机器遗忘赋能安全智能教育应用实践

吴伊雨，刘毓，邝瑶雯，王涛*
华中师范大学人工智能教育学部 数字教育湖北省重点实验室 湖北 武汉 430079
*tmac@ccnu.edu.cn

【摘要】“人工智能+教育”的深度融合依赖于对教育数据的无感采集与融合计算，这种数据密集型的研究范式极大地推动了智能教育的发展与进步。然而，随着基于人工智能的教育模型的广泛应用，一系列安全风险也随之浮现，这些风险覆盖了从数据采集、模型训练到输出结果的全过程。本研究通过系统分析，揭示了在这一过程中可能遇到的安全隐患：数据采集阶段存在数据正确性与安全性的双重挑战；模型训练阶段则面临算法偏见与模型安全性的风险；而在数据应用阶段，结果的准确性、公平性及安全性同样令人担忧；这些风险实际根源于数据问题。因此，本研究致力于探索先进技术手段以应对这些挑战，特别分析了机器遗忘技术在缓解安全风险方面的潜力，并通过人脸识别模型保护的实例进行了实践验证，从而为确保模型安全提供了切实可行的方法。
【关键词】人工智能教育模型；安全风险；机器遗忘；应用实践

1.引言

2024 年，全国教育工作会议提出“要不断开辟教育数字化新赛道，以智能化赋能教育治理，引领教育变革创新”(教育部, 2024)。人工智能以其广泛且强大的数据驱动应用的能力，不断与教育深度融合，通过对学生等教育参与者数据的无感采集、融合计算和及时反馈，在教育过程表征、教育主体刻画、教育规律挖掘、教育服务变革等方面发挥重要的作用(杨宗凯 & 李卿, 2020)，并衍生出学习行为分析、教育认知诊断、教育知识追踪、教育情感计算、学习投入识别等研究领域(王一岩 & 郑永和, 2022)；利用数据密集型科学的研究范式探究教育教学的内在机理和演变规律，助力智能教育发展与变革(郑永和 et al., 2021)。

国务院《“十四五“数字经济发展规划(国务院, 2022)中明确提出要提升数据安全保障水平，规范数据从采集、存储到处理、销毁的全生命周期管理。然而，基于人工智能的教育模型的应用在带来诸多便利和优势的同时，也引发了诸多安全风险和挑战，尤其是 AI 大模型的应用，更是将安全伦理问题推向了公众视野的中心，这些挑战广泛涵盖了数据隐私保护、算法偏见等关键问题(褚乐阳 et al., 2024)。这些安全风险不仅限于教育大模型，而是普遍存在于所有基于人工智能的教育模型中，贯穿于数据采集、模型训练、结果输出的全过程，对教育数据的安全性、教育公平性和伦理性构成了威胁。

因此，本研究旨在深入探讨基于人工智能的教育模型应用过程中可能产生的安全风险，探索有效的应对策略，旨在在保护教育数据安全的同时，充分利用深度学习技术的潜力，推动教育的数字化转型和现代化进程，为教育技术学的学术研究和实践应用提供参考。

2. 人工智能在教育领域应用的安全风险

人工智能在教育领域的应用虽然为教学活动、学习过程、教育管理以及教学评价等多个关键环节带来了显著的促进作用，在提升教育效率与质量方面展现出了积极的效能，但也引发了隐私泄露、算法偏见、教育不公平等一系列伦理问题。王佑镁等基于 4R 危机管理模型将教育人工智能的伦理问题归纳为软件相关风险、硬件相关风险、环境相关风险和人件相关风险四类(王佑镁 et al., 2024)，吴砥等从教育安全、教育公平、教育关系三个方面分析应用潜在伦理风险(吴砥 & 吴河江, 2024)，吴河江等从技术、内容、数据、算法四个角度进行了风险分析(吴河 & 江吴砥, 2024)，杨俊锋等从数据、算法和应用三层面对伦理问题进行了总结及成因分析(褚娟, 2024)。教育人工智能伦理风险问题正受到学界的广泛关注和积极研究。本研究旨在聚焦人工智能教育模型生成的全过程，深入剖析这一完整过程中潜藏的安全风险，具体涵盖模型数据采集、模型训练，模型输出三个环节。

