future internet

IMPACT FACTOR 2.8

CITESCORE 7.1

Review

A Survey on MLLMs in Education: Application and Future Directions

Weicheng Xing, Tianqing Zhu, Jenny Wang and Bo Liu

Special Issue

ICT and AI in Intelligent E-systems

Edited by

Prof. Dr. Ergun Gide and Dr. Robert M. X. Wu

MDPI

https://doi.org/10.3390/fi16120467

future internet

MDPI

Review

A Survey on MLLMs in Education: Application and Future Directions

Weicheng Xing [1], Tianqing Zhu [2], Jenny Wang [3] and Bo Liu [1,*]

[1] School of Computer Science, Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW 2007, Australia; weicheng.xing@student.uts.edu.au
[2] Faculty of Data Science, City University of Macau, Macau 999078, China; tqzhu@cityu.edu.mo
[3] Australia Education Management Group, Melbourn 3001, Australia; jenny.wang@aemg.edu.au
* Correspondence: bo.liu@uts.edu.au

Abstract: This survey paper examines the applications, methodologies, and future prospects of multimodal large language models (MLLMs) within the educational landscape. MLLMs, which integrate multiple data modalities such as text, images, and audio, offer innovative solutions that enhance learning experiences across various educational domains, including language acquisition, STEM education, interactive content creation, and medical training. The paper highlights how MLLMs contribute to improved engagement, personalized learning paths, and enhanced comprehension by leveraging their ability to process and generate contextually relevant content. The key findings underscore the transformative potential of MLLMs in modern education, suggesting significant improvements in both learner outcomes and pedagogical strategies. The paper also explores emerging trends and technological advancements that could shape the future of education, advocating for continued research and collaboration among stakeholders to fully harness the capabilities of MLLMs. As the integration of MLLMs into educational settings progresses, addressing ethical considerations and ensuring equitable access remain critical to maximizing their benefits.

Keywords: multimodal large language models (MLLMs); AI in education; educational technology (EdTech); multimodal integration in learning; computer vision in education

Citation: Xing, W.; Zhu, T.; Wang, J.; Liu, B. A Survey on MLLMs in Education: Application and Future Directions. *Future Internet* **2024**, *16*, 467. https://doi.org/10.3390/fi16120467

Academic Editor: Paolo Bellavista

Received: 14 November 2024
Revised: 6 December 2024
Accepted: 11 December 2024
Published: 13 December 2024

1. Introduction

The integration of technology in education has undergone significant evolution over the past few decades, transforming learning landscapes and pedagogical practices. From early computer-assisted learning programs to sophisticated AI-driven educational applications, technology's impact on education has been profound and multifaceted [1,2]. These innovations have provided educators and learners with unprecedented access to information, personalized learning experiences, and tools that enhance engagement and comprehension.

Recently, artificial intelligence (AI) has emerged as a pivotal force in education, revolutionizing how content is delivered and consumed. AI-driven educational applications, such as adaptive learning systems and intelligent tutoring, have demonstrated the potential to tailor educational experiences to individual learner needs, promoting greater accessibility and inclusivity [3]. The historical context of these technologies highlights their significance in addressing diverse educational challenges, from bridging knowledge gaps to fostering critical thinking skills.

Among the most transformative developments in AI are large language models (LLMs), such as OpenAI's GPT-4o and GPT-o1, which have exhibited remarkable capabilities in understanding and generating human-like text [4]. Trained on vast amounts of data, these models can perform a wide array of language tasks, including translation, summarization, question-answering, and creative writing. The introduction of LLMs into

educational settings has opened new avenues for enhancing teaching and learning processes. They offer personalized feedback, generate educational content, support language learning, and even assist in administrative tasks [1].

Advancing beyond LLMs, multimodal large language models (MLLMs) extend these capabilities by integrating multiple data modalities such as text, images, audio, and video [5]. This integration enables MLLMs to process and generate content across different formats, making them particularly valuable for educational applications that require a holistic understanding of multimodal information. For instance, MLLMs can assist in creating interactive learning materials, facilitating visual explanations of complex concepts, and supporting accessibility for learners with diverse needs [6].

The convergence of LLMs and multimodal processing represents a significant shift in educational technology, promising to further personalize and enrich learning experiences. As these models become increasingly sophisticated and accessible, understanding their potential applications and implications in education becomes crucial.

This survey aims to explore the current state of MLLMs in education, examining their applications, challenges, and future directions. The main contributions of this paper are threefold: (1) an exploration of the current applications of MLLMs in education, highlighting their potential to transform traditional learning methods; (2) an analysis of the challenges and ethical considerations associated with implementing MLLMs in educational environments, such as data privacy and security; and (3) a discussion of emerging trends and future directions for MLLMs in education [7], with recommendations for further research and collaboration.

The overall structure of the survey can be viewed in Figure 1. The Introduction sets the stage by placing multimodal large language models (MLLMs) into an educational context, examining their growing importance in teaching and learning. Preliminaries of MLLMs then offers a foundational exploration of their core components and underlying technologies, clarifying how they integrate multiple modalities. Applications of MLLMs in Education illustrates their current roles in enhancing adaptive learning, providing virtual tutoring, automating content creation, and streamlining educational management systems. Discussion and Future Directions addresses ethical issues, technological hurdles, and the need for improved personalization, engagement, and accessibility, suggesting pathways to strengthen MLLMs' contributions to the educational landscape. Finally, Conclusions synthesizes the key insights and findings, highlighting the paper's main contributions and implications for future practice and research.

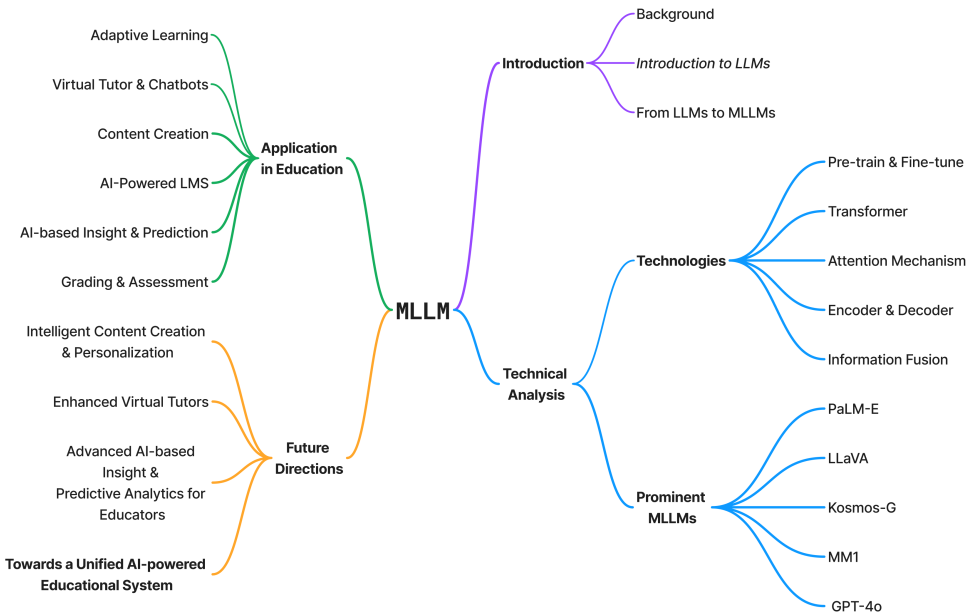

Figure 1. The overall structure of the survey.

2. Preliminaries of MLLMs

2.1. Overview of Large Language Models (LLMs)

Large language models (LLMs) have revolutionized the field of artificial intelligence, particularly in natural language processing (NLP) tasks. These models are designed to understand, generate, and manipulate human language with a high degree of fluency and coherence. Notable examples include OpenAI's GPT-4o and Anthropic's Claude3.5, which are trained on vast amounts of textual data and can perform a wide array of language-related tasks such as translation, summarization, question answering, and content creation [8–10]. LLMs leverage the transformer architecture [11], utilizing self-attention mechanisms to capture complex dependencies in language, enabling them to generate human-like text that is contextually relevant.

The transformer architecture has been pivotal in the success of LLMs. It allows models to process input data in parallel rather than sequentially, significantly improving computational efficiency and performance on large datasets. Self-attention mechanisms within transformers enable the models to weigh the importance of different parts of the input data, capturing long-range dependencies and nuances in language [6]. This capability has led to breakthroughs in tasks that were previously challenging for AI, such as coherent essay writing and intricate dialogue generation.

Despite their impressive capabilities, traditional LLMs are limited to processing textual data [12]. This constraint restricts their applicability in contexts where information is conveyed through multiple modalities, such as images, audio, and video. Human communication and learning are inherently multimodal, often involving the integration of visual cues, auditory signals, and textual information. The inability of LLMs to process non-textual data limits their effectiveness in applications that require a holistic understanding of diverse information sources.

2.2. Evolution to Multimodal Large Language Models (MLLMs)

To address the limitations of traditional LLMs, researchers have developed multimodal large language models (MLLMs), which extend the abilities of LLMs by integrating and processing diverse data modalities alongside text [6,12,13]. MLLMs represent a significant advancement in AI, enabling the fusion of textual and non-textual data to create models that can comprehend and generate contextually relevant outputs across different modalities. This integration allows MLLMs to perform complex tasks that mirror human-like understanding and communication, such as interpreting images with accompanying descriptions or generating detailed visual content from textual prompts.

The development of MLLMs has led to several notable models that demonstrate the seamless integration of multiple modalities. One such model is OpenAI's CLIP (Contrastive Language–Image Pre-training), which learns visual concepts from natural language supervision [14]. CLIP aligns images and textual descriptions in a shared embedding space, allowing it to perform tasks such as image classification, object recognition, and image-text retrieval without explicit task-specific training data. By leveraging a contrastive learning approach, CLIP can understand the relationship between images and text, making it highly versatile in handling multimodal data.

Similarly, DALL·E and DALL·E 2 are models capable of generating high-quality images from textual prompts by utilizing diffusion models conditioned on text embeddings [15,16]. These models showcase the ability to merge text and visual understanding seamlessly, creating novel images that correspond to detailed textual descriptions. This capability has profound implications for creative industries, design, and educational content creation, where visualizations generated from textual concepts can enhance comprehension and engagement.

Another significant development is DeepMind's Flamingo, a visual language model that can perform few-shot learning across multiple modalities [13]. Flamingo integrates vision and language modalities in a single model, enabling it to handle tasks like visual question answering, image captioning, and dialogue with minimal task-specific training. By

leveraging large-scale pre-training and attention mechanisms [17], Flamingo can adapt to new tasks with limited data, highlighting the potential for efficient and scalable deployment of MLLMs in various applications.

In educational contexts, MLLMs support diverse learning needs by facilitating adaptive and inclusive educational experiences. For example, they can generate interactive learning materials, produce visual explanations for complex topics, and provide accessibility features, such as audio descriptions for visual content. The transition from text-based LLMs to MLLMs thus marks a significant shift in educational technology, offering educators and students new opportunities to interact with AI in more dynamic, immersive, and effective ways [18].

As MLLMs continue to evolve, they will further redefine the educational landscape, enabling a deeper level of personalization, engagement, and accessibility in learning. This integration of multiple modalities aligns well with the diverse learning styles and needs of today's students, enhancing both teaching and learning experiences.

2.3. Key Technologies and Architectures

The development of multimodal large language models (MLLMs) relies on several advanced technologies and architectures, each playing a crucial role in enabling these models to process and integrate data from multiple modalities effectively. Figure 2 shows the general model architecture of MLLMs. This subsection highlights the core components, including the transformer architecture, fusion techniques, and specialized training methods that empower MLLMs to generate coherent and contextually relevant multimodal outputs.

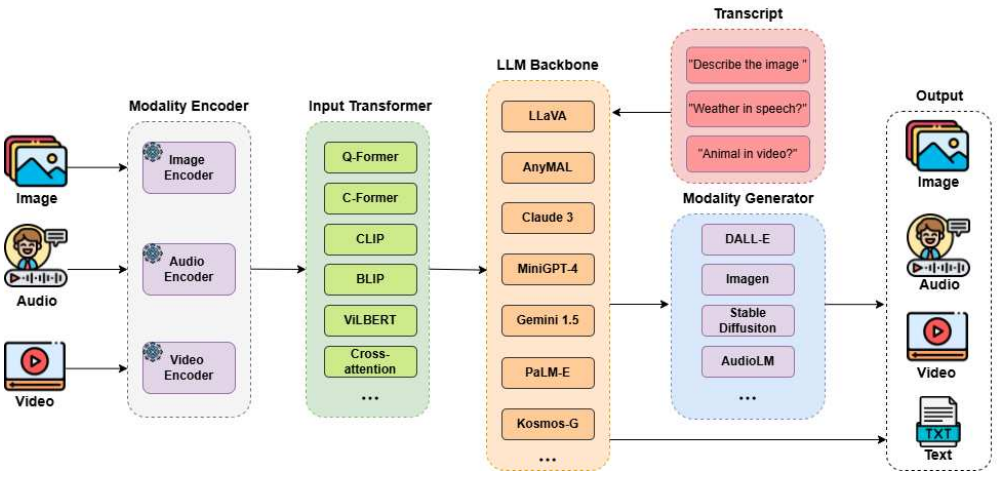

Figure 2. The general model architecture of MLLMs.

