https://www.amitrakshar.co.in/journal : www.amitrakshar.co.in
Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

Multimodality of AI for Education: Towards Artificial General Intelligence

Dr. Aniruddha Saha[1]

Abstract

Multimodal artificial intelligence (AI) is changing the learning process through unifying the various forms of data that can be integrated into a single system: text, speech, images, gestures and sensor signals that replicate human cognition. Multimodal AI represents a transitional stage to the wider objective of Artificial General Intelligence (AGI) between specialized AI models and their use in education. We examine how it is being used in personalized tutoring, virtual classrooms that have immersive aspects, and accessible learning to students with diverse needs. The unusual role of multimodal models in facilitating contextual comprehension, cross domain reasoning and adaptive pedagogy. We also respond to critical issues and those are ethical considerations, privacy threats, data bias, and resources-intensive infrastructures. To leverage multimodal AI to achieve inclusive and interactive educational experience that promotes creativity, critical thinking, and lifelong learning, we suggest a roadmap to these outcomes that will enable human-like AGI to think, reason, and respond like human tutors as education systems.

Keywords: *Multimodal AI, Artificial General Intelligence, Personalized Learning, Educational Technology, Human-Computer Interaction.*

AIJITR - Volume - 2, Issue - VI, Nov-Dec 2025

Introduction:

The concept of artificial intelligence (AI) has turned out to be a disruptor in various fields, and in the education industry the possibility is immense. Artificial intelligence-based tutoring, predictive analytics systems, and machine-generated feedback systems are transforming the interaction with material by the learners and the organization of the instruction by teachers. However, when carefully incorporated, AI can offer more personalized, adaptive, and scalable assistance that can decrease disparities in access and be able to adjust learning to individual paths (Latif et al., 2023). However the current state of AI in education is still limited: it is analyzing one mode of communication at a time, text (automated essay scoring, chatbots), speech (Speech recognition, language tutoring), or vision (image based assessment, gesture recognition) separately. The future direction is multimodal AI - systems capable of perceiving, reasoning, and producing multiple modalities at once: text, speech, vision, gestures, and sensor (e.g. eye-tracking, motion) inputs. Multiple modal AI has the prospect of more interactive response: a learning agent can listen to a spoken query made by a student, observe a gesture, a drawn drawing, perceive the expression, and react with a visual, auditory and gestural display. There is a gap between the existing multimodal AI applications and the vision of the grand AI in education known as artificial general intelligence (AGI). Narrow AI is fragile and problem-solved, with a lack of deep common sense, transferability, and the ability to independently self-develop (Fei et al., 2022). In the case of education, AGI would require reasoning across domains, ability to read between the lines through flexibility to constantly learn of multimodal struggles as a human tutor. Making the abstraction between theoretical explanation and practical learning is not merely the gap

[1] ***Assistant Professor and Head, Department of Education, Asannagar Madan Mohan Tarkalankar College, Email.- aniruddha.saha11@gmail.com***

DOI Link (Crossref) Prefix: https://doi.org/10.63431/AIJITR/2.VI.2025.90-99
AIJITR, Volume 2, Issue –VI, November – December, 2025, PP. 90-99
Received on 19th, December 2025 & Accepted on 27th, December, 2025, Published: 30th December, 2025

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

between the models (in model architectures or data) but also the gap between the concepts (how to integrate ethical, explainable, context-aware intelligence that works in students) or the gap between the senses (how to teach students to perceive all aspects holistically) (Lee et al., 2023). Multimodality of AI in education may be a preliminary step to AGI, but not a goal, but a path. Our mission is to describe both the promise and the challenges, between present systems and the future aspirations.

Objectives:

1. To multimodal AI can be useful in changing traditional and digital education by adapting and personalizing the learning experience.
2. To the contribution of multimodal integration to human-like reasoning and communication towards advancement towards AGI in education.
3. To determine obstacles, the ethical issues, and the future perspectives of applying the concept of multimodal AI to design a highly inclusive and effective learning process.

3. Foundations of Multimodal AI

Definition Multimodality in AI Systems Multimodality is a term used in artificial intelligence systems to define the ability of computers or robots to perceive and understand multiple different inputs.

In artificial intelligence multimodality is something able to receive, combine or generate data through more than a single channel of sensory or communication (modality), including text, speech, image, video, gesture or even a tactile/haptic signal (Li et al., 2024; Jin et al., 2025). Unlike a unimodal AI that can only process a single type of input (e.g. text only or a sight only), a multimodal AI tends to bypass heterogeneous data streams so that they can be whole and more contextually aware in their thinking (Jin et al., 2025). The point is that human communication and perception are always multimodal: we do not only read either write, we gesture and look and gesture, listen and scribble and so on. Multimodal AI attempts to reflect that sophistication.

