REVIEW ARTICLE

Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review

Lixiang Yan[1] | Lele Sha[1] | Linxuan Zhao[1] | Yuheng Li[1] | Roberto Martinez-Maldonado[1] | Guanliang Chen[1] | Xinyu Li[1] | Yueqiao Jin[1] | Dragan Gašević[1]

[1]Centre for Learning Analytics at Monash, Faculty of Information Technology, Monash University, Clayton, Victoria, Australia

Correspondence
Lixiang Yan, Centre for Learning Analytics at Monash, Faculty of Information Technology, Monash University, 20 Exhibition Walk, Clayton, VIC 3800, Australia
Email: jimmie.yan@monash.edu

Funding information
This research was at least in part funded by the Australian Research Council (DP210100060) and Jacobs Foundation (Research Fellowship).

Educational technology innovations leveraging large language models (LLMs) have shown the potential to automate the laborious process of generating and analysing textual content. While various innovations have been developed to automate a range of educational tasks (e.g., question generation, feedback provision, and essay grading), there are concerns regarding the practicality and ethicality of these innovations. Such concerns may hinder future research and the adoption of LLMs-based innovations in authentic educational contexts. To address this, we conducted a systematic scoping review of 118 peer-reviewed papers published since 2017 to pinpoint the current state of research on using LLMs to automate and support educational tasks. The findings revealed 53 use cases for LLMs in automating education tasks, categorised into nine main categories: profiling/labelling, detection, grading, teaching support, prediction, knowledge representation, feedback, content generation, and recommendation. Additionally, we also identified several practical and ethical challenges, including low technological readiness, lack of replicability and transparency, and insufficient privacy and beneficence considerations. The findings were summarised into three recommendations for

future studies, including updating existing innovations with state-of-the-art models (e.g., GPT-3/4), embracing the initiative of open-sourcing models/systems, and adopting a human-centred approach throughout the developmental process. As the intersection of AI and education is continuously evolving, the findings of this study can serve as an essential reference point for researchers, allowing them to leverage the strengths, learn from the limitations, and uncover potential research opportunities enabled by ChatGPT and other generative AI models.

KEYWORDS

large language models, pre-trained language models, artificial intelligence, education, systematic scoping review, GPT-3, BERT, ChatGPT

Practitioner notes

What is currently known about this topic

- Generating and analysing text-based content are time-consuming and laborious tasks.
- Large language models are capable of efficiently analysing an unprecedented amount of textual content and completing complex natural language processing and generation tasks.
- Large language models have been increasingly used to develop educational technologies that aim to automate the generation and analysis of textual content, such as automated question generation and essay scoring.

What this paper adds

- A comprehensive list of 53 different educational tasks that could potentially benefit from LLMs-based innovations through automation.
- A structured assessment of the practicality and ethicality of existing LLMs-based innovations from seven important aspects using established frameworks.
- Three recommendations that could potentially support future studies to develop LLMs-based innovations that are practical and ethical to implement in authentic educational contexts.

Implications for practitioners

- Updating existing innovations with state-of-the-art models may further reduce the amount of manual effort required for adapting existing models to different educational tasks.
- The reporting standards of empirical research that aims to develop educational technologies using large language models need to be improved.

- Adopting a human-centred approach throughout the developmental process could contribute to resolving the practical and ethical challenges of large language models in education. 1

1 | INTRODUCTION 2

Advancements in generative artificial intelligence (AI) and large language models (LLMs) have fueled the development of many educational technology innovations that aim to automate the often time-consuming and laborious tasks of generating and analysing textual content (e.g., generating open-ended questions and analysing student feedback survey) (Kasneci et al., 2023; Wollny et al., 2021; Leiker et al., 2023). LLMs are generative artificial intelligence models that have been trained on an extensive amount of text data, capable of generating human-like text content based on natural language inputs. Specifically, these LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Generative Pre-trained Transformer (GPT) (Brown et al., 2020), utilise deep learning and self-attention mechanisms (Vaswani et al., 2017) to selectively attend to the different parts of input texts, depending on the focus of the current tasks, allowing the model to learn complex patterns and relationships among textual contents, such as their semantic, contextual, and syntactic relationships (Min et al., 2021; Liu et al., 2023). As several LLMs (e.g., GPT-3 and Codex) have been pre-trained on massive amounts of data across multiple disciplines, they are capable of completing natural language processing tasks with little (few-shot learning) or no additional training (zero-shot learning) (Brown et al., 2020; Wu et al., 2023). This could lower the technological barriers to LLMs-based innovations as researchers and practitioners can develop new educational technologies by fine-tuning LLMs on specific educational tasks without starting from scratch (Caines et al., 2023; Sridhar et al., 2023). The recent release of ChatGPT, an LLMs-based generative AI chatbot that requires only natural language prompts without additional model training or fine-tuning (OpenAI, 2023), has further lowered the barrier for individuals without technological background to leverage the generative powers of LLMs. 3