2.1. 人工智能教育模型数据采集阶段的安全风险

第一，数据安全性面临诸多挑战。一方面，数据匿名化仍可能被识别，随着技术的发展以及恶意攻击者手段的不断升级，原本旨在保护用户隐私的数据处理方式可能失效，个人或相关主体的信息有暴露的风险(Ji et al., 2016)。另一方面，数据还可能遭受攻击注入错误的威胁，外部攻击者可能利用网络漏洞，蓄意往采集的数据中注入错误信息，破坏数据的完整性和准确性(Wang et al., 2024)。同时，存在植入利益集团的需求和价值取向数据的情况，某些商业利益集团可能为了推广自身的教育产品或理念，可能会通过不正当手段，将带有偏向性的数据混入正常采集的数据集中，这会使采集到的数据带有偏向性，从而影响后续模型的客观性与公正性。

第二，数据正确性方面存在隐患。就内容准确性而言，采集的数据本身可能由于多种原因与实际情况不符。例如，在采集教师教学评价数据时，若调查问卷设计不合理，问题表述模糊，就容易导致教师或学生对问题理解出现偏差，进而反馈的评价数据不能真实反映教学实际情况。而且，采集设备的故障、人工录入的失误等也会造成数据内容错误。而且标注错误也较为常见，例如对于学习者情感状态的判定，存在着主观的偏差，错误的标注会干扰模型对数据特征的准确学习，进而影响整个教育模型应用的效果与可靠性。此外，数据的更新如果滞后于实际情况的变化，这也会导致数据的准确性下降。

2.2. 人工智能教育模型训练阶段的安全风险

第一，算法偏见问题凸显，尤其在深度学习自动化决策过程中。由于训练数据可能存在的不平衡、不全面等情况，若训练数据本身存在不平衡的情况，例如在采集用于智能教育资源分配模型的数据时，优质教育资源地区的数据占比过多，而偏远地区的数据过少，就会导致模型在后续决策时更倾向于为资源丰富地区的学生分配资源，忽视了其他地区学生的需求，产生地域上的偏见。或者算法本身设计的不合理，容易导致模型在决策时产生偏向性，不能公平、客观地对待不同的输入情况，影响教育模型应用在不同场景和用户群体中的适用性。第二，模型安全性方面，模型信息泄露与攻击的风险时刻存在(Tramèr et al., 2016)。经过深度学习和训练的数据，实际上会被模型以某种形式“记忆”并存储在其参数和结构中。这意味着，一旦模型的关键信息被泄露，可能会被不法分子利用，进行恶意篡改、破坏等操作，而针对模型的攻击行为更是会直接影响其正常训练以及后续的应用效果，给教育模型应用带来极大的安全威胁。

2.3. 人工智能教育模型结果输出阶段的安全风险

第一，结果的准确性存在问题，深度学习输出的错误与误导情况时有发生。这可能是因为模型在训练过程中没有充分学习到准确的模式，或者受到了异常数据的干扰等原因，导致

输出的结果与实际应有的正确结果存在偏差，误导教育相关人员的判断与决策。随着时间的推移，数据可能变得过时或不再具有代表性。使用过时的数据进行模型训练或决策分析，可能会导致不准确的结果。此外，深度学习输出的偏差现象较为突出。不同的输入条件或者外部环境变化时，模型输出的结果可能出现较大波动，缺乏稳定的输出性能，不利于在教育场景中持续、可靠地发挥作用。

第二，结果的公平性方面也面临挑战，深度学习输出的歧视问题不容忽视。可能由于训练数据中隐含的偏见等因素，使得模型对不同群体输出的结果存在不公平对待的情况，这在教育公平性要求较高的应用场景中是极为不利的，会破坏教育模型应用的公正性和合理性。第三，结果的安全性，模型的输出本质上是对输入样本进行预测或分类后得到的类别标签或概率向量。这些输出结果看似简单直接，然而近年来的研究却揭示了一个令人担忧的现象，模型输出结果隐含一定的数据信息，攻击者可以利用模型输出在一定程度上窃取相关数据(Zhu et al., 2022)。即使不直接访问模型的内部参数或训练数据，攻击者仍然有可能通过分析模型的输出来窃取相关的敏感信息。