Multimodal systems tend to create three associated learning environments (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2014):

Multimodal fusion: each and every modality can be trained and inferred and they are merged to predict (Ngiam et al., 2011).

Cross-modality learning: In training, more than one modality is applied, and only one is available at inference hence the model is required to generalize (Ngiam et al., 2011).

Shared representation / congruence: dissimilar modalities have a common latent embedding space in such a way that the system can among modalities translate or align (e.g. image-to-text, text-to-video) (Ngiam et al., 2011; Jin et al., 2025).

Key Input/Output Modalities

This is why a wide range of modalities is taken into consideration in order to make such systems useful. Among the most important aspects are those associated with education and they are as follows:

1. Speech / audio: the speech, the tone, the prosody, the identity of the speaker, pauses.
2. Text / writing Typed text, written text, natural language, symbolic notation.
3. Handwriting / drawings: free hand typing, drawings, mathematical artifacts.
4. Vision / pictures / video: object recognition, scene recognition, gesture recognition, eyeglass, diagram recognition.
5. Augmented Reality / Virtual Reality (AR/VR): 3D world, integrated reality.
6. Haptics / tactile feedback: sense of touch, word of force, touch.
7. Brain-computer interface (BCI): neurological database or EEG based brain-machines input (only developed in education).

AI systems will not only be able to read, write, hear, or read, but can also read, write, gaze, or move the hands, gauge frustration or immersive VR or react to it with haptic feedback.

Multimosal Data Records Fusion & Learning

One of the key problems is how to make the information in various modalities useful. This is representation learning (learning features) and fusion (merge features). The major strategies are (Jin et al., 2025; Li et al., 2024; Baltrusaitis et al., 2019):

Early fusion: This form of fusion applies raw or low-level features across all modalities (e.g. feature vectors by concatenating them) before they are presented to a joint model (Baltrusaitis et al., 2019).

Late fusion: Performing exactly the same thing to each modality individually as though they were predictions, and combining them together (ensemble-style) (Baltrusaitis et al., 2019).

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

Intermediate / hybrid fusion: whereby the encoders are modality-specific but the resultant embeddings are then combined at an intermediate layer (Baltrusaitis et al., 2019).
Joint embedding / alignment: projecting the various modalities into a common latent space (embedding) in which semantically similar material across modalities are proximate (e.g. CLIP maps images and text) (Radford et al., 2021).
Attention / transformer-based fusion: an attention mechanism to distort the contribution of various characteristics of the modality and dynamically determines such modality emphasis (Li et al., 2024).
Graph-based fusion & cross-modal translation: modalities are represented as nodes and edges in graphs, or can be trained to learn to perform a translation between modalities in order to address modality gaps (Mai, Hu, and Xing, 2019).
Representation learning so as to ensure that unique statistical structure and noise are modeled and fusion deals with interactions as well as complementarities between modalities. Such advanced methods as mutually information maximization, information bottleneck, and cross-modal knowledge distillation are applied to ensure that fused embeddings are informative but redundant to only a small extent (Mai, Zeng, and Hu, 2022; Chen et al., 2023; Nguyen et al., 2023).
Benefits of Multimodal over Unimodal AIs
Multimodality has a number of important advantages over unimodality systems:
Complementarity and redundancy.
Several modalities tend to be complementary (e.g. tone of voice + facial expression + words) and can support each other in instances when one of the modalities is either noisy or absent (Chen et al., 2025). This would enhance strength.
Decontextualization and less contextualization
The meaning of words can be ambiguous, visual clues or gestures can disambiguate the meaning. A unimodal one may not capture an environment that a multimodal model can capture.
Greater generalization and transfer
Multimodal models are able to be cross-modal, so they are able to be useful in tasks such as text-to-image generation, or text-to-visual description. They are also capable of aiding in the low-data regimes moving knowledge across modalities (Li and Tang, 2024).
Easier to the intelligence of humans
Multi-modal AI is closer to natural thinking and a step to Artificial General Intelligence (AGI) since humans think and reason multi-modally (Lee, Shi, Latif et al., 2023).
Better execution of complicated tasks
Research demonstrates that multimodal models have been shown to be more successful in tasks that require reasoning, commonsense inference, or more multi-modal QA (i.e. vision and text question answering) (Li et al., 2024; Li and Tang, 2024).
Graceful degradation and dealing with missing modalities
Good multimodal systems can gracefully degrade in the event that other modalities are not present, or that they are noisy, and intelligently utilise the other modalities present.
Multimodality allows more comprehensive, strong, contextualized AI- which is essential in an educational context where student feedback is multidimensional.

