Nâng cao Năng lực Giảng viên Giáo dục Đại học trong Kỷ nguyên Trí tuệ Nhân tạo Tạo sinh: Tích hợp Khả năng Giải thích, Khả năng Tranh luận và Thực hành Phản tư

TÓM TẮT

Sự xuất hiện nhanh chóng của trí tuệ nhân tạo tạo sinh (AI) và các mô hình ngôn ngữ lớn (LLMs) mang đến cả cơ hội lẫn thách thức cho giáo dục đại học. Dù công nghệ này hứa hẹn nâng cao hiệu quả giảng dạy và nghiên cứu, bản chất "hộp đen", xu hướng "ảo giác" và thiếu minh bạch của chúng đặt ra vấn đề về niềm tin, trách nhiệm và tính toàn vẹn sư phạm. Bài viết đề xuất nhu cầu cấp thiết về một khung năng lực giúp giảng viên ứng dụng AI tạo sinh có trách nhiệm. Chúng tôi giới thiệu TECTRA (Trust through Explainability, Contestability, and Reflective Application), khung tiếp cận lấy con người làm trung tâm, kết hợp AI có khả năng giải thích (XAI) và AI có thể phản biện (CAI) như cơ chế nền tảng để xây dựng niềm tin trong giáo dục. Khung gồm bốn trụ cột: Đạo đức, dựa trên tính minh bạch của XAI; Tích hợp sư phạm, qua cấu trúc đối thoại của CAI; Hiểu biết kỹ thuật, thông qua khả năng giải thích của XAI; và Thực hành phản tư, duy trì bằng phản hồi của hai hệ thống. Bài viết xác định năng lực, hoạt động và công cụ phát triển cho từng trụ cột, cùng lộ trình ba giai đoạn: đánh giá, xây dựng năng lực và mở rộng bền vững. Ngoài ra, khuyến nghị chính sách nhấn mạnh sự linh hoạt, minh bạch, giám sát con người và đạo đức. Bằng cách xem XAI và CAI là yếu tố tương tác thực tiễn, TECTRA đưa AI trở thành đối tác minh bạch, phản biện, thúc đẩy năng lực giảng viên thích ứng, dựa trên bằng chứng và có nền tảng đạo đức.

Từ khóa: *AI Có thể Giải thích, AI Có thể Tranh luận, AI Tạo sinh, Mô hình Ngôn ngữ Lớn, Giáo dục Đại học*

Enhancing Higher Education Faculty Competencies in the Age of Generative AI: Integrating Explainability, Contestability, and Reflective Practice

ABSTRACT

The rapid emergence of generative artificial intelligence (AI) and large language models (LLMs) has created both unprecedented opportunities and significant challenges in higher education. While these technologies promise to enhance teaching effectiveness and research productivity, their "black-box" nature, tendency toward hallucinations, and opacity raise critical concerns about trust, accountability, and pedagogical integrity. This paper addresses the urgent need for a comprehensive framework to enhance faculty competencies in leveraging generative AI responsibly and effectively. We propose TECTRA (Trust through Explainability, Contestability, and Reflective Application), a novel human-centered framework that integrates Explainable AI (XAI) and Contestable AI (CAI) as foundational mechanisms for trustworthy AI adoption in education. The framework is structured around four interdependent pillars: Ethical Grounding, enabled by XAI's transparency; Pedagogical Integration, activated through CAI's dialogic structure; Technical Literacy, developed through XAI's interpretable explanations; and Reflective Practice, sustained through combined feedback loops from both mechanisms. We detail specific, measurable faculty competencies mapped to each pillar and provide concrete development activities and tools. Furthermore, we present a phased implementation strategy roadmap spanning assessment, capacity building, and sustainable scaling, alongside comprehensive policy recommendations that emphasize flexibility, transparency, human oversight, and ethical principles. By positioning XAI and CAI as active, functional elements rather than separate technical considerations, TECTRA transforms generative AI from an opaque tool into a transparent, contestable partner for critical inquiry, ultimately fostering enhanced faculty competencies that are adaptive, evidence-based, and ethically grounded in an increasingly AI-driven educational landscape.

Keywords: *Explainable AI, Contestable AI, Generative AI, Large Language Models, Higher Education*

1. INTRODUCTION

The digital era has brought transformative technologies that are reshaping higher education worldwide. Among these, generative artificial intelligence (AI), large language models (LLMs) such as OpenAI's ChatGPT, Anthropic's Claude, Meta's Llama, or Google's Gemini, have rapidly emerged as a powerful tool with the potential to enhance the teaching and research capacities of university and college lecturers. Since the public release of LLMs, educators have been exploring how such AI systems can revolutionize teaching practices and academic research[1,2]. These AI tutors and assistants provide opportunities to innovate pedagogical approaches, personalize learning, automate routine tasks, and support scholarly work. For instance, generative AI can help instructors generate lesson plans, produce quizzes or simulations, provide conversational tutoring to students, and even aid in literature recommending, grading, feedback, or interview preparation[3-6]. Likewise, researchers can leverage LLMs to summarize literature, draft manuscripts, or brainstorm research ideas. In theory, these capabilities promise to enhance lecturers' teaching effectiveness and research productivity in the digital age.

However, along with immense potential, generative AI introduces significant challenges and uncertainties in educational contexts. Modern LLMs are often "black boxes", where they produce answers without revealing clear reasoning or sources[7,8]. This opacity can erode trust, as educators and students may question how or why an AI arrived at a given answer. Additionally, LLMs are prone to hallucinations (i.e., generating incorrect or fabricated information confidently). In an academic setting, such undetected errors or falsehoods can mislead learners and undermine learning outcomes. Educators also worry about issues of academic integrity (e.g., plagiarism or uncritical use of AI in student work), biases in AI outputs, data privacy, and the broader ethical implications of delegating educational tasks to AI. These challenges highlight that while AI can assist humans, it cannot be blindly trusted, particularly in educational and academic settings, where rigor and accuracy are crucial[9,10].

To fully realize the benefits of generative AI in education, it is crucial to address these challenges. Two emerging approaches in AI research hold promise in this regard: Explainable AI (XAI)[11-21] and Contestable AI (CAI)[22-27]. XAI aims to make AI systems more transparent by providing human-understandable explanations for their outputs or decisions. In an educational context, XAI could enable a lecturer or student to see the reasoning behind an AI-generated answer, thereby improving trust and facilitating error diagnosis[13,18-21,28]. Recent regulations, including the General Data Protection Regulation (GDPR)[29] and the EU AI Act[30], establish legal requirements for AI interpretability and explainability, fundamentally grounded in the principle of *contestability*, ensuring individuals can meaningfully challenge automated decisions. CAI goes a step further by enabling users to question, dispute, and engage in dialogue with the AI's decisions or reasoning[22-26]. A CAI system would not only explain itself, but also allow teachers or students to challenge its responses and have the system revise its answers when justified[27]. Together, these approaches aim to transform AI into a more interactive and accountable assistant, rather than an opaque box.

Hence, this paper presents a comprehensive research study on how explainability and contestability can foster trustworthy use of generative AI in higher education, ultimately enhancing faculty competencies in the digital era. Our contributions are as follows:

1. **We propose a novel framework, TECTRA (Trust through Explainability, Contestability, and Reflective Application)**, which integrates XAI and CAI to create a human-centered, trustworthy framework for generative AI in education. We detail how this approach can fill current gaps and empower faculty across disciplines.

2. **We design a phased strategy roadmap for implementation and policy recommendations**, ensuring that universities can harness the benefits of generative AI while maintaining academic integrity, equity, and human oversight.

By adopting the principles of explainability and contestability for generative AI and LLMs, we aim to provide a practical framework with a phased roadmap, implementation guidelines, and policy recommendations to enhance the trust and fairness of leveraging these technologies in higher education.