3. 应对策略

人工智能教育模型生成的各个阶段紧密相连，其中存在的风险问题相互交织、彼此影响，在模型数据采集阶段，数据的质量、来源以及处理方式等因素，会对后续的算法训练及最终结果产生直接且深远的影响。而算法本身作为连接数据与结果的关键环节，同样起着至关重要的作用，其自身的特性和设计缺陷也会给结果带来诸多风险。隐私泄露这一风险隐患更是如影随形，贯穿始终，需要我们全面、深入地加以重视并积极探寻有效的应对策略，以保障教育模型应用的健康、稳定与安全发展。

3.1. 机器遗忘的应用潜力与实践

3.1.1. 机器遗忘的概念

在计算机领域，“机器遗忘”作为一种新的技术范式，旨在从机器学习模型中删除先前学习过的信息，将其影响从训练好的模型中消除，就像它从未存在过一样，用并保证遗忘后的机器学习模型性能几乎保持不变(Xu et al., 2023)，如图 12 所示。从本质上讲，它与机器学习正好相反：机器学习的目的是训练模型来进行识别、分类和预测，而机器遗忘的目的则是撤销或逆转这一过程，确保在特定情况下，能够有选择性地从模型中移除不再需要或敏感的信息。机器遗忘分为精确遗忘与近似遗忘，精确遗忘包括重训练和分片训练，即从数据集中直接删除数据进行重训练或将数据划分为分别训练子模型，删除数据后对子模型进行重训练，但这种方法计算成本较高，尤其是当数据集庞大时；近似遗忘，通过对模型参数的修改实现相近与重训练性能，也在一定程度上牺牲了准确性。

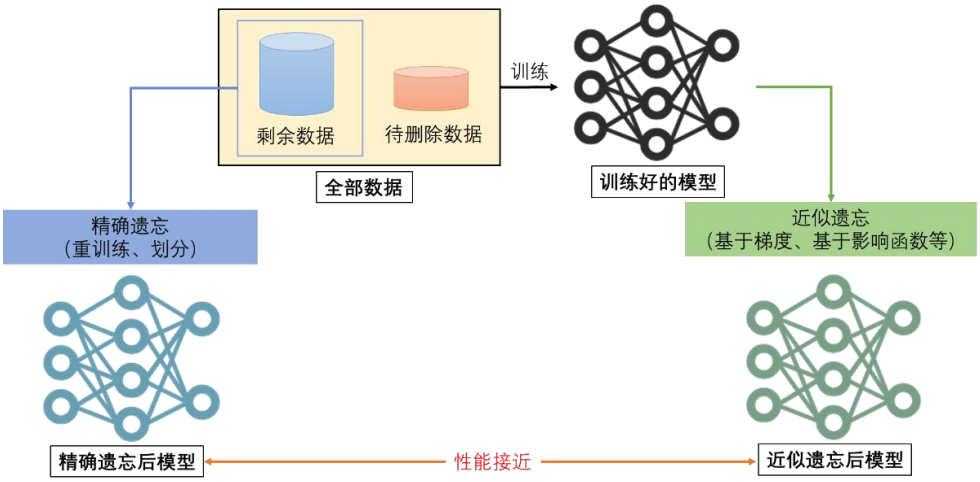

图 1 机器遗忘概念图

3.1.2. 机器遗忘解决人工智能教育模型安全风险的潜力分析

全过程的安全风险问题的根源深植于数据安全之中，若能有效销毁那些被玷污或涉及个人隐私的数据，将能在很大程度上规避相关风险。然而，必须明确的是，仅仅依赖数据库中的简单删除操作是远远不够的。在模型训练的复杂过程中，数据往往已被模型深度“学习”并“铭记”。这意味着，即便数据在数据库中被清除，其影响仍可能潜藏于模型的内部结构中，持续构成威胁。为了彻底清除这些敏感信息，必须诉诸更为先进的技术手段。在此背景下，机器遗忘技术展现出了应对教育模型应用中的安全风险的无限潜力。

