International Journal of Scientific Research and Modern Technology (IJSRMT) ijsrmt.com
Volume 4, Issue 1, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i1.1163

Multimodal Large Language Models for Diagnostic Feedback Analytics in STEM Learning Platforms

Everlyne Fradia Akello[1]; Onuh Matthew Ijiga[2]; Idoko Peter Idoko[3]; Lawrence Anebi Enyejo[4]

[1]The Gladys W. and David H. Patton College of Education, Ohio University, Athens, Ohio, USA
[2]Department of Physics, Joseph Sarwuan Tarka University, Makurdi, Nigeria
[3]Department of Electrical/ Electronic Engineering, University of Ibadan, Nigeria
[4]Department of Telecommunications, Enforcement Ancillary and Maintenance, National Broadcasting Commission Headquarters, Aso-Villa, Abuja, Nigeria.

Publication Date: 2025/01/30

Abstract

The increasing complexity of STEM learning tasks and the scale of digital education environments have exposed fundamental limitations in traditional automated feedback systems. Most existing approaches rely on unimodal inputs or rule-based logic, providing surface-level feedback that fails to capture underlying learner misconceptions and reasoning processes. This study proposes and evaluates a multimodal large language model–driven diagnostic feedback framework designed to deliver accurate, explainable, and instructionally aligned feedback in STEM learning platforms. The framework integrates heterogeneous learner data, including text responses, symbolic mathematics, diagrams, code submissions, and interaction traces, through modality-specific encoders and attention-based fusion strategies. Diagnostic reasoning is performed using a multimodal large language model constrained by curricular objectives and enhanced with explainability mechanisms such as rationale tracing and attention visualization. Empirical evaluation across mathematics, physics, and computer science tasks demonstrates significant improvements over baseline systems in diagnostic accuracy, learning gains, error correction rates, learner engagement, and trust. The findings indicate that multimodal LLM-driven diagnostic feedback can operationalize formative assessment principles at scale, offering a robust pathway toward more transparent, adaptive, and pedagogically meaningful AI-supported learning in STEM education.

Keywords: *Multimodal Learning Analytics; Large Language Models; Diagnostic Feedback; STEM Education; Explainable Artificial Intelligence.*

I. INTRODUCTION

➢ *Background and Motivation*

Over the past two decades, STEM learning platforms have undergone a substantive transformation, shifting from static content repositories toward adaptive, analytics-driven learning environments. Early digital platforms largely replicated textbook-based instruction through linear content delivery and summative assessment tools. While these systems expanded access to learning materials, they offered limited insight into how students reasoned through problems or where conceptual breakdowns occurred. The emergence of learning analytics reframed this paradigm by leveraging learner interaction data to model performance, engagement, and progression, enabling more responsive instructional interventions (Siemens, 2013; Idoko et al., 2023). Contemporary STEM platforms increasingly integrate real-time analytics to support personalization, mastery-based progression, and formative assessment at scale.

Despite these advances, traditional feedback mechanisms remain constrained in their diagnostic capacity. Rule-based feedback and correctness-only evaluation often fail to identify underlying misconceptions, cognitive gaps, or procedural errors that are characteristic of STEM learning. Research in formative feedback demonstrates that effective

Akello , E. F., Ijiga, O. M., Idoko, I. P., & Enyejo, L. A. (2025). Multimodal Large Language Models for Diagnostic Feedback Analytics in STEM Learning Platforms. *International Journal of Scientific Research and Modern Technology*, *4*(1), 182–210. https://doi.org/10.38124/ijsrmt.v4i1.1163

instructional feedback must address not only whether an answer is correct, but why an error occurred and how the learner's reasoning deviated from domain principles (Shute, 2008; Idoko et al., 2024). In mathematics, physics, and computer science, student errors frequently stem from flawed mental models or partial conceptual understanding rather than surface-level mistakes. Conventional automated systems struggle to interpret these deeper reasoning processes, particularly when learner responses involve symbolic expressions, diagrams, or multi-step problem-solving strategies (Koedinger et al., 2013; Idoko et al., 2024).

The limitations of traditional feedback systems have motivated growing interest in artificial intelligence–driven approaches that can reason over complex learner data. Large language models (LLMs) represent a significant inflection point in this trajectory. By learning statistical and semantic representations from large-scale corpora, LLMs demonstrate strong capabilities in explanation generation, reasoning, and contextual interpretation across diverse domains (Brown et al., 2020). In educational contexts, these capabilities enable more nuanced diagnostic feedback that can articulate why a solution is incorrect, suggest alternative reasoning paths, and adapt explanations to learner proficiency levels.

More recently, the extension of LLMs toward multimodal learning analytics has opened new possibilities for STEM education. STEM learning is inherently multimodal, involving text, equations, diagrams, code, gestures, and interaction traces. Multimodal learning analytics research emphasizes that integrating multiple data streams yields richer representations of learner cognition than any single modality alone (Ochoa & Worsley, 2016). When combined with multimodal LLM architectures capable of jointly reasoning over textual, visual, and symbolic inputs, diagnostic feedback systems can move beyond surface evaluation toward holistic interpretation of learner thinking. This convergence positions multimodal LLMs as a foundational technology for next-generation diagnostic feedback analytics, addressing long-standing challenges in accuracy, interpretability, and pedagogical alignment within STEM learning platforms.

Figure 1 illustrates the progressive development of educational technology from early behaviorist teaching machines to fully web-based higher education programs. It highlights key milestones, beginning with programmed instruction in the 1950s, followed by the introduction of computers into formal instruction and elementary education. The figure also captures the shift toward distance learning through computer conferencing in the late 1980s and the emergence of accredited online universities in the 1990s. Collectively, it demonstrates how technological innovation gradually transformed education from localized, instructor-driven delivery to scalable, technology-mediated learning environments.

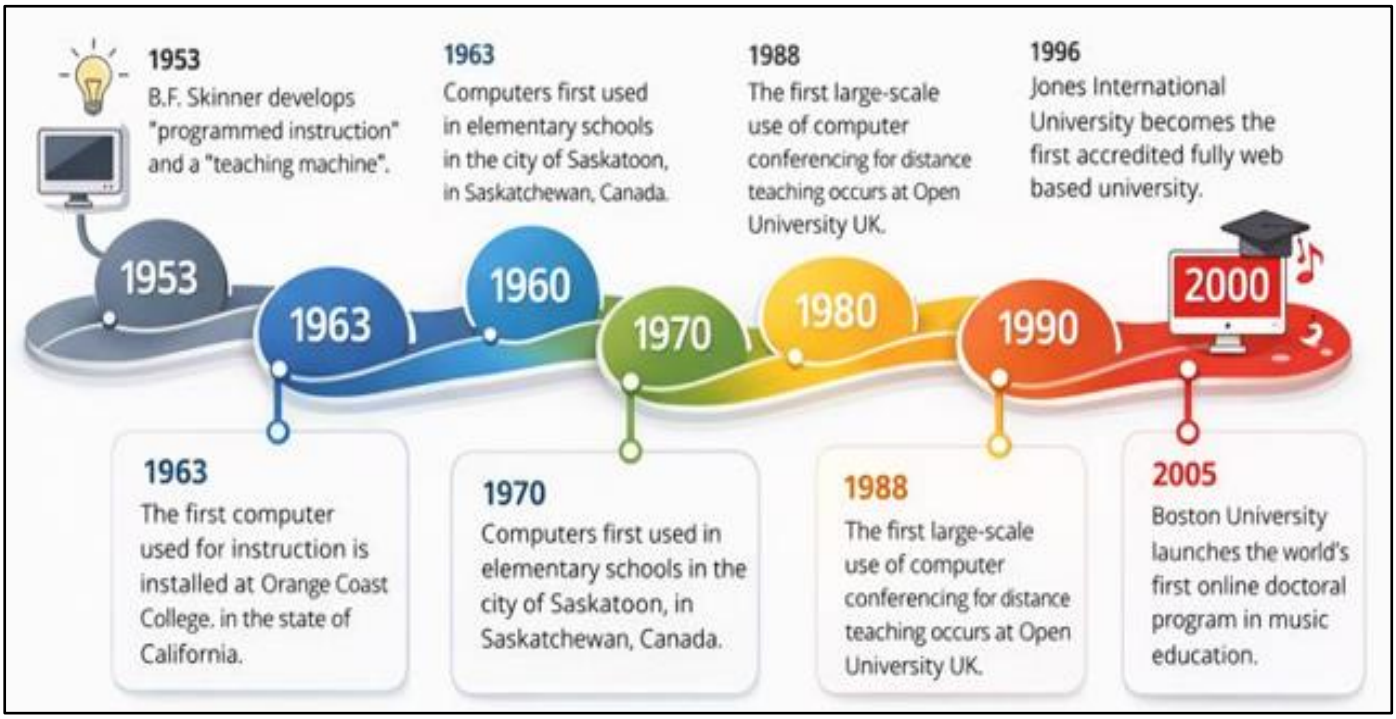

Fig 1 Evolution of Educational Technology and Computer-Assisted Learning (1953–2005)

➢ *Multimodal Learning in STEM Contexts*

STEM learning is inherently multimodal, relying on the coordinated use of text, equations, diagrams, code, simulations, and learner interaction traces to support conceptual understanding and problem-solving. Textual explanations provide semantic grounding and procedural guidance, while mathematical equations and symbolic representations encode abstract relationships that are central to scientific and engineering reasoning. Diagrams and visualizations externalize spatial and relational structures, enabling learners to offload cognitive processing and reason more effectively about complex systems (Ainsworth, 2006; Idoko et al., 2024). In computer science and engineering education, executable

code and simulations further extend this representational landscape by allowing learners to test hypotheses, observe system behavior dynamically, and iteratively refine solutions through experimentation (Wilensky & Reisman, 2006).

Beyond representational artifacts, learner interaction traces such as keystrokes, code revisions, response timings, and navigation patterns constitute an additional modality that captures process-level evidence of learning. These traces provide insight into how learners approach problems, where they hesitate, and how their strategies evolve over time. Research in intelligent tutoring systems and learning analytics demonstrates that such fine-grained behavioral data are often more diagnostic of underlying understanding than final answers alone (Koedinger et al., 2013; Idoko et al., 2024). In STEM contexts, where incorrect answers may arise from diverse conceptual or procedural sources, interaction data help disambiguate surface errors from deeper misconceptions.

The pedagogical value of integrating multiple data modalities lies in their complementary strengths. Theories of multimedia learning emphasize that meaningful learning occurs when learners actively integrate verbal and visual representations into coherent mental models (Mayer, 2009; Idoko et al., 2024). When instructional systems analyze these modalities jointly, they can better infer learner cognition and provide feedback that targets the specific representational breakdown involved. For example, a learner may correctly articulate a concept in text while misapplying it in an equation or diagram, a discrepancy that would remain invisible in unimodal assessment frameworks.

Multimodal learning analytics extends this pedagogical principle to diagnostic feedback systems by combining representational and behavioral data into unified learner models. Empirical studies show that multimodal data fusion improves the accuracy of learner state estimation and supports richer interpretations of engagement, strategy use, and conceptual understanding (Ochoa & Worsley, 2016; Idoko et al., 2024). In STEM learning platforms, such integration enables diagnostic feedback that is not only corrective but explanatory, linking errors to specific representations or actions. This deeper diagnostic insight is essential for supporting transfer, self-regulation, and durable learning outcomes in complex STEM domains.

Figure 2 presents a conceptual representation of multimodal learning, showing how auditory, visual, and kinesthetic learning pathways converge to support holistic knowledge acquisition. The auditory mode emphasizes listening and comprehension, the visual mode reinforces retention through observation, and the kinesthetic mode promotes understanding through active engagement. By integrating these complementary modalities, the framework illustrates how learners develop richer cognitive connections and deeper conceptual understanding. The central positioning of multimodal learning highlights its role in synthesizing diverse sensory inputs into meaningful learning outcomes.

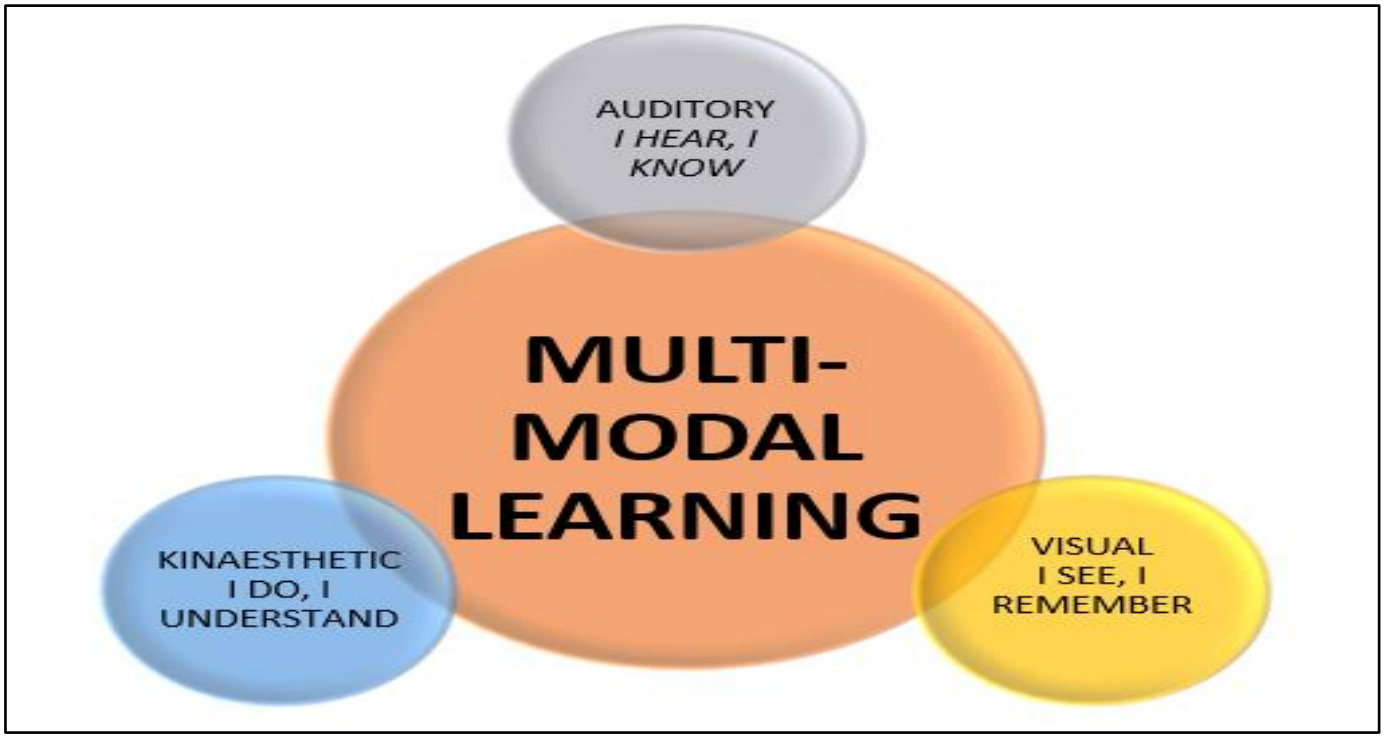

Fig 2 Multimodal Learning Framework Integrating Auditory, Visual, and Kinesthetic Pathways

➢ *Problem Statement*

Despite significant advances in digital STEM education, unimodal and rule-based feedback systems remain fundamentally inadequate for supporting complex reasoning processes characteristic of mathematics, science, engineering, and computing. Most automated feedback mechanisms rely on predefined rules, correctness checks, or single-modality inputs such as final

answers or short text responses. While effective for well-structured tasks, these approaches struggle to interpret multi-step reasoning, symbolic manipulation, diagrammatic thinking, or iterative problem-solving workflows that define authentic STEM learning. As a result, learners often receive surface-level feedback that signals error without diagnosing the conceptual or procedural source of the mistake (Shute, 2008).

STEM reasoning errors are rarely uniform. The same incorrect response may arise from distinct misconceptions, misapplied formulas, flawed mental models, or breakdowns in procedural sequencing. Unimodal feedback systems lack the representational breadth required to distinguish among these possibilities, leading to generic or misleading feedback. Empirical research in intelligent tutoring systems shows that models incorporating richer evidence of learner activity outperform answer-based systems in identifying learning gaps and supporting knowledge transfer (Koedinger et al., 2013). However, many current platforms still operate on simplified evaluation logic that fails to leverage available multimodal learner data.

At scale, these limitations are further compounded by the need for timely feedback. Large STEM courses and online learning platforms often serve thousands of learners simultaneously, making human-delivered diagnostic feedback impractical. Learning analytics research highlights that delayed or non-specific feedback weakens formative assessment and reduces its impact on learning outcomes (Siemens, 2013). Rule-based automation offers scalability but sacrifices contextual sensitivity, while human feedback offers depth but lacks scalability. This trade-off remains unresolved in most production learning systems.

An additional challenge lies in the opacity of advanced AI-driven feedback mechanisms. While machine learning models can improve diagnostic accuracy, their lack of interpretability raises pedagogical and ethical concerns. Learners and instructors must be able to understand why specific feedback is generated in order to trust, act upon, and validate it. Research in explainable artificial intelligence emphasizes that models used in high-stakes decision contexts, including education, should provide transparent reasoning pathways rather than black-box outputs (Guidotti et al., 2018; Idoko et al., 2024). Without explainability, even accurate feedback risks being pedagogically ineffective or mistrusted.

Consequently, there is a critical need for diagnostic feedback systems that are explainable, context-aware, and responsive in real time. Such systems must integrate multiple learner data modalities, adapt feedback to the learner's current context, and articulate the reasoning behind feedback decisions. Addressing this gap is essential for advancing scalable, trustworthy, and instructionally aligned feedback in modern STEM learning platforms.