2. BACKGROUND

2.1. Generative AI in Education

Generative AI refers to a class of AI models that can produce new content (e.g., text, images, music, or code) by learning patterns from existing data. A prominent example is the LLM, which can engage in human-like dialogue and create texts in response to prompts. OpenAI's ChatGPT, Anthropic's Claude, Meta's Llama, or Google's Gemini, and similar LLMs are generative AI models consisting of billions of parameters trained on massive textual databases and can perform tasks like answering questions, writing essays, summarizing documents, and creating lesson plans. In essence, these models predict text based on learned patterns, stringing together words that are statistically likely to follow a given prompt.

In education, generative AI has swiftly found diverse applications[3-6,31-35]. Students can utilize AI chatbots to explain complex concepts, generate ideas, or receive feedback on their writing drafts. Faculty members are exploring the use of generative AI to enhance teaching materials and workflows. For instance, an instructor can prompt an AI to generate quiz questions, example problems, or even first drafts of lesson summaries, which can save time on preparation[4-6]. Generative AI can also assist with research by summarizing scholarly articles or suggesting new research directions[3,32]. From a learning perspective, these tools offer personalized support, where an AI tutor can converse with a student, quiz them on course content, and adjust the difficulty of questions to the student's level in real-time[33,34]. This ability to provide instant, tailored feedback and access to information has led many to view generative AI as a powerful aid for both instructors and learners. Early implementations have demonstrated that generative AI can benefit a diverse range of learners. For example, translating or simplifying content for non-native speakers can spark creativity, providing new examples or analogies that enrich the learning experience[36,37].

However, the use of generative AI in education also raises pedagogical questions. Because LLMs generate text based on patterns rather than genuine understanding, educators must consider how students' use of AI affects learning outcomes. Some studies suggest that while generative AI can increase productivity on certain tasks, it might reduce cognitive effort or lead to more homogeneous student work if over-relied upon[35,38,39]. Thus, a consensus is emerging that generative AI should augment teaching and learning, serving as a smart assistant or tutor, rather than replacing the essential human elements of creativity, critical thinking, and mentorship[5,40]. To realize this vision, faculty members need to guide students in the proper use of AI, and importantly, they themselves must have a clear understanding of how generative AI works and where its outputs can or cannot be trusted. This sets the stage for examining issues of transparency and fairness in generative AI systems.

2.2. Transparency and Fairness: The "Black-Box" and Hallucination Problems

2.2.1. Black-box Problem

Despite their impressive capabilities, most state-of-the-art AI models operate as "black boxes." In a black-box model, the internal reasoning leading to any given output is hidden or too complex for users and even developers to interpret. Users see both the input and the AI's output, but not the decision-making process that occurs in between[11,41-44]. This opacity poses a serious problem in education, where trust and accountability are critical. Educators are understandably uneasy when an AI provides an answer or assessment without any explanation, especially knowing that even the engineers who built the model cannot fully explain how a particular result was generated. The lack of transparency makes it difficult to judge the correctness or bias of AI outputs, and it hinders the ability to contest or audit those outputs. As several studies of AI in education noted[45-49], the black-box nature of some AI algorithms makes it challenging for stakeholders to understand or challenge AI-driven decisions, raising ethical concerns about their use in education. In other words, if a generative AI system gives flawed information or an unfair recommendation, its inscrutable logic means educators might not realize the error or have the means to dispute it. This undermines confidence and can erode the educational integrity of AI-assisted processes.

2.2.2. Hallucination of Generative AI

Compounding the transparency issue is the tendency of generative AI models to produce hallucinations. Hallucination refers to the well-documented phenomenon where AI systems, especially generative AI and LLMs, generate incorrect, misleading, or entirely fabricated information or responses that do not accurately reflect the data they were trained on or the input provided to them[9,10,50,51]. For example, generative AI might confidently generate a citation for a non-existent scientific article or incorrectly explain a concept in a way that appears authoritative[52]. These tools do not truly understand the facts or possess a comprehensive

knowledge model of reality. Instead, they base their responses on statistical patterns in the training data. Therefore, an AI might assert false facts because those word combinations seem likely. Research has highlighted that general-purpose LLMs often draw on poor-quality or incorrect data absorbed during training, which can lead to incorrect outputs, and they typically lack the ability to verify information or cite specific sources for their statements. In academic contexts, such hallucinations can mislead students or propagate misinformation if unchecked. A faculty member using AI to generate lecture notes or a student relying on it for research may be misled by a confidently stated but false piece of information. Instructors also expressed concern about the use of generative AI in education, citing inaccurate or unreliable outputs as a top concern.

2.3. Explainability and Contestability for Generative AI in Higher Education

2.3.1. Explainable AI (XAI)

XAI refers to a set of methods and techniques that make the decisions or outputs of an AI system understandable to human users. Traditional AI models, particularly deep learning (DL) and LLMs, often function as black boxes, producing results without a clear explanation of how they were derived. XAI aims to open this box by providing human-interpretable insights into the model's reasoning, thereby enhancing trust, transparency, and accountability. Common XAI approaches include saliency mapping[15,16,42-44,53-55], feature attribution[56-58], and counterfactual explanations[59-61].

For generative AI within the education sector, especially those used in natural language applications with LLMs, XAI techniques increasingly involve prompt engineering for rationale generation, rubric-based evaluations and play a critical role in supporting student learning, assisting teachers, and ensuring the fairness of AI-driven assessments[18-21]. XAI technique can provide a justification in natural language, highlight the factors (input features) that most influenced its decision, or produce confidence scores and evidence (e.g, source citations) to support its output. For example, an explainable generative AI writing assistant might underline which parts of its answer were drawn from which reference texts, or an AI used for student evaluation might show the rubric criteria and how it applied them. Recent works, such as ExASAG[18], CourseEvalAI[19], and QwenScore+[20], incorporate explainable reasoning strategies into rubric-aligned evaluation to break down automated scoring into human-understandable criteria, enabling clear explanations for grades.

By opening up the AI's reasoning, XAI helps users verify correctness, detect errors or biases, and ultimately decide when to trust the AI and when to be skeptical. This is particularly important in education, where teachers need to understand an AI's suggestion in order to confidently act on it or explain it to students. If an AI can show its work, for instance, by citing the source of a fact or explaining the steps in a solution it provided, the instructor can more easily validate the result. For example, advanced tutoring systems like LPITutor[21] use retrieval-augmented generation (RAG) and user-adaptive prompts to ensure that the content and explanations are grounded in course materials and personalized to the learner's level. Indeed, research confirms that increasing a system's transparency via explainability can strengthen teachers' trust and willingness to accept AI recommendations.

These developments illustrate how XAI enhances learning experiences by providing context-aware, pedagogically useful feedback, turning AI systems into interpretable educational partners. Explainability also aids AI literacy, as educators interpret AI outputs, they learn more about the AI's limits and behavior, becoming better equipped to integrate it appropriately. In short, XAI is a foundation for responsible AI adoption in education, as it addresses the black-box issue by illuminating the AI's decision logic and thereby fosters user understanding and trust.