Although educational research that leverages LLMs to develop technological innovations for automating educational tasks is yet to achieve its full potential (i.e., most works have focused on improving model performances (Kurdi et al., 2020; Ramesh and Sanampudi, 2022)), a growing body of literature hints at how different stakeholders could potentially benefit from such innovations. Specifically, these innovations could potentially play a vital role in addressing teachers' high levels of stress and burnout by reducing their heavy workloads by automating punctual, time-consuming tasks (Carroll et al., 2022) such as question generation (Kurdi et al., 2020; Bulut and Yildirim-Erbasli, 2022; Oleny, 2023), feedback provision (Cavalcanti et al., 2021; Nye et al., 2023), scoring essays (Ramesh and Sanampudi, 2022) and short answers (Zeng et al., 2023). These innovations could also potentially benefit both students and institutions by improving the efficiency of often tedious administrative processes such as learning resource recommendation, course recommendation and student feedback evaluation, potentially (Zawacki-Richter et al., 2019; Wollny et al., 2021; Sridhar et al., 2023). 4

Despite the growing empirical evidence of LLMs' potential in automating a wide range of educational tasks, none of the existing work has systematically reviewed the practical and ethical challenges of these LLMs-based innovations. Understanding these challenges is essential for developing responsible technologies as LLMs-based innovations (e.g., ChatGPT) could contain human-like biases based on the existing ethical and moral norms of society, such as inheriting biased and toxic knowledge (e.g., gender and racial biases) when trained on unfiltered internet text data (Schramowski et al., 2022). Prior systematic reviews have focused on investigating these issues related to one specific application scenario of LLMs-based innovations (e.g., question generation, essay scoring, chatbots, or automated feedback) (Kurdi et al., 2020; Cavalcanti et al., 2021; Wollny et al., 2021; Ramesh and Sanampudi, 2022). The practical and ethical 5

challenges of LLMs in automating different types of educational tasks remain unclear. Understanding these challenges is essential for translating research findings into educational technologies that stakeholders (e.g., students, teachers, and institutions) can use in authentic teaching and learning practices (Adams et al., 2021). 1

The current study is the first systematic scoping review that aimed to address this gap by reviewing the *current state of research* on using LLMs to automate educational tasks and identify the *practical* and *ethical* challenges of adopting these LLMs-based innovations in authentic educational contexts. A total of 118 peer-reviewed publications from four prominent databases were included in this review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021) protocol. An inductive thematic analysis was conducted to extract details regarding the different types of educational tasks, stakeholders, LLMs, and machine learning tasks investigated in prior literature. The practicality of LLMs-based innovations was assessed through the lens of technological readiness, model performance, and model replicability. Lastly, the ethicality of these innovations was assessed by investigating system transparency, privacy, equality, and beneficence. 2

The contribution of this paper to the educational technology community is threefold: 1) we systematically summarise a comprehensive list of 53 different educational tasks that could potentially benefit from LLMs-based innovations through automation, 2) we present a structured assessment of the practicality and ethicality of existing LLMs-based innovations based on seven important aspects using established frameworks (e.g., the transparency index (Chaudhry et al., 2022)), and 3) we propose three recommendations that could potentially support future studies to develop LLMs-based innovations to be practically and ethically implement in authentic educational contexts. As the intersection of LLMs and education is continuously evolving, the findings of this systematic scoping review can serve as an essential reference point for researchers, allowing them to leverage the strengths, learn from the limitations, and uncover potential opportunities for novel LLMs in supporting educational research and practice. Specifically, emerging works should carefully consider the practical and ethical challenges identified in this study while exploring the research opportunities enabled by ChatGPT and other generative AI models. 3

2 | BACKGROUND 4

In this section, we first establish the definitions for the key terminologies, specifically the definitions of practicality and ethicality in the context of educational technology. We then provided an overview of prior systematic reviews on LLMs in education. Then, we present the research questions based on the gaps identified in the existing literature. 5