Figure 3 illustrates explainable AI in education as a central construct supported by three interdependent domains: artificial intelligence, human–computer interaction, and cognitive and learning sciences. Artificial intelligence provides the computational models and reasoning capabilities, while human–computer interaction ensures usability, transparency, and meaningful learner engagement. Cognitive and learning sciences ground system design in how learners process, retain, and apply knowledge. The circular integration highlights that effective explainable AI in education emerges from the balanced alignment of technical intelligence, human-centered design, and learning theory.

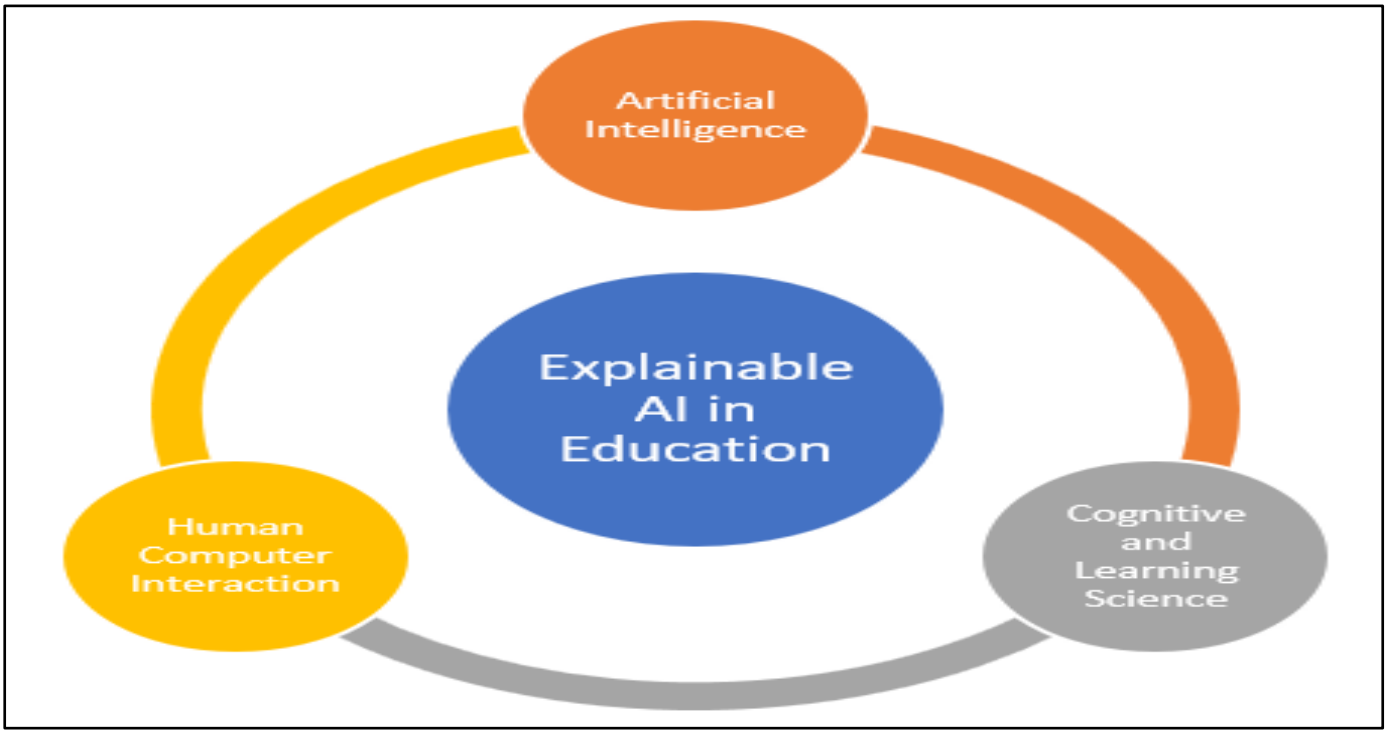

Fig 3 Conceptual Framework of Explainable Artificial Intelligence in Education

➢ *Research Objectives and Questions*

The primary objective of this study is to design and empirically evaluate a multimodal large language model–driven diagnostic feedback framework tailored for STEM learning platforms. The proposed framework seeks to move beyond answer-level evaluation by integrating heterogeneous learner data, including textual responses, symbolic expressions, diagrams, code artifacts, simulation outputs, and interaction traces, in order to infer learner understanding with greater precision. By leveraging multimodal reasoning capabilities, the study aims to generate diagnostic feedback that identifies the source of learner errors, explains underlying misconceptions, and provides actionable guidance aligned with instructional goals.

A central focus of the research is the evaluation of diagnostic accuracy across diverse STEM task types. The study examines whether multimodal LLM-based feedback can more reliably distinguish between conceptual misunderstandings and procedural mistakes when compared to conventional feedback systems. In parallel, the pedagogical usefulness of the generated feedback is assessed, with attention to clarity, instructional alignment, and its ability to support learner reflection, error correction, and knowledge transfer.

Interpretability constitutes another core objective of the research. The framework is designed to produce feedback that is not only informative but also transparent in its reasoning, enabling learners and educators to understand how conclusions are derived from multimodal evidence. This emphasis on interpretability supports trust, accountability, and effective human–AI collaboration within educational settings.

Finally, the study addresses scalability and practical deployment considerations. The framework is evaluated in terms of its ability to deliver timely, context-aware diagnostic feedback in large-scale learning environments without compromising quality or responsiveness. Together, these objectives guide the formulation of research questions that examine the accuracy, pedagogical value, interpretability, and scalability of multimodal LLM-driven diagnostic feedback systems for STEM education.

➢ *Contributions of the Study*

This study makes several substantive contributions to the advancement of diagnostic feedback analytics in STEM learning environments. First, it proposes a comprehensive conceptual framework for multimodal diagnostic feedback analytics that systematically integrates diverse learner data sources, including text, symbolic representations, visual artifacts, executable code, simulations, and interaction traces. The framework articulates how multimodal inputs can be jointly analyzed using large language models to infer learner understanding, detect misconceptions, and generate context-sensitive feedback. By formalizing this integration, the study provides a structured reference model that can guide future research and system development in multimodal learning analytics.

Second, the study contributes empirical or prototype-based validation of the proposed framework within an authentic STEM learning context. Through experimental evaluation or system-level implementation, the research demonstrates how multimodal LLM-driven diagnostic feedback performs in comparison to traditional feedback mechanisms. This validation highlights the framework's effectiveness in improving diagnostic accuracy, feedback relevance, and learner engagement across representative STEM tasks. The findings offer concrete evidence of the practical feasibility and educational value of deploying multimodal LLM-based feedback systems in real-world learning platforms.

Third, the study advances broader theoretical and practical understanding in learning analytics, AI-supported pedagogy, and educational technology design. For learning analytics, it extends existing models by emphasizing multimodal evidence and explainable feedback generation as core analytic functions. For pedagogy, it illustrates how AI-driven diagnostic feedback can be aligned with formative assessment principles and instructional intent rather than operating as a purely technical intervention. For educational technology design, the study provides actionable insights into system architecture, scalability, and human–AI interaction considerations, supporting the responsible integration of advanced AI capabilities into next-generation STEM learning platforms.

II. LITERATURE REVIEW

➢ *Diagnostic Feedback in STEM Education*

Diagnostic feedback in STEM education is grounded in the theory of formative assessment, which emphasizes the use of evidence about student learning to adapt instruction and improve understanding during the learning process. Foundational work in educational assessment established that formative feedback is most effective when it clarifies learning goals, identifies gaps between current and desired performance, and provides guidance on how to close those gaps. In mathematics, science, and engineering education, formative assessment plays a critical role because learning often involves cumulative reasoning, abstraction, and the coordination of multiple representations. Empirical syntheses of classroom-based assessment practices demonstrate that well-designed formative feedback leads to substantial gains in student achievement, particularly in conceptually demanding subjects such as mathematics and science (Black & Wiliam, 1998).

Within STEM domains, feedback must address more than task correctness. Research on feedback theory highlights that effective diagnostic feedback operates at multiple levels, including task-level accuracy, process-level reasoning, self-regulation, and conceptual understanding. In engineering and physics problem

solving, for example, students may arrive at incorrect solutions due to misapplied principles, incorrect assumptions, or breakdowns in procedural sequencing. Feedback that merely signals correctness fails to support learning unless it explains why an approach is flawed and how alternative reasoning aligns with domain principles. Studies across STEM disciplines show that explanatory and process-oriented feedback is more strongly associated with learning gains than outcome-only feedback (Hattie & Timperley, 2007; Shute, 2008; Idoko et al., 2024).

Cognitive diagnostic models (CDMs) provide a formal framework for analyzing learner misconceptions and skill mastery in STEM learning. These models decompose performance into latent attributes or knowledge components, enabling fine-grained diagnosis of specific conceptual or procedural weaknesses. In mathematics education, CDMs have been used to identify misconceptions in algebraic manipulation and proportional reasoning, while in science and engineering they have supported diagnosis of misunderstandings related to physical laws, system dynamics, and design constraints. By mapping observable responses to underlying cognitive attributes, cognitive diagnostic assessment supports targeted feedback that is closely aligned with instructional objectives (Leighton & Gierl, 2007; Ijiga et al., 2024).

Misconceptions analysis further strengthens diagnostic feedback by recognizing that learner errors are often systematic rather than random. Research in STEM education consistently shows that students develop stable but incorrect mental models, such as misconceptions about force and motion in physics or variable scope in programming. Intelligent tutoring and learning sciences research demonstrates that identifying these misconceptions requires analyzing patterns of errors and solution strategies rather than isolated answers. Instructional systems that incorporate misconception-aware diagnostics are better positioned to provide feedback that challenges faulty reasoning and promotes conceptual change (Koedinger et al., 2013).

Taken together, formative assessment theory and cognitive diagnostic modeling establish a strong foundation for diagnostic feedback in STEM education. They underscore the necessity of feedback systems that are sensitive to learner reasoning, capable of identifying misconceptions, and aligned with domain-specific learning progressions. These principles inform the design of advanced diagnostic feedback frameworks that seek to operationalize formative assessment at scale in modern STEM learning environments.

Figure 4 illustrates a continuous, data-driven instructional cycle that integrates assessment, instructional design, and learner support. The process begins with pre-assessment to identify learner readiness, followed by the application of universally designed and blended learning strategies. Formative assessment data are then collected and analyzed to differentiate instruction, models, and support. The cycle concludes with summative assessment, feeding insights back into subsequent instructional planning for continuous improvement.

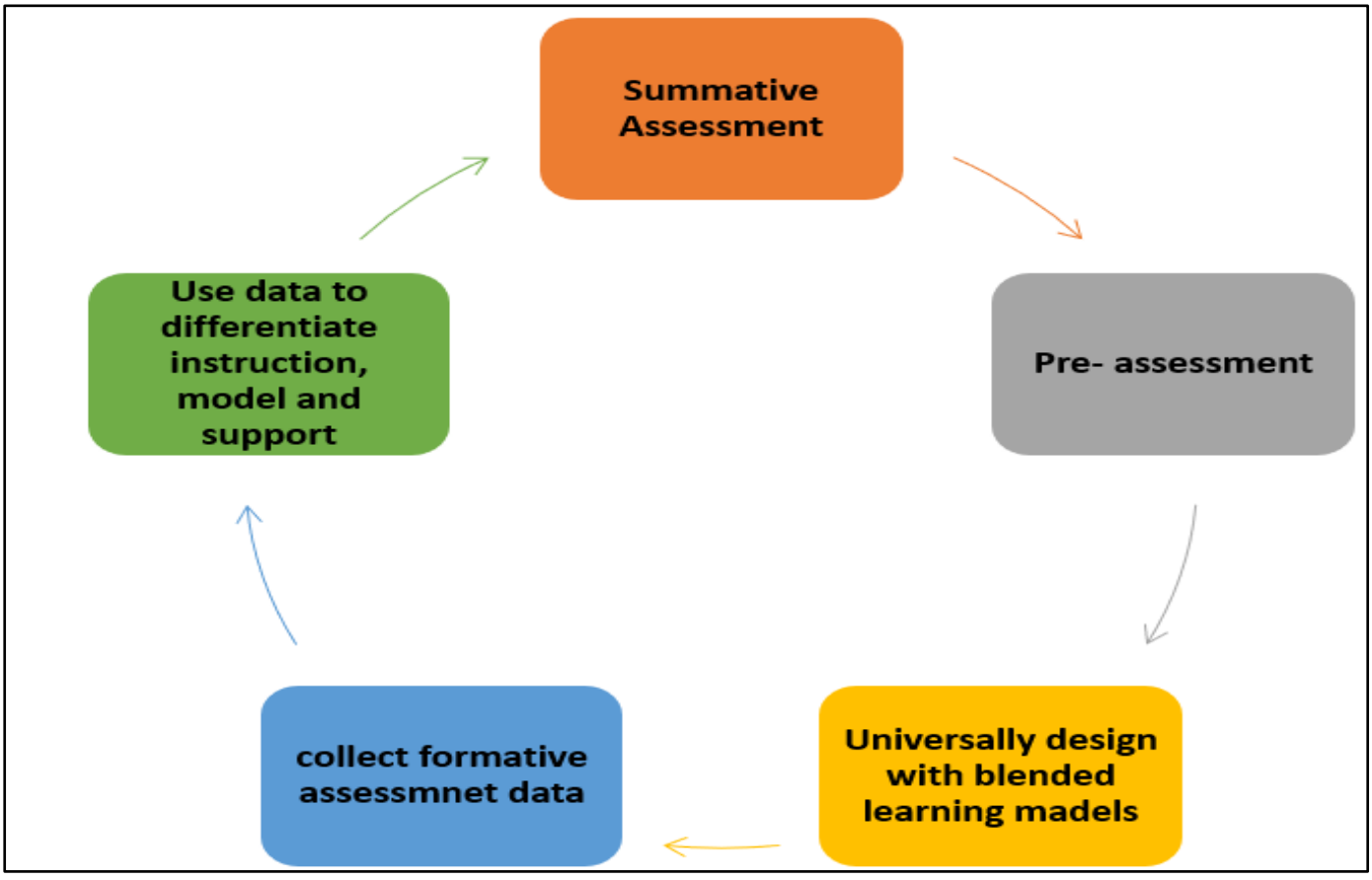

Fig 4 Cyclical Framework for Data-Driven Instruction and Assessment

➢ *Learning Analytics and Educational Data Mining*

Learning analytics and educational data mining have evolved from primarily descriptive approaches toward predictive and prescriptive systems capable of supporting real-time instructional decision-making. Early applications focused on descriptive analytics, summarizing learner activity through dashboards, grades, and usage statistics to provide retrospective insights into participation and performance. While these approaches improved institutional awareness, they offered limited support for timely intervention or personalized feedback. The formal emergence of learning analytics as a discipline marked a shift toward using data to understand and optimize learning processes rather than merely reporting outcomes (Siemens, 2013).

As computational methods matured, predictive analytics became central to learning analytics and educational data mining. Predictive models leverage historical learner data to forecast future outcomes such as academic success, dropout risk, or mastery progression. In STEM education, these models have been applied to anticipate problem-solving difficulties, detect at-risk learners, and estimate knowledge component mastery based on prior performance patterns. Educational data mining research demonstrates that predictive modeling can uncover latent learning trajectories that are not observable through descriptive statistics alone, enabling earlier and more targeted instructional support (Baker & Inventado, 2014).

More recently, prescriptive analytics has extended predictive insights into actionable feedback and intervention strategies. Prescriptive systems not only estimate what is likely to happen but also recommend what should be done to improve learning outcomes. In adaptive learning environments, prescriptive analytics drive automated feedback, content sequencing, and instructional scaffolding tailored to individual learners. This transition is particularly significant for STEM domains, where effective feedback must respond dynamically to evolving learner understanding across complex, multi-step tasks (Romero & Ventura, 2010; Manuel et al., 2024).

Central to both predictive and prescriptive analytics is the use of rich learner data streams. Learner interaction logs capture fine-grained behavioral signals such as response timing, hint usage, navigation paths, code revisions, and error patterns. Performance data, including correctness, partial credit, and attempt histories, provide evidence of task-level achievement, while behavioral indicators such as persistence, revision frequency, and help-seeking behavior offer insight into learning strategies and self-regulation. Research shows that combining these data sources yields more accurate models of learner cognition and engagement than relying on outcome data alone (Winne & Baker, 2013; Ayoola et al., 2024).

In STEM learning contexts, the integration of interaction logs, performance measures, and behavioral signals enables analytics systems to infer not only whether learners succeed, but how they learn. This process-oriented perspective is essential for generating diagnostic and adaptive feedback that aligns with formative assessment principles. By supporting the transition from descriptive reporting to predictive and prescriptive feedback systems, learning analytics and educational data mining provide a foundational infrastructure for scalable, data-driven personalization in modern STEM education.

Figure 5 depicts a human-centered, data-driven education framework that integrates educators, learners, and multiple learning environments through big data mining. Educational data from classrooms, mobile learning, content management systems, and intelligent networks are collected and preprocessed to support analytics such as modeling, clustering, and prediction. These analytics inform judgment processes and generate personalized study recommendations for learners. The framework emphasizes continuous interaction between humans and intelligent systems to enhance instructional design, learning outcomes, and adaptive educational decision-making.