2.3.2. Contestable AI (CAI)

However, simply explaining an AI's output does not fully resolve the power imbalance between human and machine. Even if a teacher knows why an AI produced a certain result, they also need

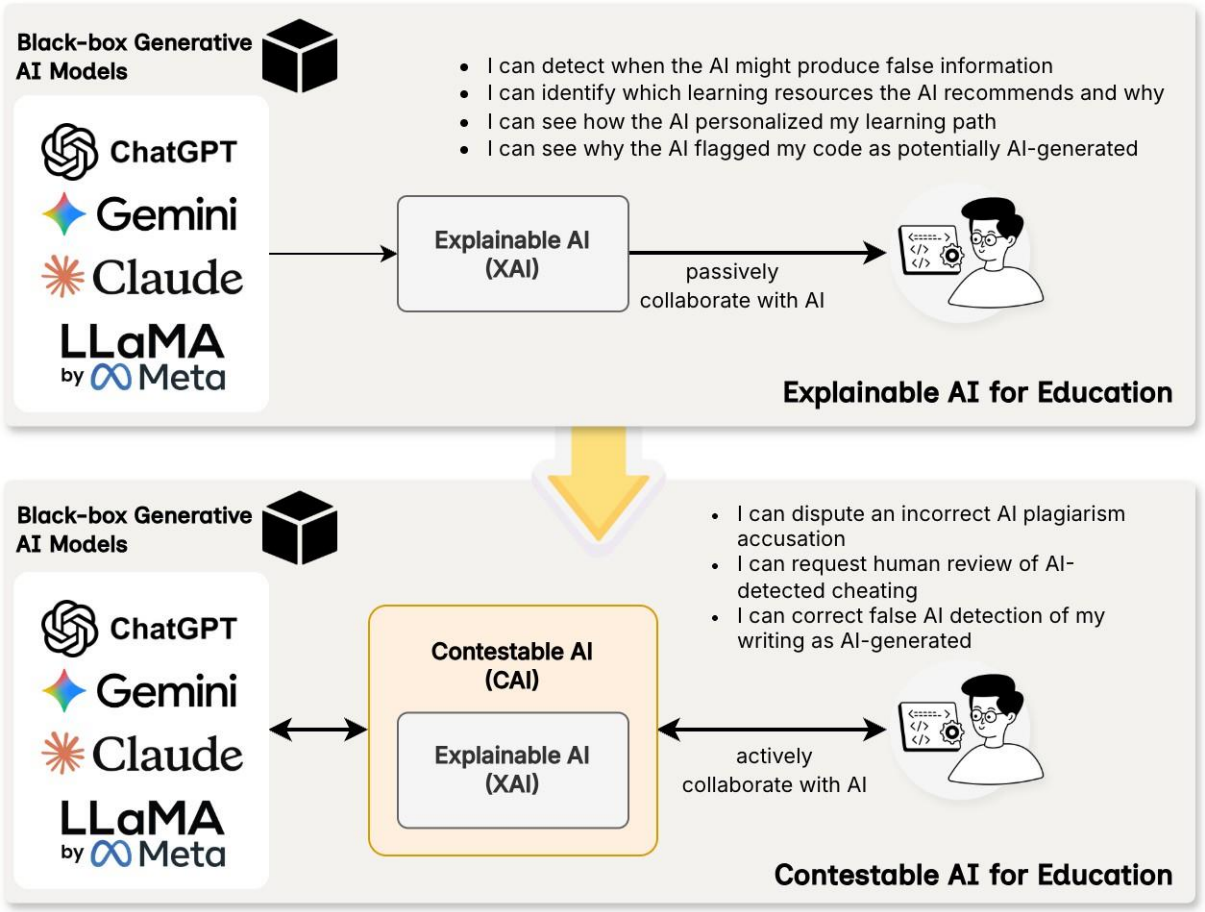

Figure 1. Moving from Explainable AI (XAI) to Contestable AI (CAI) in Enhancing Faculty and Student Competencies in Higher Education.

the ability to say "I disagree" or override the AI when it is wrong. This is where the concept of CAI emerges as significant.

Contestability in AI systems means that users have the right and means to challenge or appeal an AI's decision, and the system can adapt or be corrected based on that human feedback[24-26,62]. In other words, a CAI system is not a one-way authority but rather allows a dialogue or iterative process with human oversight. According to AI ethics frameworks, contestability implies that those affected by an AI outcome are provided with "plain and easy-to-understand information" about how the decision was made, enabling them to challenge the outcome if necessary. Contemporary regulatory frameworks, including the GDPR[29] and the EU AI Act[30], highlight the critical role of interpretability and contestability. The GDPR's Articles 13 and 14 establish that individuals subjected to profiling are entitled to receive "meaningful information about the logic involved." As clarified in official EU guidance, these provisions extend to healthcare settings. The explainability mandate within the GDPR should be interpreted as establishing a framework for contestability, requiring that AI-driven decisions be sufficiently transparent to enable individuals to challenge them. Article 22 specifies that when automated decision-making, including profiling, is legitimately applied to individuals, data controllers must protect their right "to express his or her point of view and to *contest* the decision"[29]. Contestability is thus increasingly recognized as a key aspect of accountability in AI deployments.

Contestability requires a few fundamental capacities: the AI must provide explanations that users can inspect; users must have clear paths to challenge the AI's output, and the system and human administrators must have the

Table 1. An Overview of Prominent Applications of Generative AI Integrated with Explainable AI (XAI) and Contestable AI (CAI) Frameworks in Education.

Method	XAI/CAI	Application	Contributions
ExASAG[18]	XAI	Automatic grading of short, free-text student answers.	Produces human-understandable rationales for each grade using XAI methods; Provides interpretable feedback, enabling teachers to understand and verify AI grading.
CourseEvalAI[19]	XAI	Transparent, rubric-based evaluation of LLMs for grading open-ended student work.	Fine-tunes an LLM using dual-layer rubrics (for answers and explanations); Evaluations and scores stored in a graph database for full traceability; Reduces bias, increases rubric fidelity, and improves the evaluability of model-generated explanations.
QwenScore+[20]	XAI	Essay scoring with formative feedback.	Applies LLMs to generate personalized, formative feedback rather than just grades; Emphasizes student trust, ethical design, and transparency in educational AI systems; Evaluates LLMs' ability to provide scaffolded, rubric-aligned responses that enhance learning outcomes.
LPITutor[21]	XAI	Intelligent tutoring across subjects (adaptive questions and answers, hints).	Increases transparency and accuracy by grounding answers in retrieved documents (students can be shown the source or at least trust the answer is curriculum-aligned); Customizable difficulty: explanations and answers suited to learner's skill level; Addresses hallucinations by tethering model to real content.
CAELF[27]	CAI	Interactive feedback on essays.	Highly robust to student push-back, maintains logical consistency and only changes grade if student's counter-argument is valid; Inherently explainable feedback (built from explicit arguments and rubric criteria); Improves LLM's reasoning and reduces susceptibility to manipulation in an educational dialogue.

ability to revise or adapt the decision based on that challenge[24-26,62]. Some researchers describe this as transitioning from static AI systems to dynamic, interactive AI, which are machines that can engage in dialogue about their reasoning and adjust when valid points of contestation are raised. It represents a shift from viewing explainability as the end goal to viewing recourse and redress as the ultimate goal. If the AI is found to be flawed, there must be a way to correct it or mitigate its effects. In education, contestability means that faculty and students are not passive recipients of AI outputs, but active participants who can question and modify those outputs. For example, consider an AI system that flags student essays for potential plagiarism or grading purposes. In a contestable design, a student could appeal that flag to a human instructor, or the instructor

could override the AI's grading suggestion if they see it's based on a misinterpretation. The AI system would then ideally learn from this correction, or at least record it, thus improving over time or avoiding repeated mistakes. A recent work introduced CAELF[27], a contestable feedback framework where multiple AI teaching assistant agents independently grade different aspects of an essay, and a teacher agent aggregates their evaluations via formal argumentation. This design allows students to query, challenge, and clarify the AI's feedback, making the grading process interactive and open to dispute.