2.1 | Practicality 6

Several theoretical frameworks have been proposed regarding the practicality of integrating technological innovations in educational settings. For example, Ertmer's (1999) first- and second-order barriers to change focused on the external conditions of the educational system (e.g., infrastructure readiness) and teachers' internal states (e.g., personal beliefs). Becker (2000) further suggested that for technological innovations to have actual benefits in supporting pedagogical practices, these innovations should be convenient to access, support constructivist pedagogical beliefs, be adaptable to changes in the curriculum, and be compatible to teachers' level of knowledge and skills. These factors were also presented in an earlier framework of the practicality index (Doyle and Ponder, 1977), which summarised three critical components for integrating educational technologies, including the degree of adoption feasibility, the cost and benefit ratio, and the alignment with existing practices and beliefs. Based on these prior theoretical frameworks and considering the recentness of LLMs-based innovations (which only emerged in the past five years), the 7

practical challenges of LLMs-based innovations in automating educational tasks can be assessed from three primary perspectives. First, evaluating the technological readiness of these innovations is essential for determining whether there is empirical evidence to support successful integration and operation in authentic educational contexts. Second, assessing the model performance could contribute valuable insights into the cost and benefits of adopting these innovations, such as comparing the benefits of automation with the costs of inaccurate predictions. Finally, understanding whether these innovations are methodologically replicable could be important for future studies to investigate their alignment with different educational contexts and stakeholders. We elaborated on the evaluation items for each challenge in Section 3.2.

2.2 | Ethicality

Ethical AI is a prevalent topic of discussion in multiple communities, such as learning analytics, AI in education, educational data mining, and educational technology communities (Adams et al., 2021; Pardo and Siemens, 2014). There are ongoing debates regarding AI ethics in education with a mixture of focuses on algorithmic and human ethics among educational data mining and AI in education communities (Holmes and Porayska-Pomsta, 2022). As such debates continue, it is difficult to identify an established definition of ethical AI from these fields. Whereas, ethicality has already been thoroughly investigated and addressed in a closed field to AI in education, namely, the field of learning analytics (Pardo and Siemens, 2014; Selwyn, 2019). Drawing on the established definition of ethicality from the field of learning analytics (Pardo and Siemens, 2014), the ethicality of LLMs-based innovations can thus be defined as the systematisation of appropriate and inappropriate functionalities and outcomes of these innovations, as determined by all stakeholders (e.g., students, teachers, parents, and institutions). For example, Khosravi et al. (2022) explained that the ethicality of AI-powered educational technology systems needs to involve the consideration of accountability, explainability, fairness, interpretability, and safety of these systems. These different domains of ethical AI are all closely related and can be addressed by considering system transparency. Transparency is a subset of ethical AI that involves making all information, decisions, decision-making processes, and assumptions available to stakeholders, which in turn enhances their comprehension of the AI systems and related outputs (Chaudhry et al., 2022). Additionally, for LLMs-based innovations, Weidinger et al. (2021) suggested six types of ethical risks, including 1) discrimination, exclusion, and toxicity, 2) information hazards, 3) misinformation harms, 4) malicious uses, 5) human-computer interaction harms, and 6) automation, access, and environmental harms. These risks can be further aggregated into three fundamental ethical issues, such as privacy concerns regarding educational stakeholders' personal data, equality concerns regarding the accessibility of stakeholders with different backgrounds, and beneficence concerns about the potential harms and negative impacts that LLMs-based innovations may have on stakeholders (Ferguson et al., 2016). These three fundamental ethical issues were considered in the analysis of the reviewed literature. Further details were available in Section 3.2.

2.3 | Related Work

Prior systematic reviews have focused primarily on reviewing a specific application scenario (e.g., question generation, automated feedback, chatbots and essay scoring) of natural language processing and LLMs. For example, Kurdi et al. (2020) have systematically reviewed empirical studies that aimed to tackle the problem of automatic question generation in educational domains. They comprehensively summarised the different generation methods, generation tasks, and evaluation methods presented in prior literature. In particular, LLMs could potentially benefit the semantic-based approaches for generating meaningful questions that are closely related to the source contents. Likewise, Cavalcanti