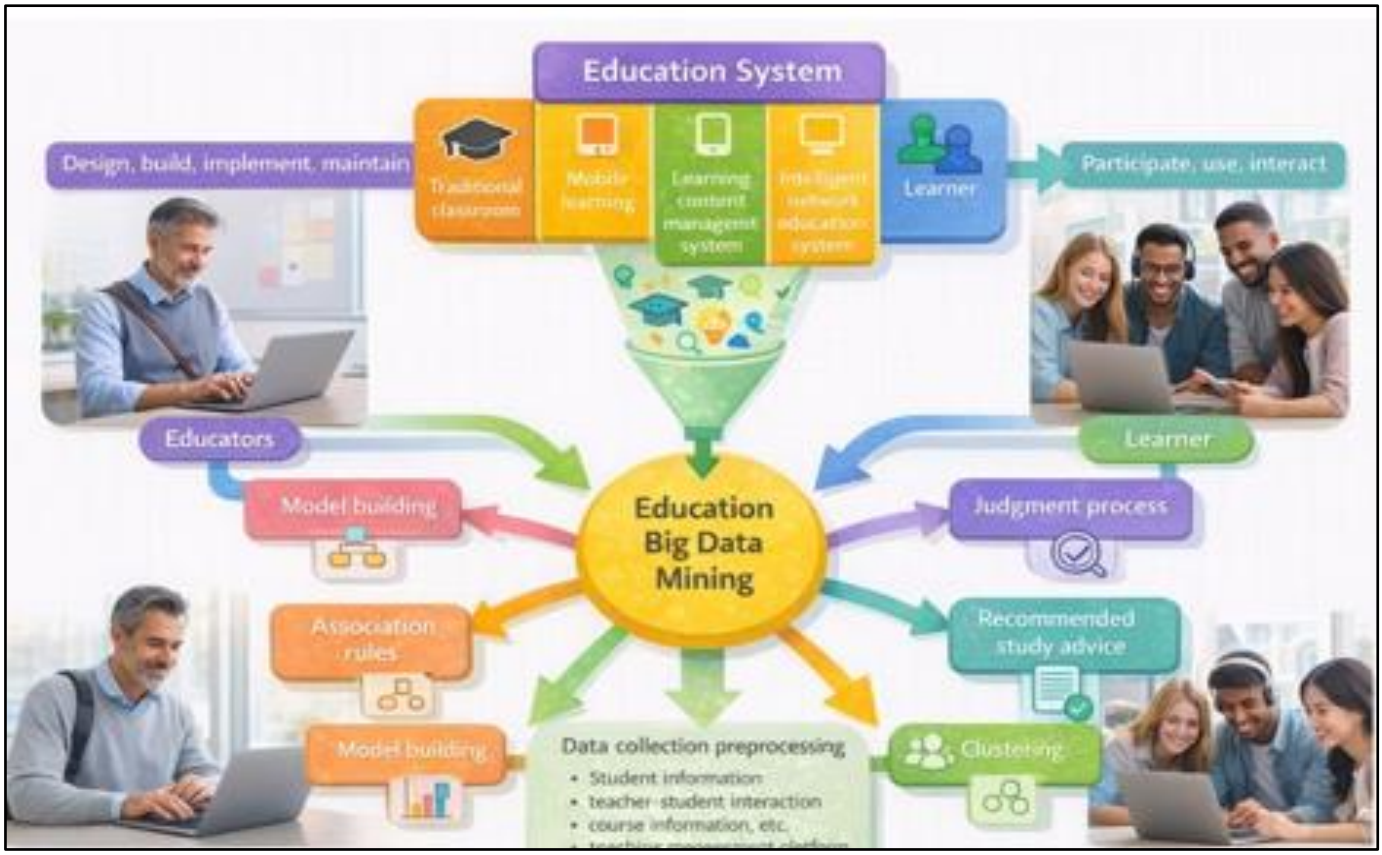

Fig 5 Human-Centered Big Data Mining Framework for Intelligent Education Systems

➢ *Large Language Models in Education*

Large language models (LLMs) have emerged as powerful computational tools with growing relevance for educational applications, particularly in STEM learning environments. Trained on large-scale textual corpora, LLMs demonstrate advanced capabilities in reasoning, explanation generation, and conversational interaction. These capabilities enable LLMs to support tasks such as step-by-step problem explanation, conceptual clarification, automated tutoring dialogues, and feedback generation across a wide range of subject domains. Empirical studies show that LLMs can generate coherent explanations, adapt responses to learner prompts, and simulate aspects of human tutoring, making them attractive for scalable instructional support in educational contexts (Brown et al., 2020; Kasneci et al., 2023; Ijiga et al., 2024).

In tutoring and feedback scenarios, LLMs are particularly effective at natural language explanation and contextualization. They can rephrase concepts using alternative analogies, respond to follow-up questions, and provide justifications for problem-solving steps. These affordances align well with formative assessment practices that emphasize explanation, reflection, and iterative learning. In STEM education, where learners often struggle to articulate reasoning or interpret abstract representations, LLMs offer a flexible interface for dialogic support and on-demand feedback. Research on foundation models highlights their capacity to generalize across tasks, allowing a single model to support multiple instructional functions without task-specific retraining (Bommasani et al., 2021; Ugbane et al., 2024).

Despite these strengths, the deployment of LLMs in education raises significant challenges. One of the most widely documented limitations is hallucination, where models generate fluent but incorrect or unsupported information. In educational settings, hallucinations pose a serious risk because learners may accept erroneous explanations as authoritative, particularly when feedback appears confident and detailed. Surveys of neural text generation systems demonstrate that hallucination remains a persistent issue, especially in reasoning-intensive tasks that require factual precision or domain-specific constraints (Ji et al., 2023; Ikedionu et al., 2025).

Bias represents another critical limitation of LLMs in educational contexts. Because these models learn from large, heterogeneous datasets, they may reproduce societal, cultural, or epistemic biases present in training data. Such biases can influence problem framing, examples, or feedback tone, potentially disadvantaging certain learner groups or reinforcing misconceptions. Foundational critiques of large language models emphasize that unexamined bias and representational imbalance can undermine fairness and inclusivity when models are applied in high-impact domains such as education (Bender et al., 2021; Eguagie et al., 2025).

Figure 6 illustrates an end-to-end framework that integrates model fine-tuning with a student practice and feedback loop for laparoscopic skill assessment. Expert-annotated surgical videos are used to fine-tune a multimodal video–language model capable of translating procedural performance into structured textual evaluations. During training sessions, student practice videos are decomposed into frames, encoded, and

analyzed by the fine-tuned model in conjunction with a large language model to generate objective performance feedback. The resulting feedback supports iterative learning by enabling students to review, reflect, and improve surgical technique through continuous AI-assisted assessment.

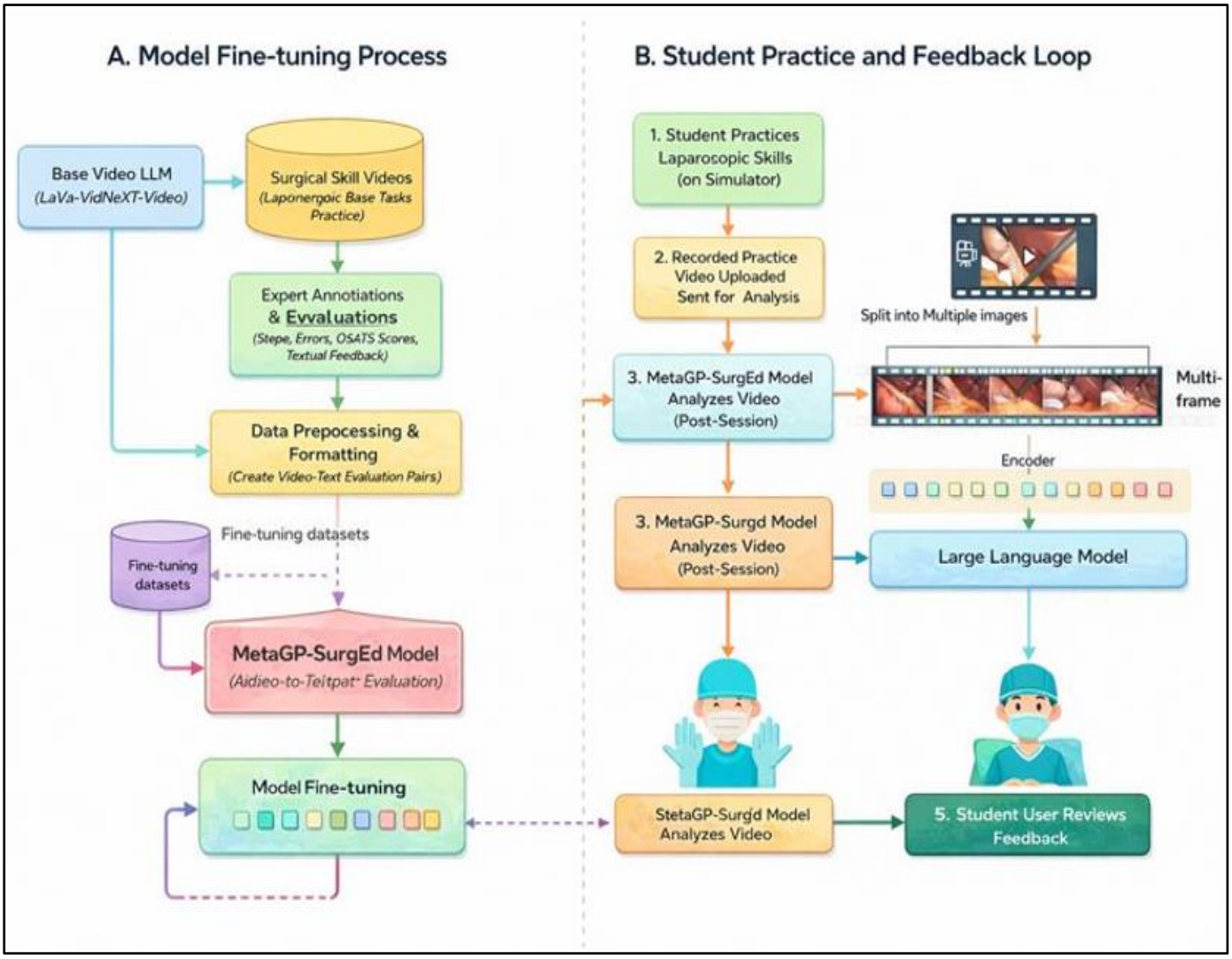

Fig 6 AI-Driven Video Analysis Framework for Surgical Skill Assessment and Adaptive Feedback Generation

A further challenge lies in pedagogical alignment. While LLMs excel at generating plausible explanations, they do not inherently reason according to curricular standards, learning progressions, or instructional intent. Without explicit constraints, LLM-generated feedback may prioritize surface coherence over instructional relevance, offer explanations that are misaligned with course objectives, or skip intermediate reasoning steps that are pedagogically important. Educational research stresses that effective tutoring requires alignment with domain models of learning and instructional design principles, which cannot be assumed from language modeling alone (Kasneci et al., 2023; Okika et al., 2025).

Collectively, these limitations underscore the need for careful integration of LLMs into educational systems. While LLMs offer substantial potential for reasoning support and tutoring, their use in STEM education must be guided by mechanisms for validation, bias mitigation, explainability, and pedagogical grounding to ensure that feedback is accurate, trustworthy, and instructionally meaningful.

Figure 7 presents a staged view of large language model evolution, beginning with foundational model construction and large-scale pre-training. Post-training fine-tuning enables the adaptation of foundation models into task-oriented systems such as classifiers and personal assistants. The framework then extends into deeper specialization, where models are optimized for distinct functions including retrieval-augmented generation, multimodal reasoning, and agent-based interaction. Emphasis is placed on reasoning models as a critical specialization layer, highlighting their role in advancing structured inference and decision-making capabilities.

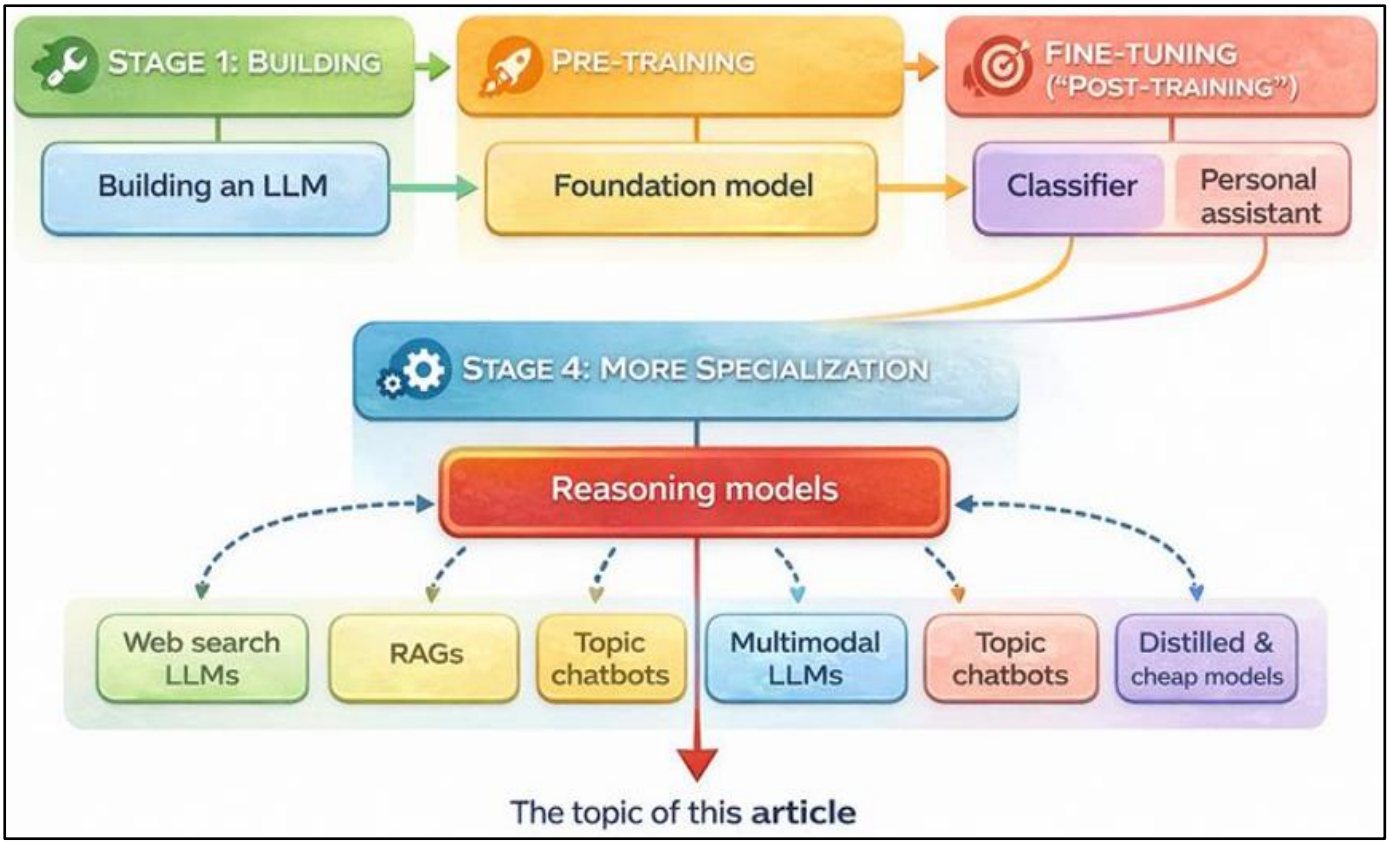

Fig 7 Progressive Stages in Large Language Model Development and Specialization

➢ *Multimodal AI Models and Representation Learning*

Multimodal artificial intelligence models are designed to learn joint representations from heterogeneous data sources such as text, visual inputs, symbolic structures, and temporal signals. At the architectural level, these models typically employ modality-specific encoders that transform raw inputs into latent representations, followed by fusion mechanisms that enable cross-modal interaction and reasoning. Early multimodal architectures relied on late or intermediate fusion strategies, while more recent approaches use shared embedding spaces and attention-based transformers to support fine-grained alignment across modalities (Baltrušaitis et al., 2019; Gaye et al., 2025). This evolution has enabled models to reason over complex combinations of linguistic, visual, and structured information within a unified representational framework.

Transformer-based multimodal architectures have become particularly influential. Models such as VisualBERT and VilBERT extend language transformers by incorporating visual embeddings derived from images, allowing bidirectional interaction between textual and visual representations (Li et al., 2019; Lu et al., 2019; Darko et al., 2025). Similarly, contrastive multimodal models align text and vision in a shared semantic space, enabling robust cross-modal retrieval and inference (Radford et al., 2021; Idogho et al., 2025). More recent general-purpose architectures, such as Perceiver IO, further generalize multimodal representation learning by supporting flexible input and output modalities, including temporal data streams, through attention-based latent bottlenecks (Jaegle et al., 2021).

Symbolic and structured representations play a critical role in extending multimodal learning to STEM domains. Equations, graphs, code syntax trees, and logical expressions encode formal relationships that differ fundamentally from natural language or images. Multimodal representation learning research increasingly emphasizes the integration of symbolic information with perceptual and textual data to support reasoning tasks that require precision and compositionality. Surveys of multimodal learning highlight that incorporating symbolic constraints improves robustness and interpretability in domains where correctness depends on formal structure rather than semantic plausibility alone (Kiela et al., 2019).

The relevance of multimodal AI models to STEM problem-solving is particularly pronounced. STEM tasks frequently involve coordinating diagrams with equations, interpreting graphical representations alongside textual problem statements, or debugging code based on execution traces and specifications. Multimodal models are well suited to these tasks because they can align visual features with symbolic expressions and linguistic explanations, enabling more holistic interpretation of learner solutions. For example, in mathematics and physics education, diagrams often encode spatial or relational information that complements algebraic formulations, while in computer science, code representations must be interpreted in conjunction with natural language problem descriptions and temporal execution behavior.

By supporting integrated reasoning across text, vision, symbols, and time, multimodal AI models provide a foundational capability for advanced diagnostic

feedback systems in STEM learning environments. Their representational flexibility enables analysis of learner work at the level of reasoning processes rather than isolated outputs, creating opportunities for feedback that is both diagnostically precise and pedagogically meaningful.

Figure 8 depicts a refined multimodal learning framework that integrates physiological signals (EDA, BVP, and temperature) through modality-specific encoders and a centralized multimodal transformer. Self-supervised pretraining enables robust shared representations, while selective freezing or fine-tuning supports efficient downstream adaptation. Modality-specific heads allow auxiliary supervision, whereas a unified emotion classification head captures cross-modal dependencies. This architecture improves generalization, interpretability, and robustness in affective computing tasks by harmonizing supervised and self-supervised learning stages.

