frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 13 November 2025
DOI 10.3389/frai.2025.1683960

OPEN ACCESS

EDITED BY
José Manuel de Amo Sánchez-Fortún,
University of Almeria, Spain

REVIEWED BY
John Domingue,
The Open University, United Kingdom
Kevin Baldrich,
University of Almeria, Spain

*CORRESPONDENCE
Seyyedali Hosseinalipour
✉ alipour@buffalo.edu

RECEIVED 11 August 2025
ACCEPTED 20 October 2025
PUBLISHED 13 November 2025

CITATION
Borazjani K, Khosravan N, Sahay R, Akram B and Hosseinalipour S (2025) Bringing multi-modal multi-task federated foundation models to education domain: prospects and challenges. *Front. Artif. Intell.* 8:1683960. doi: 10.3389/frai.2025.1683960

Bringing multi-modal multi-task federated foundation models to education domain: prospects and challenges

Kasra Borazjani[1], Naji Khosravan[2], Rajeev Sahay[3], Bita Akram[4] and Seyyedali Hosseinalipour[1]*

[1]Department of Electrical Engineering, University at Buffalo—SUNY, Buffalo, NY, United States, [2]Adobe Research, Seattle, WA, United States, [3]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, United States, [4]Department of Computer Science, NC State University, Raleigh, NC, United States

Multi-modal multi-task (M3T) foundation models (FMs) have recently shown transformative potential in artificial intelligence, with emerging applications in education. However, their deployment in real-world educational settings is hindered by privacy regulations, data silos, and limited domain-specific data availability. We introduce M3T Federated Foundation Models (FedFMs) for education: a paradigm that integrates federated learning (FL) with M3T FMs to enable collaborative, privacy-preserving training across decentralized institutions while accommodating diverse modalities and tasks. Subsequently, this perspective paper aims to unveil M3T FedFMs as a promising yet underexplored approach to the education community, explore its potentials, and reveal its related future research directions. We outline how M3T FedFMs can advance three critical pillars of next-generation intelligent education systems: *(i) privacy preservation*, by keeping sensitive multi-modal student and institutional data local; *(ii) personalization*, through modular architectures enabling tailored models for students, instructors, and institutions; and *(iii) equity and inclusivity*, by facilitating participation from underrepresented and resource-constrained entities. We finally identify various open research challenges, including studying of (i) inter-institution heterogeneous privacy regulations, (ii) the non-uniformity of data modalities' characteristics, (iii) the unlearning approaches for M3T FedFMs, (iv) the continual learning frameworks for M3T FedFMs, and (v) M3T FedFM model interpretability, which must be collectively addressed for practical deployment.

KEYWORDS

AI-assisted education, foundation models, multi-modal learning, multi-task learning, federated learning

1 Introduction

The modern era has witnessed a surge in the use of artificial intelligence (AI) and machine learning (ML) to support a range of education-related tasks. These include predicting student learning outcomes and success (Ofori et al., 2020), analyzing peer-to-peer collaboration patterns in online/in-person classrooms (Hridi et al., 2025), monitoring students with behavioral or neurodevelopmental needs (Barua et al., 2022), designing curricula for diverse educational settings (Ball et al., 2019),

improving the students' mental health (Ebrahimi et al., 2025), and enabling personalized learning experiences in self-regulated learning environments (Ingkavara et al., 2022). With the expansion of AI/ML applications in education, two parallel trends have emerged. On one hand, leveraging multiple data modalities (e.g., text, audio, video, image) collected in educational environments to train *multi-modal ML models*, capable of outperforming their uni-modal counterparts, has become a vibrant area of research (Xie et al., 2025; Griol et al., 2014). On the other hand, the use of these diverse modalities to train multi-task ML models that serve a variety of downstream educational tasks has also attracted growing attention (An et al., 2022; Geden et al., 2020). For example, video input in a humanoid robot can simultaneously support gesture tracking, object identification, and enhance speech understanding. As a result, the convergence of these trends has positioned *multi-modal multi-task (M3T) learning* at the forefront of AI/ML applications in education (Küchemann et al., 2025; Xu et al., 2024).

In parallel, the broader AI/ML community has undergone a significant transformation with the rise of M3T ML models. Initially popularized as foundation models (FMs) in the form of large language models (LLMs)—such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), LLaMA (Touvron et al., 2023), and PaLM (Chowdhery et al., 2023), which focused primarily on text-based tasks—they have now evolved into M3T FMs, such as ChatGPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), Llama-3 (Grattafiori et al., 2024), and CLIP (Radford et al., 2021). These emerging M3T FMs are capable of simultaneously processing multiple input modalities and capturing contextual relationships across multiple modalities and tasks. They have demonstrated remarkable generalization abilities, which is a result of (pre-)training on massive data. Despite their promise, M3T FMs remain largely underexplored in the education domain mostly due to their recent emergence. In particular, there exists a growing body of literature on the use-cases of AI/ML in the education domain that focuses on training/fine-tuning centralized LLMs for various pedagogical means, while M3T FMs remain to be rather unexplored. For example, the researchers in Moon et al. (2024) used LLMs to process multi-modal data for knowledge extraction and tracing which enhances the instructors' capabilities for automated assessment. Also, the authors of Hoq et al. (2025) proposed using LLMs to assist instructors in designing programming problems for students, leveraging both the novel ideas generated by the (human) instructor and the LLMs' ability to formulate an idea into a well-described problem. Further, the proposed method in Valverde-Rebaza et al. (2024) experimented with participants of non-computational fields on tackling a data analytics task and showed improved performance in the group of participants that employed an LLM to learn how to proceed with their task. Moreover, the work Rao et al. (2025) integrated an LLM in providing assistance to the students in abstraction, algorithmic thinking, and generalization when being taught about a new concept. They then evaluated the performance of the LLM-assisted framework with concepts from mathematics, biology, and networking to show how LLMs perform in teaching scientific problem-solving to middle school STEM students. Additionally, the researchers in Whitehead et al. (2025) experimented with multi-modal LLMs in analyzing instructors' non-verbal signals, such as posture, for determining collaborative learning effectiveness. Other existing works on LLMs in the education domain have focused on a variety of tasks, such as automated feedback generation and essay grading (Jia et al., 2024), question answering (Mitra et al., 2024), and intelligent tutoring systems (Molina et al., 2024). Following the aforementioned growing body of literature, the influence of FMs and LLMs is now being extended beyond traditional classroom contexts: they are being integrated into remote learning platforms, used to provide real-time feedback, and applied to educational research by offering analytical insights derived from classroom data (Küchemann et al., 2025).

Although the use of M3T FMs in education has been proposed only in a few recent studies (Küchemann et al., 2025; Xu et al., 2024), a critical and largely unresolved question remains: *Where does the data come from to train or fine-tune these data-hungry models in educational settings?* Specifically, educational tasks require domain-specific data, which is typically *siloed* across multiple infrastructure layers, ranging from school-level and departmental servers to college and university data repositories. A major obstacle in utilizing this data lies in stringent data-sharing restrictions, including privacy regulations on both institutional and regional levels (e.g., FERPA) (Zeide and Nissenbaum, 2018), ethical considerations, and student consent requirements (Prinsloo and Slade, 2018), all of which prohibit the transfer of sensitive educational data to external servers for model training. As a result, the *conventional centralized training/fine-tuning of M3T FMs* becomes infeasible for deploying them in many real-world educational environments. Even if centralized access to the above siloed data were possible (e.g., a statewide institution that uses its own data and aims to train a unified model for all students), the issue of data scarcity persists: high-quality, task-relevant educational data is often limited and fragmented across the isolated data sources (e.g., institutions across the nation). This challenge is further compounded by equity concerns, where models trained primarily on data from a single institution or demographically skewed population risk amplifying bias and marginalizing underrepresented or under-resourced learners. Without addressing these fundamental barriers, the deployment of M3T FMs in education, despite their theoretical promise, remains largely aspirational. In this paper, we propose a path forward by leveraging federated learning (FL) (McMahan et al., 2017), a pioneering distributed learning paradigm that enables collaborative model training without sharing raw data, for the training/fine-tuning of M3T FMs. Specifically, we give our perspective on *M3T Federated Foundation Models (FedFMs) for education*, a novel direction that opens up an untapped research space at the intersection of M3T FMs, FL, and privacy-preserving human-centered AI/ML.