et al. (2021) have systematically reviewed different automated feedback systems regarding their impacts on improving students' learning performances and reducing teachers' workloads. Despite half of their reviewed studies showing no evidence of reducing teachers' workloads, as these automated feedback systems were mostly rule-based and required extensive manual efforts, they identified that using natural language generation techniques could further enhance such systems' generalisability and potentially reduce manual workloads. On the other hand, Wollny et al. (2021) have systematically reviewed areas of education where chatbots have already been applied. They concluded that there is still much to be done for chatbots to achieve their full potential, such as making them more adaptable to different educational contexts. A systematic review has also investigated the various automated essay scoring systems (Ramesh and Sanampudi, 2022). The findings have revealed multiple limitations of the existing systems based on traditional machine learning (e.g., regression and random forest) and deep learning algorithms (e.g., LSTM and BERT). In sum, these previous systematic reviews have identified room for improvement that can be potentially addressed using state-of-the-art LLMs (e.g., GPT-3 or Codex). However, none of the prior systematic reviews has investigated the practical and ethical issues related to LLMs-based innovations in education generally rather than particularly (e.g., limited to a specific task). 1

The recent hype around one of the latest LLMs-based innovations, ChatGPT, has intensified the discussion about the practical and ethical challenges related to using LLMs in education. For example, in a position paper, Kasneci et al. (2023) provided an overview of some existing LLMs research and proposed several practical opportunities and challenges of LLMs from students' and teachers' perspectives. Likewise, Rudolph et al. (2023) also provided an overview of the potential impacts, challenges, and opportunities that ChatGPT might have on future educational practices. Although these studies have not systematically reviewed the existing educational literature on LLMs, their arguments resonated with some of the pressing issues around LLMs and ethical AI, such as data privacy, bias, and risks. On the other hand, Sallam (2023) systematically reviewed the implications and limitations of ChatGPT in healthcare education and identified potential utility around personalisation and automation. However, it is worth noting that most papers reviewed in Sallam's study were either editorials, commentaries, or preprints. This lack of peer-reviewed empirical studies on ChatGPT is understandable as it has only been released since late 2022 (OpenAI, 2023). None of the existing work has systematically reviewed the peer-reviewed literature on prior LLMs-based innovations. Such investigations could provide more reliable and empirically-based evidence regarding the potential opportunities and challenges of LLMs in educational practices. Thus, the current study aimed to address this gap in the literature by conducting a systematic scoping review of prior educational research on LLMs. Specifically, the following research questions were investigated to guide this review: 2

- **RQ1**: What is *the current state of research* on using LLMs to automate educational tasks, specifically through the lens of educational tasks, stakeholders, LLMs, and machine-learning tasks[1]? 3
- **RQ2**: What are the *practical* challenges of LLMs in automating educational tasks, specifically through the lens of technological readiness, model performance, and model replicability?
- **RQ3**: What are the *ethical* challenges of LLMs in automating educational tasks, specifically through the lens of system transparency, privacy, equality, and beneficence?

[1] Such as classification, prediction, clustering, etc.

3 | METHODS 1

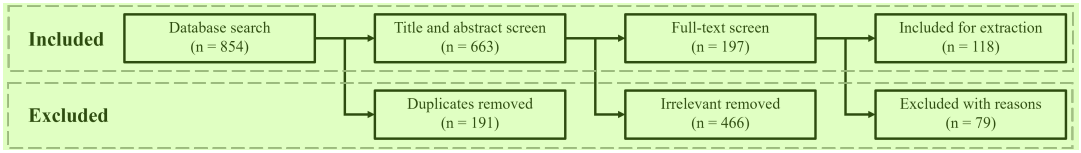
A systematic scoping review was conducted in this study as this method has been frequently used in emerging and rapidly evolving research areas to scope a body of literature and identify the key concepts, methods, evidence, and challenges (Munn et al., 2018). Consequently, the quality of the included studies was often not assessed as the aim is to provide a boarder picture of an emerging field. 2

3.1 | Review Procedures 3

We followed the PRISMA (Page et al., 2021) protocol to conduct the current systematic scoping review of LLMs. We searched four reputable bibliographic databases, including Scopus, ACM Digital Library, IEEE Xplore, and Web of Science, to find high-quality peer-reviewed publications. Additional searches were conducted through Google Scholar and Education Resources Information Center (ERIC) to identify peer-reviewed publications that have yet to be indexed by these databases, either recently published or not indexed (e.g., Journal of Educational Data Mining; prior to 2020). Our initial search query for the title, abstract, and keywords included terms such as "large language model", "pre*trained language model", "GPT-*", "BERT", "education", "student*", and "teacher*". A publication year constraint was also applied to restrict the search to studies published since 2017, specifically from 01/01/2017 to 12/31/2022, as the foundational architecture (Transformer) of LLMs was formally released in 2017 (Vaswani et al., 2017). Only peer-reviewed publications were considered to enhance the scientific credibility of this review. The initial database search was conducted by two researchers independently. Any discrepancies between the search results were resolved through further discussion or consulting the librarian for guidance. 4