对数据，一方面，它能够精准地去除不正确、带有偏见、受到攻击或被植入不合理价值观的数据信息，避免对模型产生不良影响，确保模型所依据的数据源头尽可能准确可靠。另一方面，机器遗忘可有效保障教育参与者的数据隐私和被遗忘权。在智能教育场景中，存在大量涉及教育参与者个人隐私的敏感信息，这些信息需要得到妥善保护。通过机器遗忘，在不影响智能教育模型结果准确性的前提下，能够有选择性地删除敏感的个人数据，使得这些数据不再留存于模型之中，从而充分尊重和保护教育参与者的隐私权益，从根本上杜绝因数据泄露等问题引发的安全风险。

对模型，机器学习模型在训练过程中容易受到历史训练数据的影响，若训练数据存在偏见，例如某些特定群体的数据占比过高或过低，就可能使模型产生算法偏见，进而在后续应用中做出不公平的决策。机器遗忘通过其独特的“撤销”或“逆转”功能，能够从模型中移除这些因历史数据偏差所带来的影响。它可以定期对智能教育模型进行分析，精准定位那些因历史训练数据中的偏见或错误而导致模型出现问题的部分，并将相关信息从模型中消除，使模型摆脱这些不良影响的束缚，恢复到更为客观、公正的状态，从而提升模型自身的准确性和公平性，保障模型在教育实践中的合理应用，降低因模型自身缺陷带来的安全风险。另外，机器遗忘技术通过从模型中移除数据信息的手段，还能有效防范利用模型参数反向还原并窃取数据的风险。这一功能进一步增强了模型的安全性，使得教育参与者的隐私和数据得到了更为周全的保护。

对结果，智能教育模型输出的结果对教育决策等有着重要的指导作用，但如果模型因数据或自身的问题产生不准确的预测或不公平的决策，会直接影响教育实践的公正性和有效性。机器遗忘通过去除数据中的错误信息以及消除模型因历史数据导致的偏差，能够间接提升智能教育模型输出结果的准确性和公平性。同时，机器遗忘使智能教育模型能及时更新、适应

新的教育情境，确保其结果始终保持有效性。在动态变化的教育环境中，过时的信息若持续影响模型，会使模型输出的结果与实际需求脱节，而机器遗忘帮助模型“忘却”这些不适用的信息，促使模型输出更贴合当下教育实际的结果，增强结果的稳定性和可靠性，更好地服务于教育实践，从结果角度有效防范各类安全风险，提升整个智能教育应用的质量与安全性。总而言之，通过机器遗忘技术，针对不同的需求移除特定数据信息，能够有效应对全过程的安全风险。

3.2. 机器遗忘的实施案例

3.2.1. 机器遗忘的实施流程

机器遗忘为遗忘算法技术的统称，在教育领域的延伸与应用，需进一步复杂化，形成一个更为系统化、综合性的概念框架，不仅触及技术层面的革新，更涉及整体设计与实施。

为了适应教育场景，本研究提出了机器遗忘的实施流程，基于法律法规、相关政策以及具体实施细则，包含三大核心阶段：审查阶段、执行阶段和反馈阶段，如图 13 所示。从数据利用与生成的视角来看，教育参与者可被分为数据使用者与数据提供者。数据使用者涵盖学校管理层、教育机构内的科研人员以及直接从事教学活动的教师，他们利用数据进行教育管理、学术研究及个性化教学等目的。而数据提供者则主要为学生和教师群体，他们在教学与学习活动中产生并提供了大量的数据，这些数据成为教育数据挖掘与分析的基础。