The remainder of the paper is organized as follows. We begin by reviewing the relevant literature on M3T FMs and FL within educational contexts. We then explore the potential of M3T FedFMs to advance education through three key dimensions: (1) privacy preservation, (2) personalization, and (3) equity enhancement. Finally, we discuss the key challenges associated with

implementing M3T FedFMs in education and outline promising future research directions.

2 Overview on FL, M3T FMs, and M3T FedFMs

2.1 Federated learning (FL)

FL is a pioneering distributed ML paradigm that enables collaborative model training across multiple clients/participants (e.g., students, educators, institutions). FL operates through a series of global aggregation rounds, each comprising four key steps: (1) each client trains a local model on its own data (e.g., via stochastic gradient descent approach), (2) the locally trained models/gradients of clients are periodically sent to the server through uplink transmissions, (3) the server aggregates (e.g., via weighted averaging) the received trained models to create an updated *global model*, (4) the server broadcasts the updated global model to the clients, synchronizing their local models and initiating the next round of local model training. FL is widely regarded as a privacy-preserving distributed ML approach, as it replaces the transmission of sensitive raw data with model/gradient parameters. Note that although prior work has shown that even such transmitted parameters are still prone to adversarial attacks, such as *reconstruction attacks* that aim to regenerate training data (Chen C. et al., 2022) or *model inversion attacks* that extract client private information (Li et al., 2022) from the transmitted parameters, several countermeasures exist, including: (1) *Differential Privacy (DP)* (El Ouadrhiri and Abdelhadi, 2022), which injects calibrated noise into transmitted parameters to obfuscate the underlying client data used to train the model, and (2) *Functional Encryption* (Fang and Qian, 2021; Chang et al., 2023), where model parameters are encrypted in a way that allows only specific FL-related computations (e.g., model aggregation) to be performed without exposing the underlying client data.

By facilitating collaboration across a diverse network of institutions, FL helps overcome two key challenges typically faced when employing ML in education domains: *(1) data scarcity,* by enabling isolated and limited datasets to contribute collectively to a shared global model, and *(2) equity and inclusion,* by incorporating data from underrepresented or marginalized groups, distributed across different institutions, into the global model. Given these promising capabilities, FL has recently gained attention in the AI-assisted education literature (Fachola et al., 2023; Guo et al., 2020; Hridi et al., 2024; Chu et al., 2022, 2024). For example, in Fachola et al. (2023), FL is applied to the learning-analytics task of student dropout prediction, showing privacy-preserving training can attain performance comparable to centralized models while avoiding raw data centralization. Also, the authors in Guo et al. (2020) proposed FEEDAN, an FL framework that enables multi-institution pedagogical data analysis. Further, the researchers in Hridi et al. (2024) articulated how FL can benefit students, classrooms, and institutions while detailing technical, logistical, and ethical challenges for sustainable FL adoption in educational settings. Moreover, the work Chu et al. (2022) introduced an attention-based, subgroup-personalized FL approach with self-supervised behavioral pretraining that mitigates model biases and improves prediction accuracy across various student demographic groups in real-world online course datasets. As a follow-up, the work Chu et al. (2024) extended subgroup personalization with a multi-layer FL strategy (by course and demographics) for knowledge tracing and outcome prediction, yielding higher average performance and lower model variance (i.e., improved fairness) across student subgroups. Despite the tremendous contributions of the above-described body of works, the majority of these studies focus on the adoption of FL for training of conventional ML models (e.g., convolutional neural networks) and have yet to explore the FL-driven training/fine-tuning of M3T FMs within the education domain.

2.2 Multi-modal multi-task foundation models (M3T FMs)

M3T FMs are typically pre-trained on massive, heterogeneous datasets using self-supervised or unsupervised learning techniques, enabling them to acquire broad contextual understanding that can be effectively adapted to a wide range of domain-specific applications (e.g., enabling the operation of humanoid robots in domestic environments and extended reality systems) (Borazjani et al., 2025b; Nadimi et al., 2025). Fundamentally, M3T FMs extend the capabilities of conventional LLMs by incorporating multiple input modalities (e.g., text, audio, image, and video) and supporting a more diverse set of tasks (e.g., video understanding, conditional image generation, and image classification) alongside traditional text-based applications like question answering and text generation. Their potential in the education domain has recently been recognized (Küchemann et al., 2025), with emerging use cases such as intelligent tutoring, automated feedback generation, and curriculum design. While the seminal work in Küchemann et al. (2025) provides an in-depth analysis of the educational impact of M3T FMs, it does not address the specific training mechanisms behind these models and implicitly assumes their centralized training/fine-tuning. We therefore refer the reader to that work for broader context and position this paper as a complementary contribution, with the main focus on introducing the education community to the novelties of distributed, privacy-preserving, FL-driven training/fine-tuning of M3T FMs under the umbrella of M3T FedFMs.

To facilitate a clear understanding of their internal mechanisms, we next present a high-level overview of the general architecture of M3T FMs. Depicted in Figure 1, M3T FMs architecturally consist of three main components: (1) *modality encoders*, (2) *backbone*, and (3) *task heads*. They also can accommodate two additional components in their architecture, more commonly used in the scenarios entailing fine-tuning a pre-trained model to new contexts or tasks: *prompt tuner*, and *adapter*. For a more detailed description of these components refer to the caption/legend of Figure 1. M3T FMs support a wide range of training regimes, offering flexibility to either *train from scratch* or *fine-tune on downstream tasks* after large-scale pretraining.

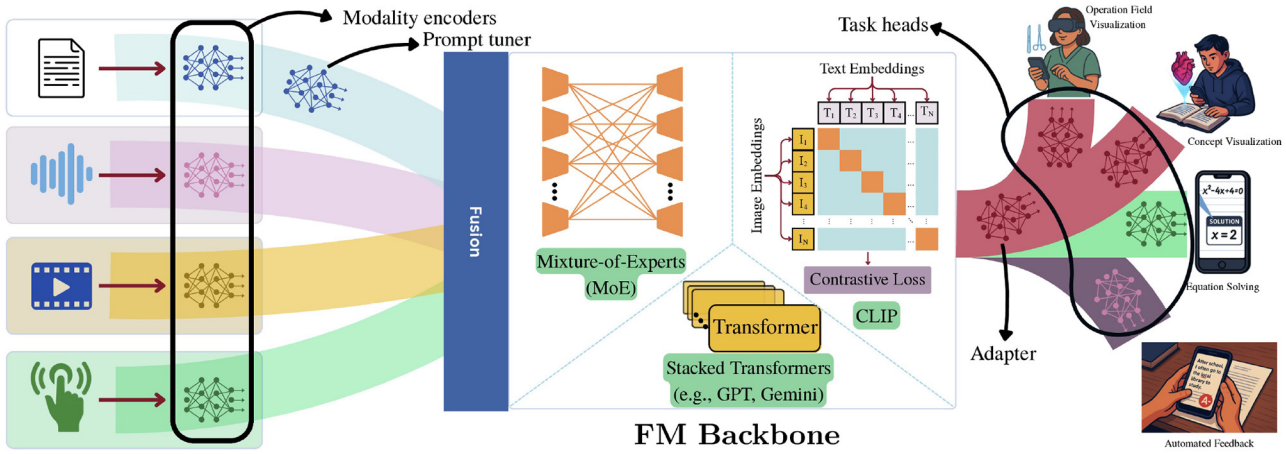

FIGURE 1
High-level architecture of an M3T FM, consisting of three major components. (1) *Modality encoders:* These modules transform raw input data from various modalities into intermediate embeddings. Fusion of modality-specific embeddings can be achieved via simple concatenation or more sophisticated mechanisms such as neural fusion blocks or attention-based integration. It can also be non-existent in some cases (e.g., CLIP). (2) *Backbone:* The backbone performs contextual reasoning, inter-modality correlation, and task generalization. It can be instantiated using various architectures, including Mixture-of-Experts (MoEs) (Chen and Zhang, 2024), dual encoders (as in CLIP; Radford et al., 2021), or stacked transformers (as in GPT models; Achiam et al., 2023). (3) *Task heads:* These are task-specific output layers that generate the results (e.g., classification labels, generated text) based on the representations produced by the backbone. M3T FMs also support lightweight fine-tuning strategies, where most of the model parameters are frozen and only a small subset is adapted. Three instances of these strategies are as follows. *(1) Prompt tuners* (Guo et al., 2024; Jia et al., 2022): Modules that condition input embeddings to align with task-specific contexts. *(2) Adapters* (Long et al., 2024; Zhang and Ré, 2022): Trainable parameter blocks inserted at different depths of the model to enable rapid adaptation to new tasks or modalities. *(3) Low-rank adaptations (LoRA)* (Yang et al., 2024; Wen and Chaudhuri, 2023): Efficient fine-tuning methods that decompose and optimize a low-rank subset of parameters or adapter weights, significantly reducing training cost while preserving performance.