Two researchers independently reviewed the titles and abstracts of eligible articles based on five predetermined inclusion and exclusion criteria. First, we included studies that used large or pre-trained language models directly or built on top of such models, and excluded studies that used general machine-learning or deep-learning models with unspecified usage of LLMs. Second, we included empirical studies with detailed methodologies, such as a detailed description of the LLMs and research procedures, and excluded review, opinion, and scoping works. Third, we only included full-length peer-reviewed papers, and excluded short, workshop, and poster papers that were less than six and eight pages for double- and single-column layouts, respectively. Additionally, we included studies that used LLMs for the purpose of automating educational tasks (e.g., essay grading and question generation), and excluded studies that merely used LLMs as part of the analysis without educational implications. Finally, we only included studies that were published in English (both the abstract and the main text) and excluded studies that were published in other languages. Any conflicting decisions were resolved through further discussion between the two researchers or consulting with a third researcher to achieve a consensus. 5

The database search initially yielded 854 publications, with 191 duplicates removed, resulting in 663 publications for the title and abstract screening (see Figure 1). After the title and abstract screening, 197 articles were included for the full-text review with an inter-rater reliability (Cohen's kappa) of 0.75, indicating substantial agreement between the reviewers during the title and abstract screening. A total of 118 articles were selected for data extraction after the full-text review with an inter-rater reliability (Cohen's kappa) of 0.73, indicating substantial agreement between the reviewers during the full-text review. Out of the initial 197 articles, 79 were excluded for various reasons, including not full paper (n=41), lack of educational automation (n=17), lack of pre-trained or LLMs (n=12), merely using pre-trained or LLMs as part of the analysis (n=3), non-English paper (n=2), and non-empirical paper (n=2). 6



1

FIGURE 1 Systematic scoping review process following the PRISMA protocol. 2

3.2 | Data Analysis 3

For the first research question (RQ1), we conducted an inductive thematic analysis to extract information regarding the current state of research on using LLMs to automate educational tasks. Specifically, we extracted four primary types of contextual information from each included paper: educational tasks, stakeholders, LLMs, and machine-learning tasks. This contextual information would provide a holistic view of the existing research and inform researchers and practitioners regarding the viable directions to explore with the state-of-the-art LLMs (e.g., GPT-3.5 and Codex). A total of seven data extraction items were developed to address the second and third research questions. These items were developed as they are directly related to the definition of practicality (RQ2: Item 1-3) and ethicality (RQ3: Item 4-7), as defined in the Background section (Section 2). The following list elaborates on the final set of items along with the corresponding guiding questions. For the thematic analysis and Items, two researchers independently coded 20 random samples of the included studies. Any conflicts were resolved through further discussion or consulting a third researcher. After reaching a Cohen's kappa of more than 0.80 (indicating almost perfect agreement), each researcher coded half of the remaining 98 studies (49 studies each) and cross-checked each other's work. The database of the studies included in this review and the extracted data for each item are available in the supplementary document. 4

1. **Technology readiness** What levels of technology readiness are the LLMs-based innovations at? We adopted the assessment tool from the Australian government, namely the Australian Department of Defence's Technology Readiness Levels (TRL) (Science and Group), which has been used to assess the maturity of educational technologies in prior SLR (Yan et al., 2022). There are nine different technological readiness levels: Basic Research (TRL-1), Applied Research (TRL-2), Critical Function or Proof of Concept Established (TRL-3), Lab Testing/Validation of Alpha Prototype Component/Process (TRL-4), Laboratory Testing of Integrated/Semi-Integrated System (TRL-5), Prototype System Verified (TRL-6), Integrated Pilot System Demonstrated (TRL-7), System Incorporated in Commercial Design (TRL-8), and System Proven and Ready for Full Commercial Deployment (TRL-9), further explained in the Result section. 5
2. **Performance**: How accurate and reliable can the LLMs-based innovations complete the designated educational tasks? For example, what are the model performance scores for classification (e.g., AUC and F1 scores), generation (e.g., BLEU score), and prediction tasks (e.g., RMSE and Pearson's correlation)? 6
3. **Replicability**: Can other researchers or practitioners replicate the LLMs-based innovations without additional support from the original authors? This item evaluates whether the paper provided sufficient details about the LLMs (e.g., open-sourced algorithms) and the dataset (e.g., open-source data). 7
4. **Transparency**: What tiers of transparency index (Chaudhry et al., 2022) are the LLMs-based innovations at? The transparency index proposed three tiers of transparency, including transparent to AI researchers and practitioners (Tier 1), transparent to educational technology experts and enthusiasts (Tier 2), and transparent to educators and parents (Tier 3). The tier of transparency increases as educational stakeholders become fully involved in developing and evaluating the AI system. These tiers were further elaborated on in the Results section. 8