2.3.1. Transformer Architectures and Attention Mechanisms

At the core of MLLMs are neural networks based on the transformer architecture [6], which are well suited for handling sequential data and learning complex patterns across different modalities. Transformers use self-attention mechanisms to weigh the significance of different parts of the input data, allowing the model to focus on the most relevant information [19]. Extensions of the transformer architecture have been developed to handle multimodal inputs, such as Vision Transformer (ViT) for images [20] and Audio Transformer for audio data [21].

Attention mechanisms are crucial for MLLMs, enabling them to prioritize certain aspects of input data over others. In multimodal settings, cross-modal attention allows the model to align and fuse information from different modalities [22]. For example, in visual question-answering tasks, the model needs to attend to relevant regions in an image based on a textual question [23].

2.3.2. Multimodal Fusion Techniques

Information fusion is a critical component in MLLMs, enabling the effective combination of data from different modalities. Fusion methods are typically categorized into the following categories:

- **Early fusion (single-stream)**: Integrates modalities at the input level, allowing the model to learn cross-modal representations from the beginning. This method processes all modalities simultaneously, enabling the model to capture interactions between different types of data early in the processing pipeline [6].
- **Late fusion (dual-stream)**: Processes each modality separately before combining their representations at a higher level. This approach allows for specialized processing of each modality, which can be beneficial when modalities have very different characteristics or when pre-trained unimodal models are used [6].

The choice between early and late fusion depends on the specific application and the nature of the data involved. Effective fusion strategies are essential for MLLMs to leverage the complementary information present in different modalities, leading to more robust and accurate models.

2.3.3. Pre-Training and Fine-Tuning

MLLMs often undergo extensive pre-training on large multimodal datasets to acquire a broad understanding of various data types [5,24]. This is followed by fine-tuning on specific tasks or domains to optimize their performance for particular applications [25–27].

2.3.4. Encoder and Decoder Architectures

The architecture of MLLMs often builds upon the transformer framework, extending it to handle multiple modalities. Two primary architectural approaches are used:

- **Encoder-only models**: Models like CLIP focus on creating embeddings for different modalities that can be compared or combined [14]. The encoder processes input data to generate a fixed-size representation, capturing the essential features of the input regardless of its modality. This approach is effective for tasks that require matching or retrieving information across modalities.
- **Encoder–decoder models**: Models used in tasks like image captioning process input data through an encoder and generate outputs via a decoder, allowing for generative tasks [5]. The encoder transforms the input data into a latent representation, which the decoder then uses to generate a sequence of outputs in another modality. This architecture is well suited for tasks that involve translation between modalities, such as generating descriptive text from images.

2.4. Examples of Prominent MLLMs

As shown in Figure 3, several notable multimodal large language models (MLLMs) have emerged in recent years, showcasing the transformative potential of this technology in various fields.

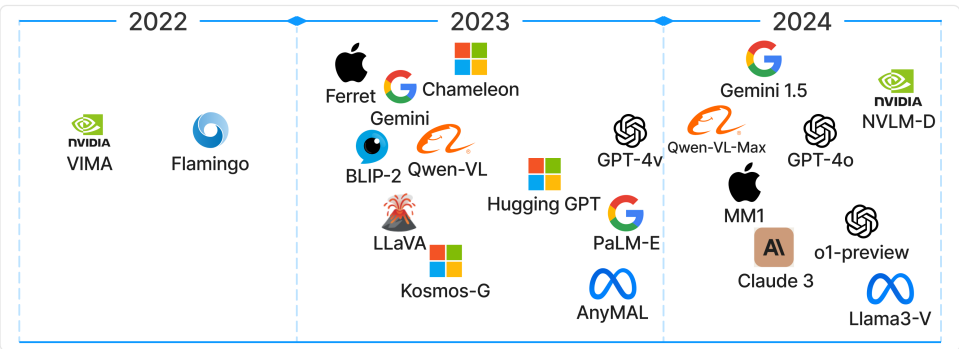

Figure 3. Timeline of major representative MLLMs, starting from 2022.

We selected some of the state-of-the-art MLLMs to analyze their technical details as well as their advantages and disadvantages in application. It is worth noting that we found the open-source MLLM technical information to be more detailed, allowing us to perform a more nuanced analysis. Table 1 provides an overview of some of the most impactful MLLMs and their key features.

Table 1. Overview of selected prominent MLLMs and their educational applications.

Model	Key Capabilities	Educational Applications	Open Source
PaLM-E	Integrates vision and language for robotics and embodied AI.	Enhances interactive, physical learning environments, especially in STEM and robotics education.	No
LLaVA	Combines vision and language for general-purpose understanding.	Visual question answering, image captioning, supporting visually enriched content in learning platforms.	Yes
Kosmos-G	Processes text and images for multimodal comprehension.	Facilitates interactive content, supports collaborative and visual learning tools.	No
GPT-4o	Extends LLM capabilities to visual data, enabling conversational responses to images.	Supports interactive learning with text and visual input, such as image-based Q&A and description generation.	No
MM1	Achieves state-of-the-art performance in multimodal tasks by combining high-resolution visual processing and language models.	Supports tasks like in-context learning, multi-image reasoning, and few-shot learning, useful for assessments and exploratory learning.	No
Llama 3-V	Combines vision, language, coding, reasoning, and tool usage with multilingual support.	Enables adaptive and multilingual learning, coding education, and collaborative projects.	Yes
NVLM 1.0	Frontier-class multimodal model with exceptional vision-language reasoning and text-only improvements.	Supports OCR tasks, multimodal math reasoning, and document analysis in educational environments.	Yes
BLIP	Pre-trained for text-image retrieval and multimodal content generation.	Facilitates creative content development and storytelling in visual education.	Yes

2.4.1. Open-Source MLLMs

NVLM 1.0: Developed by NVIDIA, NVLM 1.0 is a frontier-class family of multimodal large language models designed to excel in vision-language tasks while maintaining or even improving text-only performance [28]. NVLM introduces three architectural variants: the decoder-only NVLM-D, the cross-attention-based NVLM-X, and the hybrid NVLM-H, each catering to different multimodal processing needs. With model sizes up to 72 billion parameters, NVLM combines state-of-the-art performance with high flexibility for handling diverse multimodal inputs.

NVLM's training process incorporates a meticulously curated dataset blend, including high-quality multimodal and text-only supervised fine-tuning (SFT) datasets. This approach preserves text-only capabilities while enhancing vision-language performance. NVLM-D processes multimodal input directly within the LLM, achieving unified reasoning capabilities. NVLM-X employs gated cross-attention for efficient high-resolution image processing, while NVLM-H combines elements of both architectures, providing superior reasoning and computational efficiency.

One of NVLM 1.0's key strengths is its performance on diverse benchmarks, where it rivals or surpasses proprietary models like GPT-4o and Claude 3.5. For instance, NVLM-D demonstrates exceptional OCR capabilities, achieving the highest scores on benchmarks such as OCRBench and VQAv2. NVLM-H leads in multimodal reasoning tasks like MathVista and multidisciplinary reasoning, proving its versatility for complex educational applications.

Despite these advancements, NVLM models also present challenges. The large-scale models require substantial computational resources for both training and deployment,

which may limit accessibility for smaller institutions. Additionally, while NVIDIA plans to open-source the training code and model weights, the complexity of implementation might pose challenges for non-specialist users without significant expertise or infrastructure.

NVLM 1.0's public release is a significant contribution to the open-source ecosystem, offering powerful tools for research and development in multimodal AI. Its focus on maintaining text-only performance during multimodal training sets a new standard, enabling its integration into educational technologies without compromising existing capabilities. By providing production-grade multimodality and robust benchmarks, NVLM supports the creation of advanced applications across various fields, including education, healthcare, and content generation.

Potential applications: NVLM 1.0 can be utilized to build OCR-based educational tools, multimodal math reasoning assistants, and adaptive content generation systems. For example, its superior OCR capabilities can support the digitization of historical documents for research purposes, while its multimodal reasoning abilities can enhance STEM education with visual problem-solving tasks.

Llama 3-V: Llama 3-V is the latest addition to Meta's family of open-source foundation models, featuring a herd of models designed to natively support multilinguality, coding, reasoning, and tool usage. The largest variant of Llama 3-V is a dense transformer model with 405 billion parameters and a context window of up to 128 K tokens [29]. The models are trained on an unprecedented corpus of 15 trillion multilingual tokens, enabling them to achieve state-of-the-art performance across a wide range of language understanding tasks. The Llama 3-V models, available under an updated open-source license, include pre-trained and post-trained versions, with the flagship Llama 3-V.1 model delivering competitive quality compared to GPT-4.

One of the major advantages of Llama 3-V lies in its versatility and performance. By supporting multilinguality and tool usage, it offers immense flexibility for global educational applications, including language instruction, cross-cultural studies, and adaptive learning environments. Its extended context window makes it particularly well suited for tasks requiring in-depth reasoning, such as literature analysis, historical document interpretation, and coding exercises. Additionally, Meta's emphasis on safety, demonstrated through the Llama Guard 3 model for input and output safety, ensures that the models can be deployed responsibly in educational contexts.

Despite its strengths, Llama 3-V presents challenges. While being open-source, its large-scale models like the 405B parameter variant require substantial computational resources, making them difficult to deploy for institutions with limited infrastructure. The extensive training and fine-tuning processes, although yielding exceptional results, also necessitate expertise and investment in hardware and personnel. Moreover, while Meta has improved pre-training data quality and post-training alignment processes, the scale of Llama 3-V may raise concerns about data privacy and ethical implications in specific use cases.

Llama 3-V's compositional approach to multimodality—integrating image, video, and speech recognition capabilities—shows promising results that rival state-of-the-art models in these domains. Although these multimodal extensions are still under development, they highlight the potential of Llama 3-V to transform education by enabling innovative applications such as visual storytelling, multilingual video-based lessons, and speech-assisted learning tools. The commitment to public release ensures transparency and fosters a collaborative research ecosystem to enhance the model further.

Overall, Llama 3-V represents a significant milestone in the evolution of open-source MLLMs, demonstrating the potential of large-scale, multimodal models to advance global education and research. Its balance between performance, openness, and safety positions it as a valuable asset for institutions aiming to integrate cutting-edge AI technologies into their educational strategies.

Potential applications: Llama 3-V can enable visual storytelling, multilingual video-based lessons, and adaptive learning platforms. Its large context window is particularly

useful for creating tools that assist in complex problem solving, such as code debugging environments or interactive literature exploration.

LLaVA (Large Language and Vision Assistant): LLaVA is an open-source model designed to combine visual encoders with large language models, enabling general-purpose visual and language understanding [30]. It supports tasks such as visual question answering, image captioning, and the development of interactive, multimodal educational content. One major advantage of LLaVA is its accessibility and adaptability. Being open-source, it allows educators and developers to customize the model to address specific learning needs, such as creating tailored visual aids or interactive lessons. Furthermore, it fosters innovation by providing a platform for experimentation and collaboration within the educational community.

LLaVA also poses significant challenges. The implementation of the model requires considerable technical expertise and computational resources, which may not be readily available to many educational institutions. Furthermore, the absence of dedicated support services means that troubleshooting and model optimization must often be handled internally, potentially delaying adoption and reducing its usability in time-sensitive scenarios.

Potential applications: LLaVA can support the creation of interactive learning environments, such as visual science tutorials or personalized art history lessons. Its adaptability makes it ideal for generating tailored visual aids for diverse subjects.

2.4.2. Proprietary MLLMs

GPT-4o (Multimodal Version): GPT-4o, developed by OpenAI, integrates visual and textual data for advanced multimodal reasoning. Its applications in education are vast, ranging from interactive learning environments to visual question answering and conversational image interpretation. One of its primary advantages is its exceptional performance and reliability [31]. The model's robust capabilities ensure accurate and consistent results across a wide range of tasks, making it a reliable tool for educators. Moreover, the professional support and regular updates provided by OpenAI enhance its usability and stability in educational settings, ensuring that institutions can deploy the model with confidence.