4. Uses of Multimodal AI in Education

Multiimodal AI is promising in the quest to create an AI that can teach almost as flexibly as a human being. An abstract development of some of the major applications in the education field is provided below:
Intelligent Tutoring Systems (ITS): Multimodal, Rich Personally-Guided Instruction
Traditional ITS tends to be dependent on dialogues based on text or scaffolding. This is expanded by the Multimodal ITS which combines the use of speech (students ask questions verbally), vision (students track facial expression, gaze, posture), handwriting/ sketches (students draw or write out the answer), and gesture recognition. As an example, a student may pause during a geometry construction, and the system may realize it and inquire further, or provide the hint verbally. By doing this, the ITS will be more responsive, aware and sensitive.
These systems are not only flexible in their teaching content but also the pedagogical process (altering the mode of explanation (visual, auditory, embodied) depending on the learner preference and situation). They are also able to scaffold transitions: between gesture or sketch into symbolic reasoning, allowing students to have a smooth transition between informal and formal ways of thinking.
Adaptive Assessment Speech and Writing Analysis: Real-Time Emotion

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

The evaluation process is usually inflexible and late; AI multimodal allows real-time evaluation. The AI has the capability to detect emotional state (through facial expression or voice stress), confidence (pauses and hesitation), process measures (following step by step how they write/sketch), and content correctness (speaking, writing, drawing) as students provide their response (i.e. speaking, writing, drawing). It can assess by altering the difficulty based on that, probing, or providing scaffolding in-the-moment support (e.g.). Thou feares to be declarie, wouldst thou fain get a hint?). Such decoding of physiological, behavioral, and digital interaction signals in education under the name of Multimodal learning analytics (MMLA) research has been explored (Zhai et al., 2025; A Comprehensive Review, 2025).

AR/VR + Voice + Gesture Immersive Learning Environments

Immersive AR/VR-based learning is one of the richest fields, in which the multimodal AI can arrange the interaction. students are able to stroll around virtually through a lab, take in objects (monitored through gesture/hand modeling), address virtual agents, get visual overlay, and experience virtual haptic feedback. The tutor or the AI environment reacts to voice queries, gestures, eye movement or even control of a handheld device. The combination of modalities can create richer involvement and embodied learning, in particular, in the fields of physics, biology, or engineering, where spatial intuition holds the key to success.

Language and Literacy Development Speech to text and handwriting recognition

Multimodal AI can also be used in speech recognition, handwriting recognition and text analysis in supporting language and literacy. A student is able to write up an essay or speak it and the system dynamically transcribes it, provides feedback and compares the two parameters. It is able to identify the problem of pronunciation, fluency, or handwriting readability and advice the student as needed. Pen-and-speech modality serves to scaffold the process of emergent writers, who have already mastered the language, but just need to learn to write it down.

Multimodal Laboratory Simulations and Experiments

The learners in STEM subjects are usually required to manipulate things, graph and calculate as well as reason all at the same time. The multimodal AI-enhanced simulation could enable the students to draw a circuit, say commands (connect a resistor here), drag digital objects, and receive feedback in real time. Gaze and gesture could be tracked by the system to identify misconceptions (you are about to short circuit), intervene or provide scaffolded prompts. The combination of the vision, gesture, speech, and simulation bridges the gap between the physical intuition and formal representation.

Assistive Technologies: Catering to Special Needs Students

Multimodal AI has a solid potential of access and inclusion. To students with hearing or visual impairments, such as deafness or hearing impairment, the AI could translate speech into sign-language animation and decipher hand gestures. Gaze control, brain computer interface (BCI) or speech-only input may be used instead of handwriting or typing by students with motor impairments. To visually challenged students, diagrams or charts can be made more accessible through the use of tactile / haptic feedback with speech / sonification of data. Multimodal AI permits personalized input and output in every instance.

Co-learning Systems: Multimedia Interaction between students and instructors

Education is social. Examples of tools used in a multimodal collaborative learning setting include voice, gesture, shared sketches/whiteboards, AR visualization, and the use of gaze where students and teachers expand on their new ideas. The AI facilitates: the AI introduces the opportunity to record the discussion (Speech + gesture), moderate turn-taking, point out important diagrams, or even scaffold group work by knowing when a group is working off course. The AI can be adaptively specified to propose breakouts, or recommend reflection, or prompts in multimodal formats (i.e. highlight a visual cue and ask a question verbally). The richness of modes assists in keeping on track, less ambiguity, and increased interaction.