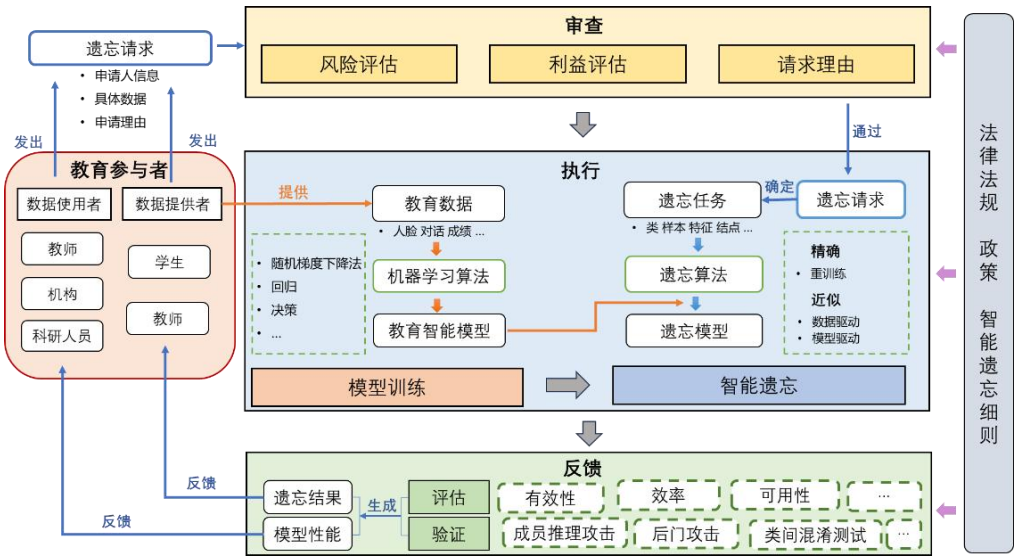

图 2 教育数据智能遗忘流程

审查阶段，考虑到数据销毁的彻底性，需要对请求进行谨慎审查。教育参与主体出于对隐私保护的需求或是为了纠正模型中的问题，依据要求提出明确的遗忘请求，包括基本信息、请求遗忘的数据及申请理由；为确保请求的合规性，需要对请求进行审查，评估请求遗忘的数据是否存在对应的安全风险，以及风险的大小，这包括对数据敏感性、数据泄露后的潜在影响以及数据保护措施的有效性进行综合考量；权衡该数据涉及的各方利益，这包括学生的个人隐私权、教育机构的教育管理权以及数据使用者的合法权益，通过综合考虑各方利益，审查团队将决定是否予以删除。以决定是否予以删除。

执行阶段，根据提出的不同级别的数据遗忘请求，确定对应遗忘任务包括遗忘样本、遗忘类集、遗忘特征。为实现数据遗忘，我们需选择合适的遗忘方法，如模型重训练、数据重组和模型修改等。这些方法的选择应基于数据的特性、模型的类型以及遗忘请求的具体要求。选择性地从原始数据集中以及训练和部署的教育智能模型（如学习分析模型、教育评测模型、

学习资源推荐模型等）中完整且迅速地删除对应的数据样本、类集和特征，实现准确或近似遗忘，并保证智能教育模型性能几乎保持不变。

反馈阶段，对遗忘模型进行评估验证，以确保模型的有效性、效率和可用性。评估模型的有效性、效率和可用性等，同时对遗忘模型进行验证，通过成员推理攻击、后门攻击、类间混淆测试等。根据评估验证的结果，生成遗忘结果报告和模型性能报告。这些报告将分别反馈给数据提供者和数据使用者，以确保信息的透明度和流程的完整性。同时，我们还将提供必要的解释和说明，以消除因信息滞后而产生的误解、担忧及损失。

3.2.2. 机器遗忘的应用实践：对人脸识别模型的保护

人脸识别是当前智能教育实现无感签到(方书雅 & 刘守印, 2020)、学生行为分析(周楠 & 周建设, 2021)、学习情感计算(徐振国 et al., 2019)等多个人工智能教育模型的核心模块。而人脸数据是一种重要的生物特征标识，与个人身份直接关联，是通用身份标识符(胡凌, 2021)，具有高度敏感性。本研究以人脸识别模型为例，实现对学生人脸数据的遗忘，以应对各环节的安全风险。