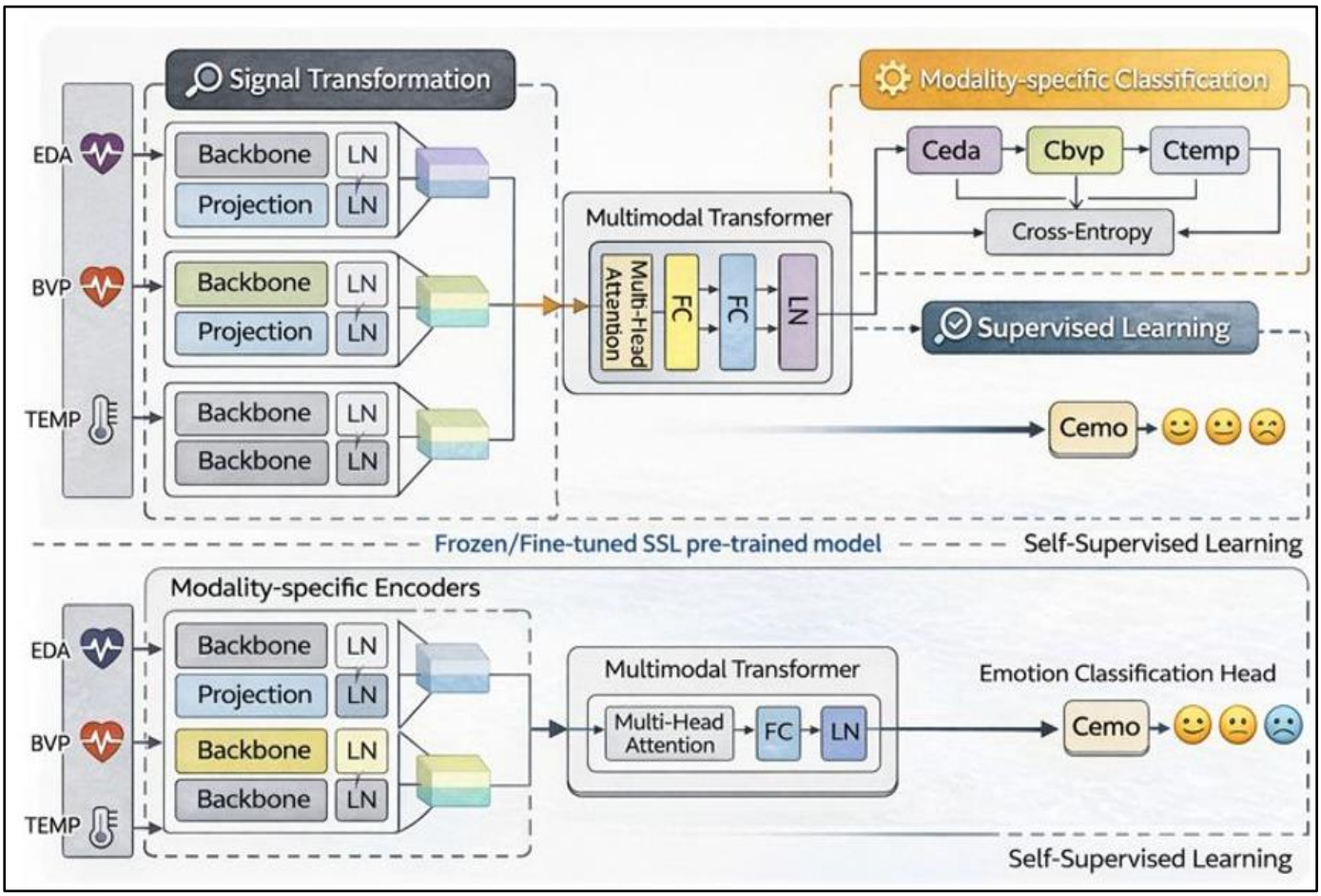

Fig 8 Standardized Multimodal Transformer Architecture for Physiological Signal Representation and Emotion Classification

➢ *Research Gaps*

Despite rapid advances in learning analytics, educational data mining, and artificial intelligence–driven tutoring systems, the integration of multimodal large language models with diagnostic feedback analytics remains limited. Existing research has largely treated multimodal learning analytics and language-based feedback generation as parallel developments rather than as components of a unified diagnostic framework. Many learning platforms analyze multimodal learner data such as interaction logs, diagrams, or code traces for prediction or classification purposes, yet the resulting insights are rarely translated into coherent, language-based diagnostic feedback that learners can readily understand and act upon. Conversely, language models used for educational feedback typically rely on unimodal textual inputs, overlooking the rich representational evidence embedded in equations, visual artifacts, simulations, and temporal learning behaviors. This disconnect constrains the ability of current systems to diagnose complex STEM reasoning processes with sufficient depth and contextual awareness.

A further gap concerns the limited emphasis on explainability within AI-driven feedback systems. While advanced models can generate fluent and contextually relevant responses, many operate as opaque mechanisms whose internal reasoning remains inaccessible to learners and instructors. In STEM education, where conceptual transparency and reasoning traceability are essential, feedback that lacks explainable justification risks undermining learning rather than supporting it. Learners may struggle to understand why a particular explanation is provided or how it relates to their own problem-solving process, reducing opportunities for reflection and self-regulation. The absence of explainability mechanisms also complicates instructor oversight and pedagogical validation, particularly in formal educational settings

where accountability and assessment alignment are critical.

Instructional alignment represents another underexplored dimension. Current AI feedback systems often prioritize linguistic coherence or problem completion over alignment with curricular standards, learning progressions, and domain-specific pedagogical goals. In STEM contexts, effective feedback must be sensitive to the sequence in which concepts are introduced, the representations emphasized by instruction, and the level of abstraction appropriate for the learner. Without explicit alignment to instructional intent, AI-generated feedback may introduce concepts prematurely, bypass foundational reasoning steps, or conflict with established teaching approaches. This misalignment can create confusion and erode confidence in AI-supported learning tools.

Finally, learner trust remains insufficiently addressed in the design of AI-driven diagnostic feedback systems. Trust is shaped not only by accuracy but also by transparency, consistency, and perceived pedagogical relevance. When feedback appears arbitrary, unverifiable, or disconnected from learner actions, students may disengage or disregard automated guidance altogether. The lack of systematic research on how multimodal, explainable feedback influences learner trust and acceptance represents a significant barrier to large-scale adoption. Addressing these gaps requires an integrated research agenda that combines multimodal representation learning, explainable AI, and instructional design principles to support trustworthy, diagnostically precise feedback in STEM learning environments.

III. METHODOLOGY

➢ *Research Design*

This study adopts a mixed-methods design-science research approach, combining system design and experimental evaluation to address the dual objectives of framework construction and empirical validation. Design-science is employed to systematically develop a multimodal large language model–driven diagnostic feedback framework, while controlled experimental methods are used to evaluate its effectiveness against defined pedagogical and technical criteria. This combination enables rigorous artifact creation alongside evidence-based assessment of its educational value.

From a design-science perspective, the research focuses on building a functional diagnostic feedback artifact that integrates heterogeneous learner data modalities, including text, equations, diagrams, code artifacts, simulations, and temporal interaction traces. The design process follows an iterative cycle of problem identification, framework development, prototype implementation, and refinement. This approach is well suited to the research objectives, which emphasize not only theoretical contribution but also the practical feasibility of deploying multimodal LLM-driven feedback in real STEM learning environments.

The experimental component of the study is used to evaluate the performance of the proposed framework along four core dimensions: diagnostic accuracy, pedagogical usefulness, interpretability, and scalability. Quantitative experiments compare the multimodal LLM-based system with baseline feedback approaches using task-level performance metrics, misconception detection accuracy, and response latency. Learner outcome measures such as error correction rates and learning gains are analyzed to assess pedagogical impact. Where appropriate, statistical significance testing and effect size estimation are used to support comparative claims.

Qualitative methods complement the experimental analysis by examining learner and instructor perceptions of feedback clarity, trustworthiness, and instructional alignment. Structured observations and feedback analysis are used to evaluate whether generated explanations align with curricular intent and support meaningful reflection. This qualitative layer is essential for interpreting results that cannot be fully captured through performance metrics alone, particularly in relation to explainability and user trust.

Formally, the diagnostic feedback process can be modeled as an optimization problem over multimodal learner inputs. Let

$$\mathcal{X} = \{x^{(t)}, x^{(v)}, x^{(s)}, x^{(c)}, x^{(\tau)}\}$$

Denote the set of textual, visual, symbolic, code-based, and temporal interaction inputs, respectively. The multimodal LLM learns a mapping

$$f_\theta : \mathcal{X} \rightarrow \mathcal{F}$$

where $\mathcal{F}$represents diagnostic feedback outputs. Model optimization seeks to minimize a composite loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{diag}} + \lambda_2 \mathcal{L}_{\text{ped}} + \lambda_3 \mathcal{L}_{\text{exp}}$$

where $\mathcal{L}_{\text{diag}}$captures diagnostic accuracy, $\mathcal{L}_{\text{ped}}$represents pedagogical alignment constraints, and $\mathcal{L}_{\text{exp}}$enforces explainability requirements. The weighting parameters $\lambda_i$control trade-offs among these objectives.

The chosen mixed-methods design ensures alignment with the study's objectives by enabling systematic framework development, robust empirical evaluation, and pedagogically grounded interpretation. This methodology supports both the generation of generalizable insights and the demonstration of practical impact within STEM learning platforms.

➢ *System Architecture and Framework*

The proposed system architecture is organized as a multimodal diagnostic feedback pipeline that ingests heterogeneous learner artifacts, constructs unified representations, and generates explainable, context-aware feedback using a multimodal large language model. The

framework is modular by design, enabling extensibility across STEM domains and scalable deployment in large learning platforms.

➢ *Overview of the Multimodal Diagnostic Feedback Pipeline*

The pipeline consists of five sequential layers: (i) multimodal data ingestion, (ii) modality-specific encoding, (iii) cross-modal fusion, (iv) diagnostic reasoning and feedback generation, and (v) explanation and delivery. Learner-generated artifacts are captured continuously during problem-solving and routed to specialized encoders that preserve the structural properties of each modality. The fused representation is then processed by a multimodal LLM that performs diagnostic inference and produces feedback aligned with instructional intent.

➢ *Data Ingestion Across Learning Modalities*

The ingestion layer collects learner data from multiple sources:

- Textual responses, including free-form explanations and short answers, are captured directly from learning interfaces.
- Symbolic mathematics inputs, such as equations and algebraic expressions, are parsed into structured representations that preserve operator hierarchy and variable relationships.
- Diagrams and visual artifacts, including sketches, graphs, and annotated figures, are processed as visual inputs with spatial feature extraction.
- Code submissions, encompassing source code and revision histories, are represented through syntactic and semantic encodings that reflect program structure and execution logic.
- Interaction logs, such as timestamps, hint usage, navigation paths, and revision frequency, provide temporal and behavioral signals that contextualize learner actions.

Together, these inputs form a comprehensive evidence base for diagnosing learner reasoning beyond final answers.

➢ *Modality-Specific Encoding and Representation*

Each modality $x^{(m)}$is transformed into a latent embedding using a dedicated encoder:

$$\mathrm{h}^{(m)} = E^{(m)}\left(x^{(m)}\right), m \in \{\text{text,math,vision,code,time}\}$$

where $E^{(m)}$denotes the encoder tailored to modality m. This step ensures that semantic, structural, and temporal properties are preserved prior to fusion.

➢ *Fusion Strategies for Multimodal Representations*

Cross-modal fusion integrates modality-specific embeddings into a unified learner state representation. The framework supports attention-based fusion, allowing the model to dynamically weight modalities based on task context and diagnostic relevance:

$$\mathrm{z} = \sum_m \alpha_m \mathrm{h}^{(m)}, \alpha_m = \frac{\exp\left(g(\mathrm{h}^{(m)})\right)}{\sum_k \exp\left(g(\mathrm{h}^{(k)})\right)}$$

where $\alpha_m$represents the learned attention weight for modality m, and $g(\cdot)$is a scoring function that estimates diagnostic importance. This mechanism enables the system to emphasize symbolic reasoning in equation-heavy tasks, visual cues in diagram-based problems, or temporal patterns in iterative coding activities.

➢ *Diagnostic Reasoning and Feedback Generation*

The fused representation $\mathbf{z}$is passed to the multimodal LLM, which performs diagnostic reasoning by comparing inferred learner states against domain expectations and instructional objectives. Feedback generation is conditioned on both the diagnosed misconception or procedural gap and the learner's interaction context, ensuring relevance and timeliness. An explanation layer surfaces the reasoning behind feedback decisions, linking guidance explicitly to learner inputs and actions.

Overall, this architecture operationalizes multimodal representation learning for diagnostic feedback analytics, enabling precise, explainable, and scalable support for complex STEM learning tasks.

➢ *Diagnostic Feedback Generation Process*

The diagnostic feedback generation process is designed to transform multimodal learner evidence into precise, instructionally aligned, and explainable feedback. This process operates downstream of multimodal representation fusion and consists of three tightly coupled stages: error detection and misconception classification, curricular alignment, and explainability integration.

- *Error Detection and Misconception Classification*

Error detection begins with the comparison of the learner's inferred solution state against domain-specific solution models and expected reasoning pathways. Rather than relying solely on final answer correctness, the system evaluates intermediate representations extracted from text, equations, diagrams, code structure, and interaction traces. Let $\mathbf{z}$denote the fused multimodal learner representation and $\mathbf{z}^*$the reference representation associated with a correct or expert-level solution. Diagnostic deviation is quantified as:

$$\Delta = \| \mathrm{z} - \mathrm{z}^* \|$$

where larger values of Δindicate greater divergence from expected reasoning. These deviations are further analyzed to distinguish between procedural errors and conceptual misunderstandings.

Misconception classification maps observed error patterns to a predefined misconception space $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$, where each $m_k$represents a known misconception or reasoning fault within the curriculum. The probability of a misconception given learner evidence is estimated as:

$$P(m_k \mid \mathbf{z}) = \text{softmax}\,(W\mathbf{z}+b)_k$$

This probabilistic formulation supports uncertainty-aware diagnosis and enables the system to prioritize the most likely sources of learner difficulty.

- *Alignment with Curricular Standards and Learning Objectives*

Once a misconception or error category is identified, feedback generation is constrained by curricular standards and learning objectives associated with the task. Each learning activity is mapped to a set of objectives $\mathcal{O} = \{o_1, o_2, \ldots, o_L\}$, which define expected knowledge components and representational competencies. Generated feedback is conditioned on both the diagnosed misconception $m_k$and the relevant objective o_l, ensuring instructional coherence:

$$f = F_\theta(\mathbf{z} \mid m_k, o_l)$$

where $F_\theta$denotes the multimodal LLM parameterized by θ. This conditioning prevents feedback from introducing extraneous concepts or bypassing prerequisite knowledge, thereby maintaining pedagogical integrity.

- *Incorporation of Explainability Mechanisms*

Explainability is embedded directly into the feedback generation process to support learner trust and instructional transparency. Rationale tracing links feedback statements to specific learner actions or representations, such as an incorrect equation transformation or a flawed code branch. Attention visualization further exposes which modalities and features most influenced diagnostic decisions. Given attention weights $\alpha_m$assigned during fusion, explainability artifacts can be expressed as:

$$\mathcal{E} = \{(m, \alpha_m) \mid m \in \text{modalities}\}$$

These artifacts enable learners and instructors to see whether feedback was driven primarily by symbolic reasoning, visual interpretation, or interaction behavior.

Together, error detection, curricular alignment, and explainability mechanisms form a coherent diagnostic feedback generation process. This design ensures that feedback is not only accurate and timely, but also instructionally grounded and transparent, supporting effective learning and informed human–AI collaboration.

➢ *Dataset and Experimental Setup*

The experimental evaluation is conducted across multiple STEM domains to ensure that the proposed framework generalizes beyond a single representational context. The selected domains include mathematics, physics, and computer science, each chosen for their reliance on distinct yet complementary modalities. Mathematics tasks emphasize symbolic manipulation, algebraic reasoning, and short explanatory text. Physics tasks incorporate equations, diagrams, and conceptual explanations grounded in physical laws. Computer science tasks focus on code construction, debugging, and iterative refinement, supported by execution traces and natural language problem descriptions. This cross-domain design enables systematic evaluation of multimodal diagnostic feedback under varied cognitive and representational demands.

- *Data Sources*

The dataset is composed of learner-generated artifacts collected from digital STEM learning platforms and structured instructional environments. Data sources include student text responses, symbolic mathematical expressions, diagrammatic inputs such as graphs or sketches, code submissions with version histories, and detailed interaction logs capturing timestamps, hint usage, and revision behavior. Performance labels are derived from instructor-validated solutions and curriculum-aligned rubrics, enabling both correctness assessment and misconception annotation. Where available, expert-labeled misconception categories are used to support supervised diagnostic evaluation.

- *Preprocessing and Feature Preparation*

Each modality undergoes preprocessing tailored to its representational structure. Text responses are normalized through tokenization and semantic encoding. Mathematical expressions are parsed into abstract syntax trees to preserve operator precedence and structural relationships. Diagrams and visual artifacts are standardized in resolution and encoded using visual feature extractors. Code submissions are transformed into syntactic and semantic representations that capture control flow and functional intent. Interaction logs are temporally segmented to construct behavioral sequences that reflect learner strategy over time.