2.3 Multi-modal multi-task federated foundation models (M3T FedFMs)

M3T FedFMs can be understood as the FL-driven training of M3T FMs across a distributed set of clients. Similar to conventional FL, M3T FedFMs operate through a series of global aggregation rounds, each comprising the standard aforementioned four steps: (1) local training, (2) uplink transmission of local models/gradients, (3) server-side model aggregation, and (4) broadcast of the updated global model back to clients. However, a key distinction between M3T FedFMs and traditional FL training of conventional ML models lies in the nature of local adaptation and aggregation. In M3T FedFMs, local training typically involves lightweight fine-tuning techniques, where only a subset of the model components, such as modality encoders, task heads, adapters, or prompt tuners, are updated. These components can then be selectively aggregated to produce a unified, fine-tuned global model that better generalizes across diverse client data distributions. This modular[1] training and aggregation approach (Chen and Zhang, 2024) enables clients to obtain local M3T FMs suitable for their own tasks or modalities.

While M3T FedFMs hold great promise for enabling high-performance, locally adapted M3T FMs across distributed clients (Chen and Zhang, 2024; Chen et al., 2024), their implementation introduces a range of challenges. These include inherited issues from conventional FL [data heterogeneity (Borazjani et al., 2025a, 2024), intermittent client connectivity (Parasnis et al., 2023), and limited client-side computational resources (Chai et al., 2019)], as well as challenges specific to M3T FMs (e.g., selecting which components or parameters to fine-tune and aggregate). Moreover, the integration of M3T FMs with FL introduces a set of unique challenges at their intersection (as will be explained later in the context of "Challenges and Open Directions"), challenges that are not fully addressed by existing work in either field alone and are unique to M3T FedFMs. It is worth mentioning that M3T FedFMs represent a highly emerging topic within the AI/ML community, with only a handful of early studies exploring their theoretical foundations (Chen et al., 2024; Chen and Zhang, 2024) and envisioning their applicability across domains such as healthcare (Li et al., 2025), embodied AI (Borazjani et al., 2025b), extended reality (Nadimi et al., 2025), and wireless edge/fog networks (Abdisarabshali et al., 2025). One promising domain still poised for breakthroughs enabled by M3T FedFMs is education, which we explore in the remainder of this paper to illuminate its unique applications and challenges.

2.4 Tailoring M3T FedFMs to education ecosystem

Here, we describe the system model envisioned for realizing a network of M3T FedFMs, as illustrated in Figure 2a. The system model follows the conventional "star topology" (Wu et al., 2024) in

1 Here, "modularity" refers to the capability of training various local M3T FM modules (e.g., encoders, task heads, adapters) independently across the clients.

FL setting which includes a *global server* interacting with a set of *clients*,[2] each described as follows:

1. **Global server**, which hosts a comprehensive M3T FedFM consisting of globally aggregated versions of each available component, including modality encoders, task-specific heads, backbone structures, and context-specific prompt tuners. This global server selectively broadcasts necessary model components to the clients.
2. **Clients**, comprising of three groups: institutions, instructors and students. The clients receive relevant subsets of the model components according to the modalities and tasks involved in the operations at each group from the global server. For instance, an instructor involved in a course might receive modality encoders and task heads corresponding to video, text, and image data for curriculum design, feedback generation, and content visualization. Also, a student in the same course may be provided with components that support video, audio, and text modalities necessary for classroom transcription, conceptual visualization, and supplementary research tasks. These clients subsequently transmit their locally trained/fine-tuned model parameters directly back to the global server for aggregation.

Note that, as depicted in Figure 2, additional modifications, such as introducing or removing connections between the individual client groups and changing the style of model aggregations, can be explored (e.g., to enhance model performance, convergence speed, and resource efficiency).

3 Unique applications of M3T FedFMs for education

Prior works (Ebrahimi et al., 2025; Küchemann et al., 2025; Hridi et al., 2024; Chu et al., 2022, 2024) have highlighted the broad benefits of adopting FL over conventional centralized ML approaches within the education domain. However, these discussions have largely focused on the application of FL to traditional ML models, such as multi-layer perceptrons and convolutional neural networks, without examining the unique potential of FL when applied to M3T FMs under the emerging M3T FedFM paradigm. In this section, we revisit three critical dimensions commonly emphasized in the aforementioned literature (i.e., privacy, personalization, and equity) and reframe them through the lens of M3T FedFMs. To solidify this discussion, we provide forward-looking examples that illustrate how the unique properties of M3T FedFMs can further advance these objectives in practical educational settings.

3.1 Dimension 1: privacy-enhanced M3T intelligence

M3T FedFMs naturally address longstanding data privacy concerns in educational ML applications by transmitting only model or gradient parameters between clients and the server, rather than raw sensitive data. By mitigating data privacy risks, M3T FedFMs enable greater participation from privacy-conscious individuals and institutions whose strict policies on raw-data sharing would have prevented them from contributing to the model training upon relying on the centralized training/fine-tuning of M3T FMs. This enhanced participation contributes to the development of models that generalize across varied educational settings. Below are three examples illustrating how M3T FedFMs can enable privacy-aware intelligence across M3T educational applications:

- **Student activity traces:** Future education systems may integrate smartphone-based learning companions and augmented reality (AR) headsets that passively collect contextual data such as study hours, geolocation patterns (e.g., time spent in libraries or study zones), ambient noise, or device interactions (Antonioli et al., 2014; Bower et al., 2014; Kleftodimos et al., 2023). These data sources span various modalities such as time-series logs, location metadata, ambient audio, and app usage sequences. Such data is often privacy-sensitive in nature as it can reveal private user information, and thus cannot be shared or transferred across the network and must remain local to its data collecting unit/device. M3T FedFMs can process such collected data across the smartphones and AR devices to support education-related downstream tasks such as predicting study burnout, recommending optimal study windows, or modeling learning motivation, all without exposing raw activity data to external servers.
- **Mental health assistance:** With the rise of Internet-of-Things (IoT) wearable biosensors, such as electroencephalogram (EEG) headbands and smartwatches equipped with heart rate variability (HRV) sensors, digital learning assistants can unobtrusively collect physiological, behavioral, and emotional cues to assess student well-being (Xu and Zhong, 2018; Kim et al., 2024; Aranberri-Ruiz et al., 2022). When combined with other privacy-sensitive modalities such as speech samples, writing patterns, and facial micro-expressions captured during classroom interactions or reflective exercises, this geo-distributed multi-modal data becomes a valuable resource for mental health analytics. Specifically, leveraging the collaborative training of M3T FedFMs over this data, digital learning assistants can be equipped with models for downstream tasks such as stress detection, mood tracking, and early intervention for depression or anxiety, all without exposing raw student data or violating privacy norms.
- **Student learning outcome prediction:** Future classrooms are expected to increasingly incorporate ambient sensors and camera-equipped devices, such as AI-driven smartboards and virtual reality (VR) learning environments, to assess student engagement in real-time (Lin et al., 2024; Grewe and Gie, 2023). In this technological realm, features such as gaze tracking, posture analysis, voice tone, and note-taking behavior, derived from privacy-sensitive modalities (e.g., audio and video), offer deep insights into cognitive and behavioral states. However, due to their sensitive nature, such data cannot be shared across institutions/classrooms.