首先本研究构建了一个涵盖 54 名学生学习期间的人脸图像数据集，图 14 中展示了部分学生的人脸图像示例（为了合规，每张图片示例进行了打码处理），每名学生的面部图像根据其身份标识分别归类存储，经过这一系列的训练和优化，人脸识别模型在测试集上的准确度高达 99.94%，充分展示了其卓越的性能。

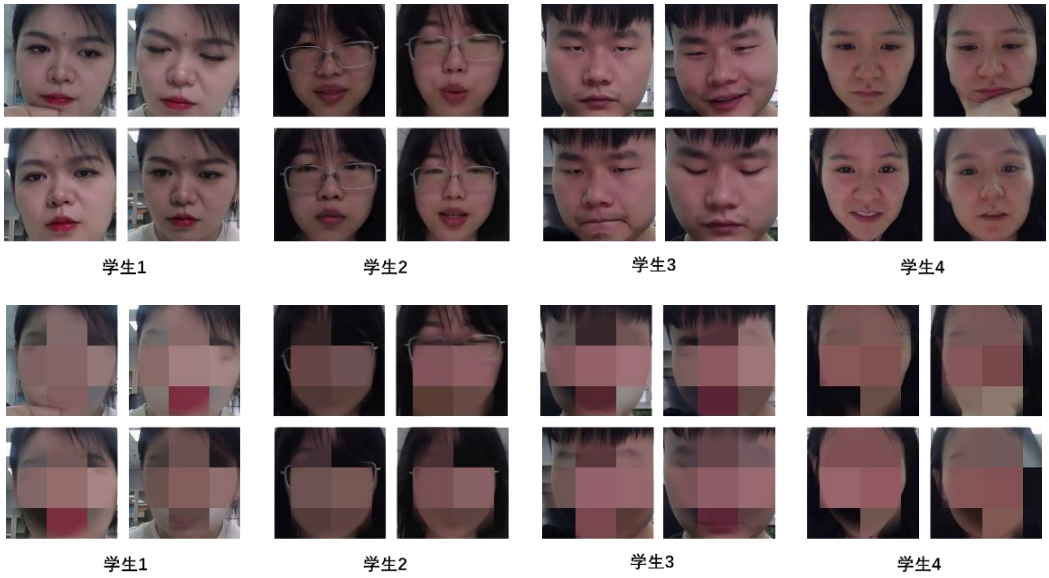

图 3 部分学生人脸图像

近似遗忘算法在遗忘过程中通常能达到约 80-90%的准确率，权衡准确确性和计算成本，近似遗忘更具成本效益(Sai et al., 2024)。基于此本文选择了基于模型修改的梯度投影遗忘方法(Hoang et al., 2024)应用于人脸数据的遗忘。

通过对数据集编号，依据学生提供的信息，定位学生对应人脸数据，以实现单位与多位学生人脸数据的智能遗忘。采用梯度投影遗忘算法进行遗忘后，分别给出了遗忘单位学生人脸数据和遗忘 5 位学生人脸数据的原始人脸识别模型、利用剩余训练集重训练的人脸识别模型、以及梯度投影遗忘后的人脸识别模型分别在原始测试集 D_{test}、剩余测试集 D_{r_test}、遗忘训练集 D_{u_train}和遗忘测试集 D_{u_test}的识别准确率以及运行的平均时间。此外，考虑到遗忘后的模型置信度，图为遗忘数据在遗忘模型与重训练模型上熵的分布。

表 1 人脸识别模型遗忘结果

模型 测试集	原始模型	重训练的模型	遗忘后的模型	重训练的模型	遗忘后的模型
		遗忘 1 位学生		遗忘 5 位学生	
D_{test}	99.94%	96.39%	98.03%	90.69%	81.94%
D_{r_test}	\	98.18%	99.88%	99.93%	89.30%
D_{u_train}	\	0.00%	0.00%	0.00%	1.07%
D_{u_test}	\	49.75%	49.75%	50.00%	49.6%
Average Time（s）	\	147.50	18.94	143.34	24.48