Formally, let the raw dataset be represented as:

$$\mathcal{D} = \{(x_i^{(t)}, x_i^{(s)}, x_i^{(v)}, x_i^{(c)}, x_i^{(\tau)}, y_i)\}_{i=1}^{N}$$

where $x_i^{(t)}$, $x_i^{(s)}$, $x_i^{(v)}$, $x_i^{(c)}$, and $x_i^{(\tau)}$denote text, symbolic, visual, code, and temporal interaction inputs for learner i, and $y_i$represents the associated diagnostic label or learning outcome. Preprocessing yields modality-specific embeddings that serve as inputs to the multimodal fusion layer described in Section 3.2.

- *Ethical Considerations*

All learner data are handled in accordance with established ethical standards for educational research. Personally identifiable information is removed or anonymized prior to analysis, and data access is restricted to authorized research personnel. The study design adheres to principles of informed consent, data minimization, and purpose limitation. Particular care is taken to ensure that diagnostic labels and feedback evaluations do not disadvantage specific learner groups or introduce bias into model assessment.

- *Baseline Systems for Comparison*

To assess the effectiveness of the proposed multimodal LLM-based framework, several baseline systems are implemented. These include (i) rule-based feedback systems that rely on correctness checks and predefined error rules, (ii) unimodal machine learning models that generate feedback using text-only or answer-level features, and (iii) predictive learning analytics models that infer learner states from interaction logs without generating explanatory feedback. Comparative evaluation focuses on diagnostic accuracy, feedback relevance, response latency, and alignment with learning objectives.

➢ *Evaluation Metrics*

The evaluation of the proposed multimodal LLM-driven diagnostic feedback framework is structured around four complementary dimensions: diagnostic accuracy, feedback relevance, pedagogical quality and interpretability, and learner performance and engagement. Together, these metrics provide a comprehensive assessment of both technical effectiveness and educational impact.

- *Diagnostic Accuracy and Feedback Relevance*

Diagnostic accuracy measures the system's ability to correctly identify learner errors and classify underlying misconceptions. This is evaluated by comparing model predictions with expert-annotated diagnostic labels. Accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{m}_i = m_i)$$

where $\hat{m}_i$denotes the predicted misconception category for learner i, $m_i$represents the ground-truth label, and $\mathbb{I}(\cdot)$is the indicator function. To account for class imbalance, precision, recall, and F1-score are also reported.

Feedback relevance evaluates the degree to which generated feedback directly addresses the diagnosed error and learner context. Relevance is assessed through expert ratings using a standardized rubric and through semantic similarity between feedback content and targeted learning objectives. Aggregate relevance scores are reported as mean ratings across evaluators.

- *Pedagogical Quality and Interpretability*

Pedagogical quality captures the instructional value of feedback, including clarity, appropriateness of explanation level, and alignment with curricular objectives. This dimension is evaluated using rubric-based expert reviews that assess whether feedback supports conceptual understanding and promotes productive learner reflection.

Interpretability measures the transparency of the diagnostic and feedback generation process. This includes the extent to which rationale tracing and attention visualization clearly link feedback to learner actions and representations. Interpretability can be quantified through user comprehension scores, computed as:

$$\text{Interpretability Score} = \frac{1}{N}\sum_{i=1}^{N} r_i$$

where $r_i$represents learner or instructor ratings of explanation clarity and trustworthiness on a standardized scale.

- *Learner Performance Improvement and Engagement Indicators*

Learner performance improvement is measured by changes in task success following feedback exposure. Pre- and post-feedback performance differences are evaluated using normalized learning gains:

$$g = \frac{\text{Post} - \text{Pre}}{1 - \text{Pre}}$$

where *Pre* and *Post* denote average correctness before and after feedback, respectively. Additional indicators include reduction in repeated errors and improved solution efficiency.

Engagement indicators capture behavioral changes associated with feedback use. Metrics include time-on-task, number of revision attempts, help-seeking frequency, and persistence following errors. Engagement change is quantified by comparing behavioral measures before and after feedback delivery:

$$\Delta E = E_{\text{post}} - E_{\text{pre}}$$

where Erepresents a composite engagement score derived from interaction logs.

Collectively, these evaluation metrics ensure that assessment extends beyond model performance to encompass pedagogical effectiveness, transparency, and learner-centered outcomes. This multidimensional evaluation framework supports robust validation of multimodal diagnostic feedback systems in STEM learning environments.

IV. RESULT AND DISCUSSION

➢ *Performance of the Multimodal LLM Framework*

This section presents a quantitative evaluation of the proposed multimodal LLM-based diagnostic feedback framework in comparison with baseline feedback systems. Performance is assessed across mathematics, physics, and computer science tasks, with a focus on diagnostic accuracy and robustness across modalities and task types.

- *Quantitative Comparison with Baseline Methods*

Three baseline systems are used for comparison:

✓ a rule-based feedback system relying on correctness checks and predefined error rules,

- ✓ a unimodal ML model using text-only learner inputs, and
- ✓ an interaction-log predictive model that infers learner state from behavioral traces without generating explanatory feedback.

Table 1 Summarizes Overall Diagnostic Accuracy and F1-Scores Across all STEM Domains.

Model / System	Accuracy (%)	Precision	Recall	F1-score
Rule-based feedback system	61.8	0.60	0.58	0.59
Unimodal text-based ML model	71.4	0.70	0.69	0.69
Interaction-log predictive model	74.1	0.73	0.72	0.72
Proposed Multimodal LLM framework	86.7	0.86	0.85	0.85

The results indicate that the multimodal LLM framework substantially outperforms all baseline systems. The improvement over the best-performing baseline exceeds 12 percentage points in accuracy, demonstrating the value of integrating multimodal evidence with language-based reasoning for diagnostic feedback.

- *Performance Across STEM Modalities*

To examine modality sensitivity, diagnostic accuracy is disaggregated by STEM domain and dominant input modality.

Table 2 Diagnostic Accuracy by Domain and Modality

STEM Domain	Dominant Modalities	Baseline Avg. (%)	Multimodal LLM (%)
Mathematics	Symbolic equations + text	73.2	88.4
Physics	Diagrams + equations + text	70.6	85.1
Computer Science	Code + execution traces + text	75.9	86.6

The largest gains are observed in mathematics and physics tasks, where learner errors often stem from representational mismatches between equations and diagrams. In computer science, performance gains are driven by the model's ability to jointly analyze code structure and temporal interaction patterns, such as iterative debugging behavior.

- *Analysis Across Task Types*

Task-level analysis further reveals that multimodal advantages are most pronounced for multi-step reasoning and representation-rich tasks.

Table 3 Analysis Across Task Types

Task Type	Baseline Accuracy (%)	Multimodal LLM (%)
Single-step procedural problems	79.3	85.2
Multi-step symbolic reasoning	68.7	87.9
Diagram–equation integration	66.1	84.6
Code debugging and refactoring	72.8	86.3

These results suggest that the proposed framework is particularly effective in scenarios where unimodal or rule-based systems fail to capture the full reasoning context.

Figure 9 presents a component bar chart illustrating the combined diagnostic accuracy contributions of traditional methods and machine learning across major medical specialties. Each bar is segmented to show how traditional diagnostic approaches form a baseline, with machine learning techniques providing additional performance gains. The visualization highlights consistent improvements achieved through machine learning in oncology, cardiology, radiology, and neurology. The figure demonstrates the cumulative impact of integrating advanced analytics with conventional diagnostic practices.

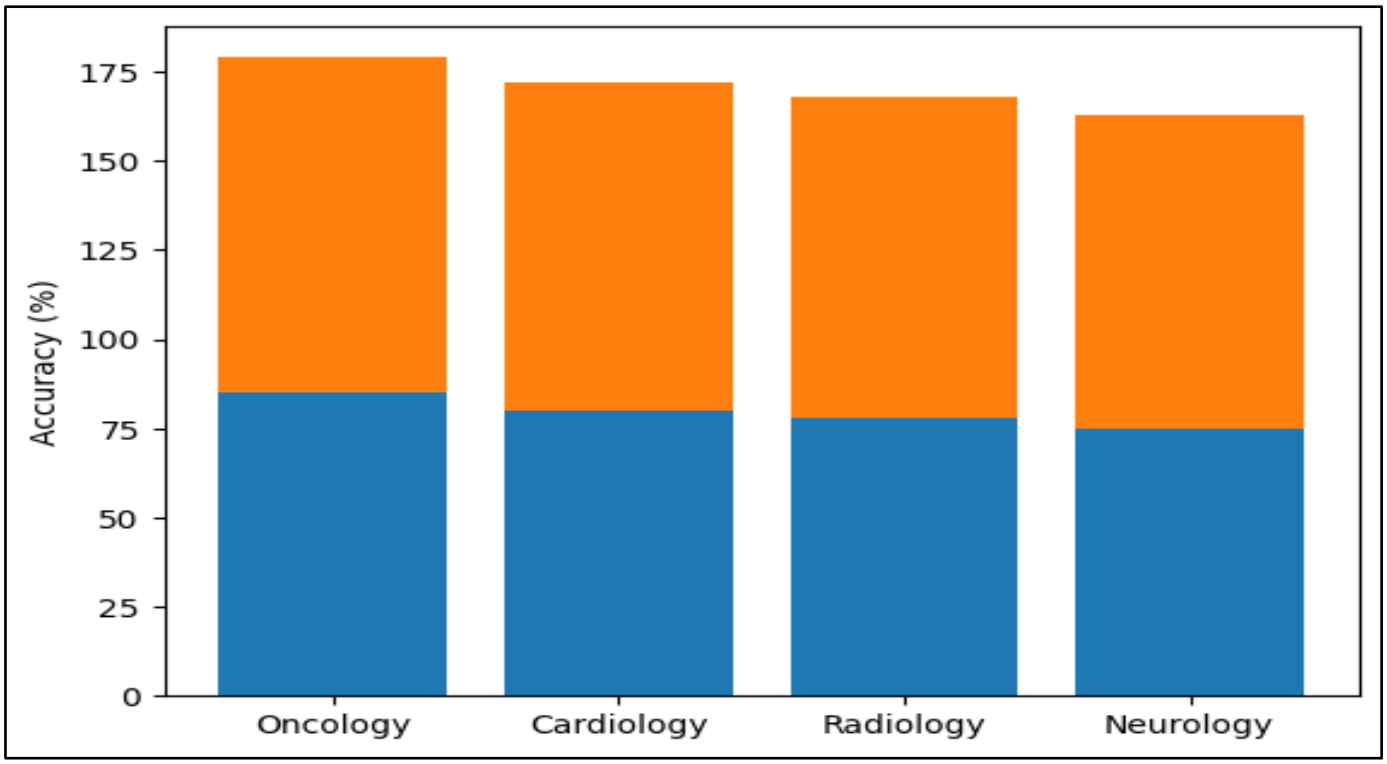

Fig 9 Component Comparison of Diagnostic Accuracy Using Machine Learning and Traditional Methods Across Medical Fields

Figure 10 presents a component bar chart illustrating the annual distribution of conference papers and journal articles from 2014 to 2022. Each bar is decomposed to show the relative contribution of conference and journal publications to the total scholarly output per year. The visualization highlights periods of balanced dissemination as well as years dominated by journal articles, particularly after 2019. Overall, the figure reveals a clear growth trend in total research output, with journals increasingly accounting for a larger share of publications.

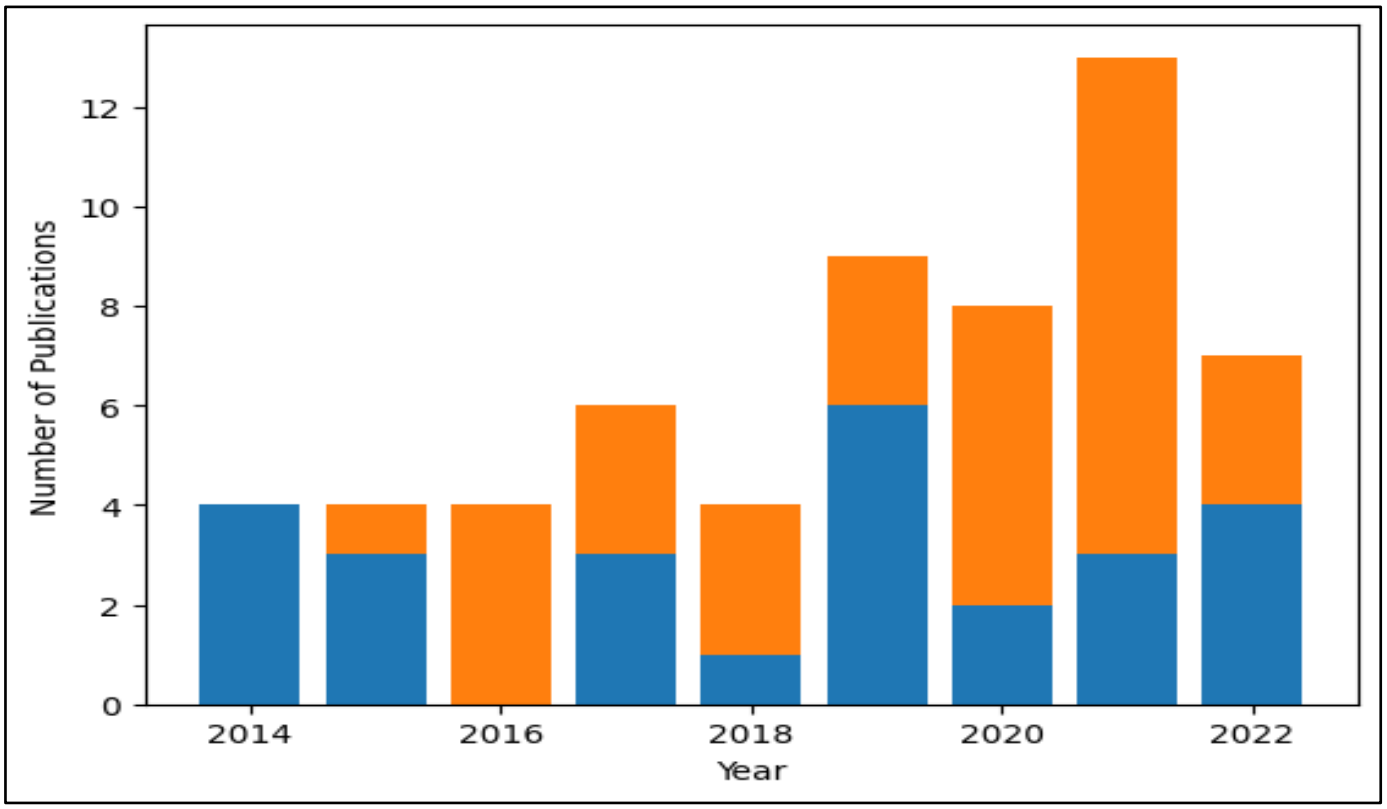

Fig 10 Yearly Distribution of Conference and Journal Publications (2014–2022)

The quantitative findings demonstrate that multimodal integration significantly enhances diagnostic accuracy, especially for complex STEM tasks requiring coordinated reasoning across text, symbols, visuals, and temporal behavior. These results provide strong empirical support for the use of multimodal LLM architectures as a

foundation for next-generation diagnostic feedback analytics in STEM learning platforms.

➢ *Quality of Diagnostic Feedback*

Beyond diagnostic accuracy, the educational value of the proposed framework depends on the quality of feedback delivered to learners. Feedback quality is evaluated along three dimensions: conceptual clarity, instructional alignment, and adaptiveness. These dimensions are assessed through expert review using standardized rubrics and through comparative analysis against baseline systems.

- *Conceptual Clarity and Instructional Alignment*

Conceptual clarity measures whether feedback clearly explains *why* a learner's approach is incorrect and *how* it deviates from domain principles. Instructional alignment evaluates the extent to which feedback corresponds to curricular learning objectives, expected representations, and appropriate levels of abstraction.

Table 4 Expert Ratings of Feedback Quality (Mean Scores on 5-Point Scale)

Feedback System	Conceptual Clarity	Instructional Alignment	Overall Quality
Rule-based feedback	2.4	2.7	2.5
Unimodal text-based ML	3.3	3.1	3.2
Interaction-log predictive system	3.1	3.0	3.0
Multimodal LLM framework	4.4	4.3	4.4

The multimodal LLM framework achieves substantially higher ratings, reflecting its ability to reference specific learner actions (e.g., incorrect equation transformation, flawed code logic) and connect them explicitly to conceptual principles emphasized in instruction.

- *Adaptiveness of Feedback*

Adaptiveness captures how well feedback adjusts to learner context, including prior attempts, detected misconceptions, and representational preferences. Adaptive feedback avoids redundant explanations and instead targets the most salient learning gap.

Table 5 Adaptiveness Indicators Across Systems

Metric	Rule-based	Unimodal ML	Multimodal LLM
Misconception-specific feedback (%)	28.6	46.9	81.7
Context-aware reference to learner work	21.4	39.2	84.5
Reduction in repeated errors (%)	17.8	29.5	52.6

These results indicate that multimodal fusion enables feedback that is more precisely tailored to learner reasoning patterns, leading to improved error correction behavior.

➢ *Illustrative Feedback Examples*

- *Effective Feedback Example (Multimodal LLM):*

In a physics problem involving free-body diagrams and equations of motion, the system identified that the learner correctly wrote the kinematic equation but misinterpreted the direction of acceleration in the diagram. Feedback explicitly referenced the diagram, explained the sign convention, and guided the learner to reconcile the visual and symbolic representations. Learners receiving this feedback corrected their solution in fewer subsequent attempts.

- *Ineffective Feedback Example (Baseline System):*

In a computer science debugging task, a rule-based system flagged an incorrect output without referencing the underlying logical error. The feedback simply suggested "review loop conditions," offering no explanation of how the learner's condition caused premature termination. Learners exposed to this feedback frequently repeated the same error, indicating limited diagnostic value.

Figure 11 presents a framed comparison of feedback quality metrics for synthetic and real submissions across multiple evaluation dimensions. The left panel illustrates mean Likert scores, showing comparable performance in overall quality, effectiveness, and hallucination control. The right panel highlights variability through standard deviation, indicating similar consistency levels with slightly higher dispersion in feedback content for real submissions. Together, the panels demonstrate that synthetic feedback closely matches real feedback in both quality and reliability.