2 Star topology refers to the usage of client-to-server links for model aggregation and broadcast.

10.3389/frai.2025.1683960

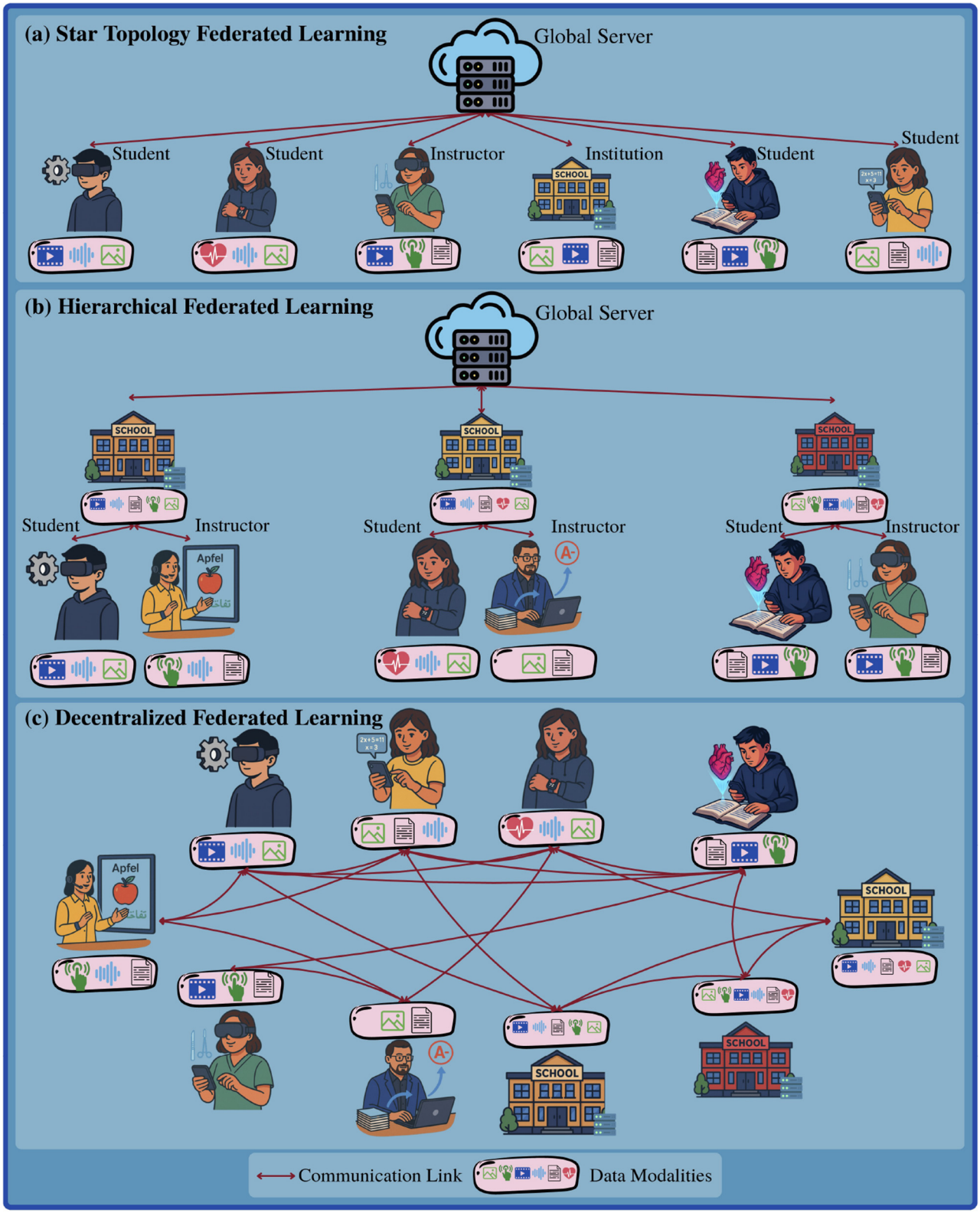

FIGURE 2
Various configurations can be adopted for M3T FedFM-enabled networks across clients in the education contexts, three of which are depicted: **(a)** Star Topology FL: A global server maintains a comprehensive global model encompassing all task and modality variations present across the system. All clients (i.e., institutions, instructors, and students) are directly connected to the global server. Client receive customized subsets of the global model tailored to their specific tasks and input modalities. Following local training/fine-tuning, clients updated models are transmitted back to the global server for aggregation. **(b)** Hierarchical FL: The global server distributes task- and modality-relevant subsets of the global model to each

(Continued)

FIGURE 2 (Continued)
institution. These subsets contain only the components required by the institution's associated instructors and students. Each institution then relays the necessary parts of the model to its end users based on their individual task and modality needs. After local training/fine-tuning, local models of end users are sent back to the institutions for aggregation. These institution-level models can then be further aggregated at the global server, combining insights across multiple institutions. **(c)** Decentralized FL: Without a global server, model aggregation is performed in a decentralized manner across clients (e.g., through consensus-based methods). Clients exchange model updates directly with their neighbors, aggregate the exchanged models on their devices, and proceed with the next round of model training/fine-tuning.

Here, M3T FedFMs provide a viable solution by enabling distributed, privacy-preserving model training across institutions/classrooms. This allows for the development of generalizable models capable of predicting attention span, learning progress, and knowledge retention without compromising student privacy.

3.2 Dimension 2: personalization of M3T intelligence

Personalization is a foundational pillar of effective education, reflecting the need to tailor the learning experience to the unique characteristics, preferences, and needs of various *educational entities*, including students, instructors, and institutions. M3T FedFMs, with their inherently modular and flexible architectures, are uniquely positioned to support model personalization. Specifically, the notion of *personalization* within M3T FedFMs can be understood from two complementary perspectives: *(1) Soft Personalization (local fine-tuning):* At the core of this perspective lies the ability to perform *local fine-tuning*, allowing clients to personalize prediction or generation tasks based on their own data distributions and contextual nuances, such as behavioral patterns, cultural or linguistic backgrounds, and interaction styles. M3T FedFMs leverage attention-based mechanisms and adaptable modules (e.g., prompt tuners and adapters) to personalize outputs dynamically, without the need for retraining the entire global model. *(2) Hard Personalization (Architectural/Component Adaptation):* In this approach, personalization is embedded in the model architecture itself. Specifically, each client (e.g., institution, instructor, or student) is served a version of the M3T FedFM that contains only the components relevant to their available data modalities and educational tasks, such as specific modality encoders (e.g., for audio or video) and task heads (e.g., for problem solving, code generation, or essay evaluation). This selective architectural deployment ensures efficiency and relevance while maintaining interoperability with the broader intelligent education ecosystem. In the following, we describe three examples of personalization across client groups (students, instructors, institutions):

- **Student personalization:** Students require personalized learning in various contexts, including concept explanation and problem-solving support. For example, some learners may benefit from visual explanations (e.g., video demonstrations or interactive visualizations), whereas others require textual or verbal guidance (e.g., spoken explanations, interactive Q&A sessions). M3T FedFMs can leverage multi-modal inputs such as handwritten notes (image modality), spoken questions (audio), and clickstream behaviors (event logs), to address these varied needs by locally fine-tuning models (e.g., via adapters) for personalized concept explanations, adaptive problem-solving assistance, and real-time feedback.
- **Instructor personalization:** Personalization for instructors often revolves around customized assessment generation and curricular support, which can differ based on the taught subject matter (e.g., essays, presentations, coding tasks, or visual projects). M3T FedFMs facilitate instructor personalization by employing specialized task heads and adapters that efficiently generate these customized assessments, minimizing preparation time and enhancing instructional quality. For instance, a language arts teacher may require automated feedback systems that assess creativity and narrative flow, whereas a computer science instructor might rely on auto-generated problem sets with dynamic test cases.
- **Institution personalization:** Institutions vary in curricular standards, target outcomes, and infrastructural capabilities. A vocational training center focused on mechanical skills may use video-based object manipulation tasks, while a liberal arts college may emphasize text analysis. M3T FedFMs support this diversity by allowing each institution to personalize the model's architecture, activating only relevant modalities (e.g., image and video for one, text and speech for another), and updating specific components (e.g., adapters or task heads) to reflect their educational mission.