从遗忘效果来看，遗忘后的模型表现出了较高的遗忘水平，与重训练结果接近；遗忘模型与重训练模型上熵的分布情况，即遗忘训练集预测的混乱程度。从图中可以看出，重训练模型和遗忘后模型在遗忘数据上的混乱程度相当一致，这进一步验证了本研究遗忘结果的精确性。这一结果深刻地说明了机器遗忘技术能够精准地根据实际需求，移除那些有毒、过时、带有偏见或涉及个人隐私的数据，从而在确保数据正确性和安全性的同时，有效地防止了因数据问题而引发的算法偏见和结果异常。

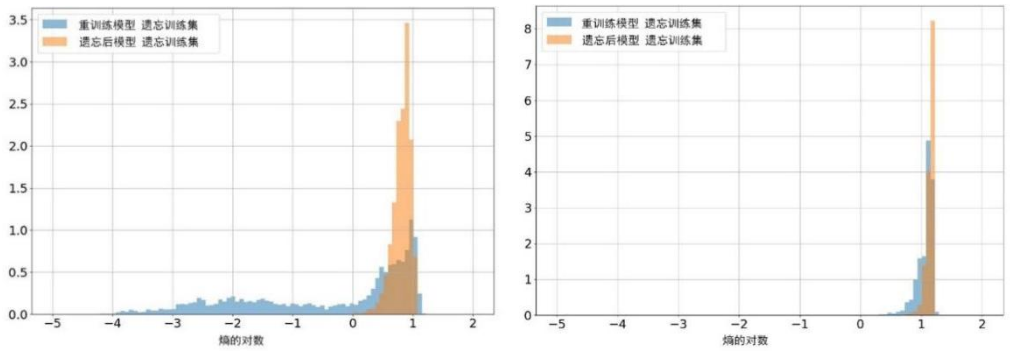

图 15 遗忘单位（左）、多位（右）学生人脸的遗忘数据在遗忘模型与重训练模型上熵的分布

从模型性能来看，遗忘单位学生模型性能没有受到显著影响，遗忘多位学生，虽然存在一定损失，但仍处于可接受范围内。这很可能是因为遗忘多位学生的人脸数据对模型的内部结构产生了更为显著的调整，从而导致了更大的性能波动。

从时间效率来看，采用近似遗忘方法显著提升了遗忘效率，重训练模型的运行时间约为近似遗忘后模型运行时间的 5-7 倍。这一优势使得数据遗忘操作在实际应用中更具实用性。

因此，通过修改模型的特定参数确实能够高效地实现数据的遗忘，在基于人工智能的教育模型的应用中，可以采用机器遗忘，移除特定数据及其在模型中的“影子”，有效应对因数据、算法、结果产生的安全风险问题，真正实现模型全流程保护。

4. 结语

人工智能在教育领域的应用已经展现出显著的潜力，它通过数据驱动的方式，提升了教学效果、实现了学习资源的个性化推荐、智能化了课堂管理，并创新了教学评价与反馈机制。然而，这些基于人工智能教育模型的应用也伴随着安全风险。这些问题不仅影响教育数据的安全性、教育公平性和伦理性，还可能对教育的数字化转型和现代化进程产生负面影响。针对这些安全风险，本研究提出了机器遗忘的概念和实施流程，旨在通过技术手段保护应对产

生的安全风险，保证数据安全，减少算法偏见，提升模型的准确性和公平性。机器遗忘技术在数据、模型和结果三个层面上都显示出了其应对安全风险的潜力。

当前，机器遗忘作为应对人工智能安全风险的技术手段，仍处于发展阶段，面临多重挑战。由于机器学习应用场景复杂，缺乏统一的遗忘算法，需根据具体教育场景和需求选择。同时，遗忘算法评价指标多样，缺乏统一标准，且可能引发数据隐私泄漏风险，数据安全性和隐私性保护成为技术难题。此外，我国法律体系内尚未明确被遗忘权，尽管《个人信息保护法》规定了个人信息删除权，但存在广泛争议，导致机器遗忘的应用缺乏法律指引，数据控制者面临法律不确定性和潜在风险。机器遗忘的实施还需权衡多方利益，特别是数据共享与个人隐私保护之间的冲突，以及不同主体间的利益平衡，以确保个人隐私得到保护的同时，合理利用数据，促进人工智能技术健康发展。