5. Multi-media Architectures and Technologies

The architecture of the desire to have a truly multimodal, perceptive AI in education must be thought through with great care in modalities, fusion, and contexts of deployment. Among the worsted threads, there are some below:

Big Vision Integration Models

The examples of modern vision + language models like GPT-V, Gemini, and Claude 3 demonstrate that the large language models (LLMs) are already being expanded to enable visual input (images or even video) in addition to textual one. These architectures functionally combine a vision encoder (e.g. a convolutional-based vision encoder or transformer-based vision encoder) with those inside the language model hence allowing grounding of textual reasoning in a visual context (Lee et al., 2023).

Within the educational environment, those integrated models may, e.g. identify diagrams or handwritten student drawings as a component of a problem-solving activity interaction, therefore facilitating a more fluent multi-modal

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

tutoring. Such perceptual foundation is a move in establishing the missing connection between embodied (visual) context and symbolic reasoning.

Fusion models of speech-language and vision

On the next level, bona fide multimodal agents should integrate speech (Audio), vision and language information on the basis of regular ingestion and reasoning. Architectures can be of various forms:

Transformer based hybridization models: In this architecture, modalities are not only tokenized (e.g. speech or image patches or text tokens), but the modalities are represented as inputs to a single transformer with cross-modal attention interactions between modalities that allow modalities to align and interact.

Hybrid RNN-CNN or CNN-Transformer layers: Ancienter or simpler models may first convert speech into RNN states (or audio CNNs) and vision into CNNs and then combine them both into an equivalent space by a transformer or any attention block.

Hierarchical or modular architecture: There are systems that store context selective modality specific experts in their form, and other layers in their structure fuse at application-specific later junctions.

This type of fusion makes possible, e.g. an educational AI listening to and following an oral explanation given by a student whilst also observing their gestures or facial expressions to provide even deeper feedback based on both verbal and nonverbal communication.

Knowledge Pictures and Thinking Machines

It cannot do raw multimodal perception; to gain profound conceptualization, the system should combine structured knowledge and reasoning powers. The knowledge graphs are scaffolded: curricular knowledge (e.g. "Pythagorean theorem," triangle, right angle) and multimodal evidences (definition of concepts in text, diagrams, worked example) are given.

These graphs are then processed by reasoning engines (symbolic, neuro-symbolic) in order to make inferences, justify student actions, identify a misconception, or provide a hint to the student. The hybrid nature, that is, neural perceptual front end and symbolic thinking, is the one that gives explanations that are easy to understand and regulate; an essential factor when teaching (Lee et al., 2023).

Edge AI of Classrooms and IoT Multimodal Sensing

Using multimodal AI in real classrooms is limited by latency and bandwidth as well as privacy bandwidth. In this way, local devices (tablets, smart whiteboards, sensor nodes) inference (and possibly even light adaptation) based on Edge AI is essential.

Multimodal sensing (microphones, cameras, gesture sensors, eye trackers, stylus input) sensing tells the virtual reality about the student behavior and context in real time, made by IoT. Embedded AI framework and the device is able to combine all of this to customize instruction, recognize confusion, or change scaffolding- with limited requirements of cloud connectivity.

Edge judicial system minimizes the latency, conserves privacy (raw streams do not necessarily have to be replicated and uploaded continuously), and enables quicker responses.

Privacy-Preserving AI Federated Learning of Education

Since the data of the educational sphere is very sensitive (student work, voice, video, and so on), the collection of information centralized is burdened with the privacy and policy issues. One of such directions is federated learning (FL): Local devices can update local model parameters (the local training), and they can be exchangeable at the central point (or peer federation) without exchanging raw information (Hridi et al., 2024).

The multimodal federated learning (MMFL) frameworks in a multimodal environment are based on extending FL and learning modalities (text, speech, vision) together in a privacy-preserving mode (Pan et al., 2024).

Recent research suggests multi-task federated foundation models (M3T FedFMs) as the new model supportive of a variety of modalities and educational activities (e.g. assessment, feedback, content generation) but being data sovereign (Ofori et al., 2025).

In this manner, federated multimodal architectures are the underlying facilitating technologies towards real world application in schools and other learning institutions.

6. Pathway to AGI in Education

The end is not necessarily more competent tutors; it is agents of the educational process that would operate on the level of human-level general intelligence (AGI). A conceptual roadmap of important transitions as well as architectural enabling mechanisms is provided below.