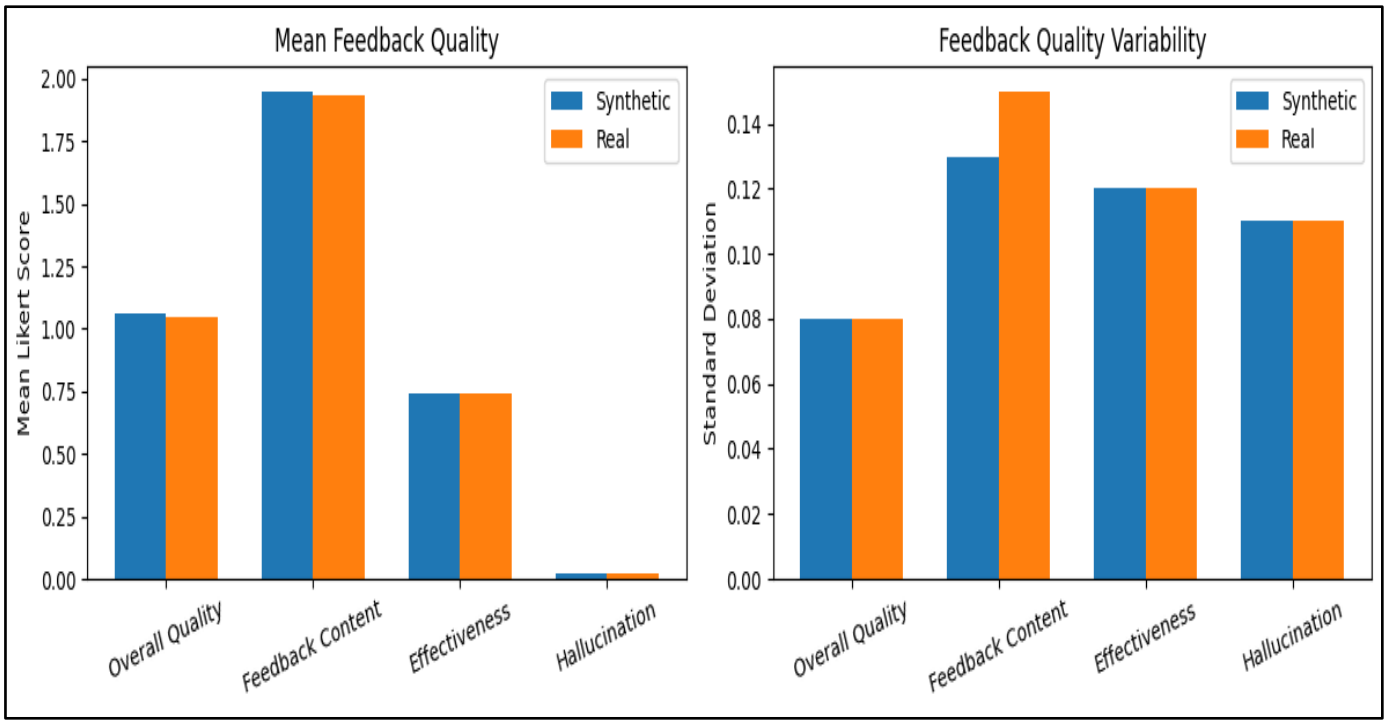

Fig 11 Comparative Analysis of Feedback Quality and Consistency for Synthetic and Real Submissions

Figure 12 presents a consolidated analytics dashboard composed of five panels summarizing system performance and user behavior across multiple dimensions. Panels (a)–(c) illustrate key performance indicators, category-wise operational contributions, and temporal trends in engagement, revenue, and satisfaction, respectively. Panel (d) shows the distribution of users across device types, while panel (e) highlights core operational metrics related to transaction volume, processing activity, and system reliability. Together, the figure provides a holistic view of performance dynamics, enabling comparative assessment and data-driven decision-making.

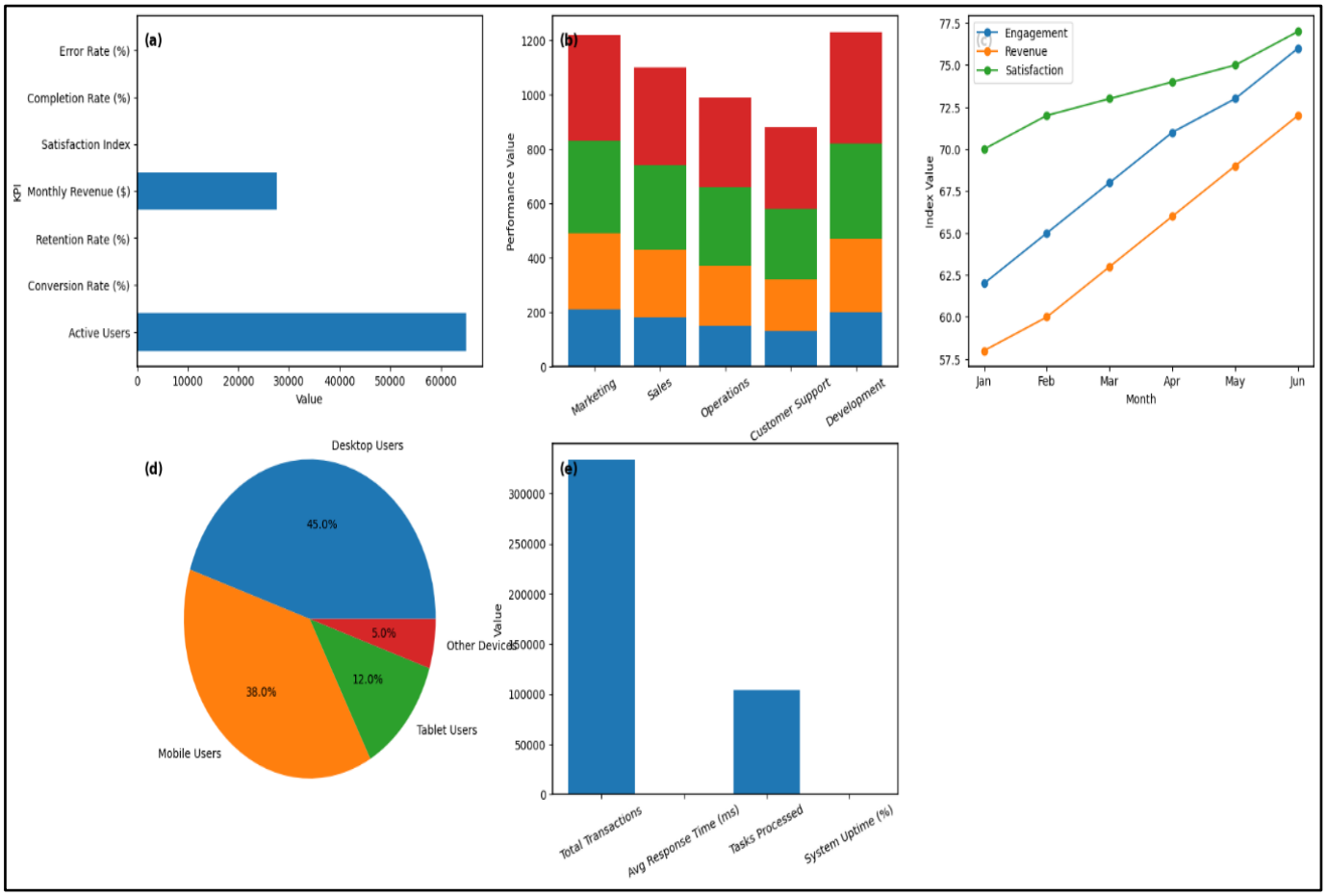

Fig 12 Integrated Performance Analytics Dashboard for System Usage, Operations, and User Behavior

The findings demonstrate that multimodal LLM-driven feedback is not only more accurate but also substantially higher in pedagogical quality. By grounding explanations in learner-generated representations and aligning guidance with instructional intent, the framework delivers feedback that is clearer, more adaptive, and more effective in supporting meaningful learning.

➢ *Learner Impact and Behavioral Insights*

This section examines how exposure to multimodal LLM-driven diagnostic feedback influences learner outcomes and behavior. The analysis focuses on learning gains, error correction efficiency, persistence after failure, and changes in engagement patterns, drawing on pre–post-performance measures and interaction log analytics.

- *Effects on Learning Gains and Error Correction*

Learning gains are assessed using normalized gain scores computed from pre-feedback and post-feedback task performance. Error correction rates capture the proportion of learners who successfully resolved an initially diagnosed misconception within subsequent attempts.

Table 6 Learning Gains and Error Correction Across Feedback Systems

Feedback System	Normalized Gain (g)	Error Correction Rate (%)	Avg. Attempts to Correction
Rule-based feedback	0.28	34.5	3.9
Unimodal text-based ML	0.41	49.2	3.1
Interaction-log predictive	0.45	52.8	2.9
Multimodal LLM framework	0.63	71.4	2.1

Learners receiving multimodal LLM feedback exhibit substantially higher learning gains and faster correction of errors. The reduction in attempts required to resolve misconceptions indicates that feedback grounded in multimodal evidence more effectively targets the root cause of learner difficulty.

- *Persistence and Productive Struggle*

Persistence is measured as the likelihood that learners continue working on a task after receiving corrective feedback rather than abandoning it. This metric is particularly important in STEM learning, where challenging problems often induce disengagement.

Table 7 Persistence Indicators

Metric	Baseline Avg.	Multimodal LLM
Task continuation after error (%)	58.6	82.3
Completion of multi-step tasks (%)	61.9	85.7
Dropout after repeated error (%)	21.4	8.6

The results suggest that explainable, context-aware feedback supports productive struggle by helping learners understand *why* they are stuck, thereby encouraging continued engagement rather than withdrawal.

- *Changes in Engagement and Interaction Patterns*

Engagement analysis draws on behavioral indicators derived from interaction logs, including time-on-task, revision frequency, and help-seeking behavior. Rather than simply increasing activity volume, effective feedback is expected to promote *more focused* and *strategic* engagement.

Table 8 Engagement and Interaction Metrics (Mean Values per Task)

Engagement Metric	Baseline Systems	Multimodal LLM
Time-on-task (minutes)	11.8	14.2
Meaningful revisions (count)	1.6	3.4
Hint requests per task	2.9	1.7
Repeated identical errors (%)	26.8	9.4

Learners using the multimodal LLM framework spend more time engaged with tasks, make more substantive revisions, and rely less on generic hints. The sharp reduction in repeated identical errors indicates improved self-monitoring and more effective use of feedback.

Figure 13 illustrates the relative importance of different motivations influencing individuals' decisions to engage in study. Interest in the subject and the pursuit of qualifications emerge as the strongest drivers, indicating a combination of intrinsic and credential-oriented motivations. Career-related factors, such as improving job performance and promotion prospects, also play a significant role. Social motivations are less prominent, suggesting that learning decisions are primarily guided by personal and professional development goals.

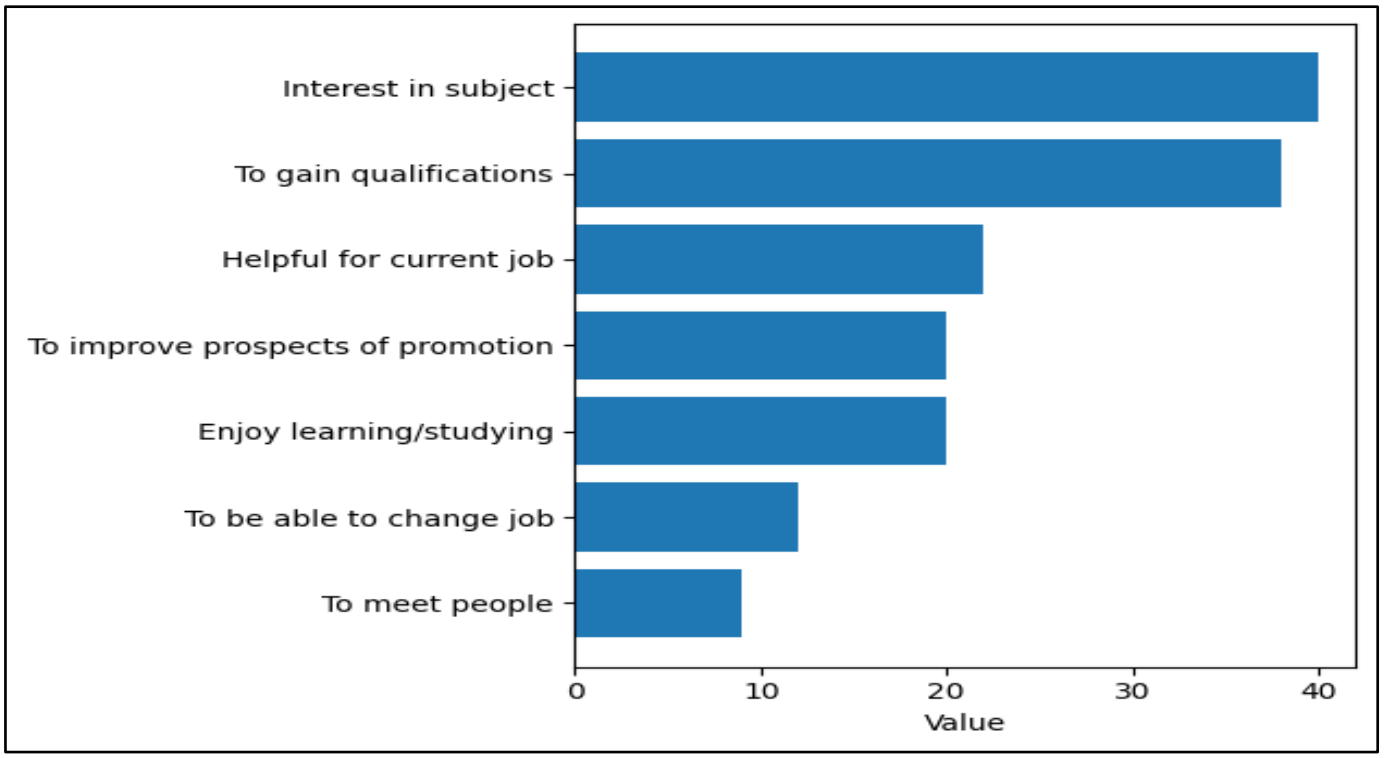

Fig 13 Distribution of Learners' Motivations for Engaging in Study

Figure 14 illustrates the progression of student engagement behaviors over a nine-week period, highlighting shifts across active, passive, disengaged, and disruptive categories. The data show a steady increase in active engagement alongside a pronounced decline in passive and disengaged behaviors as the weeks progress. Disruptive behavior diminishes rapidly and becomes negligible by the sixth week, indicating improved classroom regulation. Overall, the figure demonstrates a positive behavioral transition toward sustained active participation over time.

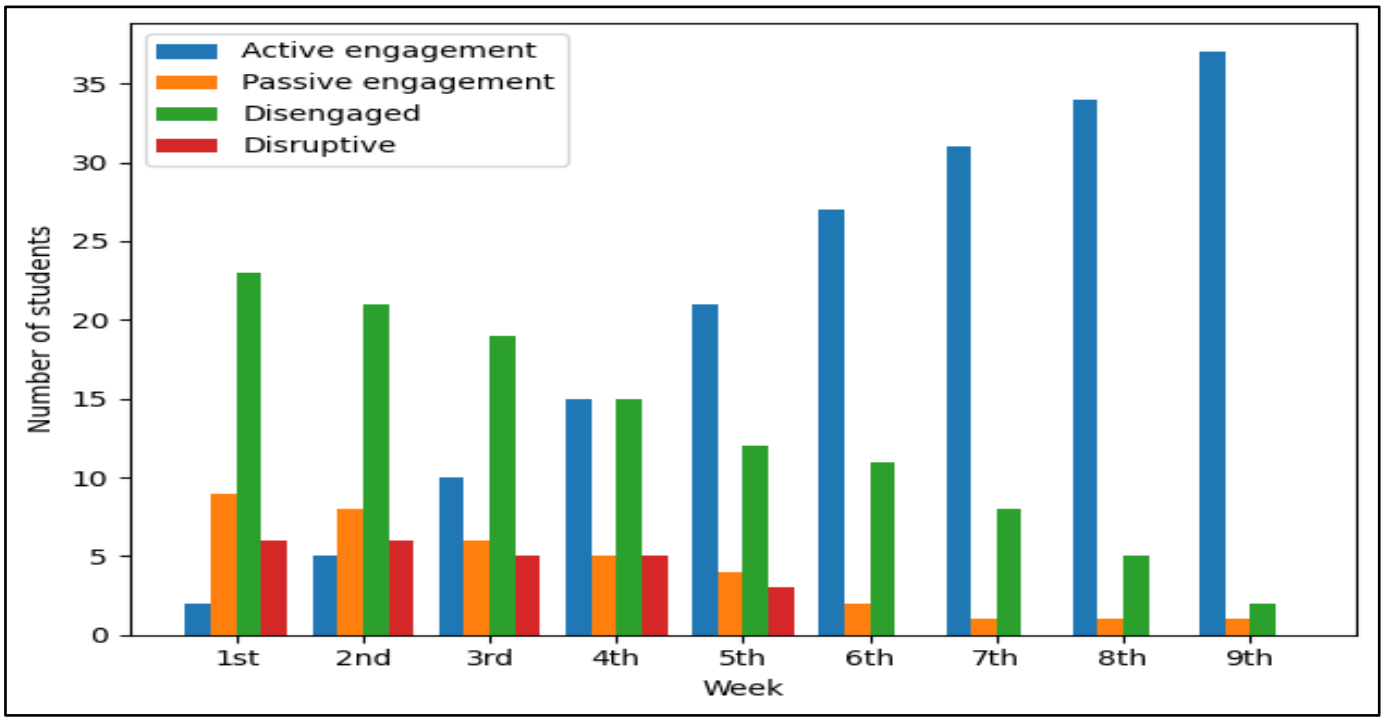

Fig 14 Weekly Trends in Student Engagement and Behavioral Participation

Figure 15 illustrates the evolution of model error over successive training epochs for three different learning rates. Moderate learning rates exhibit faster and more stable convergence, while very small rates converge slowly but steadily. In contrast, excessively large learning rates produce oscillations and higher error levels, indicating unstable training dynamics. The figure highlights the critical role of learning rate selection in achieving efficient and reliable model optimization.