3.3 Dimension 3: equitable and inclusive M3T intelligence

While *personalization* focuses on tailoring educational models to the individual preferences, behaviors, or needs of specific users (e.g., students, instructors, or institutions), equity and inclusivity emphasize *system-wide fairness and representation* across diverse social, cultural, linguistic, and infrastructural contexts. More specifically, personalization ensures that each user receives an optimized experience; equity and inclusivity ensure that every type of user (regardless of region, resources, identity, or participation patterns) is fairly represented in the training and utility of AI models. In this context, M3T FedFMs offer a practical pathway toward fostering a more equitable and inclusive educational ecosystem by accommodating variations in both curricular content and hardware/computation infrastructure across diverse learners and institutions. Specifically, unlike centralized models that often reflect dominant languages, curricula, or well-resourced environments/institutions, M3T FedFMs empower geographically distributed institutions to collaboratively train M3T FMs using locally relevant data while maintaining data sovereignty. Below, we

present three examples that highlight how equity and inclusivity are advanced through M3T FedFMs in education:

- **Cultural and linguistic representation:** In distributed educational systems, curriculum content, language, and cultural references may vary significantly across regions. Centralized models often fail to capture this diversity, especially for languages belonging to low-resource communities or locally relevant subjects. M3T FedFMs enable institutions to train/fine-tune local models using culturally specific data, such as textbooks in indigenous languages or region-specific historical texts. As a result, the global model becomes more representative of diverse educational needs and fosters inclusivity in its learned knowledge.
- **Infrastructure/hardware-aware participation:** Educational entities participating in M3T FedFM training or fine-tuning often differ widely in their computational resources. While some (e.g., universities or research centers) may possess high-performance servers capable of full-scale model training, others (e.g., individual students or small schools) may rely on resource-constrained devices such as smartphones or tablets. To accommodate such disparities in computation capabilities, M3T FedFMs support *modular engagement*, allowing resource-constrained education entities to contribute via lightweight computations (e.g., training only task heads or prompts) or to perform inference using relevant pre-trained components. This flexibility ensures that both high-end and low-resource education entities can benefit from and contribute to the collective learning process.
- **Gender bias and fairness mitigation:** Centralized training of M3T FMs often risks amplifying existing gender biases, particularly when the underlying data disproportionately represents one gender over others. Such imbalance can lead to models that perform better for overrepresented groups while exhibiting reduced accuracy, relevance, or responsiveness for underrepresented genders, ultimately reinforcing inequality in educational outcomes. M3T FedFMs offer a more equitable alternative compared to centralized M3T FM training/fine-tuning by enabling distributed, gender-diverse participation in model training. Institutions and users across different regions and demographics can contribute model/gradient parameters reflecting balanced or marginalized gender identities without compromising privacy.

3.4 Toward the implementation of M3T FedFMs in educational environments

Although M3T FedFMs have yet to be explored within the education domain, they are already gaining significant attention across other fields. Notably, the implementation strategies developed in other domains can serve as a basis for adapting M3T FedFMs to educational contexts. In particular, publicly available implementations from recent works (Borazjani et al., 2025b; Chen et al., 2024; Abdisarabshali et al., 2025; Fang et al., 2025) offer practical starting points for such adaptations. However, to evaluate the performance of these models in education-specific scenarios, access to publicly available datasets curated within educational settings is essential. In this regard, while existing works such as Xu et al. (2025) and Huang et al. (2025) have employed centralized training of LLMs and M3T FMs, their datasets can be repurposed for federated training by employing standard FL data partitioning techniques, such as those demonstrated in the aforementioned M3T FedFM implementations. Furthermore, given that M3T FedFMs are inherently capable of training across pooled datasets, a broader datalake can be constructed by integrating various education-relevant datasets. For example, datasets from Mathew et al. (2022) and Hiippala et al. (2021) on visual question answering, Lu et al. (2023) and Wang et al. (2024) on visual mathematical reasoning, and Sabuncuoglu and Sezgin (2023) on multimodal classroom analytics, although not originally designed for M3T FedFMs, can be *collectively* leveraged to support the development and evaluation of M3T FedFMs in realistic educational environments.

Despite the above-described implementation pathways, advancing M3T FedFM research in education will benefit from the development of a unified corpus that offers multi-modal, multi-task data spanning diverse user roles, i.e., students and teachers. Such a curated dataset should be designed to support federated evaluation by incorporating site-, school-, or device-level partitions, along with consented and de-identified metadata, to enable the study of non-IID (non-independent and identically distributed) data distributions and realistic client (e.g., student or institution) participation dynamics commonly encountered in educational settings. Moreover, realizing the practical deployment of M3T FedFMs in education entails addressing a range of domain-specific challenges, which we detail in Section 4.

4 Challenges and open directions of federated foundation models (FedFMs) in education

Despite their potential, M3T FedFMs face unique deployment challenges in education. Below, we formulate overarching research questions aimed at addressing them.

4.1 Inter-institution heterogeneous privacy regulations and their impact on data availability

As educational institutions across the globe adopt AI-driven systems, the deployment of M3T FedFMs in practice will become increasingly constrained by diverse and evolving privacy regulations. Specifically, legal frameworks such as the General Data Protection Regulation (GDPR) in the European Union and the Family Educational Rights and Privacy Act (FERPA) in the United States impose different requirements on how sensitive educational data, such as video, voice recordings, and physiological signals, can be used or shared across the education systems. These jurisdictional differences introduce a significant barrier to uniform collaboration in model training across institutions. In FL, where raw data remains local and only model/gradient parameters

are shared, the variability in legal constraints manifests as DP budgets or encryption standards applied to local updates. For example, an institution in a stricter jurisdiction may be obligated to inject stronger DP noise into model updates derived from video or speech data, reducing their informativeness relative to updates from regions with less strict regulations in parameter sharing. As a result, the aggregation process in M3T FedFMs becomes non-trivial: updates now vary not only in content and modality but also in privacy-induced distortion levels. This privacy heterogeneity is especially problematic in educational contexts, where certain tasks (e.g., affect recognition or engagement tracking) heavily rely on privacy-sensitive modalities. In particular, a naïve aggregation of differentially distorted model updates from clients can inadvertently amplify the influence of under-regulated clients while marginalizing updates from more privacy-conscious clients, leading to skewed global model behavior.

So far, the study of regulation-aware model aggregation in multi-modal FL has been limited to Liu X. et al. (2024), which introduces a multi-modal gradient inversion attack and defense framework for conventional multi-modal ML models (e.g., multi-encoder neural networks) that exploits cross-modal correlations to reconstruct multi-modal inputs. This leaves regulation-driven privacy mechanisms in M3T FedFMs an unexplored area, raising an urgent research question: *How can trust-aware aggregation mechanisms be designed in M3T FedFMs to fairly and effectively integrate updates subject to heterogeneous privacy regulations, while preserving convergence, modality balance, and cross-jurisdictional equity in global model behavior?*

4.2 Modality-specific characteristics and transmission overhead

While the above-discussed jurisdictional differences constrain data handling policies across institutions, an orthogonal and equally critical dimension arises from the inherent privacy sensitivity and computational demands associated with *different input modalities* themselves. Specifically, in educational settings, the multi-modal nature of M3T FedFMs introduces distinct privacy risks across different input streams. While modalities such as text logs or quiz responses are generally considered lower-risk, others such as eye gaze, facial expressions, EEG signals, or audio recordings carry higher privacy sensitivity due to their biometric nature and potential to reveal deeply personal information. As ambient sensing technologies become more prevalent in classrooms, ensuring appropriate protection for these high-risk modalities is essential. Compounding this challenge is the asymmetric contribution of modalities to different downstream educational tasks supported by M3T FedFMs. For example, facial expressions and vocal tone might be pivotal for engagement estimation, whereas textual responses are more relevant for concept mastery or personalized feedback. This variation makes uniform privacy-preserving strategies infeasible. Instead, techniques such as DP must be selectively applied based on each modality's sensitivity and its utility for specific learning objectives.

