global Education Monitoring Report Team 2024). Static educational systems struggle in such settings, especially when learner needs and delivery contexts shift faster than curriculum or staffing can adapt.

AI-driven resilience in education requires systems that can (1) handle unpredictable interactions without degrading, (2) adapt strategies in real time to individual needs, (3) scale across contexts without extensive retraining, (4) maintain trust through transparent grounding, and (5) support human-AI collaboration that augments rather than replaces people (Amershi et al. 2019; Shneiderman 2020). In operational terms, we define resilience through at least three measurable indicators: (1) *performance stability under noisy or incomplete inputs* (e.g., grounded action-reference accuracy), (2) *latency variation tolerance* (e.g., acceptable response stability with timeout fallback), and (3) *behavioral robustness to user shifts* (e.g., support quality across changing mission phases, modalities, or help-seeking patterns).

Our Position: *Multimodal context fusion*, the real-time integration of environmental, behavioral, and task signals into AI decision-making, provides a practical architecture for resilient educational AI (Driess et al. 2023; Liu et al. 2023). We demonstrate this through a Unity-based Meta Quest 3 virtual reality (VR) system where learners explore a virtual environment to solve algebra problems (Shapiro 2019; Lindgren and Johnson-Glenberg 2013). This is based on an argument that resilience benefits derive not from VR alone but from disciplined fusion of task-relevant signals, as the same architectural logic appears in other fusion-heavy domains, including healthcare classification, smart-grid defense, biometric identification, and aerial infrastructure monitoring (Shah et al. 2026; Shabbir et al. 2025; Rehman et al. 2025; Qayyum et al. 2021).

Theoretical Grounding and State of the Art: Our argument combines embodied learning (Shapiro 2019; Lindgren and Johnson-Glenberg 2013), human-centered AI (Amershi et al. 2019; Shneiderman 2020), and resilience-oriented system design. Relative to current educational LLM work, our contribution is not a new foundation model; it is a deployment pattern that keeps an instruction-tuned LLM auditable by binding it to compact multimodal state rather than relying on generic dialogue alone (Ouyang et al. 2022; Kasneci et al. 2023; Wang et al. 2025). This places the work between generic educational chatbots and fully embodied multimodal agents (Driess et al. 2023; Liu et al. 2023; Mon-Williams et al. 2025): more grounded than text-only tutoring, but lighter weight and more deployable than end-to-end embodied foundation models.

Architecture for Resilience

The Context-Fusion Pipeline

When users query the system (voice or button), a resilience-oriented pipeline executes:

Robust State Capture (<100ms): Query scene for object parameters, retrieve transformation history (last 10 actions), capture gesture sequences (grab/release, pinch, rotation), include mission context. *Graceful degradation*: fall-

back to mission-only if capture fails.

Lightweight Serialization (1-2KB JSON): Context intentionally compact for low-bandwidth scenarios (rural deployment, emergency networks):

```
{"mission":{
  "id":"M04",
  "problem":"Match y=2^x+1"
 },
 "currentObject":{
  "type":"exponential",
  "parameters":{"base":2.0,"shift_y":1.0}
 },
 "transformationHistory":[{
  "action":"change_base",
  "delta":2.0,
  "ts":1737212345678
 }],
 "recentGestures":[
  {"type":"grab","hand":"right"}
 ]}
```

Context-Grounded Prompting: The teammate uses an instruction-tuned OpenAI GPT-family model; project deployments have used GPT-4o and GPT-5 in VR pilots. The recommended configuration is intentionally simple: a compact 1–2KB context block, an explicit teammate system prompt (*Ground responses in provided context. Describe physical changes before abstractions. Ask guiding questions. Suggest specific actions.*), a 2–4 sentence output constraint, and few-shot examples for action grounding (e.g., *You doubled the base, so the curve steepens. Try dropping below 1 to see decay*) (Yin et al. 2023). We emphasize prompt constraint and observable grounding over task-specific fine-tuning because these settings are easier to audit, port, and maintain.