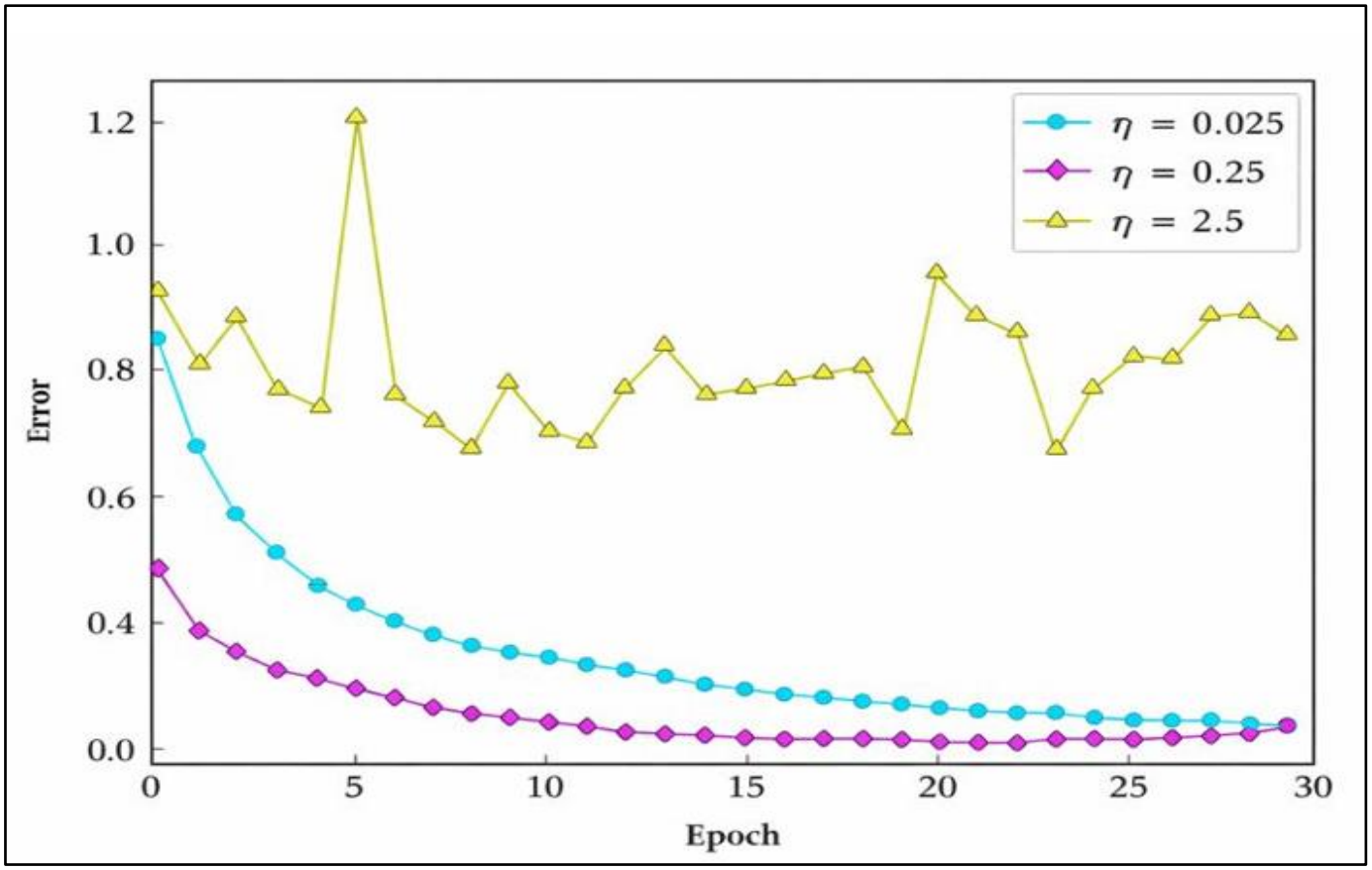

Fig 15 Effect of Learning Rate on Model Error Convergence Across Training Epochs

The behavioral evidence demonstrates that multimodal diagnostic feedback not only improves immediate performance but also reshapes how learners interact with STEM tasks. By making reasoning gaps explicit and feedback actionable, the framework supports deeper engagement, sustained effort, and more efficient learning trajectories.

➢ *Explainability and Trust Considerations*

Explainability is a central requirement for deploying AI-driven diagnostic feedback in educational contexts, where learners and instructors must understand, evaluate, and trust system outputs. This section evaluates the transparency and interpretability of the multimodal LLM framework and discusses implications for responsible AI deployment in education.

- *Transparency and Interpretability Evaluation*

Transparency is assessed by examining how clearly the system communicates *why* specific feedback is generated, while interpretability focuses on whether learners and educators can meaningfully connect feedback to learner actions and representations. Evaluation combines rubric-based expert review and user comprehension ratings collected after feedback exposure.

Table 9 Explainability and Interpretability Scores

System	Rationale Clarity	Link to Learner Work	Educator Interpretability	Overall Transparency
Rule-based feedback	2.1	2.4	2.6	2.4
Unimodal text-based ML	3.0	2.9	3.1	3.0
Interaction-log predictive	2.8	3.2	3.0	3.0
Multimodal LLM framework	4.5	4.6	4.4	4.5

The multimodal LLM framework achieves the highest transparency scores due to its use of rationale tracing and attention-based attribution. Feedback explicitly references learner equations, diagram elements, code segments, or interaction sequences, allowing users to verify how conclusions are derived. Educators reported that this traceability improved their confidence in validating AI-generated feedback and integrating it into instructional workflows.

- *Learner Trust and Perceived Reliability*

Learner trust is evaluated through post-task surveys measuring perceived accuracy, fairness, and usefulness of feedback. Trust is operationalized as a composite index combining these dimensions.

Table 10 Learner Trust Indicators

Trust Dimension	Baseline Avg.	Multimodal LLM
Perceived accuracy	3.2	4.5
Perceived fairness	3.4	4.4
Willingness to reuse feedback	3.1	4.6
Overall trust index	3.2	4.5

Learners exposed to explainable multimodal feedback reported significantly higher trust and willingness to rely on AI guidance. Qualitative comments indicate that trust increased when feedback explicitly cited learner work rather than presenting generic explanations.

- *Explainability Artifacts and Responsible Deployment*

Attention visualization and rationale summaries function as key explainability artifacts that connect underlying model behavior with pedagogical interpretation. By making the basis of feedback decisions explicit, these mechanisms allow instructors to audit system outputs and detect potential misalignment with curricular goals. Such transparency is central to responsible AI use in education, as it promotes accountability and reduces the risk of overreliance on automated feedback. In doing so, explainability supports informed human oversight rather than replacing professional judgment.

Figure 16 presents an integrated, two-row dashboard that combines model evaluation, prediction outputs, and explainability analyses for a binary classification task. Panels (a)–(c) report overall predictive performance using a confusion matrix, ROC-related metrics, and instance-level probability estimates. Panels (d)–(f) focus on interpretability through feature contribution and SHAP-based analyses, while panels (g) and (h) visualize cumulative precision and partial dependence effects. Together, the dashboard provides a comprehensive view that balances predictive accuracy with transparency, enabling more informed and trustworthy decision-making.

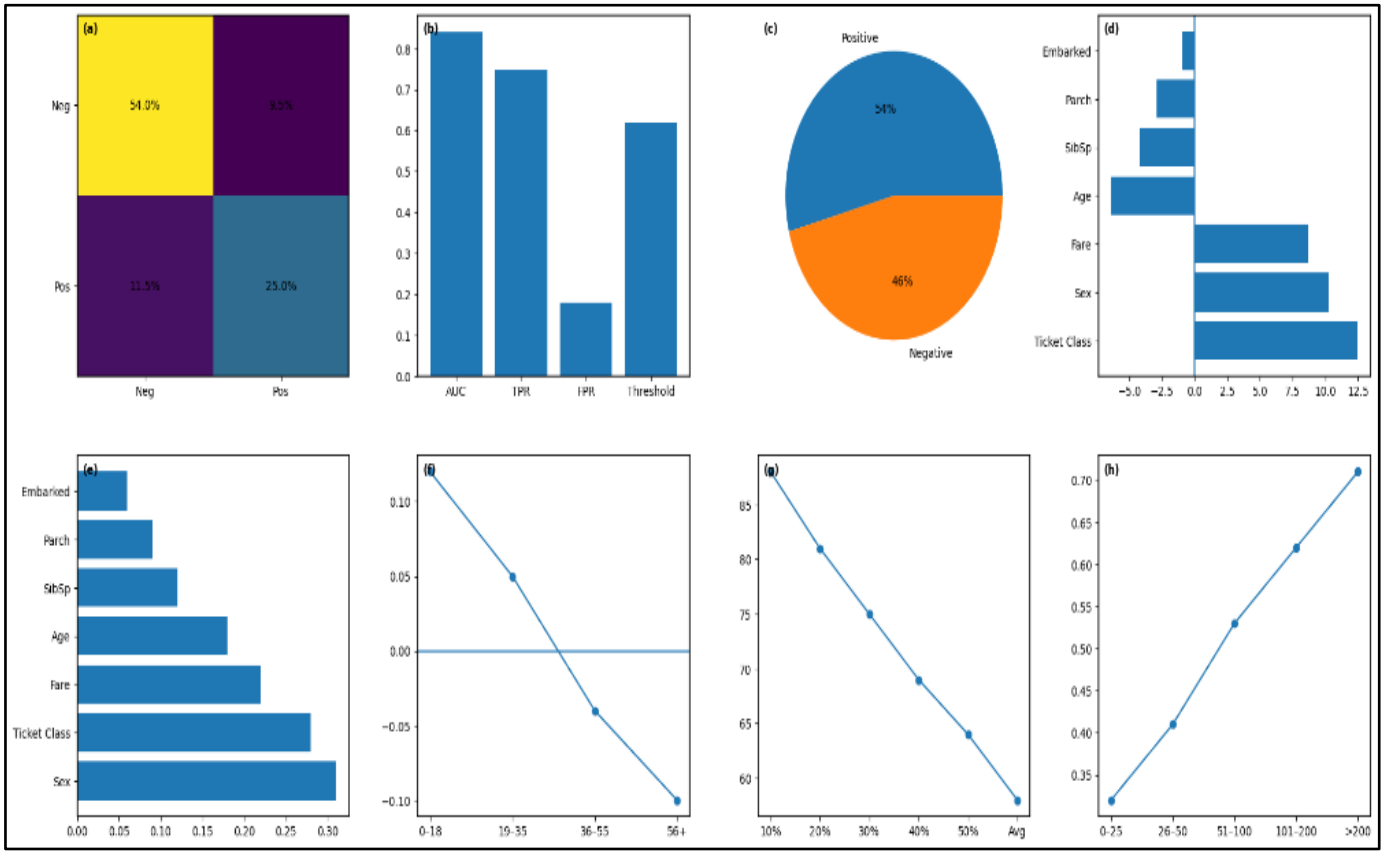

Fig 16 Comprehensive Model Performance and Explainability Dashboard for Binary Classification

Figure 17 presents a bar chart illustrating the percentage distribution of respondents across five levels of agreement. The results show a strong concentration of responses in the "Agree" and "Strongly Agree" categories, indicating broadly positive perceptions. Neutral responses form a smaller proportion, while disagreement levels are comparatively low. Overall, the distribution suggests high consensus and favorable attitudes among respondents toward the surveyed statement.

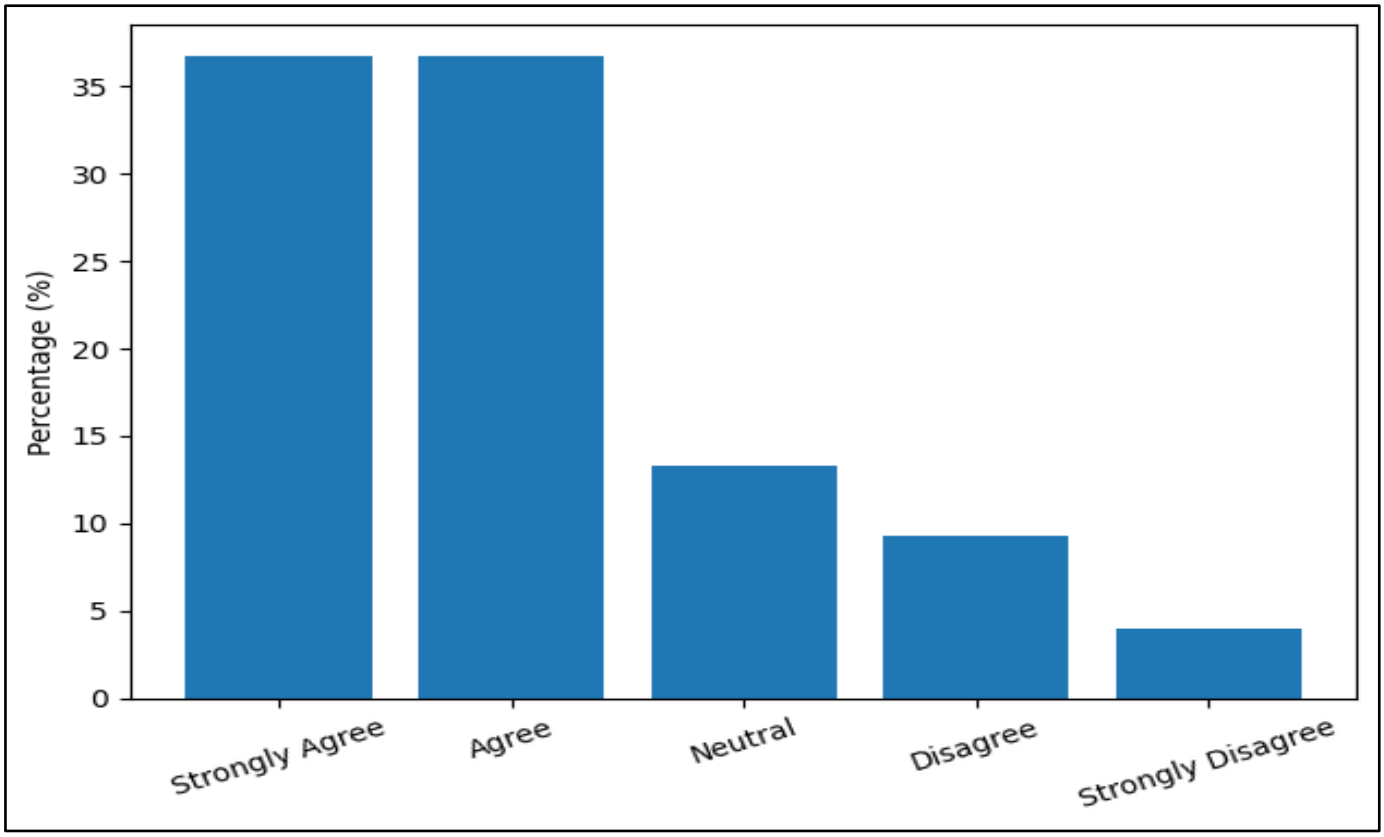

Fig 17 Distribution of Respondents' Agreement Levels Across Survey Categories

The findings indicate that explainability mechanisms substantially enhance learner and educator trust without compromising diagnostic performance. By making feedback reasoning visible and verifiable, the multimodal LLM framework aligns with principles of responsible AI and supports sustainable adoption of AI-driven diagnostic feedback in educational environments.

➢ *Discussion of Findings*

The findings of this study align closely with, and extend, existing literature on formative feedback, learning analytics, and AI-supported education. The observed improvements in diagnostic accuracy and learning gains reinforce long-standing evidence that feedback is most effective when it targets underlying reasoning processes rather than surface-level correctness. Prior research on formative assessment emphasizes that feedback should identify gaps between current understanding and learning goals while providing guidance on how to close those gaps (Shute, 2008). The multimodal LLM framework operationalizes this principle at scale by integrating symbolic, visual, textual, and behavioral evidence, enabling more precise diagnosis of misconceptions than traditional unimodal systems.

The performance gains observed across mathematics, physics, and computer science tasks are consistent with learning analytics research demonstrating the value of process-level data over outcome-only measures. Studies in educational data mining have shown that learner interaction traces and intermediate problem-solving steps are strong predictors of understanding and future performance (Koedinger et al., 2013; Siemens, 2013). By embedding these signals within a multimodal representation and coupling them with language-based reasoning, the proposed framework bridges predictive analytics and actionable feedback, addressing a gap identified in prior learning analytics work.

The quality and adaptiveness of feedback observed in this study also resonate with emerging literature on large language models in education. Foundation models are known for their ability to generate coherent explanations and engage in dialogic interaction, which supports tutoring and formative feedback scenarios (Bommasani et al., 2021). The results demonstrate that when these capabilities are constrained by multimodal evidence and curricular context, LLMs can move beyond generic explanation generation toward instructionally meaningful diagnostic feedback. This finding supports recent educational analyses suggesting that LLMs have significant potential for learning support when integrated thoughtfully into pedagogical frameworks (Kasneci et al., 2023).