The twin notions of XAI and CAI are complementary and together promise a more trustworthy AI ecosystem in education, as illustrated in Figure 1. Explainability provides transparency *("Why did the AI say that?")* and contestability provides agency *("What can we do if the AI is wrong?")*. By embracing both, we address the earlier challenges: an explainable system reduces the fear of black boxes and helps identify errors or biases, while a contestable system ensures that those errors can be corrected and biases mitigated through human intervention. Importantly, contestability reinforces human authority and accountability. It is seen as a means to facilitate accountability, preventing blind reliance on AI and ensuring that decisions can be audited and overturned if needed. In an educational context, this aligns perfectly with the ethos that teachers (and, where appropriate, students) should have the final say in teaching and assessment decisions where AI can assist, but not autonomously control outcomes without recourse. To summarize, XAI and CAI are key pillars for integrating generative AI in a way that faculty can trust and leverage effectively. An AI system that can explain its outputs and accept human feedback aligns with educational values of transparency, critical inquiry, and continuous improvement. These principles set the foundation for the framework we propose, which aims to enhance faculty competencies and confidence in working with generative AI by embedding explainability and contestability into both technology and practice.

3. PROPOSED FRAMEWORK

To address the faculty competency gap and guide education institutions toward a more responsible and pedagogically sound integration of generative AI, this paper proposes the **TECTRA (Trust through Explainability, Contestability, and Reflective Application)** framework, which is a comprehensive, actionable model designed to cultivate the specific skills and dispositions faculty require to build and sustain a trustworthy AI ecosystem. The framework moves beyond reactive, tool-based training to foster a deeper, more critical engagement with AI, grounding its use in enduring ethical and pedagogical principles. It integrates insights from established, trustworthy AI guidelines with foundational educational technology models, including TPACK (Technological Pedagogical Content Knowledge)[63] to create a comprehensive framework for faculty development. Central to this framework is the recognition that XAI and CAI are foundational mechanisms that operationalize trust in educational AI systems. Rather than treating these as separate technical considerations, TECTRA embeds them as active, functional elements within each pillar, providing faculty with concrete tools and processes to develop and exercise critical AI competencies.

3.1. Core Principles

The TECTRA framework is constructed upon four interdependent pillars, which together form a comprehensive approach to faculty competency in education. These pillars are designed to be mutually reinforcing, ensuring that technical skills are always linked to pedagogical purpose and ethical considerations. Each pillar is strengthened by the integration of explainability and contestability mechanisms. As visualized in Figure 2, the TECTRA framework can be conceptualized as an integrated ecosystem focused on creating trustworthy AI in education. Rather than positioning XAI and contestability as external mechanisms, the framework recognizes them as intrinsic to each pillar's function.

(1) Ethical Grounding is made actionable through XAI's revelatory power; (2) Pedagogical Integration is activated through the dialogic

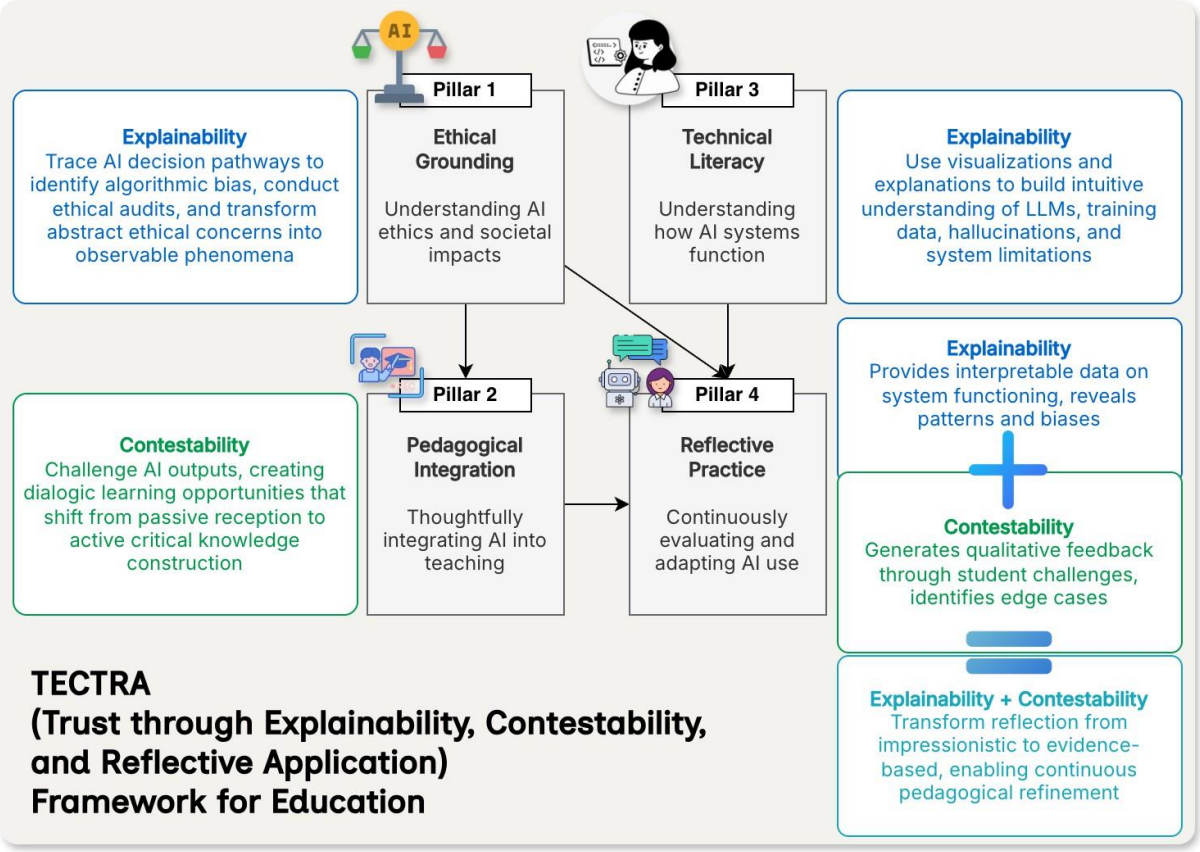

Figure 2. Overview of TECTRA (Trust through Explainability, Contestability, and Reflective Application) Framework for Higher Education.

structure of contestability; (3) Technical Literacy is developed through XAI's explanations; and (4) Reflective Practice is sustained through the combined feedback loops created by both mechanisms.

3.1.1. Pillar 1: Ethical Grounding (Enabled by Explainability)

This foundational pillar moves the conversation about AI ethics beyond the narrow confines of plagiarism and academic misconduct. It requires faculty to develop a robust understanding of the broader ethical landscape of AI, including the ability to critically analyze issues of algorithmic bias, data privacy, intellectual property, and the potential for AI to perpetuate societal inequities.

The Role of Explainability: XAI serves as the primary mechanism through which faculty can develop and demonstrate ethical competency. When AI systems make their decision-making processes transparent, faculty gain the ability to conduct meaningful ethical audits of the technology they employ. Through XAI dashboards and interpretation tools, faculty can trace how an AI system arrives at specific outputs, identifying the data points, features, or patterns that influence results. This visibility transforms abstract ethical concerns into concrete, observable phenomena that can be analyzed, discussed, and addressed. Competency in this pillar means faculty can not only articulate ethical risks but can also use XAI tools to actively investigate and demonstrate them. For instance, a faculty member might use XAI techniques to reveal how an automated essay grading system disproportionately penalizes certain writing styles or linguistic patterns associated with multilingual learners. This capability enables faculty to design learning environments and institutional policies that proactively mitigate harm, ensuring that the use of AI aligns with the core values of fairness, accountability, and transparency. Furthermore,

XAI empowers faculty to transform ethical instruction into active learning. Rather than lecturing about algorithmic bias in the abstract, faculty can use XAI as a pedagogical tool, demonstrating in real-time how bias manifests, helping students develop critical evaluation skills, and fostering a more ethically informed user base. This approach creates a culture of ethical vigilance that extends beyond compliance to genuine understanding and advocacy.