Specialized Tutoring Agents to Generalized Reasoning AI

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

The existing systems are extremely task/domain-focused (e.g. math problem tutor, language correction bots). The path to AGI will be to increase the sophistication of these specialized agents to an explicit general purpose reasoning engine that canualistically specialize and transpose in domains like science, history, language as a human teacher would. This requires the system to learn how to generalize between tasks, know when to transfer knowledge and self-reconfigure according to context.

The Multimodal perception plays a role in human-like understanding

Human Teachers do not think only through the text, but they see gestures, face expressions, tones, diagrams, demonstrations and so on. It is necessary to have a strong multimodal perception interface to provide AI agents with human like situational awareness, such as the recognition of student confusion based on body language, or understanding of a scribbled diagram on the paper by a student. This kind of perceptual base prevents the brittle reasoning that is based strictly on text and favors the emergent skills that are a transition between concrete experience and abstract cognition.

Brain Motivated Cognitive Architectures

Designers aiming to attain AGI have approached it based on cognitive architectures (e.g. Soar, ACT-R, OpenCog) which organise modules around perception, memory, learning, planning and control (Langley, 2006).

Individually, OpenCog Prime, refers to an effort to achieve an emergent AGI by the interaction of knowledge representations, probabilistic reasoning and pattern mining (Goertzel et al.).

Within the education domain a cognitive architecture may hold a working memory (of the subject matter currently being taught), episodic memory (of historical student interactions), a planning component to structure scaffolding and a metacognitive controller to keep track of interactions.

Learning to Learn and Meta-Learning styles

The most important aspect of adaptability is meta-learning (learning how to learn). By taking advantage of previous experiences to be able to change rapidly to a new field or pedagogical approach, an AI is nearer to AGI. An example is that it may be highly adaptive to a new curriculum or get accustomed to the idiosyncratic pattern of reasoning of a particular student with very limited examples. It necessitates meta-learning layers (e.g. MAML, adaptive optimizers) which adjust the update rules of the system across tasks, but not just within tasks.

Lifelong Agent Continual Learning and Memory Integration

The human teachers and learners constantly construct knowledge, without forgetting that is catastrophic. Educational agents concerned with AGI need to assist the constant learning process: the newly acquired knowledge should not override the previously known skills. The memory systems (episodic, semantic, procedural) have to coordinate such that the agent would be able to remember the strategies they used in the past, the history of the students, and the knowledge about the domain, and reach adaptation at the same time. Such mechanisms as replay buffers, regularization, and dynamic architecture modulation are necessitated.

An intelligent AGI would therefore improve over the years, strengthening its ties with a large number of students, recollections of likes and dislikes, and enhance its teaching methods.

7. Opportunities and Advantages of Multimodal AI in Education

The emergence of multimodal artificial intelligence (AI), which is able to read and write both text and speech, images, video, and sensor input, has potential to revolutionize education when societies will bring closer to artificial general intelligence (AGI). Such technologies have generated unmatched possibilities to customize learning, promote equity, classrooms that are diverse, creative and critical in learners and educator feedback loops.

Individualized learning experiences

Multimodal AI facilitates adaptable learning experience addressing strength, weakness, and learning styles of students on the fly. With the speech recognition, gesture recognition, eye-movement analytics, and text support features, AI tutors will be able to regulate the pacing, give the explanation more modality, and increase the complexity of problems in real time (Luckin et al., 2022). This personalization is no longer the adaptive test, but it takes into account the emotions and interest of a learner by building experiences comparable to personal mentorship (Roll & Wylie, 2016). Such personalized routes have been associated with side effects such as increased retention, intrinsic motivation and increased conceptual comprehension.

Equal education opportunities around the world

With an adequate infrastructure in place, AI-powered educational systems can bring good-quality educational material to students in underserved and remote areas. Multimodal tutoring can be provided in the form of cloud-based, used to provide students with a lack of local qualified teachers with STEM simulation, language training, or sign-language lessons (UNESCO, 2023). Mobile-first AI systems improve the reduced entry threshold by utilizing smartphones as

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

opposed to costly desktops (West and Chew, 2014). This would expand access to widely disparaged achievement, especially in poor neighborhoods and in humanitarian settings.