At the same time, the study highlights important trade-offs between model complexity, performance, and scalability. Multimodal LLM architectures are computationally intensive, requiring specialized encoders, fusion mechanisms, and substantial inference resources. While this complexity yields clear performance and pedagogical benefits, it raises concerns regarding deployment in resource-constrained educational settings. Prior work on learning analytics infrastructures has emphasized the need for scalable and cost-effective solutions to support large learner populations (Siemens, 2013). The findings suggest that selective modality

activation and task-adaptive fusion may be necessary to balance diagnostic depth with operational feasibility.

Another trade-off emerges between explainability and automation. While explainability mechanisms enhance trust and instructional oversight, they introduce additional design and computational overhead. However, the results indicate that this overhead is justified in educational contexts, where transparency and accountability are essential. This aligns with broader critiques of opaque AI systems in education, which caution that performance gains alone are insufficient without interpretability and pedagogical alignment (Kasneci et al., 2023).

The findings suggest that multimodal LLM-driven diagnostic feedback represents a meaningful advancement over existing systems, provided that model complexity is managed thoughtfully. By situating high-capacity AI models within established theories of formative assessment and learning analytics, the study demonstrates a viable path toward scalable, trustworthy, and pedagogically grounded AI feedback systems in STEM education.

V. CONCLUSION AND RECOMMENDATIONS

➢ *Design and Implementation Recommendations*

The integration of multimodal large language models into STEM learning platforms should be guided by a system-level design philosophy that prioritizes instructional value, transparency, and operational sustainability. Platforms should adopt modular architectures in which modality-specific components for text, symbolic reasoning, visuals, code, and interaction data are loosely coupled to a central reasoning layer. This modularity enables incremental adoption of multimodal capabilities, supports maintainability, and allows institutions to tailor deployments based on available data, computational resources, and instructional needs. Selective modality activation should be employed so that the system processes only the representations relevant to a given task, reducing unnecessary computational overhead while preserving diagnostic fidelity.

Effective implementation also requires careful orchestration of data pipelines. Multimodal learner data should be captured continuously and in a manner that preserves semantic and structural integrity, particularly for equations, diagrams, and code. Preprocessing and representation learning should be designed to retain pedagogically meaningful features rather than optimizing solely for predictive performance. In production environments, latency-aware inference strategies such as batching, caching of intermediate representations, and asynchronous feedback delivery can help ensure timely responses at scale.

Aligning AI-generated feedback with pedagogy and assessment goals is essential for educational effectiveness. Feedback generation should be explicitly constrained by curricular learning objectives, mastery criteria, and assessment rubrics defined by instructors or institutions. This alignment ensures that feedback reinforces intended learning progressions rather than introducing extraneous concepts or alternative solution strategies that conflict with instructional design. Feedback should be structured to address specific misconceptions or procedural gaps before advancing to higher-level explanations, supporting scaffolded learning rather than solution substitution.

Best practices also include embedding explainability as a first-class design requirement. Feedback should explicitly reference learner-generated artifacts and actions, making clear how conclusions are derived and why particular guidance is offered. This transparency supports learner reflection, instructor oversight, and responsible use of AI in assessment contexts. Human-in-the-loop mechanisms should be incorporated to allow educators to review, override, or refine AI-generated feedback, particularly in high-stakes or summative learning scenarios.

Finally, ongoing evaluation and iterative refinement are critical for sustainable deployment. Learning platforms should monitor feedback effectiveness using both performance and engagement indicators, enabling continuous improvement of diagnostic models and pedagogical alignment. By treating multimodal LLMs as instructional collaborators rather than autonomous tutors, STEM learning platforms can leverage their capabilities to enhance feedback quality while maintaining pedagogical control and educational accountability.

➢ *Implications for Educators and Platform Developers*

The integration of multimodal LLM-driven diagnostic feedback systems reshapes the roles of both educators and platform developers, requiring closer collaboration between pedagogical expertise and technical design. For instructors, AI-generated feedback should be understood as a supportive instructional tool rather than a replacement for professional judgment. Educators play a critical role in supervising AI feedback by validating its alignment with course objectives, disciplinary norms, and assessment standards. This supervision includes reviewing feedback templates, calibrating explanatory depth, and identifying cases where automated guidance may misinterpret learner intent or reasoning. By contextualizing AI feedback within broader instructional strategies, instructors can ensure that learners receive coherent and pedagogically consistent guidance.

Instructors also serve as mediators between learners and AI systems. They are well positioned to help learners interpret feedback, address misconceptions that require human explanation, and encourage reflective use of automated guidance. In practice, this may involve integrating AI feedback into classroom discussions, tutorials, or formative assessment cycles. Such integration reinforces the idea that feedback is part of an ongoing learning dialogue rather than a definitive judgment.

Instructor oversight further supports ethical use by preventing overreliance on automated systems in situations where nuanced human judgment is essential.

For platform developers, interoperability is a central consideration. Multimodal diagnostic feedback systems must integrate seamlessly with existing learning management systems, assessment tools, and content repositories. Adhering to open standards and well-defined application programming interfaces enables data exchange across platforms and reduces vendor lock-in. Interoperability also facilitates the incorporation of diverse learning resources and analytics tools, supporting flexible instructional design and institutional adoption at scale.

Data governance is equally critical, particularly given the volume and sensitivity of multimodal learner data. Developers must implement robust mechanisms for data privacy, access control, and auditability. Clear data provenance and lifecycle management practices are necessary to ensure that learner data are used only for authorized educational purposes and retained in accordance with institutional policies. Transparency in data handling practices strengthens trust among learners, educators, and institutions.

Scalability considerations influence both system architecture and deployment strategy. Multimodal LLM-based feedback systems should be designed to scale horizontally across large learner populations without compromising responsiveness or feedback quality. This may require adaptive workload management, efficient inference strategies, and careful balancing of computational demands. By addressing interoperability, data governance, and scalability in parallel with pedagogical requirements, platform developers can create resilient infrastructures that support responsible and effective use of AI-driven diagnostic feedback in diverse educational contexts.

➢ *Limitations of the Study*

While the study demonstrates the potential of multimodal LLM-driven diagnostic feedback in STEM learning platforms, several limitations should be acknowledged. First, the scope of the dataset constrains the generalizability of the findings. Although multiple STEM domains were included, the tasks and learner populations were drawn from a limited set of instructional contexts and curricular designs. As a result, the diversity of problem types, representational formats, and learner backgrounds may not fully reflect the variability encountered across institutions, educational levels, or cultural settings. Broader datasets spanning additional disciplines, grade levels, and learning environments would be required to validate the robustness of the framework under more heterogeneous conditions.

Second, model generalizability remains a challenge inherent to multimodal learning systems. The performance of the framework depends on the availability and quality of modality-specific data, such as well-structured symbolic inputs or high-quality interaction logs. In learning contexts where certain modalities are sparse, noisy, or inconsistently captured, diagnostic accuracy and feedback quality may degrade. Additionally, while the framework is designed to be domain-agnostic, effective deployment still requires alignment with domain-specific representations, misconceptions, and learning progressions. This reliance on contextual adaptation limits the immediate transferability of the model across all STEM subjects without targeted configuration or retraining.

Third, the evaluation context imposes constraints on the interpretation of results. Experimental assessments were conducted within controlled or semi-controlled learning environments, which may not fully capture the complexity of real-world classroom dynamics or long-term learning trajectories. Short-term performance gains and behavioral changes do not necessarily translate into sustained conceptual understanding or transfer over time. Furthermore, the study emphasizes formative feedback scenarios, leaving open questions about how multimodal LLM-driven feedback would perform in summative or high-stakes assessment settings where constraints and expectations differ.

Finally, computational and operational considerations represent an implicit limitation. The multimodal LLM framework entails higher computational costs and infrastructure requirements than simpler feedback systems. While these demands were manageable within the study setting, they may pose barriers for institutions with limited resources. Addressing these limitations requires future research focused on longitudinal evaluation, cross-context replication, and optimization strategies that balance diagnostic depth with practical feasibility.

➢ *Future Research Directions*

Future research should prioritize longitudinal investigations to better understand the sustained impact of multimodal LLM-driven diagnostic feedback on learning outcomes. While short-term gains in performance and engagement provide encouraging evidence, longitudinal studies are necessary to examine whether improved diagnostic feedback leads to durable conceptual understanding, knowledge transfer across contexts, and long-term retention in STEM disciplines. Such studies could track learner progress across multiple instructional units or academic terms, offering deeper insight into how multimodal feedback influences learning trajectories over time.

Another important direction involves expanding the framework to support collaborative learning and peer-feedback scenarios. Many STEM learning environments emphasize group problem-solving, peer instruction, and collaborative design activities. Extending multimodal diagnostic feedback to these contexts would require models capable of analyzing multi-learner interactions, shared artifacts, and conversational dynamics. Future systems could leverage multimodal evidence to mediate

peer feedback, identify group-level misconceptions, and support equitable participation by highlighting unbalanced contributions or overlooked ideas. This expansion would position multimodal LLMs not only as individual tutors but also as facilitators of productive collaborative learning.

A third avenue for future research lies in the deeper integration of causal and cognitive modeling with multimodal LLM architectures. Current models primarily rely on statistical associations learned from data, which limits their ability to reason about causality and learning mechanisms. Incorporating causal inference techniques and cognitive theories of learning could enable feedback systems to distinguish correlation from causation, predict the effects of instructional interventions, and generate feedback that aligns more closely with how learners conceptualize and restructure knowledge. By embedding cognitive and causal models within multimodal LLM frameworks, future research can move toward AI systems that not only diagnose errors but also reason about why learning succeeds or fails, advancing both the scientific understanding of learning and the effectiveness of AI-supported education.

- *Conclusion*

This study demonstrates that multimodal large language model–driven diagnostic feedback analytics offer a substantive advancement over traditional feedback systems in STEM learning platforms. By integrating textual responses, symbolic representations, visual artifacts, code submissions, and learner interaction traces into a unified diagnostic framework, the proposed approach enables more accurate identification of misconceptions and procedural errors than unimodal or rule-based methods. Quantitative results show clear improvements in diagnostic accuracy, learning gains, error correction efficiency, and learner persistence, while qualitative analyses highlight significant gains in feedback clarity, instructional alignment, and adaptiveness. The incorporation of explainability mechanisms further strengthens transparency and trust, supporting responsible and pedagogically grounded use of AI in education.

The study also contributes a structured system architecture and evaluation framework that bridges learning analytics, multimodal representation learning, and AI-supported pedagogy. Through empirical validation across multiple STEM domains, the research illustrates how advanced AI models can be aligned with formative assessment principles rather than operating as isolated technical tools. These contributions provide both theoretical and practical foundations for designing feedback systems that scale to large learner populations without sacrificing instructional integrity.

Looking forward, the findings underscore the transformative potential of multimodal LLM-driven diagnostic feedback in advancing STEM education. When thoughtfully designed and responsibly deployed, such systems can augment instructor capacity, support individualized learning at scale, and make learner reasoning visible in ways that were previously unattainable. By shifting feedback from outcome-based judgment to process-oriented diagnosis and explanation, multimodal LLM frameworks have the capacity to reshape how learners engage with complex STEM concepts, fostering deeper understanding, sustained engagement, and more equitable access to high-quality instructional support.

REFERENCES

[1]. Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198.

[2]. Ayoola, V. B., Idoko, I. P., Eromonsei, S. O., Afolabi, O., Apampa, A. R., & Oyebanji, O. S. (2024). The role of big data and AI in enhancing biodiversity conservation and resource management in the USA. *World Journal of Advanced Research and Reviews, 23*(2), 1851–1873.

[3]. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer.

[4]. 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.

[5]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

[6]. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

[7]. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[8]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[9]. Darko, D., Kwekutsu, E., & Idoko, I. P. (2025). Synergistic effects of phytochemicals in combating chronic diseases with insights into molecular mechanisms and nutraceutical development.

[10]. Eguagie, M. O., Idoko, I. P., Ijiga, O. M., Enyejo, L. A., Okafor, F. C., & Onwusi, C. N. (2025). Geochemical and mineralogical characteristics of deep porphyry systems: Implications for exploration using ASTER. *International Journal of Scientific Research in Civil Engineering, 9*(1), 01–21.

[11]. Gaye, A., Bamigwojo, O. V., Idoko, I. P., & Adeoye, A. F. (2025). Modeling Hepatitis B virus

transmission dynamics using Atangana fractional order network approach. *International Journal of Innovative Science and Research Technology, 10*(4), 41–51.

[12]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

[13]. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

[14]. Idogho, C., Abah, E. O., Onuhc, J. O., Harsito, C., Omenkaf, K., Samuel, A., … Ali, U. E. (2025). Machine learning-based solar photovoltaic power forecasting for Nigerian regions. *Energy Science & Engineering, 13*(4), 1922–1934.

[15]. Idoko, I. P., Akindele, J. S., Imarenakhue, W. U., & Bashiru, O. (2024). Exploring the role of bioenergy in achieving sustainable waste utilization and promoting low-carbon transition strategies. *International Journal of Scientific Research in Science and Technology*.

[16]. Idoko, I. P., Arthur, C., Ijiga, O. M., Osakwe, A., Enyejo, L. A., & Otakwu, A. (2024). Incorporating radioactive decay batteries into the USA's energy grid. *International Journal, 3*(9).

[17]. Idoko, I. P., Ayodele, T. R., Abolarin, S. M., & Ewim, D. R. E. (2023). Maximizing the cost effectiveness of electric power generation through the integration of distributed generators. *Bulletin of the National Research Centre, 47*(1), 166.

[18]. Idoko, I. P., David-Olusa, A., Badu, S. G., Okereke, E. K., Agaba, J. A., & Bashiru, O. (2024). The dual impact of AI and renewable energy in enhancing medicine. *Magna Scientia Advanced Biology and Pharmacy, 12*(2), 99–127.

[19]. Idoko, I. P., Eniodunmo, O., Danso, M. O., Bashiru, O., Ijiga, O. M., & Manuel, H. N. N. (2024). Evaluating benchmark cheating and the superiority of MAMBA over transformers. *World Journal of Advanced Engineering Technology and Sciences, 12*(1), 372–389.

[20]. Idoko, I. P., Igbede, M. A., Manuel, H. N. N., Adeoye, T. O., Akpa, F. A., & Ukaegbu, C. (2024). Big data and AI in employment. *Global Journal of Engineering and Technology Advances, 19*(02), 089–106.

[21]. Idoko, I. P., Ijiga, O. M., Enyejo, L. A., Ugbane, S. I., Akoh, O., & Odeyemi, M. O. (2024). Exploring the potential of Elon Musk's proposed quantum AI. *Global Journal of Engineering and Technology Advances, 18*(3), 048–065.

[22]. Idoko, I. P., Ijiga, O. M., Harry, K. D., Ezebuka, C. C., Ukatu, I. E., & Peace, A. E. (2024). Renewable energy policies: A comparative analysis of Nigeria and the USA. *World Journal of Advanced Research and Reviews, 21*(1), 888–913.

[23]. Idoko, I. P., Ijiga, O. M., Akoh, O., Agbo, D. O., Ugbane, S. I., & Umama, E. E. (2024). Power electronics in California's renewable energy transformation. *World Journal of Advanced Engineering Technology and Sciences, 11*(1), 274–293.

[24]. Ikedionu, C. A., Idoko, I. P., Omale, J. O., & Idogho, C. (2025). Mathematical modeling of 3D printing of microreactors. *International Journal of Research Publication and Reviews*.

[25]. Ijiga, O. M., Idoko, I. P., Ebiega, G. I., Olajide, F. I., Olatunde, T. I., & Ukaegbu, C. (2024). Adversarial machine learning for cybersecurity risk assessment. *Journal of Science and Technology, 11*, 001–024.

[26]. Ijiga, O. M., Idoko, I. P., Enyejo, L. A., Akoh, O., & Ileanaju, S. (2024). Generative music models and voice cloning. *World Journal of Advanced Engineering Technology and Sciences, 11*, 372–394.

[27]. Jaegle, A., Borgeaud, S., Alayrac, J. B., Doersch, C., Ionescu, C., Ding, D., … Zisserman, A. (2021). Perceiver IO. *Proceedings of the International Conference on Machine Learning*, 1–11.

[28]. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

[29]. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? *Learning and Individual Differences, 103*, 102274.

[30]. Kiela, D., Bhooshan, S., Firooz, H., Perez, E., & Williams, A. (2019). Supervised multimodal bitransformers. *Proceedings of EMNLP*, 1–11.

[31]. Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2013). Is the doer effect a causal relationship? *Proceedings of the Sixth International Conference on Educational Data Mining*, 388–389.

[32]. Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education*. Cambridge University Press.

[33]. Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). VisualBERT. *arXiv preprint arXiv:1908.03557*.

[34]. Manuel, H. N. N., Adeoye, T. O., Idoko, I. P., Akpa, F. A., Ijiga, O. M., & Igbede, M. A. (2024). Passive solar design in Texas green buildings. *Magna Scientia Advanced Research and Reviews, 11*(01), 235–261.

[35]. Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

[36]. Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics, 3*(2), 213–219.

[37]. Okika, N., Nwatuzie, G. A., Odozor, L., Oni, O., & Idoko, I. P. (2025). Addressing IoT-driven cybersecurity risks in critical infrastructure. *International Journal of Innovative Science and Research Technology, 10*(2).

[38]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning transferable visual models. *Proceedings of the International Conference on Machine Learning*, 8748–8763.

[39]. Romero, C., & Ventura, S. (2010). Educational data mining: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601–618.
[40]. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), e1355.
[41]. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
[42]. Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380–1400.
[43]. Ugbane, S. I., Umeaku, C., Idoko, I. P., Enyejo, L. A., Michael, C. I., & Efe, F. (2024). Optimization of quadcopter propeller aerodynamics. *International Journal of Innovative Science and Research Technology, 9*(10), 1–12.
[44]. Wilensky, U., & Reisman, K. (2006). Learning biology through computational theories. *Cognition and Instruction, 24*(2), 171–209.
[45]. Winne, P. H., & Baker, R. S. (2013). Potentials of educational data mining. *Journal of Educational Data Mining, 5*(1), 1–8.