Resilient Delivery: Delivery includes an API call with 1.38s average latency, a timeout fallback after 3s to a static hint, multimodal output (text + optional TTS), and logging for continual improvement.

Three Resilience Properties

These three properties summarize how the teammate remains useful under uncertainty, domain shifts, and resource constraints.

Robustness Through Grounding: Context injection prevents hallucination by anchoring explanation to observable state. The comparison baseline is the same teammate pipeline with mission text only, with gesture and parameter fields removed. Table 1 shows 92% action reference accuracy with full context versus 38% for that mission-only baseline; hallucinations were reduced 86% (14→2 events across 30 traces), which is critical when learners or operators need to trust system references to recent actions.

Adaptability Through Interoperability: Architecture is domain-agnostic. JSON schema modifications allow rapid reconfiguration across domains. In education, they support transitions from algebra to calculus and physics. In crisis management, gesture inputs can be replaced with sensor data such as temperature or crowd density, and missions can be replaced with emergency protocols. In community support, the same structure can support mental health check-ins and resource allocation. No model retraining is required.

Metric	Context	No Context
Relevance (0–5)	4.6	2.9
Action Accuracy (%)	92	38
Hallucination Events	2	14
Pedagogical Value	4.3	2.7
Latency (s)	1.38	1.21

Table 1. Context Fusion Impact on Resilience (30 scripted traces).

Scalability Across Resources: High-resource deployments can use cloud APIs with full multimodal support; mid-resource deployments can use regional endpoints with text-only delivery; low-resource deployments can rely on cached responses and on-device small models. The 1-2KB context remains suitable for low-bandwidth settings (e.g., satellite, rural cellular).

Evaluation Protocol: The quantitative comparison used 30 scripted traces representing common algebra missions, gesture sequences, and learner questions. Each trace was run through the full pipeline and through the mission-only ablation. Responses were scored with a fixed rubric for relevance, action-reference accuracy, hallucination events, and pedagogical value; latency was measured end-to-end. Average latency was measured over the same controlled pipeline, with 50 scripted requests used to characterize round-trip timing.

Stress Testing and Boundary Conditions

The current evidence is strongest on grounding accuracy and weakest on systematic stress testing across off-nominal conditions. Resilience engineering emphasizes sustaining useful performance under disturbance, surprise, and resource shifts rather than optimizing only for average-case behavior (Woods 2015). For educational AI, that means evaluating not only whether the teammate answers well when context is complete, but also whether it degrades predictably when state capture is partial, latency spikes occur, or learners abruptly change strategy.

Table 2 sharpens what a stronger resilience claim would require. A resilience-oriented evaluation should report not only mean response quality, but also whether failure handling is observable, bounded, and recoverable. This matters in educational settings where trust can erode quickly if the teammate references nonexistent actions, stalls during a teachable moment, or gives ungrounded advice after a context drop. These tests are feasible with scripted perturbations around the current pipeline and would strengthen the case for generalization beyond VR algebra.

A practical next evaluation cycle follows directly from this matrix. First, a scripted perturbation benchmark can vary state completeness, network delay, and gesture noise while holding task demands constant. Second, a live deployment study can test whether fallback behavior remains legible to learners and instructors rather than merely whether

Stressor	Risk if unhandled	Expected containment	Primary metric
Partial state capture	The teammate refers to missing actions or stale state.	Fall back to mission-only context and expose reduced observability.	Action accuracy; hallucinations
Latency spike	Help arrives after the teachable moment or conversational turn.	Return a static hint after timeout and preserve turn continuity.	Timeout rate; response stability
Gesture noise	Transient sensor noise is mistaken for meaningful learner action.	Weight object state and action history more than transient gestures.	Grounded reference quality
Learner strategy shift	Guidance remains anchored to an outdated mission phase.	Reinterpret requests from mission phase and recent actions.	Support quality across phases
Bandwidth loss	Cloud dependency creates assistance gaps in low-resource settings.	Use cached hints or an on-device small model.	Availability; recovery time

Table 2. Stressors that should complement the current ablation.

the model returns an answer. Third, a transfer study can apply the same prompt-and-context discipline to a second domain, such as introductory physics or emergency planning, to separate architecture-level resilience gains from algebra-specific scaffolds. Together, these additions would move the paper from a strong proof-of-concept toward a more convincing resilience methodology.