Given that privacy calibration across modalities remains underexplored in M3T FedFMs, this raises a key open research question: *How can privacy-preserving techniques in M3T FedFMs be dynamically adapted across modalities to balance privacy risks and task-specific utility, especially when different modalities contribute asymmetrically to various tasks?*

4.3 User-initiated data removal and the need for federated unlearning

A critical challenge in privacy-aware educational systems is enabling users and institutions to revoke their data contributions after participation, an increasingly important right under regulations such as GDPR and FERPA. In the context of M3T FedFMs, this necessitates the development of effective *federated unlearning* mechanisms: methods that can selectively remove the influence of a client's data from the global model without requiring model retraining from scratch. Unlike traditional centralized models, M3T FedFMs present unique obstacles for unlearning due to their modular structure, multi-modal data inputs, and decentralized training process. Specifically, client contributions are distributed across various components, such as modality encoders, adapters, and task heads, making their influence deeply entangled within the global model parameters. This makes it difficult to (1) accurately isolate and remove a client's impact, and (2) maintain the global model's utility, adaptation capability, and fairness for the remaining participants.

Federated unlearning has begun to receive attention in classical FL settings (Halimi et al., 2022; Liu Z. et al., 2024). Particularly, Halimi et al. (2022) proposed a client-level federated unlearning method that performs local unlearning at the departing client and then runs a few additional FL rounds with the remaining clients, removing the necessity for the server to have access to historical updates. Yet, federated unlearning remains entirely unexplored in the context of M3T FedFMs. This gap raises a crucial and timely research question: *How can we design scalable, component-aware unlearning techniques for M3T FedFMs that ensure efficient, verifiable removal of user-contributed knowledge, while preserving model performance, fairness, and adaptability across heterogeneous and privacy-sensitive educational environments?*

4.4 Continual learning

Educational systems are inherently dynamic: new subjects are introduced, pedagogical approaches evolve, institutional priorities shift, and the nature of data modalities continually changes. In such an environment, static M3T FedFMs can quickly become misaligned with emerging learning objectives or newly introduced input modalities, limiting their adaptability and long-term relevance. This issue is amplified in federated/distributed settings, where decentralized and asynchronous model updates from various clients can complicate continual model adaptation. A central challenge arising from this distributed evolution is *federated catastrophic forgetting*: as local model updates from clients are integrated sequentially or asynchronously, newly learned patterns (often specific to a subset of clients)

can inadvertently overwrite previously acquired knowledge encoded in the global model. This forgetting effect is especially detrimental in education contexts, where preserving knowledge pertaining to personalized pedagogical methods (e.g., a student's learning history or an institution's domain-specific curriculum) is critical for long-term model effectiveness. Specifically, without robust mechanisms to manage incremental learning and protect previously acquired knowledge, M3T FedFMs risk deteriorating in performance over time, especially for clients whose data distributions are no longer active but remain pedagogically important. This compromises the global model's reliability and generalization, reducing its value over time.

Continual learning remains an emerging topic in the FM literature (Ostapenko et al., 2022; Yi et al., 2023; Yang et al., 2025). For instance, in Ostapenko et al. (2022), the authors have benchmarked the use of pre-trained vision encoders as feature extractors for continual learning, showing that latent-space replay can achieve strong performance at low compute costs while highlighting how encoder characteristics and pre-training data influence knowledge forgetting and transfer. Also, the researchers in Yi et al. (2023) have explored vision-language models as medical FMs and demonstrated that rehearsal-based continual learning substantially improves cross-domain and cross-task generalization. Despite these efforts, continual learning remains to be highly unexplored in the context of M3T FedFMs, which raises a pressing research question: *How can continual learning strategies for M3T FedFMs be designed to balance asynchronous client updates with prior knowledge retention, effectively mitigating federated catastrophic forgetting while supporting evolving educational tasks and modalities?*

4.5 Model interpretability

In educational settings, transparency and trust are paramount. Specifically, AI models that influence high-stakes decisions, such as grading, personalized feedback, skill assessment, or behavioral monitoring, must be explainable/interpretable to a wide range of stakeholders, including students, instructors, administrators, and parents. When the reasoning behind model outputs is unclear or opaque, it can undermine confidence, hinder adoption, and raise critical ethical concerns. Interpretability becomes especially challenging in the context of M3T FedFMs, which introduce a compounded layer of complexity. First, their multi-modal nature involves inputs such as text, audio, video, and physiological signals, each varying in semantics, structure, and abstraction. Second, their modular architecture, comprising independently functioning components, such as prompt tuners, adapters, and task heads, makes it difficult to attribute predictions to specific modules or modalities. Third, the federated training paradigm adds further opacity: models are updated across decentralized clients with non-IID data distributions, meaning that the global model's behavior emerges from a combination of locally trained, heterogeneous data sources. As a result, interpretability tools applied to the global model may fail to capture client-specific nuances or may produce misleading explanations when generalized across participants. Together, compared to traditional centralized ML models, these challenges make it substantially harder to trace how specific input modalities or data features influence a given prediction, to identify hidden biases, or to justify model decisions for each separate client.

Model interpretability is a rather nascent area in FMs (Chen J. et al., 2022; Rajendran et al., 2024; Fu et al., 2024), with the notable work (Chen J. et al., 2022) introducing RNA-FM, which is a large-scale self-supervised FM that learns interpretable sequential and evolutionary features for improvement in function prediction tasks. Nevertheless, interpretability remains almost unexplored in M3T FedFMs, which raises a foundational open research question: *How can we design inherently interpretable M3T FedFMs that not only provide accurate outputs but also generate actionable, role-sensitive explanations aligned with the pedagogical and ethical demands of education systems?*

5 Conclusion

In this perspective paper, we examined the emerging convergence of federated learning and foundation models within the education domain, framing the concept of *multi-modal, multi-task federated foundation models (M3T FedFMs)* as a transformative step toward next-generation intelligent educational systems. We outlined the architectural structure of M3T FedFMs and discussed how their modular and distributed design offers a framework to address core needs in education, specifically: preserving privacy in learning processes, enabling model personalization, and promoting equity and inclusivity. We also identified a set of open challenges and articulated key research questions designed to guide future inquiry in this nascent area.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KB: Methodology, Visualization, Conceptualization, Writing – original draft, Investigation, Writing – review & editing. NK: Writing – review & editing, Writing – original draft. RS: Writing – original draft, Conceptualization, Writing – review & editing. BA: Writing – original draft, Conceptualization, Writing – review & editing. SH: Methodology, Writing – original draft, Conceptualization, Writing – review & editing, Investigation, Supervision.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors acknowledge support from the U.S. National Science Foundation (NSF) under Grants ECCS 2512911 and IIS 2426837.

Conflict of interest

NK was employed at Adobe.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abdisarabshali, P., Nadimi, F., Borazjani, K., Liwang, M., Langberg, M., and Hosseinalipour, S. (2025). Hierarchical federated foundation models over wireless networks for multi-modal multi-task intelligence: Integration of edge learning with D2D/P2P-enabled fog learning architectures. *arXiv* [preprint] arXiv:2506.05683. doi: 10.1109/MWC.009.2300501

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). GPT-4 technical report. *arXiv* [preprint] arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774

An, S., Kim, J., Kim, M., and Park, J. (2022). "No task left behind: Multi-task learning of knowledge tracing and option tracing for better student assessment" in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vancouver, BC), 4424–4431.

Antonioli, M., Blake, C., and Sparks, K. (2014). Augmented reality applications in education. *J. Technol. Stud.* 40:96–107. doi: 10.21061/jots.v40i2.a.4

Aranberri-Ruiz, A., Aritzeta, A., Olarza, A., Soroa, G., and Mindeguia, R. (2022). Reducing anxiety and social stress in primary education: a breath-focused heart rate variability biofeedback intervention. *Int. J. Environm. Res. Public Health* 19:10181. doi: 10.3390/ijerph191610181

Ball, R., Duhadway, L., Feuz, K., Jensen, J., Rague, B., and Weidman, D. (2019). "Applying machine learning to improve curriculum design" in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (New York: ACM), 787–793.