GPT-4o's proprietary nature introduces challenges. The high licensing fees can be prohibitive for many educational institutions, particularly those operating on limited budgets. Additionally, its closed design restricts customization, preventing institutions from tailoring the model to specific educational contexts or integrating it with existing systems beyond the predefined parameters.

Potential applications: GPT-4o can be used to create advanced virtual tutors capable of visual and textual reasoning. For example, it could support biology lessons with real-time microscopic image analysis or assist in art classes with detailed image composition feedback.

Kosmos-G: Developed by Microsoft, Kosmos-G offers enhanced multimodal capabilities, including advanced visual and text comprehension. It is well suited for collaborative and visual learning tools, integrating seamlessly with Microsoft's ecosystem, such as Teams and Office 365, to provide a unified platform for education [32]. This integration simplifies its adoption and enhances user experience, particularly for institutions already utilizing Microsoft products.

Nevertheless, Kosmos-G's advantages are accompanied by significant limitations. The model's cost can be a major barrier for adoption, as it requires both licensing fees and infrastructure investment. Moreover, its reliance on Microsoft's ecosystem may limit its flexibility, making it less suitable for institutions that use diverse or non-Microsoft technologies.

Potential applications: Kosmos-G can enhance educational platforms like Teams by integrating multimodal features, such as document summarization and diagram understanding, making collaborative projects more efficient.

MM1: MM1 is a family of multimodal large language models (MLLMs) developed by Apple, designed to achieve state-of-the-art performance in multimodal tasks by carefully analyzing architectural components and data choices [33]. MM1 includes both dense models with up to 30 billion parameters and mixture-of-experts (MoE) models with up to 64 billion parameters. Through comprehensive ablation studies, MM1 identifies crucial design principles for integrating visual and textual data, such as the importance of image resolution, visual encoder capacity, and the mix of pre-training data. The models exhibit enhanced capabilities like in-context learning, multi-image reasoning, and few-shot chain-of-thought prompting. MM1 demonstrates competitive performance across a range of established multimodal benchmarks, highlighting the effectiveness of large-scale multimodal pre-training and offering valuable insights for future MLLM development.

MM1's proprietary status limits its accessibility. The high costs associated with its deployment, both in licensing and required infrastructure, restrict its use to well-funded institutions. Additionally, its closed system design precludes customization, which may limit its adaptability for diverse educational needs.

Potential applications: MM1 can support advanced assessments, such as AI-driven exam proctoring or complex data interpretation tasks in STEM fields, making it a valuable tool for high-level education.

3. Applications of MLLMs in Education

The advent of multimodal large language models (MLLMs) has ushered in a new era in educational technology, offering innovative solutions that enhance both teaching and learning experiences. Using the ability to process and generate content across multiple modalities, such as text, images, audio, and video, MLLMs are transforming the educational landscape [34]. This section explores the diverse applications of MLLMs in education, highlighting how these advanced AI models are being integrated into various educational tools and platforms to address longstanding challenges and meet the evolving needs of learners and educators.

Table 2 summarizes several key areas where MLLMs are making a significant impact:

- **Adaptive learning platforms**: Examining how MLLMs enable personalized learning experiences by dynamically adjusting instructional content to meet individual learners' needs, preferences, and performance levels.
- **Virtual tutors and chatbots**: Exploring the role of MLLMs in developing intelligent virtual assistants that provide personalized support, guidance, and interactive learning opportunities through natural language conversations.
- **Intelligent content creation**: Investigating how MLLMs automate and enhance the development of educational materials, including textbooks, lesson plans, assessments, and multimedia resources, thereby increasing efficiency and accessibility.
- **AI-powered learning management systems (LMSs)**: Analyzing the integration of MLLMs into LMS platforms to enhance content delivery, personalize learning paths, and facilitate more engaging and interactive educational experiences.
- **AI-based insight and predictive analytics for educators**: Discussing how MLLMs provide educators with actionable insights by analyzing multimodal educational data, enabling early identification of at-risk students and informed decision making.
- **Grading and assessment tools**: Assessing the application of MLLMs in automating grading processes, providing objective evaluations, and delivering detailed, personalized feedback across various types of student work.

Through these explorations, the section aims to demonstrate the transformative potential of MLLMs in education. By presenting case studies, technological insights, and practical applications, we seek to highlight how MLLMs not only enhance learning outcomes and engagement but also support educators in delivering high-quality, personalized instruction. The objective is to provide a comprehensive understanding of how MLLMs are shaping the future of education and to inspire further research and implementation in this promising field.

Table 2. Applications of MLLMs in education.

Application Area	Technologies Used	Examples of Applications	Case Study
Adaptive Learning Platforms	MLLMs with transformer architectures (ViT, AST)	Cognii, Carnegie Learning's MATHia, Knewton, Duolingo, Smart Sparrow	Integration of MLLMs in Duolingo
Virtual Tutors and Chatbots	Transformer-based language models (GPT-3, GPT-4), VisualGPT, DeepSpeech, Tacotron 2, ADS	Squirrel AI Learning, Duolingo, Watson Tutor, Woebot [35]	Implementation of LOVA^3 in Virtual Tutoring Systems
Intelligent Content Creation	Transformer-based multimodal models (GPT-4, BLIP-2), DALL·E and DALL·E 2, NLG (GPT-3, T5)	Automated textbook generation, quiz and assessment generation, interactive simulations, multimedia content creation, language translation and localization	Automated Quiz Generation Using GPT-3
AI-powered LMS	GPT-4 and CLIP, natural language interfaces, Wav2Vec 2.0, Tacotron 2	Coursera, edX, Udemy, Knewton's Alta	Integration of MLLMs in Coursera
Insight and Predictive Analytics	GPT-4, CLIP, computer vision models (OpenFace), speech and audio processing models (Wav2Vec 2.0)	Early warning systems, sentiment and emotion analysis, adaptive feedback generation, curriculum and instructional design insights, collaborative skills assessment	Early Warning System Using MLLMs
Grading and Assessment Tools	NLP and NLU (GPT-4, BERT), computer vision and image recognition (CLIP, ViT), speech and audio processing (Wav2Vec 2.0)	E-Rater by ETS, MOSS, CodeRunner, Duolingo English Test, tools for multimodal assignment evaluation	Reducing the Cost of Short-Answer Scoring with MLLMs

The table summarizes key applications of MLLMs in education.

3.1. Adaptive Learning Platforms

3.1.1. Introduction to Adaptive Learning Platforms

Adaptive learning platforms are educational systems designed to deliver personalized learning experiences by dynamically adjusting instructional content based on individual learners' needs, preferences, and performance [36,37]. These platforms leverage data analytics and artificial intelligence to monitor learner interactions, assess understanding, and tailor content to optimize engagement and learning outcomes [38]. By providing customized pathways, adaptive learning aims to address the diverse abilities and learning styles present in educational settings.

The advent of multimodal large language models (MLLMs) has significantly enhanced the capabilities of adaptive learning platforms. MLLMs enable the integration of various data modalities—such as text, images, audio, and video—allowing for richer, more interactive learning experiences [12]. This multimodal approach aligns with the understanding that learning is a complex process involving multiple senses and cognitive functions [39].

3.1.2. Technology Used: MLLMs in Adaptive Learning

The incorporation of MLLMs into adaptive learning platforms represents a convergence of advanced AI technologies and educational methodologies. MLLMs utilize transformer-based architectures [11], which employ self-attention mechanisms to process sequential and non-sequential data across different modalities [19]. Extensions like Vision Transformer (ViT) [20] and Audio Spectrogram Transformer (AST) [20] enable processing of visual and auditory data, respectively.

Techniques for integrating data from multiple modalities are crucial. Models like CLIP [14] and UNITER [24] align visual and textual representations in a shared embedding space, facilitating cross-modal understanding. Adaptive platforms use algorithms that analyze learner data to adjust content delivery. MLLMs enhance this by interpreting complex multimodal inputs, such as facial expressions, speech patterns, and handwriting [5].

Moreover, MLLMs can generate contextually relevant feedback and explanations in natural language, improving learner comprehension and engagement [8]. By leveraging these technologies, adaptive learning platforms can offer more nuanced and effective personalization, catering to individual learner profiles.

3.1.3. Examples of Applications

Several educational platforms and research initiatives have begun integrating MLLMs into adaptive learning systems. *Cognii* [40] is an AI-based virtual learning assistant that uses NLP and MLLMs to provide personalized tutoring and open-response assessments in natural language. *Carnegie Learning's MATHia* [41] incorporates AI to adapt mathematics instruction, potentially enhanced by MLLMs to interpret handwritten mathematical expressions and provide step-by-step feedback. *Knewton* [42] is an adaptive learning technology that could utilize MLLMs to process multimodal learner data, such as interaction patterns and facial expressions, to adjust instructional strategies.

Duolingo [43,44], a language learning app, employs AI for personalized lesson plans. The integration of MLLMs allows for speech recognition and dialogue practice, enhancing language acquisition through multimodal interactions. *Smart Sparrow* [45] is an adaptive e-learning platform that enables the creation of interactive and adaptive courseware, potentially leveraging MLLMs for richer content delivery and learner analytics.

3.1.4. Case Study: Integration of MLLMs in Duolingo

Duolingo, a widely used language learning platform, provides an illustrative case of integrating MLLMs into adaptive learning. Aiming to make language learning accessible and engaging through gamified lessons and personalized learning paths [44,46], the platform adapts to learners' proficiency levels, focusing on areas that need improvement.

Duolingo incorporates MLLMs to enhance its adaptive capabilities. It uses speech recognition and synthesis technologies, employing models such as WaveNet [47] for high-quality speech synthesis and integrating speech recognition to assess pronunciation and fluency. Using MLLM, Duolingo generates and interprets textual, auditory, and visual content, providing a comprehensive learning experience [43]. It uses data from the students across the modalities to tailor the difficulty of the lesson, the types of content, and the feedback [48].

For example, MLLMs enable Duolingo to create interactive exercises that involve listening, speaking, reading, and writing, catering to different learning styles. By processing speech and written input, the platform provides immediate, personalized feedback, helping learners correct errors in real time. Analyzing learner performance allows the platform to adjust the sequence of lessons and introduce new vocabulary or grammar concepts when appropriate.

A study examined how Duolingo leverages AI and MLLM to improve language learning [43]. The findings indicated improved engagement: the integration of multimodal exercises increased user engagement and time spent on the platform. The learners showed enhanced proficiency gains due to personalized feedback and adaptive content. MLLMs allowed Duolingo to generate content across multiple languages efficiently, supporting a diverse user base.

However, challenges such as data privacy emerged, as handling multimodal learner data raised concerns about privacy and data security. In addition, resource intensity was a concern, as training and deployment of MLLMs require significant computational resources.

3.2. Virtual Tutors and Chatbots

3.2.1. Introduction to Virtual Tutors and Chatbots

Virtual tutors and chatbots are AI-driven systems designed to simulate human tutoring and conversational interactions in educational settings. They provide learners with personalized assistance, feedback, and guidance, often available on-demand and scalable to large numbers of users [49]. These systems aim to enhance learning experiences by offering interactive dialogues, answering queries, and adapting to individual learner needs.

The integration of multimodal large language models (MLLMs) has significantly advanced the capabilities of virtual tutors and chatbots. By leveraging MLLMs, these systems can process and generate content across multiple modalities—such as text, speech, and images—enabling richer interactions and more effective communication [50]. This multimodal capability aligns with the diverse ways in which learners perceive and engage with educational content.

Recent advancements in MLLMs, such as the development of LOVA^3 (*Learning to Visual Question Answering, Asking, and Assessment*) [51], have further enhanced the potential of virtual tutors and chatbots. LOVA^3 equips MLLMs with additional capabilities beyond traditional question answering, enabling them to ask questions and assess responses. This aligns closely with human learning mechanisms and supports deeper multimodal understanding, making virtual tutors and chatbots more effective in educational contexts.

3.2.2. Technology Used: MLLMs in Virtual Tutors and Chatbots

The deployment of MLLMs in virtual tutors and chatbots involves several key technologies. Transformer-based language models, such as GPT-4 [4,9], serve as the foundation for generating coherent and contextually relevant textual responses. Their capacity to understand and produce human-like language makes them suitable for conversational agents.

Multimodal integration allows MLLMs to extend traditional language models by incorporating additional modalities. For example, models like LOVA^3 [51] integrate visual information into language understanding and generation, enabling chatbots to reference and generate content based on images. This is particularly valuable in educational settings where visual content is integral to learning.