Socio-Technical Resilience

Resilience in deployment is ultimately socio-technical: it depends on how people interpret, trust, and govern AI support under real constraints.

Human-AI Collaboration: Teammate framing preserves human agency (Shneiderman 2020). Learners maintain exploration autonomy; in crisis contexts, responders retain decision authority while AI provides grounded situational awareness.

Trust Through Transparency: Because context grounding makes each response auditable, educators can inspect the environmental state used to produce it. This matters in high-stakes decisions such as evacuation or resource allocation. By contrast, opaque AI can undermine resilience by introducing uncertainty.

Qualitative Learner-Use Signals: A complementary pilot deployment produced 5,317 time-stamped events from 27 learners, including 776 complete conversational request chains. While this dataset did not include direct trust or usability surveys, it provides bounded qualitative evidence relevant to collaboration: learners repeatedly initiated help-seeking, response timing remained stable (0.99s mean in the live pilot), and prompting dipped in Mission 2 then rebounded in Mission 3 rather than collapsing, suggesting continued willingness to use the teammate as tasks changed. We therefore treat engagement and usability as *supported but not conclusively measured*: the logs indicate sustained use and low-friction interaction, while perceived trust should be tested directly in the next study.

Global Scalability: Lightweight design enables deployment in resource-restricted regions: pandemic-disrupted education (immersive learning on standalone headsets without cloud), climate adaptation (farmers use similar interfaces for crop planning with local sensor grounding), emergency response (first responders receive context-grounded guidance from environmental sensors, crowd dynamics).

Evidence Gaps and Evaluation Agenda

The current paper makes a strong case for context grounding, but two evidence gaps remain before resilience claims can be considered mature. First, the live pilot demonstrates sustained use but not calibrated trust. Trust in automation is not the same as frequency of use; the stronger question is whether learners and instructors rely on the teammate appropriately, challenge it when it is wrong, and recover smoothly after fallback conditions (Lee and See 2004). Future studies should therefore combine trust or usability scales with behavioral indicators such as abandonment after a timeout, willingness to ask follow-up questions, correction of incorrect system references, and instructor override patterns.

Second, generalization should be established at the workflow level rather than inferred from architectural similarity alone. A persuasive next step is a multi-condition evaluation in which the same context-schema discipline is ported to a second STEM domain and to a non-educational scenario with constrained resources. Reporting the same measures (action grounding, fallback frequency, response stability, and continued help-seeking) would clarify which gains come from multimodal grounding itself and which depend on domain-specific scaffolds. That kind of evidence would make the broader resilience claim more testable and more useful to other researchers building compact, auditable AI assistants.

Design Principles for Resilient AI

Based on this work, we propose five principles:

1. **Ground in Observable Context**: Anchor responses in real-time observable state to reduce hallucinations.
2. **Prioritize Lightweight Interoperability**: Use compact contexts (about 1–2 KB) that remain effective in limited bandwidth and changing settings.
3. **Enable Graceful Degradation**: Provide fallback modes (cached hints, simpler outputs) under resource or service failure.
4. **Preserve Human Agency**: Keep AI in a collaborative support role rather than a replacement role.
5. **Design for Transparency**: Keep reasoning inspectable through explicit context grounding, especially in high-stakes decisions.

These principles extend beyond education to emergency response, climate adaptation interfaces, community support platforms, and other domains requiring resilient adaptive technologies.

Future Directions and Call to Action

Moving from promising prototype to durable impact requires coordinated technical, social, and policy advances.

Technical: On-device LLM inference (eliminate cloud dependency for privacy/latency), proactive anomaly detection (identify struggle/risks proactively), multi-agent collaboration (teams), continual learning pipelines, and cross-domain validation (crisis management, healthcare, climate).