3.1.2. Pillar 2: Pedagogical Integration (Enabled by Contestability)

This pillar addresses the practical application of AI in teaching and learning. Drawing inspiration from the TPACK model[63], it emphasizes the crucial intersection of technological knowledge, pedagogical strategy, and subject-matter expertise. Competency here is not merely about using an AI tool but about thoughtfully integrating it into the curriculum to enhance learning outcomes through designs that promote active engagement, critical thinking, and student agency. Pedagogical integration involves redesigning assessments to prioritize process and critical thinking over simple content generation, creating AI-scaffolded activities that support rather than supplant student effort, and leveraging contestable AI systems to create more personalized and adaptive learning experiences. The ability to implement contestable frameworks requires faculty to ask not merely *"What can this tool do?"* but *"How can this tool create productive cognitive friction and dialogic opportunities that help my students achieve our learning objectives more effectively and deeply?"*

The Role of Contestability: Contestability serves as a natural mechanism for formative assessment. When students challenge AI outputs, they reveal their understanding, misconceptions, and the reasoning processes behind their responses. Faculty can use these contestation patterns as diagnostic data, identifying where additional instruction or support is needed and adapting their teaching accordingly. CAI transforms pedagogical integration from passive tool adoption to active, dialogic learning. When students are empowered to challenge, question, and debate AI-generated outputs (e.g., grades, feedback, content recommendations, or analytical interpretations), the learning dynamic fundamentally shifts from information consumption to critical knowledge construction. Faculty who integrate CAI into their pedagogy create structured opportunities for metacognitive development. Consider an AI-assisted writing tutor that provides feedback on student essays: in a traditional implementation, students might passively accept the AI's suggestions. However, when the system is designed to be contestable, students must engage in reasoned argumentation to challenge feedback they believe is inappropriate. They must gather evidence from their work, present their reasoning, and engage in structured dialogue with both the AI system and their instructor. This process inherently develops higher-order thinking skills (e.g., analysis, evaluation, and creation), positioning AI as a sophisticated learning partner.

3.1.3. Pillar 3: Technical Literacy (Enabled by Explainability)

While deep technical expertise is not required, a foundational understanding of how generative AI systems work is essential for responsible use. This pillar aims to equip faculty with a conceptual understanding of core AI principles, including the nature of LLMs, the role of training data, the statistical foundations of AI outputs, and the inherent limitations of the technology.

The Role of Explainability: XAI serves as both a teaching tool and a competency-building mechanism for technical literacy. Rather than requiring faculty to understand complex machine learning mathematics, XAI systems provide accessible visualizations and explanations that interpret AI behavior. Through interaction with XAI tools, faculty develop an intuitive understanding of how models process information, why they produce certain outputs, and where their limitations lie. A technically literate faculty member, supported by XAI, understands why AI models hallucinate, recognizes the statistical and probabilistic nature of their outputs, can identify when a model is operating outside its training domain, and can explain these concepts to students using concrete examples drawn from XAI explanations.

For instance, when an LLM produces a factual error, XAI tools can help faculty trace the error to limitations in training data, revealing that the model is generating plausible-sounding but unverified content based on statistical patterns rather than verified knowledge. This literacy extends to understanding the source and characteristics of training data. XAI techniques that reveal which aspects of training data most influence specific outputs help faculty recognize potential blind spots, biases, or domain limitations in AI systems. This knowledge prevents both uncritical acceptance and unfounded fear, forming the basis for a more nuanced and effective pedagogical approach. Moreover, XAI supports the development of technical literacy in faculty through guided exploration. Faculty development programs can utilize XAI dashboards as learning environments where educators experiment with different inputs, observe how the AI's explanations evolve, and develop mental models of system behavior. This hands-on, explanation-guided approach to technical literacy is more accessible and pedagogically effective than traditional technical training, making AI competency achievable for faculty across all disciplines.

3.1.4. Pillar 4: Reflective Practice (Enabled by Explainability & Contestability)

This final pillar encourages faculty to adopt a critical, evidence-based, and iterative approach to their use of AI. Competency in this area involves continuously evaluating the impact of AI tools on their own teaching workflows, as well as on student learning and engagement. This pillar also establishes the practice of systematically gathering feedback through both XAI analytics and contestation records, reflecting on successes and failures revealed through transparent system behavior, and adapting pedagogical strategies in response to evidence.

The Role of Explainability and Contestability: Both XAI and CAI function as structured systems for generating the evidence and insights necessary for meaningful reflective practice. XAI provides faculty with interpretable data on how AI systems are functioning in their courses, revealing patterns in automated feedback, highlighting which students are receiving what types of AI support, and making visible any systematic biases or limitations in AI-mediated interactions. This transparency transforms reflection from subjective impression to data-informed inquiry. Meanwhile, CAI creates natural feedback loops that drive reflection. When students challenge AI outputs, they generate rich qualitative data about the AI system's performance, revealing edge cases, misconceptions, and areas where the AI may be falling short of pedagogical goals. Faculty who systematically analyze these contestations, tracking which AI outputs are most frequently challenged, what types of student arguments are most compelling, and how challenges correlate with learning outcomes, gain actionable insights for refining their AI integration strategies.

For example, a faculty member might notice through XAI analysis that an AI tutoring system consistently provides less detailed explanations to students who initially struggle with a concept. This insight, combined with student contestations arguing that the feedback is insufficient, prompts the instructor to reconfigure the system or supplement it with additional human support. The integration of XAI and contestability ensures that reflection is not a solitary, impressionistic activity but a collaborative, evidence-based practice. It positions reflective practice as an ongoing process of inquiry and improvement, keeping pedagogical practice aligned with a rapidly evolving technological landscape. Faculty members become action researchers in their own classrooms, using the transparency and dialogic nature of well-designed AI systems to continuously refine their practice.

3.2. Defining Faculty Competencies

The TECTRA framework is operationalized through a set of specific, measurable faculty competencies, which provide a clear roadmap for professional development. Table 2 maps these core competencies to the four pillars of the framework, now explicitly incorporating XAI and CAI mechanisms, and suggesting development activities and tools that can be used to support them. This structure translates the abstract

Table 2. Core Faculty Competencies within the TECTRA Framework.

TECTRA Pillar	Core Competency	Description	Development Activities & Tools
Ethical Grounding	**Critical AI Evaluation**	The ability to systematically assess AI-generated outputs for accuracy, veracity, and potential algorithmic bias, and to understand the broader societal and ethical implications of AI use in academia.	Use XAI tools in a sandbox environment to deconstruct biased outputs and understand their origins.
	Ethical AI Pedagogy	The ability to design and articulate clear course policies, assignments, and learning environments that promote the responsible, equitable, and transparent use of AI tools by students.	+ Develop departmental AI usage and citation style guides; engage in case study analysis of complex ethical dilemmas in AI-assisted education. + Craft adaptable syllabus statements.
Pedagogical Integration	**AI-Informed Curriculum Design**	The ability to strategically redesign learning objectives, activities, and assessments to leverage AI for fostering higher-order thinking skills, rather than allowing AI to circumvent them.	+ Receive training on advanced prompt engineering for educational purposes. + Use generative AI to create diverse and complex case studies or problem sets. + Redesign assessments to focus on process, reflection, and application.
	Adaptive Teaching with AI	The ability to use AI-driven analytics and tools to personalize learning pathways, provide timely and targeted feedback, and offer differentiated support to students based on their individual needs.	Use XAI-enhanced Learning Management System (LMS) dashboards to interpret student performance predictions and identify areas for intervention.
Technical Literacy	**Foundational AI Principles**	A conceptual understanding of the basics of LLMs, the importance of training data, and the technical reasons for inherent limitations like hallucinations and bias.	Engage in hands-on, guided experimentation with a variety of generative AI tools to understand their capabilities and failure modes.
	XAI Interpretation	The ability to read and interpret the explanations provided by XAI systems to understand a model's behavior, identify key influencing factors, and assess the reliability of its outputs.	Participate in training sessions on interpreting outputs from specific XAI techniques, applying them to educational datasets.
Reflective Practice	**AI-Mediated Critical Dialogue**	The ability to facilitate and model critical, dialogic interactions with AI systems, encouraging students to question, challenge, and verify AI-generated information.	+ Implement assignments that use CAI frameworks for automated feedback. + Require students to maintain structured reflection journals on their AI usage and findings.
	Continuous Self-Improvement	A commitment to ongoing professional development, staying current with the rapid advancements in AI technology and pedagogy, and actively participating in the institutional conversation around AI.	+ Engage in interdisciplinary faculty learning communities and institutional forums on AI. + Contribute to the work of institutional AI task forces.