未来，将不断扩展研究的深度与广度，一方面，积极探索应对基于人工智能教育模型安全风险更有效的策略，深入研究机器遗忘以实现更高性能的安全保障；另一方面扩展机器遗忘在其他教育应用方面的作用，例如数据治理，不断促进智能教育的健康、稳定与安全发展。

参考文献

褚娟, 杨. A. (2024). 人工智能教育应用的伦理风险和规范原则. 中国教育学刊(11).

方书雅, & 刘守印. (2020). 基于学生人体检测的无感知课堂考勤方法. 计算机应用, 40(9), 6.

国务院. (2022). 国务院关于印发“十四五”数字经济发展规划的通知. 中华人民共和国国务院公报, 3, 5-18.

胡凌. (2021). 刷脸:身份制度、个人信息与法律规制. 法学家(2), 41-55.

教育部. (2024). 年全国教育工作会议召开. In.

王一岩, & 郑永和. (2022). 多模态数据融合: 破解智能教育关键问题的核心驱动力. 现代远程教育研究, 34(2), 93-102. https://doi.org/10.3969/j.issn.1009-5195.2022.02.011

王佑镁, 利朵, 王旦, & 柳晨晨. (2024). 基于 4R 危机管理模型的教育人工智能伦理风险防范体系构建. 中国电化教育(9), 32-40,85. https://doi.org/10.3969/j.issn.1006-9860.2024.09.006

吴砥, & 吴河江. (2024). 通用大模型教育应用的潜在风险及其规避——基于技术伦理的视角. 华东师范大学学报 (教育科学版), 42(8), 64.

吴河, & 江吴砥. (2024). 教育领域通用大模型应用伦理风险的表征, 成因与治理. Tsinghua Journal of Education, 45(2).

徐振国, 张冠文, 孟祥增, 党同桐, & 孔玺. (2019). 基于深度学习的学习者情感识别与应用. 电化教育研究, 40(2), 87-94.

杨宗凯, & 李卿. (2020). 计算教育学: 内涵与进路. 教育研究, 41(3), 152-159.

褚乐阳, 潘香霖, & 陈向东. (2024). AI 大模型在教育应用中的伦理风险与应对. Journal of Soochow University Educational Science Edition, 12(1).

郑永和, 王杨春晓, & 王一岩. (2021). 智能时代的教育科学研究:内涵、逻辑框架与实践进路. 中国远程教育（综合版）(6), 1-10,17.

周楠, & 周建设. (2021). 基于深度学习的学生行为分析与教学效果评价.

Hoang, T., Rana, S., Gupta, S., & Venkatesh, S. (2024). Learn to unlearn for deep neural networks: Minimizing unlearning interference with gradient projection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,

Ji, S., Mittal, P., & Beyah, R. (2016). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, *19*(2), 1305-1326.

Sai, S., Mittal, U., Chamola, V., Huang, K., Spinelli, I., Scardapane, S.,…Hussain, A. (2024). Machine un-learning: An overview of techniques, applications, and future directions. *Cognitive Computation*, *16*(2), 482-506.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. 25th USENIX security symposium (USENIX Security 16),

Wang, F., Wang, X., & Ban, X. J. (2024). Data poisoning attacks in intelligent transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, *165*, 104750.

Xu, H., Zhu, T., Zhang, L., Zhou, W., & Yu, P. S. (2023). Machine Unlearning: A Survey. *ACM Comput. Surv.*, *56*(1), Article 9. https://doi.org/10.1145/3603620

Zhu, T., Ye, D., Zhou, S., Liu, B., & Zhou, W. (2022). Label-only model inversion attacks: Attack with the least information. *IEEE Transactions on Information Forensics and Security*, *18*, 991-1005.