Advanced dialogue systems utilize MLLMs to manage context, maintain conversation flow, and handle multi-turn dialogues [52]. These systems enable virtual tutors to provide coherent and contextually appropriate responses over extended interactions. LOVA^3, in particular, enhances these capabilities by introducing tasks that foster the skills of asking and assessing questions in the context of images, thereby enriching the interactive experience.

Personalization and adaptive learning are enhanced by MLLMs, which can analyze user inputs and learning patterns to tailor responses and instructional strategies [53]. The ability to ask and assess questions, as enabled by frameworks like LOVA^3, allows virtual tutors to engage learners more effectively, encouraging deeper engagement with the material and fostering critical thinking skills.

3.2.3. Examples of Applications

Several applications demonstrate the integration of MLLMs into virtual tutors and chatbots. Educational platforms are increasingly incorporating advanced MLLMs to provide personalized learning experiences. For instance, LOVA^3's capabilities can be leveraged to develop virtual tutors that not only answer student queries but also pose relevant questions and assess student responses, thereby simulating a more interactive and engaging learning environment.

Language learning assistants can benefit from MLLMs that handle multiple modalities. By integrating LOVA^3's approach, these assistants can incorporate visual prompts and assessments, enhancing language acquisition through multimodal interactions.

In STEM education, virtual tutors equipped with MLLMs like LOVA^3 can assist students in subjects like mathematics and science by providing detailed explanations,

posing challenging problems, and assessing student solutions. This mirrors the human tutoring process more closely and can lead to better learning outcomes.

3.2.4. Case Study: Implementation of LOVA^3 in Virtual Tutoring Systems

The implementation of LOVA^3 in virtual tutoring systems offers valuable insights into how MLLMs can enhance educational experiences. LOVA^3 is an innovative framework designed to equip MLLMs with the abilities to answer, ask, and assess questions in the context of images [51]. This triad of skills aligns with human learning mechanisms and is crucial for understanding the world and acquiring knowledge.

Technologically, LOVA^3 utilizes two supplementary training tasks, GenQA and EvalQA, to foster the skills of asking and assessing questions. GenQA focuses on enabling the model to generate diverse question–answer pairs from a single input image, thereby equipping the MLLM with the capability to ask questions. This ability encourages learners to engage more deeply with information, enhancing problem-solving skills. EvalQA involves tasking the MLLM to predict the correctness of a given visual–question–answer triplet. This assessment capability allows virtual tutors to evaluate student responses, provide feedback, and guide learners toward a deeper understanding of the material.

The model architecture of LOVA^3 integrates a vision encoder with a large language model through a simple MLP adapter, allowing the system to process and generate multimodal content efficiently. Training involves a mixture of tasks, including traditional visual question answering (VQA), GenQA, and EvalQA, enhancing the model's comprehensive understanding and interactive capabilities.

In virtual tutoring systems, implementing LOVA^3 can lead to several benefits. By enabling the virtual tutor to ask questions and assess responses, the interaction becomes more dynamic and engaging, resembling a human tutor's approach. The tutor's ability to generate questions and assess answers encourages students to think critically and engage more deeply with the content. Incorporating visual elements into questioning and assessment aligns with diverse learning styles and can improve comprehension.

For example, a virtual tutor using LOVA^3 can present an image related to a biology lesson and generate questions that prompt the student to identify structures, explain functions, or predict outcomes. The tutor can then assess the student's responses, provide feedback, and adjust subsequent interactions based on the student's understanding.

A study by Zhao et al. [54] demonstrated that training MLLMs using the LOVA^3 framework improved performance across various multimodal datasets and benchmarks. The results underscored the critical role of these additional tasks in fostering comprehensive intelligence in MLLMs. The model showed consistent performance gains, highlighting its effectiveness in enhancing multimodal comprehension.

While LOVA^3 enhances virtual tutoring systems significantly, challenges exist. Handling a wide range of topics and queries requires extensive training data. LOVA^3 addresses this by compiling a comprehensive set of multimodal foundational tasks, but scaling this to cover all educational content remains a challenge. Ethical considerations are paramount: privacy, data security, and ensuring appropriate responses are essential, especially when dealing with sensitive information. Educational virtual tutors need to handle complex and varied content, and ensuring the accuracy and relevance of generated questions and assessments requires continuous updates and validations of the model.

The implementation of LOVA^3 in virtual tutors showcases the potential for delivering personalized and effective educational experiences. It highlights how advancements in MLLMs can address the complexities of content and provide scalable solutions that adapt to individual learner needs. By equipping virtual tutors with the abilities to answer, ask, and assess, LOVA^3 brings AI-driven education closer to the nuanced and interactive nature of human tutoring.

3.3. Intelligent Content Creation

3.3.1. Introduction to Intelligent Content Creation

Intelligent content creation refers to the utilization of artificial intelligence, particularly multimodal large language models (MLLMs), to automate and enhance the development of educational materials. This includes generating textbooks, lesson plans, assessments, interactive simulations, and multimedia resources tailored to specific learning objectives and individual learner needs [55]. By leveraging MLLMs, educators and content developers can produce high-quality, engaging, and personalized educational content more efficiently and effectively.

Traditional content creation in education is often time consuming and requires significant expertise. Intelligent content creation addresses these challenges by automating parts of the content development process and enabling the creation of materials that can adapt to diverse learning styles and preferences [1]. The integration of multiple modalities—text, images, audio, and video—enhances the accessibility and inclusivity of educational resources, catering to a wider range of learners.

3.3.2. Technology Used: MLLMs in Intelligent Content Creation

The implementation of MLLMs in intelligent content creation involves several key technologies. Transformer-based multimodal models, such as GPT-4 [9] and BLIP-2 [5], process and generate content across multiple modalities. GPT-4, for example, accepts both text and image inputs, enabling the generation of rich, contextually relevant content that combines textual explanations with visual elements.

Image generation models like DALL·E and DALL·E 2 [15,16] can generate high-quality images from textual descriptions. These models facilitate the creation of visual content such as diagrams and illustrations, enhancing comprehension of complex concepts.

Natural language generation (NLG) is another crucial technology. Advanced language models like GPT-4 [9] and T5 [56] generate coherent and contextually appropriate textual content, including explanations, summaries, and questions.

MLLMs also enable multimodal content synthesis by integrating various data types to create interactive and multimedia educational materials. Models that combine audio and visual data can generate instructional videos or interactive simulations [13].

Adaptive content personalization is achieved by analyzing learner data, allowing MLLMs to generate content personalized to the learner's proficiency level, interests, and learning style [57]. This involves adapting the difficulty, presentation style, and content focus to suit individual needs.

3.3.3. Examples of Applications

Several applications demonstrate the use of MLLMs in intelligent content creation. *Automated textbook generation* utilizes AI systems to summarize and organize information from various sources to generate textbooks and study materials. Models like GPT-3 can produce explanatory text on a given topic, compiled into educational resources.

Quiz and assessment generation involves MLLMs creating assessment items such as multiple-choice questions, short answers, and problem-solving exercises tailored to specific learning objectives [58]. This automation supports educators in developing formative and summative assessments efficiently.

Interactive simulations and virtual labs are created by integrating textual descriptions with visual and interactive elements. MLLMs help create virtual laboratory environments and simulations, allowing students to explore concepts hands-on [59].

Multimedia content creation is enhanced by MLLMs enabling the creation of educational videos, animations, and presentations by generating scripts, visual content, and even voice-overs [60].

Language translation and localization are facilitated by AI models that translate educational content into multiple languages, ensuring accessibility for non-native speakers and supporting multilingual education [61].

3.3.4. Case Study: Automated Quiz Generation Using GPT-3

Assessments are crucial for evaluating learner understanding and progress. Creating high-quality assessment items is resource intensive and requires subject matter expertise. Automated quiz generation using MLLMs like GPT-3 offers a solution by generating diverse and contextually appropriate questions based on educational content [58].

GPT-3 is leveraged for its advanced language generation capabilities to create questions and answers. It can generate various types of questions, including multiple-choice, true/false, and open-ended formats. *Prompt engineering* involves crafting specific prompts to guide GPT-3 in generating questions that align with learning objectives and appropriate difficulty levels. *Content filtering and quality assurance* algorithms evaluate the generated questions for correctness, relevance, and potential biases, ensuring the quality of assessments.

Educational platforms integrate automated quiz generation to provide immediate assessments after lessons, enhancing engagement and retention. Adaptive learning systems generate quizzes tailored to learner performance, allowing the system to adjust the difficulty and focus of subsequent content. Teacher support tools enable educators to use AI-generated quizzes as a starting point, saving time in test preparation and allowing them to focus on instructional design.

A study by Lu et al. (2022) investigated the use of GPT-3 for automated question generation in educational settings [58]. The findings indicated that GPT-3 generated questions of acceptable quality, covering key concepts with varying difficulty levels. Automation significantly reduced the time required to create assessments, and the model allowed for generating questions tailored to specific topics and learning outcomes.

Challenges included content accuracy, as some generated questions contained inaccuracies or ambiguities, necessitating human review. Bias and fairness were also concerns, as the model occasionally produced content reflecting biases present in the training data, highlighting the need for oversight. Ensuring questions are directly relevant to specific instructional content remains a challenge, emphasizing the importance of contextual alignment.

MLLMs like GPT-3 have the potential to automate educational assessment creation. While offering significant efficiencies, human oversight is essential to ensure accuracy, fairness, and alignment with educational objectives. Ongoing advancements in MLLMs are expected to improve the quality and reliability of automated content creation.

3.4. AI-Powered Learning Management Systems (LMSs)

3.4.1. Introduction to AI-Powered Learning Management Systems

AI-powered learning management systems (LMSs) are platforms that incorporate artificial intelligence technologies, including multimodal large language models (MLLMs), to enhance the delivery, management, and personalization of educational content [62]. Traditional LMS platforms facilitate the administration, documentation, tracking, reporting, and delivery of educational courses or training programs. The integration of AI transforms these systems into intelligent platforms capable of providing personalized learning experiences, adaptive content, and interactive engagement [63].

MLLMs enable AI-powered LMSs to process and generate content across multiple modalities such as text, images, audio, and video. This multimodal capability allows for richer, more engaging educational experiences that cater to diverse learning styles and needs [12]. By analyzing vast amounts of learner data, AI-powered LMSs can adapt to individual learners' progress, preferences, and performance, thereby enhancing the overall effectiveness of the educational process [55].

3.4.2. Technology Used: MLLMs in AI-Powered LMSs

The integration of MLLMs into LMS platforms involves several key technologies. MLLMs like GPT-4 [9] and CLIP [14] can process and understand content in various formats, enabling the LMS to handle textual, visual, and auditory data effectively. Advanced language models facilitate interactions between the LMS and users through natural lan-

guage interfaces, allowing for conversational queries, personalized feedback, and adaptive content delivery [8]. AI algorithms analyze learner behaviors and preferences to recommend tailored content, with MLLMs enhancing this by interpreting complex patterns in multimodal data [64]. Furthermore, MLLMs can generate and grade assessments, provide detailed explanations, and offer suggestions for improvement across multiple modalities. Incorporating models like Wav2Vec 2.0 [65] and Tacotron 2 [66] allows the LMS to support voice interactions and auditory content delivery.

3.4.3. Examples of Applications

Several AI-powered LMS platforms and initiatives demonstrate the application of MLLMs in education. *Coursera* is an online learning platform that offers massive open online courses (MOOCs), specializations, and degrees [67]. Coursera utilizes AI to personalize course recommendations and enhance the learning experience [68]. Similarly, *edX* incorporates AI to provide adaptive learning experiences, including personalized content and assessments [69]. *Udemy* uses AI algorithms to analyze learner data and provide course recommendations tailored to individual interests and needs [70]. *Knewton's Alta* is an AI-powered adaptive learning platform that uses data analytics and MLLMs to personalize learning paths and provide real-time feedback [71].

3.4.4. Case Study: Integration of MLLMs in Coursera

Coursera is one of the world's leading online learning platforms, partnering with universities and organizations to offer courses, specializations, certificates, and degree programs [72]. The platform serves millions of learners globally, providing access to a wide range of educational content. Coursera has integrated AI technologies to enhance its platform, aiming to improve learner engagement, personalize learning experiences, and increase the effectiveness of online education [68].

Coursera leverages various AI technologies, potentially including MLLMs, to enhance its platform. Natural language processing (NLP) is used for processing course content, generating summaries, and facilitating search and recommendation systems [68]. Machine learning algorithms analyze learner data to personalize course recommendations and adapt content delivery based on individual performance and preferences [73]. AI algorithms provide instant feedback on assignments, particularly in programming and technical courses [74]. While the specific use of MLLMs is not publicly detailed, Coursera handles various content types, including video lectures, readings, quizzes, and interactive assignments.