Socio-Technical: Community co-design with climate-vulnerable, resource-restricted regions; longitudinal trust and usability studies; equity analysis ensuring no digital divide exacerbation; and policy frameworks for resilient AI governance.

Collaboration: We invite participants to: (1) extend this architecture to other resilience domains, (2) develop shared evaluation frameworks for resilient AI, (3) create open-source toolkits for multimodal context fusion in resource-constrained settings, and (4) establish ethical guidelines for deploying adaptive AI in vulnerable communities.

Conclusion

Building resilient communities requires adaptive AI that is robust to behavioral diversity, scalable to resource constraints, and trustworthy through transparent grounding. We argued for an operational view of resilience in educational AI, specified by stability under noisy inputs, tolerance to latency variation, and robustness to shifting user behavior. Within that frame, multimodal context fusion provides a practical architecture: in our controlled ablation, it reduced hallucinations by 86% and maintained 92% action reference accuracy against a mission-only baseline; in complementary live VR logs, learners continued to initiate interaction under varying mission demands. The broader contribution is therefore both conceptual and technical: a position on how educational LLM systems can be made more resilient, and a deployable pattern for doing so across education and adjacent resilience domains. The next steps include direct trust/usability studies, cross-domain validation, and community co-design to ensure that resilient AI truly serves the needs of vulnerable populations in a rapidly changing world.

Acknowledgements

This work was supported by the National Science Foundation (#2405599 and #2405600) and Panther Research & Innovation for Scholarly Excellence (#230487).

References

Amershi, S.; Weld, D.; Vorvoreanu, M.; et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 8469–8488.

Kasneci, E.; Sessler, K.; Kuechemann, S.; et al. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103: 102274.

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1): 50–80.

Lindgren, R.; and Johnson-Glenberg, M. 2013. Emboldened by Embodiment: Six Precepts for Research on Embodied Learning and Mixed Reality. *Educational Researcher*, 42(8): 445–452.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Mon-Williams, R.; Li, G.; Long, R.; Du, W.; and Lucas, C. G. 2025. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 7(4): 592–601.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *arXiv preprint arXiv:2203.02155*.

Qayyum, A.; Razzak, I.; Malik, A. S.; and Anwar, S. 2021. Fusion of CNN and Sparse Representation for Threat Estimation Near Power Lines and Poles Infrastructure Using Aerial Stereo Imagery. *Technological Forecasting and Social Change*, 168: 120762.

Rehman, T. U.; Alruwaili, M.; Siddiqi, M. H.; Alhwaiti, Y.; Anwar, S.; Halim, Z.; and Alam, M. 2025. Advancing EEG-Based Biometric Identification Through Multi-Modal Data Fusion and Deep Learning Techniques. *Complex & Intelligent Systems*, 11(9): 398.

Shabbir, A.; Manzoor, H. U.; Zoha, A.; and Halim, Z. 2025. Smart Grid Security Through Fusion-Enhanced Federated Learning Against Adversarial Attacks. *Engineering Applications of Artificial Intelligence*, 157: 111169.

Shah, K. A.; Halim, Z.; Anwar, S.; Hsu, C.-C.; and Rida, I. 2026. Multi-Sensor Data Fusion for Smart Healthcare:

Optimizing Specialty-Based Classification of Imbalanced EMRs. *Information Fusion*, 125: 103503.

Shapiro, L. 2019. *Embodied cognition*. Routledge.

Shneiderman, B. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6): 495–504.

UNESCO Global Education Monitoring Report Team. 2024. 2024/5 Global Education Monitoring Report, Leadership in education: Lead for learning. Technical report, United Nations Educational, Scientific and Cultural Organization (UNESCO). Accessed March 18, 2026.

Wang, S.; Xu, X.; Li, J.; Zhang, L.; Liang, J.; Tang, J.; Yu, P. S.; and Wen, Q. 2025. Large Language Models for Education: A Survey and Outlook. *IEEE Signal Processing Magazine*, 42(6): 51–63.

Woods, D. D. 2015. Four Concepts for Resilience and the Implications for the Future of Resilience Engineering. *Reliability Engineering & System Safety*, 141: 5–9.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.