Formative feedback on learners and educators in real-time

The other advantage would be instantaneous feedback loops. Multimodal analytics, such as automated speech evaluation, handwriting recognition, and computer-vision based lab- skills tracking, enable students to access the comparison of their performance with the learning goals (Rosé et al., 2019). According to the teacher, tools such as a dashboard summing up misconceptions in a class as well as engagement rates can be used to make timely instructional decisions. It has been found that rapid formative feedback leads to better metacognition and self-regulated learning strategies (Hattie and Timperley, 2007).

Multilingual and multicultural classrooms

The provision of education in language diverse environments is usually characterized by low levels of translation and culturally sensitive supplies. Many popular languages are now learning at human level accuracy in multimodal translation models and speech-to-speech systems and low resource languages are improving at an even faster rate (Guzmán et al., 2019). These tools can allow teachers to communicate well through the language barrier and assist the students to learn in their native language as they learn the global languages. In addition, culturally sensitive AI avatars are able to customize the examples, stories, and metaphors to suit a local context, which makes the teaching process more inclusive (Floridi and Cowls, 2022).

Increased creativity, critical thinking skills and problem solving

Multimodal AI can play the role of co-creators - can provide students with suggestions on how to code projects, visualize their data, compose a musical piece or an engineering prototype. Using the delegation of lower-order cognitive load, students are able to concentrate on higher-level critical thinking and problem solving (Scardamalia, and Bereiter, 2014). Co-creative relationships of human learners and the AI agent foster innovation capabilities needed in the 21 st -century workforce.

8. Issues and Constraints of Multimodal AI in the Education

Regardless of these opportunities, multimodal AI adoption to AGI in education challenges are complicated in terms of technical, ethical, psychological, socioeconomic, and safety aspects.

Technical issues: scalability, data sparsity, complexity Fusion

To enable high-quality multimodal learning, huge and balanced datasets of diverse contexts, often not available with many topics and minority languages, are required (Baltrušaitis et al., 2019). Combining data streams (text, audio, gesture) of heterogeneous data presupposes highly complex architectures, which are computationally distances and resistant to the real-life classrooms (Khan et al., 2023). There is an unsolved engineering challenge of scaling these models to offer low latency and offline capabilities to millions of concurrent learners.

Ethical concerns: partiality, privacy, transparency, explainability

Multimodal AI is progressive and occasionally exaggerates biases in its data used to train - which influences grading, admissions and content recommendation (Binns, 2018). Constant video or biometric surveillance increases the issue of privacy and data security, particularly in the case of minors (Tuomi, 2021). In addition, the non-explainability of deep multimodal networks contributes to the impossibility of relying on algorithmic decisions by teachers and parents, since such networks cannot be easily explained. It is important that there is a transparent reporting and compliance to frameworks like the EU AI Act.

Psychological effects among students and teachers

It can lead to AI tutors being excessively dependent on automation, less human-less socialization, as well as new types of test anxiety associated with constant observation by cameras or sensors (Williamson and Piattoeva, 2022). The use of AI to take up major instructional roles may lead to de-professionalization of teachers and possibly impact on their morale and professional identity.

Uneven access to the high-tech AI devices

Although AI has the potential to enlarge access, it will also increase the digital divide when hi-tech multimodal systems are concentrated in well-equipped schools with an intensive internet and hardware infrastructure (Eubanks, 2018). Low-income families may not afford subscription-based platforms and expensive devices except when subsidized or as the good of the community. Every roll-out should be done in a fair manner, which will require the coordination of investments in the connectivity, training of teachers and usage of open-source tools.

AGI and educational safety and alignment

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

Looked at AGI, scholars caution about incompatible goals to take as a strong tutor agent favours test-score achievement perceptions at the cost of holistic studying or wellbeing (Russell, 2019). Before scale deploying near-AGI systems in the classroom, then such techniques as robust alignment research, content-safety layers that are verifiable and contingency protocols to emergent behaviors will be essential.

9. Future Directions

Co-Teaching between human and AI

It is expected that in the future, co-teaching symbiotic relationships between the educator and AI agent will be possible due to the complementary roles of each partner. Whereas teachers can emphasize on empathy, contextual insight and encouragement, AI can deliver real-time analytics, one on one instruction and multimodal assistance (text, speech, gestures and visual representations) which is individualized to particular learners. The gray solution can help optimize the learning process by integration of human creativity and ethical leadership with data-driven analytics and scaling it to the AI.

A Multisensory and embodied AI in the classroom: Cognition and Physiology

Advancing AI with robotics, haptics, and the multisensory and sensor-rich features make learning to be an experience of multiple senses at once, including sight, hearing, touch, and even kinesthetic feedback. Embodied AI tutors may act as puppets and control tangible objects, direct laboratory work, and simulate historical experiences in an immersive AR/VR experience. This multisensory interaction enhances concept learning and enables active learning particularly in STEM education and learners with special needs.