Coursera uses AI to recommend courses and content to learners based on their interests, past activity, and learning goals [64]. AI-powered translation services enable course content to be accessible in multiple languages, enhancing inclusivity [75]. For certain subjects like computer science, AI algorithms grade assignments and provide feedback, improving efficiency and scalability [74]. Additionally, AI analyzes learner progress to suggest skill improvements and potential career paths [68].

Coursera's AI-driven recommendation system increases learner engagement by suggesting relevant courses and content [64]. AI enables Coursera to manage a vast number of learners and courses, providing consistent quality of education at scale [68]. Personalized learning paths and instant feedback contribute to better learner performance and satisfaction [73]. Language translation and multimodal content delivery make education accessible to a global audience with diverse needs [75].

Challenges include data privacy and security: handling sensitive learner data requires robust security measures and compliance with regulations such as GDPR [76]. Ensuring that AI algorithms do not perpetuate biases is critical, and models must be trained on diverse datasets to mitigate this risk [77]. Learners and educators may require understanding of how AI makes recommendations or grades assessments, highlighting the need for transparency and explainability [78].

While specific details on Coursera's use of MLLMs are not publicly disclosed, the potential integration of MLLMs could enhance several aspects of the platform. MLLMs can analyze multimodal learner data to provide more accurate and personalized content recommendations. Incorporating models like GPT-4 [9] could enable conversational interfaces for learner support and engagement. MLLMs could assist in generating and curating content across text, images, and videos, enriching the learning materials available. Moreover, multimodal translation and content adaptation can make courses more accessible to learners with disabilities or those speaking different languages.

Coursera's integration of AI technologies demonstrates the significant impact of AI-powered LMS platforms in scaling education and personalizing learning experiences. While explicit use of MLLMs is not detailed, the potential for incorporating such technologies offers avenues for further enhancing the platform's capabilities. Addressing challenges related to data privacy, bias, and transparency is essential for the ethical and effective implementation of AI in education.

3.5. AI-Based Insight and Predictive Analytics for Educators

3.5.1. Introduction to AI-Based Insight and Predictive Analytics for Educators

AI-based insight and predictive analytics involve the use of artificial intelligence algorithms and machine learning models to analyze educational data, extract meaningful insights, and predict future trends or learner outcomes [79]. These tools empower educators by providing data-driven decision support, enabling them to identify at-risk students, personalize instruction, and improve curriculum design. The integration of multimodal large language models (MLLMs) enhances these capabilities by processing and interpreting data from multiple modalities—such as text, images, audio, and video—offering a more comprehensive understanding of learner behaviors and needs [12].

Traditional educational analytics often rely on structured data and predefined metrics, which may not capture the full spectrum of learner interactions and experiences. MLLMs can analyze unstructured data, such as discussion forum posts, assignment submissions, speech recordings, and facial expressions, providing deeper insights into student engagement, comprehension, and emotional states [6]. This multimodal approach enables educators to make more informed decisions and interventions to enhance learning outcomes.

3.5.2. Technology Used: MLLMs in AI-Based Insight and Predictive Analytics

The deployment of MLLMs in educational analytics involves several key technologies. Multimodal data analysis allows models like GPT-4 [9] and CLIP [14] to process and interpret data from various modalities, enabling the extraction of rich features and patterns from diverse data sources. Advanced language models perform natural language processing (NLP) and understanding (NLU) on textual data from student interactions—such as discussion forums, essays, and feedback—to identify topics, sentiments, and comprehension levels [19].

In addition, MLLMs with computer vision capabilities analyze images and videos, such as classroom recordings or student-submitted media, to assess engagement, participation, and emotional states through facial expression recognition [80]. Tools like OpenFace provide robust frameworks for facial behavior analysis, enabling educators to gauge student emotions and engagement. Speech and audio processing models, such as Wav2Vec 2.0 [65], transcribe and analyze audio data, offering insights from student presentations, oral exams, or verbal feedback.

Furthermore, MLLMs contribute features to predictive models that forecast student performance, dropout risks, and learning trajectories based on historical and real-time data [81]. Techniques for integrating data from multiple modalities enhance the robustness and accuracy of predictive analytics [6].

3.5.3. Examples of Applications

The use of MLLMs in providing AI-based insights and predictive analytics for educators is demonstrated in several applications. Early warning systems analyze multimodal data to identify students at risk of underperforming or dropping out. For instance, text analysis of student messages combined with engagement metrics can provide early indicators of disengagement [82]. Sentiment and emotion analysis processes textual and visual data to gauge student sentiments and emotions, helping educators understand the classroom climate and address issues promptly [83].

Adaptive feedback generation is another application where MLLMs analyze student submissions to provide personalized feedback, highlighting areas of strength and improvement [55]. By aggregating and analyzing data on student interactions with course materials, educators gain curriculum and instructional design insights that inform adjustments and teaching strategies [84]. Additionally, analyzing multimodal data from group projects, discussions, and peer interactions allows for the assessment of collaborative skills and group dynamics [85]. Evaluating student presentations using audio and visual data provides feedback on delivery, clarity, and engagement [86].

3.5.4. Case Study: Early Warning Systems Using MLLMs

Early warning systems (EWSs) aim to identify students who are at risk of academic failure or dropping out, allowing educators to intervene proactively [82]. Traditional EWSs rely on structured data such as grades, attendance, and demographic information. Integrating MLLMs enables the analysis of unstructured and multimodal data, providing a more comprehensive risk assessment.

In this context, data are gathered from various sources, including textual data like discussion posts, emails, and assignment submissions; visual data such as video recordings of classes or online interactions; and audio data from recordings of student speeches or discussions. NLP models like BERT [19] are utilized to analyze textual data for sentiment, topic modeling, and engagement indicators. Computer vision models, such as OpenFace [80], are employed for facial expression recognition to assess emotions and engagement levels. Speech recognition models like Wav2Vec 2.0 [65] transcribe and analyze audio data for speech patterns and indicators of confusion or hesitation.

Features extracted by MLLMs are combined with machine learning algorithms—such as random forests or neural networks—to predict at-risk students. For example, in university settings, institutions implement EWSs to monitor student engagement in online courses, using data from learning management systems (LMSs), discussion forums, and assignment submissions [87]. In K-12 education, schools use multimodal data, including classroom observations and student interactions, to identify students needing additional support [88]. Massive open online courses (MOOCs) analyze participant data to predict dropout rates and tailor interventions [89].

A study by Giannakos et al. (2019) explored the use of multimodal data and MLLMs to enhance early warning systems in online education [90]. The findings indicated that incorporating multimodal data significantly improved the accuracy of predicting at-risk students compared to models using only traditional data. Key predictive features included textual engagement, where the frequency and sentiment of discussion posts correlated with student success; emotional indicators, where facial expression analysis provided insights into student frustration or confusion during video interactions; and speech patterns, where hesitation and speech rate in audio submissions were linked to comprehension difficulties. Early identification enabled timely interventions such as personalized support, tutoring, or counseling.

However, challenges were identified, including data privacy and ethics, as collecting and analyzing sensitive data requires strict adherence to privacy regulations and ethical considerations. Technical complexity is also a concern, as implementing MLLMs and processing multimodal data demand significant computational resources and expertise.

Interpretability is another challenge, as complex models may lack transparency, making it difficult for educators to understand the basis of predictions.

The case study demonstrates the potential of MLLMs in enhancing early warning systems through the analysis of multimodal data. By providing more accurate and comprehensive risk assessments, educators can intervene effectively to support at-risk students. Addressing challenges related to privacy, technical implementation, and model interpretability is essential for practical deployment.

3.6. Grading and Assessment Tools

3.6.1. Introduction to Grading and Assessment Tools

Grading and assessment are fundamental components of the educational process, providing feedback to learners and informing instructional decisions. Traditional grading methods often involve manual evaluation of student work, which can be time consuming, subject to human bias, and inconsistent across different evaluators [91]. The advent of artificial intelligence, particularly multimodal large language models (MLLMs), has opened new avenues for automating and enhancing the grading and assessment process [92]. MLLMs can process and evaluate complex student outputs across various modalities—such as essays, short answers, presentations, code, and multimedia projects—providing timely, objective, and detailed feedback [93,94].

Automated grading systems leveraging MLLMs aim to improve efficiency, consistency, and fairness in assessments. By analyzing student submissions using advanced natural language processing (NLP) and computer vision techniques, these systems can assess not only the correctness but also the quality and depth of understanding demonstrated in student work [95]. Moreover, MLLMs enable formative assessments by providing personalized feedback and recommendations for improvement, supporting adaptive learning paths [96].

3.6.2. Technology Used: MLLMs in Grading and Assessment

The implementation of MLLMs in grading and assessment tools involves several key technologies. Natural language processing (NLP) and understanding (NLU) are employed in MLLMs like GPT-4 [9] and BERT [19] to analyze and understand student-written text, such as essays and short answers, code comments, and other textual data. These models can evaluate grammatical correctness, coherence, argumentation quality, and adherence to rubrics.

For short-answer scoring (SAS), models need to understand concise responses and compare them against expected answers or rubrics. Techniques such as semantic similarity assessment and entailment recognition are crucial in accurately evaluating short answers [97].

Computer vision and image recognition are utilized for assignments involving visual components, such as diagrams, handwritten responses, or design projects. Models like CLIP [14] and ViT [20] process and assess visual data.

Speech and audio processing is applied in assessments requiring oral presentations or language proficiency. Models like Wav2Vec 2.0 [65] and Speech Transformers [98] transcribe and analyze spoken language for fluency, pronunciation, and content accuracy.

Multimodal data integration allows MLLMs to integrate data from multiple modalities to assess complex assignments that combine text, visuals, and audio [99]. This integration enables a holistic evaluation of student work.

Rubric-based and criterion-referenced assessment involve training AI models to align with specific grading rubrics and criteria, ensuring that evaluations are aligned with learning objectives [100]. For SAS, aligning models with rubrics is particularly important due to the variability in acceptable short answers.

Feedback generation is another critical aspect, where MLLMs generate personalized feedback, highlighting strengths and areas for improvement, and suggesting resources for further learning [97].

3.6.3. Examples of Applications

Several applications and systems illustrate the use of MLLMs in grading and assessment. Automated essay scoring (AES) systems, such as E-Rater by ETS, employ NLP techniques to assess essays in standardized tests, evaluating features such as grammar, usage, mechanics, style, and development. Short-answer scoring (SAS) systems assess brief responses to questions, requiring models to understand and evaluate the correctness and relevance of concise student inputs [97]. Code assessment platforms, including systems like MOSS (Measure of Software Similarity) [101] and CodeRunner, utilize AI to evaluate programming assignments, checking for correctness, efficiency, and plagiarism [102]. Handwritten response grading involves AI models processing scanned images of handwritten math or short-answer responses, interpreting handwriting and assessing correctness. Oral language proficiency testing is seen in applications like Duolingo English Test, which use speech recognition and MLLMs to evaluate language proficiency through spoken responses [43]. Multimodal assignment evaluation includes tools that assess student presentations or projects combining text, visuals, and audio, using MLLMs to analyze each component and the overall coherence.

3.6.4. Case Study: Reducing the Cost of Short-Answer Scoring with MLLMs

Automated short-answer scoring (SAS) is the task of automatically evaluating brief student responses to prompts based on predefined rubrics and reference answers [20,41,103]. SAS is particularly challenging due to the variability in acceptable answers and the need for models to understand nuanced student inputs. A significant barrier to the widespread adoption of SAS systems is the cost associated with preparing training data for each new prompt, as rubrics and reference answers differ between prompts [23].

A recent study by Funayama et al. (2024) addressed this challenge by proposing a two-phase approach to reduce the cost of training SAS models [104]. The approach involves pre-fine-tuning a language model on existing rubrics and answers with gold score signals from annotated prompts and then fine-tuning it on a new prompt with limited data. By utilizing key phrases—representative expressions that answers should contain to increase scores—the model learns the relationship between key phrases and student answers across different prompts.

In the pre-fine-tuning phase, the model is trained on cross-prompt data, enabling it to learn general scoring principles shared across prompts. Specifically, the model uses the key phrases from rubrics to understand what constitutes a high-quality answer in a general sense. During the fine-tuning phase, the model is adapted to a new prompt using limited data, benefiting from the knowledge acquired during pre-fine-tuning without requiring access to the proprietary cross-prompt data.

The study utilized BERT [19] as the base language model and conducted experiments on a dataset enriched with a large number of prompts, rubrics, and answers. The findings indicated that fine-tuning on existing cross-prompt data with key phrases significantly improves scoring accuracy, especially when the training data for the new prompt is limited. The model demonstrated improved generalizability and reduced the amount of in-prompt data required for effective scoring.