principles of the framework into a concrete guide for designing effective faculty training programs, drawing upon competency models proposed by leading educational organizations. Our integrated approach for TECTRA ensures that faculty development is not fragmented into disconnected technical skills, but rather unified around the central goal of building trust in AI systems, pedagogical decisions, and in students' capacity to engage critically with technology. The ultimate outcome of this ecosystem is the cultivation of enhanced faculty competencies that are adaptive, evidence-based, and ethically grounded, which in turn foster greater student agency in an increasingly generative AI educational landscape.

4. IMPLEMENTATION GUIDELINES & POLICY RECOMMENDATIONS

The transformation from a theoretical model to institutional practice requires a deliberate and

structured approach. The TECTRA framework, while providing the conceptual architecture, must be supported by a practical roadmap that guides educational institutions through the complex process of implementation and policy development. This section outlines a phased strategy for integrating the TECTRA framework into the university's fabric, providing concrete recommendations for crafting agile, ethical, and enabling AI policies that support rather than hinder innovation, as visualized in Figure 3.

4.1. A Phased Strategy Roadmap

A successful institution-wide integration of trustworthy AI practices requires a strategic, multi-phase approach that builds momentum, secures stakeholder buy-in, and allows for iterative learning and adaptation.

4.1.1. Phase 1: Assess and Align

The foundational phase is dedicated to establishing a shared understanding and a common vision. The first step is to form a cross-functional AI task force comprising administrators, faculty from diverse disciplines, instructional designers, IT professionals, legal experts, and, crucially, students. This group's initial mandate is to conduct a comprehensive, campus-wide AI literacy audit to gauge the current knowledge, practices, and attitudes of both faculty and students. Concurrently, institutional leaders must work with this task force to align the university's strategic goals for AI with the core principles of the TECTRA framework. This phase is not about deploying technology but about fostering dialogue, engaging stakeholders, and collaboratively defining what a successful, human-centric AI integration will look like for the institution.

4.1.2. Phase 2: Build Capacity

The second phase focuses on building the necessary human and technical infrastructure. The institution should develop and roll out a portfolio of targeted professional development programs, workshops, and resources explicitly designed to cultivate the faculty competencies outlined in the TECTRA framework (see Table 2). To move from theory to practice, the university should fund and support pilot programs in various departments, encouraging faculty to experiment with innovative generative AI integrations and assessment redesigns in a low-stakes environment. During this phase, it is also crucial for the institution to vet, procure, and provide secure and equitable access to a suite of approved AI tools, ensuring that faculty and students have a safe and supported environment for exploration.

4.1.3. Phase 3: Scale and Sustain

The final phase transitions from pilot programs to systemic integration. Successful strategies and resources developed during the capacity-building phase should be scaled across the institution. A key element of this phase is the formal integration of AI literacy and ethical AI use into the general education or core curriculum for all students, ensuring that graduates are prepared for an AI-driven world. To sustain this effort, institutions should establish permanent support structures, such as a Center for AI in Teaching and Learning, to provide ongoing consultation, training, and resources. Finally, a continuous feedback loop must be established, allowing the AI task force to regularly gather input from faculty and students on the effectiveness of policies and support services, ensuring that the institution's approach remains agile and responsive to the rapid pace of technological change.

4.2. Agile and Ethical AI Policies

Effective institutional AI policy and robust faculty competency are not separate initiatives. They are deeply interconnected and mutually reinforcing. A well-crafted policy provides the necessary guidance and security for faculty to innovate, while competent faculty are essential for effectively implementing any policy. An enabling policy cannot be enacted by an unprepared faculty, and faculty cannot develop competency in a policy vacuum. Therefore, policy development must proceed together with capacity building, guided by the following principles:

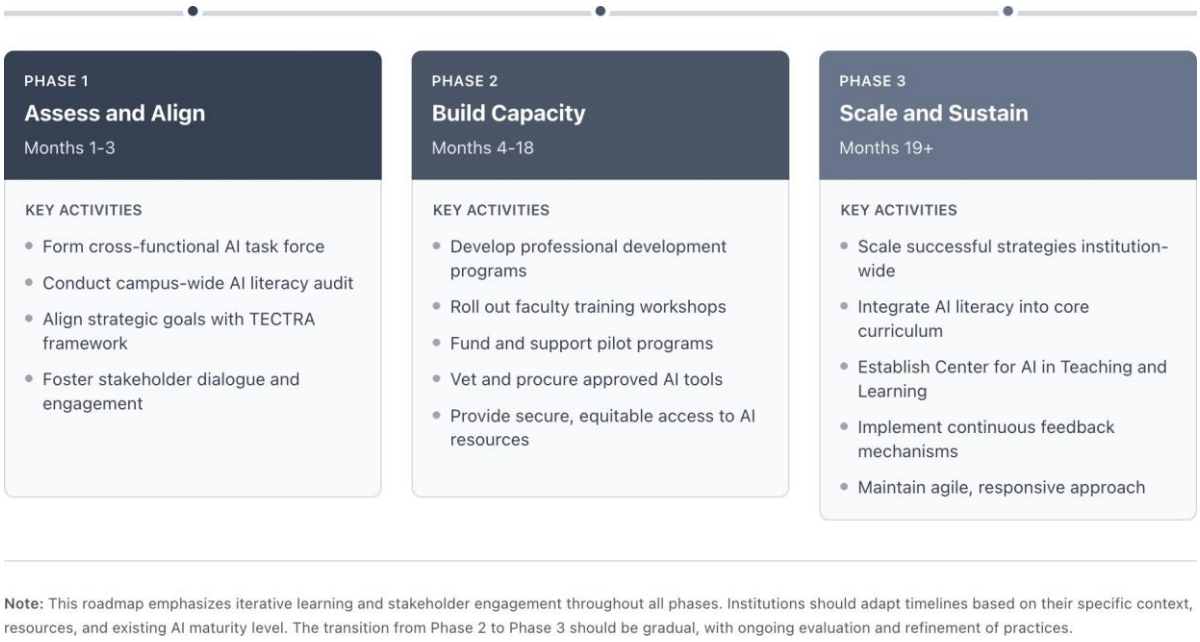

Figure 3. Phased Strategy Roadmap for Integrating the TECTRA Framework into the Higher Education Institutions' Fabric.

Define, Not Just Prohibit. Any effective AI policy must begin with a clear and comprehensive definition of generative AI and related terminology, ensuring that all stakeholders are operating from a shared understanding. Rather than resorting to blanket prohibitions, which are often unenforceable and pedagogically counterproductive, policies should provide nuanced guidelines on permissible and impermissible uses. These guidelines should be flexible and adaptable to the diverse contexts of different academic disciplines.