Friendly Educational Ecosystems that will run on AGI

Educational mediums should also be turned into interoperable systems and not isolated tools in order to maximize the potential of AGI. APIs and common standards of data will be able to permit the seamless integration of multimodal learning platforms, assessment, and student information systems. The interoperability enables lifetime, adaptive education pathways, dubbing efforts to follow the learner throughout the schools, across workplaces and within informal situations as well as guarantee data confidentiality and fairness.

Safe Adoption AGI Governance and Policy towards Education

It needs robust governance structures in order to introduce the use of AGI in schools. Transparency of algorithms, fair access to resources, and algorithms safeguards against prejudice or misuse of the learner data should be taken care of by policy. Guidelines on the national and international level will be essential in order to strike the right balance between innovation and moral accountability, making sure that AGI cannot and will not substitute human educators.

Benchmarking AGBI Educational AGI

The modern standards of AI implementation in education tend to assess a limited number of tasks, e.g., speech recognition or test scores prediction. As it becomes again, the more generic capabilities will need to be measured in the future, including contextual reasonableness, responsiveness to heterogeneous students, and multimodality in all communication, and moral judgment. The development of these AGI-compatible aims will be benchmarked to direct the allocation of research funds, policy decisions as well as customer confidence.

10. Conclusion:

A multimodal AI (a combination of text, speech, vision, gesture, and embodied interaction) is providing previously unknown possibilities to innovate education and bring closer to artificial general intelligence. Multimodality allows AI systems to perceive and react in the course of sensory channels, lessening the divide between machine and human learning functions. In classrooms, these capabilities can scale personalisation of instruction, deliver real time formative assessment, and deliver an experience that is multisensory and promotes curiosity, interest, and retention. Multimodal AI promise goes beyond technical innovation: it is a vehicle to the AGI goal proper vast and flexible and context-sensitive intelligence, which can adapt to various learners and surroundings. This vision requires the interrelationship between fields. AI research results need close collaboration with educators to approach models which resonate to classroom demands; ethicists and sociologists need to ask questions on issues of bias, equity, and fairness; policy makers should develop governance structures which protect privacy, transparency and equity. An AGI-crafted inclusive education pathology will enable the teachers and not push them off, in an extension of their ability to support creativity, empathy, and critical thinking. This vision would create relational and interoperable ecosystems in which lifelong learning will be open to everyone and which have an AI that is transparent and responsible and responsive to human values. The international community should spend on open standards, good benchmarking and interdisciplinary communication that can speed up the process of innovation without sacrificing on safety and confidence. With the adoption of multimodal AI today, and cross-sector integration, we will be able to establish the basis of an educational

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

environment where both human and artificial intelligences coexist and collaborate effectively to bring about adaptative, equitable, and valuable learning to all learners.

References:

"A Comprehensive Review of Multimodal Analysis in Education." (2025). Applied Sciences, 15(11), 5896.

"Multimodal learning." (n.d.). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Multimodal_learning

A Comprehensive Review of Multimodal Analysis in Education. (2025). Applied Sciences, 15(11), 5896. https://doi.org/10.3390/app15115896

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149-159. https://doi.org/10.1145/3287560.3287572

Chen, M., Xing, L., Wang, Y., & Zhang, Y. (2023). Enhanced Multimodal Representation Learning with Cross-Modal Knowledge Distillation. Proceedings of CVPR.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., & Wen, J.-R. (2022). Towards artificial general intelligence via a multimodal foundation model. Nature Communications, 13, Article 3094. https://doi.org/10.1038/s41467-022-30761-2

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., & Wen, J.-R. (2021). Towards artificial general intelligence via a multimodal foundation model. arXiv.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in education. Philosophy & Technology, 35(3), 1-18. https://doi.org/10.1007/s13347-022-00582-4

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., & Chaudhary, V. (2019). The FLORES evaluation datasets for low-resource machine translation. Proceedings of the 2019 Conference on Machine Translation, 62-72.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Heilala, V., Araya, R., & Hämäläinen, R. (2024). Beyond Text-to-Text: An Overview of Multimodal and Generative Artificial Intelligence for Education Using Topic Modeling. arXiv.

Hridi, A. P., Sahay, R., Hosseinalipour, S., & Akram, B. (2024). Revolutionizing AI-Assisted Education with Federated Learning: A Pathway to Distributed, Privacy-Preserving, and Debiased Learning Ecosystems. Proceedings of the AAAI Symposium Series, 3(1), 297–303.