Barua, P. D., Vicnesh, J., Gururajan, R., Oh, S. L., Palmer, E., Azizan, M. M., et al. (2022). Artificial intelligence enabled personalised assistive tools to enhance education of children with neurodevelopmental disorders—a review. *Int. J. Environm. Res. Public Health* 19:1192. doi: 10.3390/ijerph19031192

Borazjani, K., Abdisarabshali, P., Khosravan, N., and Hosseinalipour, S. (2025a). Redefining non-iid data in federated learning for computer vision tasks: Migrating from labels to embeddings for task-specific data distributions. *arXiv* [preprint] arXiv:2503.14553.

Borazjani, K., Abdisarabshali, P., Nadimi, F., Khosravan, N., Liwang, M., Wang, X., et al. (2025b). Multi-modal multi-task (M3T) federated foundation models for embodied AI: Potentials and challenges for edge integration. *arXiv* [preprint] arXiv:2505.11191.

Borazjani, K., Khosravan, N., Ying, L., and Hosseinalipour, S. (2024). Multi-modal federated learning for cancer staging over non-iid datasets with unbalanced modalities. *IEEE Trans. Med. Imag.* 44, 556–573. doi: 10.1109/TMI.2024.3450855

Bower, M., Howe, C., McCredie, N., Robinson, A., and Grover, D. (2014). Augmented reality in education-cases, places and potentials. *Educ. Media Int.* 51, 1–15. doi: 10.1080/09523987.2014.889400

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33* (Vancouver, BC), 1877–1901.

Chai, Z., Fayyaz, H., Fayyaz, Z., Anwar, A., Zhou, Y., Baracaldo, N., et al. (2019). "Towards taming the resource and data heterogeneity in federated learning" in *2019 USENIX Conference on Operational Machine Learning (OpML 19)* (Santa Clara, CA), 19–21.

Chang, Y., Zhang, K., Gong, J., and Qian, H. (2023). Privacy-preserving federated learning via functional encryption, revisited. *IEEE Trans. Inform. Forens. Security* 18, 1855–1869. doi: 10.1109/TIFS.2023.3255171

Chen, C., Lyu, L., Yu, H., and Chen, G. (2022). Practical attribute reconstruction attack against federated learning. *IEEE Trans. Big Data.* 10, 851–863. doi: 10.1109/TBDATA.2022.3159236

Chen, H., Zhang, Y., Krompass, D., Gu, J., and Tresp, V. (2024). FedDAT: an approach for foundation model finetuning in multi-modal heterogeneous federated learning. *Proc. AAAI Conf. Artif. Intellig.* 38, 11285–11293. doi: 10.1609/aaai.v38i10.29007

Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., et al. (2022). Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv* [preprint] arXiv:2204.00300. doi: 10.1101/2022.08.06.503062

Chen, J., and Zhang, A. (2024). On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning. *Proc. AAAI Conf. Artif. Intellig.* 38, 11311–11319. doi: 10.1609/aaai.v38i10.29010

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al. (2023). Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* 24, 1–113.

Chu, Y.-W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A., et al. (2022). "Mitigating biases in student performance prediction via attention-based personalized federated learning" in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3033–3042.

Chu, Y.-W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A. S., et al. (2024). Multi-layer personalized federated learning for mitigating biases in student predictive analytics. *IEEE Trans. Emerg. Topics Comput.* 10, 851–863. doi: 10.1109/TETC.2024.3407716

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint] arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Ebrahimi, M., Sahay, R., Hosseinalipour, S., and Akram, B. (2025). The transition from centralized machine learning to federated learning for mental health in education: A survey of current methods and future directions. *arXiv* [preprint] arXiv:2501.11714. doi: 10.48550/arXiv.2501.11714

El Ouadrhiri, A., and Abdelhadi, A. (2022). Differential privacy for deep and federated learning: A survey. *IEEE Access* 10:22359–22380. doi: 10.1109/ACCESS.2022.3151670

Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., and Fariello, M. I. (2023). Federated learning for data analytics in education. *Data* 8:43. doi: 10.3390/data8020043

Fang, H., and Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet* 13:94. doi: 10.3390/fi13040094

Fang, W., Han, D.-J., Yuan, L., Hosseinalipour, S., and Brinton, C. G. (2025). Federated sketching lora: On-device collaborative fine-tuning of large language models. *arXiv* [preprint] arXiv:2501.19389. doi: 10.48550/arXiv.2501.19389

Fu, S., Chen, Y., Wang, Y., and Tao, D. (2024). On championing foundation models: From explainability to interpretability. *arXiv* [preprint] arXiv:2410.11444. doi: 10.48550/arXiv. 2410.11444

Geden, M., Emerson, A., Rowe, J., Azevedo, R., and Lester, J. (2020). Predictive student modeling in educational games with multi-task learning. *Proc. AAAI Conf. Artif. Intellig.* 34, 654–661. doi: 10.1609/aaai.v34i01.5406

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 herd of models. *arXiv* [preprint] arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783

Grewe, M., and Gie, L. (2023). Can virtual reality have a positive influence on student engagement? *South Afri. J. Higher Educ*. 37, 124–141. doi: 10.20853/37-5-5815

Griol, D., Molina, J. M., and De Miguel, A. S. (2014). Developing multimodal conversational agents for an enhanced e-learning experience. *ADCAIJ* 3, 13–26. doi: 10.14201/ADCAIJ2014381326

Guo, S., Zeng, D., Dong, S. (2020). Pedagogical data analysis via federated learning toward education 4.0. *Am. J. Educ. Inform. Technol*. 4, 56–65. doi: 10.1145/3404709.3404751

Guo, W., Li, S., and Yang, J. (2024). "Scattering prompt tuning: A fine-tuned foundation model for sar object recognition" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 3056–3065.

Halimi, A., Kadhe, S., Rawat, A., and Baracaldo, N. (2022). Federated unlearning: How to efficiently erase a client in fl? *arXiv* [preprint] arXiv:2207.05521. doi: 10.48550/arXiv. 2207.05521

Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., et al. (2021). AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Eval*. 55, 661–688. doi: 10.1007/s10579-020-09517-1

Hoq, M., Vandenberg, J., Jiao, S., Lee, S., Mott, B., Norouzi, N., et al. (2025). Facilitating instructors-LLM collaboration for problem design in introductory programming classrooms. *arXiv* [preprint] arXiv:2504.01259. doi: 10.48550/arXiv. 2504.01259

Hridi, A. P., Hoq, M., Gao, Z., Lynch, C., Sahay, R., Hosseinalipour, S., et al. (2025). "Privacy-preserving distributed link predictions among peers in online classrooms using federated learning" in *Conference on Educational Data Mining (EDM)* (Palermo).

Hridi, A. P., Sahay, R., Hosseinalipour, S., and Akram, B. (2024). "Revolutionizing AI-assisted education with federated learning: A pathway to distributed, privacy-preserving, and debiased learning ecosystems" in *Proceedings of the AAAI Symposium Series*, 297–303.

Huang, C., Zhu, J., Ji, Y., Shi, W., Yang, M., Guo, H., et al. (2025). A multi-modal dataset for teacher behavior analysis in offline classrooms. *Scientific Data* 12:1115. doi: 10.1038/s41597-025-05426-6

Ingkavara, T., Panjaburee, P., Srisawasdi, N., and Sajjapanroj, S. (2022). The use of a personalized learning approach to implementing self-regulated online learning. *Comp. Educ.: Artif. Intellig*. 3:100086. doi: 10.1016/j.caeai.2022.100086

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., et al. (2022). "Visual prompt tuning" in *European Conference on Computer Vision* (Cham: Springer), 709–727.

Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., et al. (2024). "On assessing the faithfulness of LLM-generated feedback on student assignments" in *Proceedings of the 17th International Conference on Educational Data Mining* (Atlanta, GA), 491–499.