Mandate Transparency and Citation. To uphold academic integrity in an AI-augmented environment, policies must establish clear and unambiguous standards for transparency and accountability. Students should be required to disclose their use of generative AI tools in all academic work. The policy should provide specific instructions for how to cite and acknowledge this use, drawing from emerging standards in various disciplinary style guides.

Prioritize Human Oversight. Policies must unequivocally affirm that faculty and students are ultimately responsible and accountable for all academic work and educational outcomes. This principle should be operationalized by ensuring that faculty maintain meaningful oversight of any AI-driven assessment or feedback processes. Furthermore, in alignment with the principle of contestability, policies must establish clear and accessible mechanisms for students to appeal or seek review of AI-generated decisions, reinforcing human agency within the system.

Embed Ethical Principles. A robust AI policy must extend beyond academic integrity to explicitly address the core ethical principles of responsible AI. This includes strong provisions for protecting student data privacy and security, particularly when using third-party tools. The policy should also prohibit the use of AI to generate biased, discriminatory, or harmful content and should align with internationally recognized frameworks for trustworthy AI, such as the OECD AI Principles.

Ensure Flexibility and Iteration. Given the breathtaking pace of AI development, any policy written currently will be outdated tomorrow. Therefore, AI policies must be designed as living documents, not as static regulations set in stone. The institutional AI task force should be charged with conducting regular, periodic reviews of all AI-related policies,

ensuring they remain relevant, effective, and aligned with both technological advancements and the evolving pedagogical needs of the university community. This commitment to agility is paramount for navigating the future of AI in education successfully.

5. CONCLUSION

This paper identifies a critical juncture in education, where the asynchronous adoption of generative AI has created a pedagogical crisis rooted in faculty competency gaps and compounded by the risks of AI's opacity, bias, and misinformation. In response, the TECTRA framework offers a proactive, human-centric solution grounded in Ethical Grounding, Pedagogical Integration, Technical Literacy, and Reflective Practice, positioning XAI and CAI as pivot mechanisms to transform AI from an opaque tool into a transparent partner for critical inquiry. Beyond mere integration, this technological moment presents an opportunity to fundamentally reimagine education as a more equitable, engaging, and effective ecosystem where AI augments rather than replaces human intellect, creativity, and critical thinking. The TECTRA framework, with its emphasis on human agency and ethical oversight, serves as a foundational step toward this vision, making faculty empowerment the most vital investment institutions can make in securing a human-centric future for education.

REFERENCES

1. S. Milano, J. A. McGrane, and S. Leonelli. Large language models challenge the future of higher education, *Nature Machine Intelligence*, **2023**, *5*(4), 333-334.
2. G. G. Wilkinson. Enhancing Generic Skills Development in Higher Education in the Era of Large Language Model Artificial Intelligence. *Journal of Higher Education Theory & Practice*, **2024**, *24*(3), 64-76.
3. X. Wang, N. Duong-Trung, R. R. Bhoyar, and A. M. Jose. LLM-based literature recommender system in higher education: A case study of supervising students' term papers, *Computers & Education: Artificial Intelligence*, **2025**, *10*, 32-41.
4. N. Hashmi and A. S. Bal. Generative AI in higher education and beyond, *Business Horizons*, **2024**, *67*(5), 607-614.
5. T. T. H. Nguyen, T. D. Q. Nguyen, H. L. Cao, T. C. T. Tran, T. C. M. Truong, and H. Cao. *SimInterview: Transforming Business Education through Large Language Model-Based Simulated Multilingual Interview Training System*, International Conference on Economics, Finance, and Management, ICEFM 2025, Ho Chi Minh City, Vietnam, 2025.
6. D. S. A. Cufuna, J. M. Duart, and G. Rangel-de Lazaro. *Augmented reality in higher education: Interactions in LLM-based teaching and learning*, The Learning Ideas Conference, TLIC 2024, New York, NY, USA, 2024.
7. R. Manche and P. K. Myakala. Explaining black-box behavior in large language models, *International Journal of Computing and Artificial Intelligence*, **2022**, *3*(2), 102-108.
8. S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering*, **2024**, *36*(7), 3580-3599.
9. H.-T. Ho, D.-T. Ly, and L. V. Nguyen. *Mitigating Hallucinations in Large Language Models for Educational Application*, 2024 IEEE International Conference on Consumer Electronics-Asia, ICCE-Asia 2024, Da Nang, Vietnam, 2024.

10. H. Elsayed. The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education, *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, **2024**, *9*(8), 11-23.

11. H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for large language models: A survey, *ACM Transactions on Intelligent Systems and Technology*, **2024**, *15*(2), 1-38.

12. H. Nguyen, H. Q. Cao, K. V. T. Nguyen, and N. D. K. Pham. *Evaluation of explainable artificial intelligence: Shap, lime, and cam*, Proceedings of the 1st FPT AI Conference, FAIC 2021, Hanoi, Vietnam, 2021.

13. H. Thai, M. Nguyen, H. T. T. Nguyen, D. T. H. Vo, B. N. Thanh, K. Nguyen, S. Ha, and T. V. A. Le. Educational technology and responsible automated essay scoring in the generative AI era. *Navigating the Circular Age of a Sustainable Digital Revolution*, IGI Global, New York, 2024.

14. H. Nguyen, L. Nguyen, and H. Cao. XEdgeAI: A human-centered industrial inspection framework with data-centric Explainable Edge AI approach, *Information Fusion*, **2025**, *116*, 102782.

15. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. *Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks*, 2018 IEEE winter conference on applications of computer vision, WACV 2018, Lake Tahoe, NV, USA, 2018.

16. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. *Grad-cam: Visual explanations from deep networks via gradient-based localization*, Proceedings of the 2017 IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 2017.

17. P. X. Nguyen, H. Q. Cao, K. V. Nguyen, H. Nguyen, and T. Yairi. Secam: Tightly accelerate the image explanation via region-based segmentation, *IEICE TRANSACTIONS on Information and Systems*, **2022**, *105*(8), 1401-1417.

18. M. Tornqvist, M. Mahamud, E. M. Guzman, and A. Farazouli. *ExASAG: Explainable framework for automatic short answer grading*, Proceedings of the 18th workshop on innovative use of NLP for building educational applications, BEA 2023, Toronto, Canada, 2023.

19. C. Anghel, M. V. Craciun, E. Pecheanu, A. Cocu, A. A. Anghel, P. Iacobescu, C. Maier, C. A. Andrei, C. Scheau, and S. Dragosloveanu. CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models, *Computers*, **2025**, *14*(10), 431.

20. W. Xu, M. Shahreeza, W. L. Hoo, and W. Yang. Explainable AI for Education: Enhancing Essay Scoring via Rubric-Aligned Chain-of-Thought Prompting, *International Journal of Modern Physics C*, **2025**, *36*, 2542013.

21. Z. Liu, P. Agrawal, S. Singhal, V. Madaan, M. Kumar, and P. K. Verma. LPITutor: an LLM based personalized intelligent tutoring system using RAG and prompt engineering, *PeerJ Computer Science*, **2025**, *11*, e2991.

22. L. P. T. Nguyen, H. T. Do, H. T. T. Nguyen, and H. Cao. *Motion2Meaning:*

A Clinician-Centered Framework for Contestable LLM in Parkinson's Disease Gait Interpretation, 9th International Symposium on Chatbots and Human-centred AI, CONVERSATIONS 2025, Lübeck, Germany, 2025.