Jin, Q., Ge, E., Xie, Y., Luo, H., Song, J., Bi, Z., Liang, C. X., Guan, J., Yeong, J., & Hao, J. (2025). Multimodal Representation Learning and Fusion. arXiv preprint.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., & Khan, F. S. (2023). Transformers in vision: A survey. ACM Computing Surveys, 55(13), 1-41. https://doi.org/10.1145/3505244

Langley, P. (2006). Cognitive architectures: Research issues and challenges. (referenced in AGI literature)

Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., Li, S., Liu, T., & Zhai, X. (2023). AGI: Artificial general intelligence for education. arXiv.

Lee, G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., Guo, S., Liu, Z., Mai, G., Liu, T., & Zhai, X. (Accepted/In press). Multimodality of AI for Education: Towards Artificial General Intelligence. IEEE Transactions on Learning Technologies.

Lee, G.-G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., Guo, S., Liu, Z., Mai, G., Liu, T., & Zhai, X. (in press). Multimodality of AI for Education: Towards Artificial General Intelligence. IEEE Transactions on Learning Technologies.

Lee, G.-G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., … Zhai, X. (2023). Multimodality of AI for Education: Towards Artificial General Intelligence. arXiv.

Li, S., & Tang, H. (2024). Multimodal alignment and fusion: A survey. arXiv.

Li, S., Tang, H. (2024). Multimodal Alignment and Fusion: A Survey. arXiv preprint.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). Artificial intelligence and the future of teaching and learning. OECD Publishing.

Lymperaiou, M. et al. (2024). A survey on knowledge-enhanced multimodal learning. Journal of Artificial Intelligence Research.

Volume – 2 | Issue – VI | Nov-Dec 2025 | ISSN: 3049-0278 (Online)

Amitrakshar International Journal
of Interdisciplinary and Transdisciplinary Research (AIJITR)
(A Social Science, Science and Indian Knowledge Systems Perspective)
Open-Access, Peer-Reviewed, Refereed, Bi-Monthly, International E-Journal

Mai, S., Hu, H., & Xing, S. (2019). Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. arXiv preprint.

Mai, S., Zeng, Y., & Hu, H. (2022). Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations. arXiv preprint.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning. ICML.

Ofori, F., Maina, E., & Gitonga, R. (2025). Multi-modal multi-task federated foundation models for education (M3T FedFMs). (Position paper)

Pan, H., Zhao, X., He, L., Shi, Y., & Lin, X. (2024). A survey of multimodal federated learning: background, applications, and perspectives. Multimedia Systems.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. OpenAI / CLIP.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26(2), 582-599. https://doi.org/10.1007/s40593-016-0110-3

Rosé, C. P., Ferschke, O., Howley, I., & Gweon, G. (2019). Dialogue-based tutoring and collaborative learning: Tracing learning outcomes in naturalistic data. International Journal of Artificial Intelligence in Education, 29(3), 293-315. https://doi.org/10.1007/s40593-019-00184-x

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Scardamalia, M., & Bereiter, C. (2014). Knowledge building and knowledge creation: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed., pp. 397-417). Cambridge University Press.

Sharma, K., & Giannakos, M. (2025). Artificial intelligence in multimodal learning analytics: A systematic review. [Journal].

Srivastava, N., & Salakhutdinov, R. (2014). Multimodal Learning with Deep Boltzmann Machines. Journal of Machine Learning Research.

Tuomi, I. (2021). The ethics of artificial intelligence in education: Promises and perils of data-driven education. European Journal of Education, 56(4), 543-557. https://doi.org/10.1111/ejed.12468

UNESCO. (2023). Guidance on generative AI in education and research. United Nations Educational, Scientific and Cultural Organization.

West, M., & Chew, H. E. (2014). Reading in the mobile era: A study of mobile reading in developing countries. UNESCO Working Paper Series on Mobile Learning.

Williamson, B., & Piattoeva, N. (2022). Education governance and datafication: Ethics, politics and practices. Learning, Media and Technology, 47(1), 1-8. https://doi.org/10.1080/17439884.2021.2011479

Zhai, X., et al. (2025). Multimodal learning analytics in education research. (See also “A Comprehensive Review of Multimodal Analysis in Education,” 2025).

Published by: Amitrakshar Publishers – Amitrakshar International Journal of Interdisciplinary and Transdisciplinary Research (AIJITR)
https://www.amitrakshar.co.in/journal/ : editor@amitrakshar.co.in DOI (Crossref) Prefix: 10.63431