Kim, H. J., Park, Y., and Lee, J. (2024). The validity of heart rate variability (HRV) in educational research and a synthesis of recommendations. *Educ. Psychol. Rev*. 36:42. doi: 10.1007/s10648-024-09878-x

Kleftodimos, A., Moustaka, M., and Evagelou, A. (2023). Location-based augmented reality for cultural heritage education: Creating educational, gamified location-based AR applications for the prehistoric lake settlement of dispilio. *Digital* 3, 18–45. doi: 10.3390/digital3010002

Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., et al. (2025). On opportunities and challenges of large multimodal foundation models in education. *NPJ Sci. Learn*. 10:11. doi: 10.1038/s41539-025-00301-w

Li, J., Rakin, A. S., Chen, X., He, Z., Fan, D., and Chakrabarti, C. (2022). "ResSFL: A resistance transfer framework for defending model inversion attack in split federated learning" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 10194–10202.

Li, X., Peng, L., Wang, Y.-P., and Zhang, W. (2025). Open challenges and opportunities in federated foundation models towards biomedical healthcare. *BioData Mining* 18:2. doi: 10.1186/s13040-024-00414-9

Lin, X. P., Li, B. B., Yao, Z. N., Yang, Z., and Zhang, M. (2024). The impact of virtual reality on student engagement in the classroom-a critical review of the literature. *Front. Psychol*. 15:1360574.

Liu, X., Cai, S., He, R., and Yuan, J. (2024). Mutual gradient inversion: Unveiling privacy risks of federated learning on multi-modal signals. *IEEE Signal Proc. Letters*. 31, 2745–2749. doi: 10.1109/LSP.2024.3453200

Liu, Z., Jiang, Y., Shen, J., Peng, M., Lam, K.-Y., Yuan, X., et al. (2024). A survey on federated unlearning: Challenges, methods, and future directions. *ACM Comp. Surv*. 57, 1–38. doi: 10.1145/3679014

Long, G., Shen, T., Jiang, J., and Blumenstein, M. (2024). "Dual-personalizing adapter for federated foundation models" in *Advances in Neural Information Processing Systems*, 39409–39433.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., et al. (2023). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv* [preprint] arXiv:2310.02255. doi: 10.48550/arXiv.2310.02255

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. (2022). "Infographicvqa" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 1697–1706.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). "Communication-efficient learning of deep networks from decentralized data" in *Artificial Intelligence and* Statistics (New York: PMLR), 1273–1282.

Mitra, C., Miroyan, M., Jain, R., Kumud, V., Ranade, G., and Norouzi, N. (2024). "RetLLM-E: retrieval-prompt strategy for question-answering on student discussion forums" in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vancouver, BC), 23215–23223.

Molina, I. V., Montalvo, A., Ochoa, B., Denny, P., and Porter, L. (2024). Leveraging LLM tutoring systems for non-native english speakers in introductory cs courses. *arXiv* [preprint] arXiv:2411.02725. doi: 10.48550/arXiv.2411.02725

Moon, H., Davis, R., Neshaei, S. P., and Dillenbourg, P. (2024). Using large multimodal models to extract knowledge components for knowledge tracing from multimedia question information. *arXiv* [preprint] arXiv:2409.20167. doi: 10.48550/arXiv.2409.20167

Nadimi, F., Abdisarabshali, P., Borazjani, K., Chakareski, J., and Hosseinalipour, S. (2025). Multi-modal multi-task federated foundation models for next-generation extended reality systems: Towards privacy-preserving distributed intelligence in AR/VR/MR. *arXiv* [preprint] arXiv:2506.05683. doi: 10.48550/arXiv.2506. 05683

Ofori, F., Maina, E., and Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. *J. Inform. Technol*. 4, 33–55.

Ostapenko, O., Lesort, T., Rodriguez, P., Arefin, M. R., Douillard, A., Rish, I., et al. (2022). "Continual learning with foundation models: An empirical study of latent replay" in *Conference on Lifelong Learning Agents* (New York: PMLR), 60–91.

Parasnis, R., Hosseinalipour, S., Chu, Y.-W., Chiang, M., and Brinton, C. G. (2023). "Connectivity-aware semi-decentralized federated learning over time-varying D2D networks" in *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing* (Washington DC), 31–40.

Prinsloo, P., and Slade, S. (2018). "Student consent in learning analytics: the devil in the details?" in *Learning Analytics in Higher* Education (London: Routledge), 118–139.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision" in *International Conference on Machine* Learning (New York: PMLR), 8748–8763.

Rajendran, G., Buchholz, S., Aragam, B., Schölkopf, B., and Ravikumar, P. (2024). Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv* [preprint] arXiv:2402.09236. doi: 10.48550/arXiv.2402. 09236

Rao, A., Piryani, K., Jiang, S., Barnes, T., Albert, J., Hill, M., et al. (2025). *Leveraging Large Language Models to Promote AI-Infused Stem Problem-Solving for Middle School Students*. Palermo; Sicily.

Sabuncuoglu, A., and Sezgin, T. M. (2023). Multimodal group activity dataset for classroom engagement level prediction. *arXiv* [preprint] arXiv:2304.08901. doi: 10.48550/arXiv.2304.08901

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv* [preprint] arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv* [preprint] arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971

Valverde-Rebaza, J., González, A., Navarro-Hinojosa, O., and Noguez, J. (2024). Advanced large language models and visualization tools for data analytics learning. *Front. Educ*. 9:1418006. doi: 10.3389/feduc.2024.1418006

Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., et al. (2024). "Measuring multimodal mathematical reasoning with math-vision dataset" in *Advances in Neural Information Processing Systems* (Vancouver, BC), 95095–95169.

Wen, Y., and Chaudhuri, S. (2023). Batched low-rank adaptation of foundation models. *arXiv* [preprint] arXiv:2312.05677. doi: 10.48550/arXiv.2312.05677

Whitehead, R., Nguyen, A., and Järvelä, S. (2025). Utilizing multimodal large language models for video analysis of posture in studying collaborative learning: a case study. *J. Learn. Analyt*. 12, 186–200. doi: 10.18608/jla.2025.8595

Wu, J., Dong, F., Leung, H., Zhu, Z., Zhou, J., and Drew, S. (2024). Topology-aware federated learning in edge computing: a comprehensive survey. *ACM Comp. Surv.* 56, 1–41. doi: 10.1145/3659205

Xie, Y., Yang, L., Zhang, M., Chen, S., and Li, J. (2025). A review of multimodal interaction in remote education: technologies, applications, and challenges. *Appl. Sci.* 15:3937. doi: 10.3390/app15073937

Xu, B., Bai, Y., Sun, H., Lin, Y., Liu, S., Liang, X., et al. (2025). Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv* [preprint] arXiv:2505.16160. doi: 10.48550/arXiv.2505.16160

Xu, J., and Zhong, B. (2018). Review on portable eeg technology in educational research. *Computers in Human Behavior* 81:340–349. doi: 10.1016/j.chb.2017.12.037

Xu, T., Tong, R., Liang, J., Fan, X., Li, H., and Wen, Q. (2024). Foundation models for education: promises and prospects. *IEEE Intellig, Syst*. 39, 20–24. doi: 10.1109/MIS.2024.3398191

Yang, M., Chen, J., Zhang, Y., Liu, J., Zhang, J., Ma, Q., et al. (2024). Low-rank adaptation for foundation models: a comprehensive review. *arXiv* [preprint] arXiv:2501.00365. doi: 10.48550/arXiv.2501.00365

Yang, Y., Zhou, J., Ding, X., Huai, T., Liu, S., Chen, Q., et al. (2025). Recent advances of foundation language models-based continual learning: a survey. *ACM Comp. Surv.* 57, 1–38. doi: 10.1145/3705725

Yi, H., Qin, Z., Lao, Q., Xu, W., Jiang, Z., Wang, D., et al. (2023). Towards general purpose medical ai: Continual learning medical foundation model. *arXiv* [preprint] arXiv:2303.06580. doi: 10.48550/arXiv.2303.06580

Zeide, E., and Nissenbaum, H. (2018). Learner privacy in moocs and virtual education. *Theory Res. Educ.* 16:280–307. doi: 10.1177/1478778518815340

Zhang, M., and Ré, C. (2022). "Contrastive adapters for foundation model group robustness" in *Advances in Neural Information Processing Systems* (New Orleans, LA), 21682–21697.