23. G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, and F. Toni. *Argumentative Large Language Models for Explainable and Contestable Claim Verification*, Proceedings of the 39th AAAI Conference on Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, 2025.
24. A. Aler Tubella, A. Theodorou, V. Dignum, and L. Michael. *Contestable black boxes*, 4th International Joint Conference on Rules and Reasoning, RuleML+RR 2020, virtual, 2020.
25. K. Alfrink, I. Keller, G. Kortuem, and N. Doorn. Contestable AI by design: Towards a framework, *Minds and Machines*, **2023**, *33*(4), 613-639.
26. T. Ploug and S. Holm. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI, *Artificial intelligence in medicine*, **2020**, *107*, 101901.
27. S. Hong, C. Cai, S. Du, H. Feng, S. Liu, and X. Fan. *"My Grade is Wrong!": A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays*, 26th International Conference on Artificial Intelligence in Education, AIED 2025, Palermo, Italy, 2025.
28. K. S. Geethanjali and N. Umashankar. *Enhancing Educational Outcomes with Explainable AI: Bridging Transparency and Trust in Learning Systems*, 2025 International Conference on Emerging Systems and Intelligent Computing, ESIC 2025, Bhubaneswar, India, 2025.
29. P. Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council, *Regulation (eu)*, **2016**, *679*, 2016.
30. R. Neuwirth. The EU Artificial Intelligence Act, *The EU Artificial Intelligence Act*, **2022**, *106*, 1689.
31. G. M. Sekli, A. Godo, and J. C. Véliz. Generative AI solutions for faculty and students: A review of literature and roadmap for future research, *Journal of Information Technology Education: Research*, **2024**, *23*, 014.
32. C. K. Y. Chan and W. Hu. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education, *International Journal of Educational Technology in Higher Education*, **2023**, *20*(1), 43.
33. E. Dickey and A. Bejarano. *Gaide: A framework for using generative ai to assist in course content development*, 2024 IEEE Frontiers in Education Conference, FIE 2024, Washington DC, USA, 2024.
34. T. Camarata, L. McCoy, R. Rosenberg, K. R. Temprine Grellinger, K. Brettschnieder, and J. Berman. LLM-Generated multiple choice practice quizzes for preclinical medical students, *Advances in physiology education*, **2025**, *49*(3), 758-763.
35. S. Haroud and N. Saqri. Generative ai in higher education: Teachers' and students' perspectives on support, replacement, and digital literacy, *Education Sciences*, **2025**, *15*(4), 396.
36. Y. Zhong and K. Zhao. *Application and Research of Large Language Model in Foreign Language Translation*, 2024 International Conference on Information

Technology, Comunication Ecosystem and Management, ITCEM 2024, Bangkok, Thailand, 2024.

37. C. L. Chen, Y. Dong, C. Castillo-Zambrano, H. Bencheqroun, A. Barwise, A. Hoffman, K. Nalaie, Y. Qiu, O. Boulekbache, and A. S. Niven. A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally, *BMC medical education*, **2025**, *25*(1), 1022.

38. R. Howe, L. Machado, and S. Sneddon. The Ethical Implications of Generative Artificial Intelligence on Students, Academic Staff, and Researchers in Higher Education. *Artificial Intelligence Applications in Higher Education*, Routledge, Oxfordshire, UK, 2024.

39. Q. Zhang, S. B. Siraj, and R. B. Abdul Razak. Effects of AI chatbots on EFL students' critical thinking skills and intrinsic motivation in argumentative writing, *Innovation in Language Learning and Teaching*, **2025**, *1*, 1-29.

40. A. J. Pereira, L. M. Queiros, A. S. Gomes, and T. T. Primo. *Strategies of Intelligent Tutoring Systems to Engage Students in Online Learning Before LLM Approaches*, Simpósio Brasileiro de Sistemas de Informac¸ão, SBSI 2025, Pernambuco, Brazil, 2025.

41. N. Gillani, R. Eynon, C. Chiabaut, and K. Finkel. Unpacking the "Black Box" of AI in education, *Educational Technology & Society*, **2023**, *26*(1), 99-111.

42. H. Nguyen, T. Clement, L. Nguyen, N. Kemmerzell, B. Truong, K. Nguyen, M. Abdelaal, and H. Cao. *LangXAI: integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks*, Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, Korea, 2024.

43. V. B. Truong, H. Nguyen, V. T. K. Nguyen, Q. K. Nguyen, and Q. H. Cao. *Towards Better Explanations for Object Detection*, Proceedings of the 15th Asian Conference on Machine Learning, ACML 2023, Istanbul, Turkey, 2024.

44. K. Nguyen, H. Nguyen, K. Nguyen, B. Truong, T. Phan, and H. Cao. *Efficient and Concise Explanations for Object Detection with Gaussian-Class Activation Mapping Explainer*, Proceedings of the 37th Canadian Conference on Artificial Intelligence, Canadian AI 2024, Guelph, Canada, 2024.

45. O. A. Arise, M. Muzuva, R. Kader, and F. H. Chohan. *Ethical Concerns of Artificial Intelligence in Student Assessments from a Higher Education Perspective*, The Focus Conference, TFC 2024, Durban, South Africa, 2024.

46. Z. N. Khlaif, A. Ayyoub, B. Hamamra, E. Bensalem, M. A. Mitwally, A. Ayyoub, M. K. Hattab, and F. Shadid. University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education, *Education Sciences*, **2024**, *14*(10), 1090.

47. M. M. Van Wyk. Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education, *Journal of applied learning and teaching*, **2024**, *7*(1), 35-45.

48. E. Đerić, D. Frank, and D. Vuković. Exploring the ethical implications of using

generative AI tools in higher education, *Informatics*, **2025**, *12*(2), 36.

49. S. Chambers. GenAI Use and Risks in Higher Education: A Preliminary Review for Research in New Zealand Contexts, *Learning & Teaching 15*, **2025**, *1*, 8.

50. J. Waldo and S. Boussard. GPTs and hallucination: why do large language models hallucinate?, *Queue*, **2024**, *22*(4), 19-33.

51. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Transactions on Information Systems*, **2025**, *43*(2), 1-55.

52. W. H. Walters and E. I. Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT, *Scientific Reports*, **2023**, *13*(1), 14045.

53. V. Petsiuk, A. Das, and K. Saenko. *RISE: Randomized Input Sampling for Explanation of Black-box Models*, British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, 2018.

54. L. P. T. Nguyen, H. T. T. Nguyen, and H. Cao. *Odexai: A comprehensive object detection explainable ai evaluation*, The 48th German Conference on Artificial Intelligence, Ku¨ nstliche Intelligenz 2025, Potsdam, Germany, 2025.

55. H. Nguyen, V. T. K. Van Binh Truong, Q. H. C. Nguyen, and Q. K. Nguyen. Towards Trust of Explainable AI in Thyroid Nodule Diagnosis, *Artificial Intelligence for Personalized Medicine: Promoting Healthy Living and Longevity*, **2023**, *1106*, 11.

56. M. T. Ribeiro, S. Singh, and C. Guestrin. *"Why should i trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD'16, San Francisco, USA, 2016.

57. S. M. Lundberg and S.-I. Lee. *A unified approach to interpreting model predictions*, Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Red Hook, NY, USA, 2017.

58. A. Shrikumar, P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*, The 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, 2017.

59. R. K. Mothilal, A. Sharma, and C. Tan. *Explaining machine learning classifiers through diverse counterfactual explanations*, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT 2020, Barcelona, Spain, 2020.

60. R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. *FACE: feasible and actionable counterfactual explanations*, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES'20, New York, USA, 2020.

61. S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, *ACM Computing Surveys*, **2024**, *56*(12), 1-42.

62. H. Nguyen, A. Rahimi, V. Whitford, H. Fournier, I. Kondratova, R. Richard, and H. Cao. *Human-Centered Explainable Psychiatric Disorder Diagnosis System Using Wearable ECG Monitors*, Proceedings of the 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2025, Sydney, Australia, 2025.

63. M. Koehler and P. Mishra. What is technological pedagogical content knowledge (TPACK)?, *Contemporary issues in technology and teacher education*, **2009**, *9*(1), 60-70.