Key insights included the importance of designing the model to learn the general properties of the scoring task and leveraging key phrases to align with the rubrics. By focusing on the relationship between key phrases and student answers, the model effectively captured the essential elements required for high-quality responses, enabling it to generalize across different prompts.

Challenges identified in the study included ensuring data accessibility, as cross-prompt data might be proprietary and not readily available for all educators. The proposed two-phase approach addresses this by allowing the pre-fine-tuned model's parameters to be shared without exposing the underlying data, thus maintaining data privacy while still benefiting from the knowledge gained during pre-fine-tuning. Additionally, the study emphasized the need for the model to effectively learn from diverse prompts to generalize

well to new prompts, suggesting that increasing the diversity of prompts during pre-fine-tuning enhances the model's performance on unseen prompts.

The incorporation of MLLMs in SAS demonstrates significant potential for enhancing the efficiency and scalability of short-answer grading. By reducing the data requirements and leveraging cross-prompt learning, educators can adopt automated scoring systems with lower costs and improved performance. Addressing challenges related to data accessibility, model generalizability, and alignment with scoring rubrics is essential for the broader adoption of such systems.

4. Discussion and Future Directions

The integration of multimodal large language models (MLLMs) into educational applications offers transformative potential for teaching and learning. This section examines the significant benefits that MLLMs bring to education, acknowledges the limitations and challenges associated with their deployment, and proposes future directions toward a unified AI-powered educational ecosystem. By understanding both the advantages and the obstacles, we can better harness MLLMs to enhance educational outcomes while addressing critical concerns.

4.1. Benefits of Using MLLMs in Educational Applications

The integration of MLLMs into educational applications is revolutionizing the landscape of teaching and learning. One of the foremost benefits is the facilitation of personalized and adaptive learning experiences. MLLMs can analyze a vast array of learner data across multiple modalities—including text, speech, images, and videos—to tailor educational content to individual students' needs, preferences, and proficiency levels. This level of personalization enhances engagement and can significantly improve academic outcomes. For instance, adaptive learning platforms powered by MLLMs adjust the difficulty and style of content delivery in real time, responding to the learner's performance and interaction patterns [62].

Moreover, MLLMs enhance engagement through their ability to process and generate content across multiple modalities. They can transform complex textual information into visual diagrams, provide auditory explanations for intricate concepts, or generate interactive simulations. This multimodal content delivery caters to diverse learning styles, making education more inclusive and effective. Research indicates that multimodal learning experiences improve comprehension and retention rates compared to traditional unimodal approaches. For example, science education can be significantly enriched by MLLMs generating interactive 3D models of molecular structures or astronomical phenomena, thereby deepening students' understanding.

Another critical benefit is the automation and enhancement of content creation. MLLMs can generate high-quality educational materials, including textbooks, lesson plans, assessments, and multimedia resources. This automation reduces the burden on educators, allowing them to allocate more time to instructional strategies and student engagement. Additionally, the scalability of MLLMs means that educational content can be rapidly produced and updated, ensuring that learning materials remain current with the latest knowledge and pedagogical practices [105].

MLLMs also play a pivotal role in improving accessibility and inclusivity in education. By offering content in various formats, they support learners with disabilities or those requiring alternative learning resources. For instance, MLLMs can provide real-time transcription and translation services, convert text to speech for visually impaired learners, or generate sign language representations for the hearing impaired. Such capabilities ensure that education becomes more equitable, aligning with universal design for learning principles [106].

Furthermore, MLLMs offer advanced support for educators through AI-powered tools. By leveraging predictive analytics and insights derived from multimodal educational data, MLLMs can identify learning gaps, predict at-risk students, and suggest targeted

interventions. Automated grading systems powered by MLLMs provide objective evaluations and detailed, personalized feedback across various types of student work, including essays, presentations, and even creative projects. This not only enhances the efficiency of educational processes but also improves the quality of feedback, which is crucial for student development.

Lastly, MLLMs enable the development of intelligent virtual tutors and chatbots that provide personalized support and interactive learning opportunities through natural language conversations. These AI agents can answer students' queries in real time, provide explanations, and guide them through problem-solving processes. The continuous availability of such support fosters learner autonomy and can bridge gaps in understanding outside traditional classroom hours [107].

4.2. Limitations of Using MLLMs in Educational Applications

Despite the significant benefits, the deployment of MLLMs in education is accompanied by several limitations that warrant careful consideration. One of the primary challenges is the technical and resource constraints associated with implementing these advanced models. MLLMs require substantial computational power and specialized technical expertise to develop, deploy, and maintain. Educational institutions, especially those in developing regions or with limited funding, may find it prohibitive to invest in the necessary infrastructure. This disparity can exacerbate existing inequalities in educational access and quality [108].

Data privacy and security concerns also pose significant obstacles. MLLMs often rely on the collection and analysis of extensive personal data to function effectively. In educational settings, this includes sensitive information about students' learning behaviors, performance, and potentially personal identifiers. Ensuring compliance with data protection regulations such as the General Data Protection Regulation (GDPR) in Europe or the Family Educational Rights and Privacy Act (FERPA) in the United States is essential. However, many institutions may lack the robust data governance frameworks required to safeguard this information adequately [109].

Bias and fairness issues inherent in MLLMs are another critical limitation [110–112]. These models learn patterns from vast datasets, which may contain historical biases and stereotypes. Consequently, MLLMs can inadvertently perpetuate or even amplify these biases in their outputs, affecting the fairness and equity of educational content and assessments. For example, language models might generate content that is culturally insensitive or skewed towards certain demographics, disadvantaging minority groups. Addressing these biases is essential to prevent the reinforcement of inequalities within the educational system [77].

Moreover, the use of MLLM applications separately, as standalone solutions, leads to fragmentation and inefficiencies. Isolated applications may not integrate seamlessly with existing educational technologies or with each other, resulting in a disjointed learning experience for students. This lack of interoperability hampers the comprehensive understanding of a learner's progress, as data and insights are siloed within individual applications. Educators may struggle to aggregate information from disparate sources to inform instructional decisions effectively [113]. Consequently, the potential of MLLMs to enhance education is not fully realized when applications operate in isolation.

Over-reliance on technology and reduced human interaction present additional concerns. While MLLMs can augment educational processes, there is a risk that excessive dependence on AI could diminish the role of educators and the value of human relationships in learning. Social interaction, mentorship, and the development of soft skills are critical components of education that AI cannot fully replicate. Students may miss out on the nuances of human communication, empathy, and collaborative learning experiences, which are essential for personal and professional development.

Finally, ethical considerations regarding the transparency and accountability of AI decision-making processes are paramount. MLLMs operate as complex black-box models,

making it challenging to interpret how they arrive at specific outputs or recommendations. This opacity can hinder trust and acceptance among educators and learners, who may be wary of relying on systems that lack explainability. Ensuring that MLLMs provide interpretable and justifiable outputs is crucial for their successful integration into educational settings [114].

4.3. Future Directions: Towards a Unified AI-Powered Educational Ecosystem

The limitations associated with standalone MLLM applications underscore the need for a cohesive and integrated approach. The future of educational technology lies in developing a unified AI-powered educational ecosystem that brings together various LLM and MLLM agents in a coordinated and interoperable framework. Such an ecosystem would address inefficiencies by enabling seamless communication and data exchange between AI agents, educational platforms, and stakeholders.

Creating this unified ecosystem involves establishing common standards and protocols for data interoperability. By adopting open architectures and APIs, different AI agents and educational technologies can interact effectively, sharing insights and providing a holistic view of learner progress. This interoperability facilitates personalized learning pathways that adapt based on comprehensive data from multiple sources, enhancing the effectiveness of educational interventions.

MLLM agents, which integrate language understanding with multimodal data processing—such as text, images, audio, and video—play a critical role in enriching educational content and interactions. For instance, models like PaLM-E [115] and LLaVA [30] demonstrate the potential of MLLMs in understanding and generating multimodal content. These agents can interpret visual inputs and generate coherent textual explanations, enabling more interactive and engaging learning experiences.

LLM agents, built upon advanced language models like GPT-4 [9], serve as conversational interfaces that guide learners through educational materials, answer complex queries, and provide real-time feedback. They enhance learner autonomy by supporting self-directed learning and offering personalized recommendations based on individual progress and preferences.

Integrating retrieval-augmented generation (RAG) technology further enhances the capabilities of these AI agents within the educational ecosystem. RAG models combine the strengths of large language models with external knowledge bases, enabling the AI to access and retrieve relevant information from vast datasets in real time [116]. This integration ensures that the AI agents provide up-to-date and accurate information, which is particularly crucial in rapidly evolving fields of study.

In the unified ecosystem (see Figure 4), MLLM and LLM agents equipped with RAG capabilities collaborate to deliver adaptive learning pathways. For example, when a learner poses a question that requires current data or specialized knowledge, the RAG-enabled agent can retrieve information from trusted educational databases, scholarly articles, or institutional repositories [117]. This approach not only enhances the depth and accuracy of responses but also allows for personalized content generation that aligns with the learner's curriculum and learning objectives.

Furthermore, these AI agents enable global connectivity among learners and educators, fostering cross-cultural exchanges and collaborative projects. Platforms incorporating MLLM, LLM, and RAG technologies can facilitate mentorship opportunities, virtual study groups, and peer feedback mechanisms, promoting peer-to-peer learning on a global scale.

Advancements in cloud computing and edge technologies offer scalable and accessible infrastructure for deploying MLLMs [54]. Utilizing cloud-based services allows institutions to leverage powerful AI capabilities without the need for extensive on-premises hardware. This democratization of technology can reduce barriers to adoption, enabling institutions of varying sizes and resources to participate in the unified ecosystem. Partnerships with technology providers and investment in shared resources can further enhance accessibility [118].

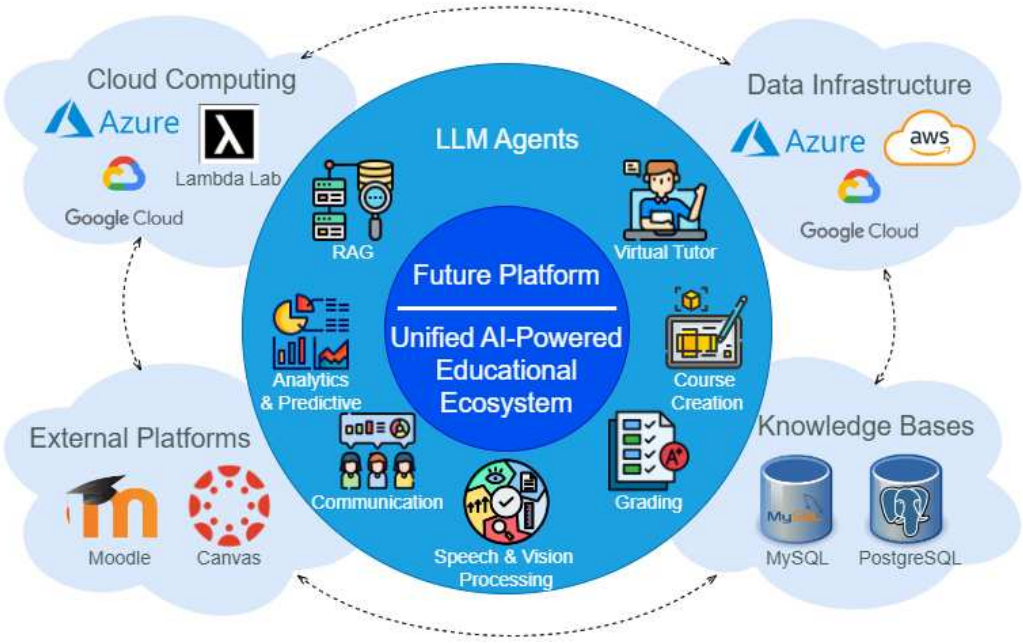

Figure 4. Future unified AI-powered educational ecosystem, supported by external modules and infrastructures.

Embedding ethical AI practices into the design and implementation of the unified ecosystem is imperative. This includes ensuring transparency in AI algorithms, implementing measures to detect and mitigate biases, and upholding stringent data privacy standards. Engaging a diverse group of stakeholders—including educators, learners, policymakers, and ethicists—in the development process can help align the technology with educational values and societal norms [119].

The unified ecosystem should be envisioned as augmenting, rather than replacing, the role of educators. AI agents can handle administrative tasks, provide supplemental instruction, and offer personalized support, freeing educators to focus on higher-order teaching activities such as facilitating critical discussions, fostering creativity, and mentoring students. This synergy between AI and human educators enhances the educational experience by combining technological efficiency with human insight and empathy [3].

Moreover, the unified ecosystem can foster global collaboration and resource sharing. By connecting educational platforms across institutions and borders, learners and educators can engage in cross-cultural exchanges, collaborative projects, and peer learning opportunities. Such global connectivity enriches the educational experience, preparing learners for participation in an increasingly interconnected world [120].

To support continuous learning and adaptability, the ecosystem should incorporate mechanisms for ongoing feedback and improvement. AI agents can learn from interactions, outcomes, and user feedback, refining their models over time to enhance effectiveness. Additionally, professional development programs for educators are essential to equip them with the skills to leverage AI tools effectively and to adapt pedagogical approaches in response to technological advancements [73].

In conclusion, transitioning towards a unified AI-powered educational ecosystem addresses the limitations of standalone MLLM applications and maximizes the benefits of AI in education. By integrating various AI agents into a cohesive platform that emphasizes interoperability, ethical practices, and collaboration between technology and human educators, the educational landscape can be transformed. This unified approach promises to create a more connected, equitable, and effective educational system that is responsive to the needs of learners in the era of AI.

4.4. Shift of Tides in Future Education

One of the most significant changes is that future education will be full of lifelong learning opportunities. Education is evolving beyond the confines of traditional institutions and specific age groups. Large language models (LLMs) have the potential to enhance lifelong learning by creating adaptive learning paths tailored to individuals at any stage of their careers. These models can curate personalized educational experiences, recommend relevant resources, and even simulate mentorship. For example, platforms like Coursera utilize AI-driven algorithms to suggest courses aligned with user preferences and identified skill gaps [121]. Additionally, Duolingo's integration of GPT-4 for personalized language instruction demonstrates how LLMs can revolutionize continuous learning [46]. Recent research by Liu et al. [122] explores how LLMs can adapt content delivery based on real-time learner feedback, further personalizing the educational experience.

At the same time, artificial intelligence is redefining the roles and responsibilities of educators. Artificial intelligence will not replace educators but will augment their roles. LLMs can handle routine tasks such as grading, administrative work, and resource generation, allowing teachers to focus on fostering critical thinking and creativity. For instance, AI assistants like Gradescope by Turnitin have already streamlined grading workflows. Future LLM-driven systems could further assist by providing insights into student behavior, recommending tailored interventions, and offering real-time teaching assistance. Projects like Khan Academy's Khanmigo, developed using OpenAI's GPT-4, exemplify how AI can support educators by providing personalized tutoring and aiding in lesson planning [123]. With proper use of LLMs tools, they can also analyze student interactions to offer educators actionable insights, enhancing the effectiveness of teaching strategies [124].

Looking ahead, it will be crucial to ensure that LLM-driven educational tools support diverse local and cultural contexts. One proposal is to encourage modular integration of LLMs within existing course management systems, allowing educators to select, adjust, and refine AI-driven resources without overhauling their entire digital infrastructure. Another avenue involves enhancing the explainability and transparency of LLM outputs, enabling teachers and learners to understand the reasoning behind suggestions and decisions. Additionally, collaborations between AI developers, educators, and policymakers could yield standards and best practices that preserve data privacy, maintain ethical rigor, and promote fair access. Such efforts may include open data frameworks that allow educational institutions to share vetted learning materials, or the development of certification programs that recognize teachers trained in effectively leveraging LLM tools.

5. Conclusions

5.1. Summary of Key Insights

This survey underscores the transformative potential of multimodal large language models (MLLMs) in education, highlighting their effectiveness in diverse fields such as language learning, STEM education, and content creation. MLLMs have demonstrated significant improvements in student engagement, comprehension, and overall educational outcomes. By leveraging their ability to integrate multiple data modalities, MLLMs provide a richer and more interactive learning experience compared to traditional educational tools.

5.2. Final Thoughts

The integration of MLLMs into educational settings marks a substantial step forward in enhancing the learning experience. However, there are several challenges that need to be addressed to ensure their responsible use. Ethical considerations, such as the potential for bias in AI-generated content, privacy concerns related to data collection, and the need for transparency in AI decision-making processes, are critical issues that must be continually evaluated [77,125]. Furthermore, it is essential to consider the implications of these technologies on educational equity, ensuring that all students have equitable access to the benefits of MLLMs, regardless of socioeconomic status or geographic location [73].

Another major ethical issue is algorithmic bias. MLLMs, like all AI systems, are only as unbiased as the data on which they are trained. If the training data reflect existing biases, these can be perpetuated or even exacerbated by the AI system, leading to unfair educational outcomes for certain groups of students. It is crucial that educational institutions and AI developers work together to ensure that the datasets used to train MLLMs are diverse and representative [77].

Further research and development are crucial to fully realize the potential of MLLMs in education. This includes not only advancing the technical capabilities of these models but also ensuring that their deployment in educational environments is guided by ethical principles and a commitment to equity. Collaboration among educators, technologists, policymakers, and ethicists will be vital in shaping the future of education in the age of AI. Developing comprehensive frameworks for the ethical evaluation of AI tools in education can help address these challenges and ensure that the integration of MLLMs benefits all learners [62].

Moreover, continuous audits and assessments of AI systems are necessary to identify and mitigate biases, ensuring that these systems remain fair, transparent, and effective. The development of MLLMs should also be inclusive, with diverse AI development teams representing different perspectives and experiences to help identify potential biases early in the development process [126].

To move forward, stakeholders must focus on creating inclusive, adaptable, and responsible educational environments that leverage the power of MLLMs while safeguarding against their potential risks. This proactive approach will help maximize the benefits of AI-driven education, preparing students for a rapidly evolving digital world and ensuring that the future of learning is bright, equitable, and effective.

Author Contributions: Formal analysis, W.X.; investigation, W.X., T.Z., B.L., J.W.; resources, W.X., J.W.; writing—original draft preparation, W.X.; writing—review and editing, W.X., T.Z., B.L.; visualization, W.X.; supervision, T.Z., B.L.; project administration, J.W.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ARC Linkage Project (LP220200808) and ARC Discovery Project (DP230100246) from the Australian Research Council, Australia.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large language model
MLLM	Multimodal large language model
ADS	Advanced dialogue system

References

1. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
2. Zhou, W.; Zhu, X.; Han, Q.-L.; Li, L.; Chen, X.; Wen, S.; Xiang, Y. The Security of Using Large Language Models—A Survey with Emphasis on ChatGPT. *IEEE/CAA J. Autom. Sin.* **2024**. [CrossRef]
3. Luckin, R.; Holmes, W.; Griffiths, M.; Forcier, L.B. *Intelligence Unleashed: An Argument for AI in Education*; Pearson Education: London, UK, 2016.
4. OpenAI. OpenAI o1 System Card. *OpenAI*, 12 September 2024. [Online]. Available online: https://openai.com/index/openai-o1-system-card/ (accessed on 3 October 2024).
5. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597.
6. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–445. [CrossRef]

7. Williamson, B.; Eynon, R. Historical threads, missing links, and future directions in {AI} in education. *Learn. Media Technol.* **2020**, *45*, 223–235. [CrossRef]
8. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
9. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
10. Anthropic. Claude 3.5 Sonnet Model Card Addendum. *Anthropic*, 2023. [Online]. Available online: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf (accessed on 5 October 2024).
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
12. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal Deep Learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.
13. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. *Flamingo: A Visual Language Model for Few-Shot Learning*; DeepMind: London, UK, 2022.
14. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
15. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8821–8831.
16. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125.
17. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. Augmented Language Models: A Survey. *arXiv* **2023**, arXiv:2302.07842.
18. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL, Florence, Italy, 28 July–2 August 2019; pp. 4171–4186.
20. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
21. Gong, Y.; Chung, Y.-A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 571–575.
22. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569.
23. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
24. Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: Universal Image-Text Representation Learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 104–120.
25. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
26. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 5100–5111.
27. Su, W.; Zhu, X.; Cao, Y. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
28. Dai, W.; Lee, N.; Wang, B.; Yang, Z.; Liu, Z.; Barker, J.; Rintamaki, T.; Shoeybi, M.; Catanzaro, B.; Ping, W. NVLM: Open Frontier-Class Multimodal LLMs. *arXiv* **2024**, arXiv:2409.11402.
29. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
30. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *arXiv* **2023**, arXiv:2304.08485.
31. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *arXiv* **2023**, arXiv:2303.04671.
32. Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Patra, B.; et al. Language Is Not All You Need: Aligning Perception with Language Models. *arXiv* **2023**, arXiv:2302.14045.
33. McKinzie, B.; Gan, Z.; Fauconnier, J.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Weers, F.; et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *arXiv* **2024**, arXiv:2403.09611.
34. Madsen, S.; Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. Evaluating the Explainability of Machine Learning Models in Education. *IEEE Trans. Learn. Technol.* **2023**, *16*, 1–14.

35. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, e19. [CrossRef]
36. Murray, T. An Overview of Intelligent Tutoring System Authoring Tools: Updated Analysis of the State of the Art. In *Authoring Tools for Advanced Technology Learning Environments*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 491–544.
37. Park, Y.; Lee, G.M. Adaptive Learning Systems. In *Encyclopedia of Education and Information Technologies*; Springer: Berlin/Heidelberg, Germany, 2019.
38. Durlach, P.J.; Lesgold, A.M. *Adaptive Technologies for Training and Education*; Cambridge University Press: Cambridge, UK, 2012.
39. Maycock, K. Multimodal Learning. In *International Handbook of the Learning Sciences*; Routledge: Oxfordshire, UK, 2019; pp. 261–271.
40. Cognii. AI and Education. 2020. [Online]. Available online: https://www.cognii.com/ (accessed on 14 October 2024).
41. Carnegie Learning. MATHia: Personalized Math Learning Software. 2020. [Online]. Available online: https://www.carnegielearning.com/mathia/ (accessed on 14 October 2024).
42. Knewton. Adaptive Learning Technology. 2018. [Online]. Available online: https://www.knewton.com/ (accessed on 14 October 2024).
43. Settles, B.; Laurel, T.; Briggs, A. Machine Learning–Driven Language Education. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 451–466.
44. Von Ahn, L. Duolingo: Learn a Language for Free While Helping to Translate the Web. In Proceedings of the International Conference on Intelligent User Interfaces, Santa Monica, CA, USA, 19–22 March 2013; pp. 1–2.
45. Smart Sparrow. Adaptive Learning Platform. 2018. [Online]. Available online: https://www.smartsparrow.com/ (accessed on 14 October 2024).
46. Duolingo Team. Introducing Duolingo Max, a Learning Experience Powered by GPT-4. *Duolingo Blog*, 2024. Available online: https://blog.duolingo.com/duolingo-max/ (accessed on 11 November 2024).
47. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499.
48. Zheng, L.; Long, M.; Zhong, L.; Gyasi, J.F. The Effectiveness of Technology-Facilitated Personalized Learning on Learning Achievements and Learning Perceptions: A Meta-Analysis. *Educ. Inf. Technol.* **2022**, *27*, 11807–11830. [CrossRef]
49. Panigrahi, S.; Rath, P.K.; Sahoo, B. Intelligent Tutoring Systems Using Large Language Models: A Review. *J. Educ. Technol. Syst.* **2023**, *51*, 5–27.
50. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 121–137.
51. Zhao, H.H.; Zhou, P.; Gao, D.; Bai, Z.; Shou, M.Z. LOVA^3: Learning to Visual Question Answering, Asking and Assessment. *arXiv* **2024**, arXiv:2405.14974v2.
52. Yi, Z.; Ouyang, J.; Liu, Y.; Liao, T.;Xu, Z.; Shen, Y. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. *arXiv* **2024**, arXiv:2402.18013.
53. Sun, X.; Panda, R.; Feris, R.; Saenko, K. AdaShare: Learning What to Share for Efficient Deep Multi-Task Learning. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020. Available online: https://cs-people.bu.edu/sunxm/AdaShare/project.html (accessed on 1 November 2024).
54. Zhao, Y.; Qu, Y.; Xiang, Y.; Uddin, M.P.; Peng, D.; Gao, L. A Comprehensive Survey on Edge Data Integrity Verification: Fundamentals and Future Trends. *ACM Comput. Surv.* **2024**, *57*, 8:1–8:34. [CrossRef]
55. Roschelle, J.; Lester, J.; Fusco, J. (Eds.) *AI and the Future of Learning: Expert Panel Report*; [Report]. Digital Promise. 2020. Available online: https://circls.org/reports/ai-report (accessed on 1 November 2024).
56. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
57. Gligorea, I.; Cioca, M.; Oancea, R.; Gorski, A.-T.; Gorski, H.; Tudorache, P. Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. *Educ. Sci.* **2023**, *13*, 1216. [CrossRef]
58. Abdelghani, R.; Wang, Y.-H.; Yuan, X.; Wang, T.; Lucas, P.; Sauzéon, H.; Oudeyer, P.-Y. GPT-3-Driven Pedagogical Agents for Training Children's Curious Question-Asking Skills. In Proceedings of the 14th International Conference on Computer Supported Education (CSEDU), Online, 22–24 April 2022.
59. Al-Ansi, A.M.; Jaboob, M.; Garad, A.; Al-Ansi, A. Analyzing Augmented Reality (AR) and Virtual Reality (VR) Recent Development in Education. *Soc. Sci. Humanit. Open* **2023**, *8*, 100532. [CrossRef]
60. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.J.; Gao, J. Unified Vision-Language Pre-Training for Image Captioning and VQA. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13041–13049.
61. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond English-Centric Multilingual Machine Translation. *J. Mach. Learn. Res.* **2021**, *22*, 1–48.
62. Holmes, W.; Bialik, M.; Fadel, C. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*; Center for Curriculum Redesign: Jamaica Plain, MA, USA, 2019.

63. Dwivedi, S.K.; Bharadwaj, A.K.; Jha, S.K. Role of Artificial Intelligence in Empowering Teaching and Learning. In Proceedings of the International Conference on Advances in Computing and Data Sciences, Ghazibad, India, 12–13 April 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 12–24.
64. Zhang, S.; Yao, L.; Sun, A.; Tay, Y. Deep Learning-Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* **2021**, *52*, 1–38. [CrossRef]
65. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
66. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.J.; et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
67. Liao, C.-H.; Wu, J.-Y. Deploying Multimodal Learning Analytics Models to Explore the Impact of Digital Distraction and Peer Learning on Student Performance. *Comput. Educ.* **2022**, *190*, 104599. [CrossRef]
68. Ouyang, F.; Wu, M.; Zheng, L.; Zhang, L.; Jiao, P. Integration of Artificial Intelligence Performance Prediction and Learning Analytics to Improve Student Learning in Online Engineering Course. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 4. [CrossRef]
69. edX Team. edX Debuts Two AI-Powered Learning Assistants Built on ChatGPT. *edX Press Release*, 12 May 2023. [Online]. Available online: https://press.edx.org/edx-debuts-two-ai-powered-learning-assistants-built-on-chatgpt (accessed on 15 October 2024).
70. Udemy. Udemy's AI-Powered Learning. 2020. [Online]. Available online: https://about.udemy.com/ (accessed on 17 October 2024).
71. Knewton. Knewton Alta. 2020. [Online]. Available online: https://japan.knewton.com/news/n2020112401.html (accessed on 17 October 2024).
72. Shah, D. By The Numbers: MOOCs in 2021. *Class Central*, 2021. [Online]. Available online: https://www.classcentral.com/report/mooc-stats-2021/ (accessed on 19 October 2024).
73. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
74. Piech, C.; Spencer, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.; Sohl-Dickstein, J. Deep Knowledge Tracing. *arXiv* **2015**, arXiv:1506.05908.
75. Lakew, S.M.; Federico, M.; Negri, M.; Turchi, M. Multilingual Neural Machine Translation for Zero-Resource Languages. *arXiv* **2018**, arXiv:1909.07342.
76. Molina, J.P.; Turchi, V.M. GDPR challenges for leveraging big data in the education and research sectors. In Proceedings of the 14th International Conference on Web Information Systems and Technologies, Seville, Spain, 18–20 September 2018; pp. 659–666.
77. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5454–5476.
78. Holzinger, A.; Biemann, C.; Pattichis, C.S.; Kell, D.B. What Do We Need to Build Explainable AI Systems for the Medical Domain? *arXiv* **2017**, arXiv:1712.09923.
79. Baker, R.S.; Siemens, G. Educational Data Mining and Learning Analytics. In *Learning Analytics*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 61–75.
80. Baltrušaitis, T.; Robinson, P.; Morency, L.-P. OpenFace: An Open Source Facial Behavior Analysis Toolkit. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1–10.
81. Alwahaby, H.; Cukurova, M.; Papamitsiou, Z.; Giannakos, M. The Evidence of Impact and Ethical Considerations of Multimodal Learning Analytics: A Systematic Literature Review. In *The Multimodal Learning Analytics Handbook*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 289–325. [CrossRef]
82. Arnold, K.E.; Pistilli, M.D. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC, Canada, 29 April 2012–2 May 2012; pp. 267–270.
83. Shoumy, N.J.; Ang, L.-M.; Seng, K.P.; Rahaman, D.M.M.; Zia, T. Multimodal Big Data Affective Analytics: A Comprehensive Survey Using Text, Audio, Visual and Physiological Signals. *J. Netw. Comput. Appl.* **2020**, *149*, 102447. [CrossRef]
84. Papamitsiou, Z.; Economides, A.A. Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educ. Technol. Soc.* **2014**, *17*, 49–64.
85. Blanchard, E.G.; Bousbia, D.; Franceschini, B. Identifying Group Dynamics and Emotion in E-Learning: An Integrated Approach. In *Intelligent Tutoring Systems*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 354–359.
86. Giannakos, M.; Cukurova, M. The Role of Learning Theory in Multimodal Learning Analytics. *Br. J. Educ. Technol.* **2023**, *54*, 1246–1267. [CrossRef]
87. Ellis, R.A.; Goodyear, P. Developing and Using a Learning Analytics Framework: A Case Study. *Teach. High. Educ.* **2019**, *24*, 394–407.
88. Çeken, B.; Taşkın, N. Multimedia Learning Principles in Different Learning Environments: A Systematic Review. *Smart Learn. Environ.* **2022**, *9*, 19. [CrossRef]
89. Romero, C.; Ventura, S. Educational Data Mining and Learning Analytics: An Updated Survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]

90. Giannakos, M.N.; Sharma, K.; Pappas, I.O.; Kostakos, V.; Velloso, E. Multimodal Data as a Means to Understand the Learning Experience. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA, 4–8 March 2019; pp. 639–640.
91. Zhao, J.; Wu, M.; Zhou, L.; Wang, X.; Jia, J. Cognitive Psychology-Based Artificial Intelligence Review. *Front. Neurosci.* **2022**, *16*, 1024316. [CrossRef]
92. Ke, Z.; Ng, V. Automated Essay Scoring: A Survey of the State of the Art. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, China, 10–16 August 2019; pp. 6300–6308. [CrossRef]
93. Prinsloo, P. Of 'Black Boxes' and Algorithmic Decision-Making in (Higher) Education—A Commentary. *Big Data Soc.* **2020**, *7*, 2053951720933994. [CrossRef]
94. Zhu, X.; Zhou, W.; Han, Q.-L.; Ma, W.; Wen, S.; Xiang, Y. When Software Security Meets Large Language Models: A Survey. *IEEE/CAA J. Autom. Sin.* **2024**, *accepted*.
95. Gierl, M.J.; Zhang, H. Automated Scoring in the Classroom. In *Handbook of Automated Essay Evaluation: Current Applications and New Directions*; Routledge: Oxfordshire, UK, 2018; pp. 136–154.
96. Gašević, D.; Dawson, S.; Siemens, G. Let's Not Forget: Learning Analytics are about Learning. *TechTrends* **2015**, *59*, 64–71. [CrossRef]
97. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *3*, 100050. [CrossRef]
98. Latif, S.; Zaidi, A.; Cuayahuitl, H.; Shamshad, F.; Shoukat, M.; Qadir, J. Transformers in Speech Processing: A Survey. *arXiv* **2023**, arXiv:2303.11607.
99. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; Chang, K.-W. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
100. Schneider, J.; Schenk, B.; Niklaus, C. Towards LLM-based Autograding for Short Textual Answers. In Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), Angers, France, 2–4 May 2024.
101. MOSS (Measure of Software Similarity). Available online: https://theory.stanford.edu/~aiken/moss/ (accessed on 13 November 2022).
102. Ramesh, D.; Sanampudi, S.K. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef] [PubMed]
103. Chen, J.; Guo, H.; Yi, K.; Li, B.; Elhoseiny, M. VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. *arXiv* **2021**, arXiv:2102.10407.
104. Funayama, H.; Asazuma, Y.; Matsubayashi, Y.; Mizumoto, T.; Inui, K. Reducing the Cost: Cross-Prompt Pre-Finetuning for Short Answer Scoring. In *Artificial Intelligence in Education (AIED 2023)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2023; Volume 13916, pp. 78–89.
105. Gupta, S.; Sharda, N. Content Generation for Serious Games in Education: The ANFIS Approach. *IEEE Trans. Learn. Technol.* **2018**, *11*, 493–507.
106. Rose, D.H.; Meyer, A. *A Practical Reader in Universal Design for Learning*; Harvard Education Press: London, UK, 2006.
107. Graesser, A.C.; Cai, Z.; Morgan, B.; Wang, L. Assessment with Computer Agents That Engage in Conversational Dialogues and Trialogues with Learners. *Comput. Hum. Behav.* **2018**, *76*, 607–616. [CrossRef]
108. Williamson, B.; Hogan, A. *Commercialisation and Privatisation in/of Education in the Context of COVID-19*; Education International: Brussels, Belgium, 2020.
109. Regan, P.M.; Jesse, J. Ethical Challenges of EdTech, Big Data and Personalized Learning: Twenty-First Century Student Sorting and Tracking. *Ethics Inf. Technol.* **2019**, *21*, 167–179. [CrossRef]
110. Tian, H.; Liu, B.; Zhu, T.; Zhou, W.; Yu, P.S. MultiFair: Model Fairness with Multiple Sensitive Attributes. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, 1–14. [CrossRef]
111. Tian, H.; Liu, B.; Zhu, T.; Zhou, W.; Yu, P.S. Distilling Fair Representations From Fair Teachers. *IEEE Trans. Big Data* **2024**, 1–14. [CrossRef]
112. Chen, H.; Zhu, T.; Zhang, T.; Zhou, W.; Yu, P.S. Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff. *ACM Comput. Surv.* **2023**, *56*, 39:1–39:37. [CrossRef]
113. Kumar, V.; Sharma, D.K.; Singh, H. Interoperability Issues in e-Learning: A Review. *Int. J. Recent Technol. Eng.* **2019**, *8*, 115–121.
114. Lipton, Z.C. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
115. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: An Embodied Multimodal Language Model. *arXiv* **2023**, arXiv:2303.03378.
116. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**, arXiv:2005.11401.
117. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. REALM: Retrieval-Augmented Language Model Pre-Training. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; pp. 3929–3938.
118. Popenici, S.A.D.; Kerr, S. Exploring the Impact of Artificial Intelligence on Teaching and Learning in Higher Education. *Res. Pract. Technol. Enhanc. Learn.* **2017**, *12*, 22. [CrossRef]

119. Villaronga, E.F.; Kieseberg, P.; Li, T. Humans Forget, Machines Remember: Artificial Intelligence and the Right to Be Forgotten. *Comput. Law Secur. Rev.* **2018**, *34*, 304–313. [CrossRef]
120. Knight, E.; Cook, S. Educating Global Citizens in a Digital Age: The Role of MOOCs. *J. Glob. Educ. Res.* **2020**, *4*, 97–111.
121. Amin, S.; Uddin, M.I.; Alarood, A.A.; Mashwani, W.K.; Alzahrani, A.; Alzahrani, A.O. Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning. 2024. Available online: https://ieeexplore.ieee.org/document/10220065 (accessed on 10 November 2024).
122. Strielkowski, W.; Grebennikova, V.; Lisovskiy, A.; Rakhimova, G.; Vasileva, T. AI-driven adaptive learning for sustainable educational transformation. *Sustain. Dev.* **2024**. [CrossRef]
123. Khan Academy. Introducing Khanmigo: AI for Education. 2023. Available online: https://www.microsoft.com/en-us/education/blog/2024/08/khanmigo-for-teachers-your-free-ai-powered-teaching-tool/ (accessed on 10 October 2024).
124. Maiti, P.; Goel, A.K. How Do Students Interact with an LLM-powered Virtual Teaching Assistant in Different Educational Settings? In Proceedings of the Seventeenth International Conference on Educational Data Mining (EDM) Workshop: Leveraging LLMs for Next Generation Educational Technologies, Atlanta, GA, USA, 14–17 July 2024.
125. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 31:1–31:36. [CrossRef]
126. Chen, L.; Chen, P.; Lin, Z. Artificial intelligence in education: A review. *IEEE Access* **2020**, *8*, 75264–75278. [CrossRef]