5. **Privacy**: Has the paper mentioned or considered privacy issues of their innovations? This item explores potential issues related to informed consent, transparent data collection, individuals' control over personal data, and unintended surveillance (Ferguson et al., 2016; Tsai et al., 2020). 1
6. **Equality**: Has the paper mentioned or considered equal access to their innovations? This item explores potential issues related to limited access for students from low-income backgrounds or rural areas and the linguistic limitation of the innovations, such as their capability to analyse different languages (Ferguson et al., 2016).
7. **Beneficence**: Has the paper mentioned or considered potential issues that violate the ethical principle of beneficence? Such violations may include the risks associated with labelling and profiling students, inadequate usage of machine-generated content for assessments, and algorithmic biases (Ferguson et al., 2016; Zawacki-Richter et al., 2019).

4 | RESULTS 2

4.1 | The Current State — RQ1 3

We identified nine different categories of educational tasks that prior studies have attempted to automate using LLMs (as shown in Table 1). Prior studies have used LLMs to automate the profiling and labelling of 17 types of education-related contents and concepts (e.g., forum posts, student sentiment, and discipline similarity), the detection of six latent constructs (e.g., confusion and urgency), the grading of five types of assessments (e.g., short answer questions and essays), the development of five types of teaching support (e.g., conversation agent and intelligent question-answering), the prediction of five types of student-orientated metrics (e.g., dropout and engagement), the construction of four types of knowledge representations (e.g., knowledge graph and entity recognition), the provision of four different forms of feedback (e.g., real-time and post-hoc feedback), the generation of four types of content (e.g., MCQs and open-ended questions), and the delivery of three types of recommendations (e.g., resource and course). Of the 118 reviewed studies, 85 studies aimed to automate educational tasks related to teachers (e.g., question grading and generation), 54 studies targeted student-related activities (e.g., feedback and resource recommendation), 20 studies focused on supporting institutional practices (e.g., course recommendations and discipline planning), and 14 studies empowered researchers with automated methods to investigate latent constructs (e.g., student confusion) and capture verbal data (e.g., speech recognition). 4

We identified five categories of LLMs used in prior studies to automate educational tasks. BERT and its variations (e.g., RoBERTa, DistilBERT, multilingual BERT, LaBSE, EstBERT, and Sentence-BERT) were the most predominant model used in 109 reviewed studies. However, they often required manual effort for fine-tuning (n=90). GPT-2 and GPT-3 have been used in five and three studies, respectively. Specifically, GPT-2 and GPT-3 have performed better than BERT-based models in content generation and evaluation tasks, such as generating university math problems (Drori et al., 2022) and evaluating the quality of student-generated short answer questions (Moore et al., 2022). OpenAI's Codex has been used in two prior studies, specifically for code generation tasks. T5 has also been used in two prior studies for classification and generation purposes. In terms of machine-learning tasks, 74 studies used LLMs to perform classification tasks. Generation and prediction tasks were investigated in 24 and 23 prior studies, respectively. In sum, LLMs-based innovations have already been used to automate a range of educational tasks, but most of these innovations were developed on older models, such as BERT and GPT-2. Although state-of-the-art models, such as GPT-3, have been introduced for over two years (Brown et al., 2020), they have yet to be widely applied to automate educational tasks. A potential reason for this lack of adoption could be these models' commercial and close-sourced nature, increasing the financial burdens of developing and operating educational technology innovations on top of 5

such models.

TABLE 1 Educational Tasks in LLMs Research

Categories	Educational Tasks
Profiling and Labelling	Forum post classification, dialogue act classification, classification of learning designs, review sentiment analysis, topic modelling, pedagogical classification of MOOCs, collaborative problem-solving modelling, paraphrase quality, speech tagging, labelling educational content with knowledge components, key sentence and keyword extraction, reflective writing analysis, multimodal representational thinking, discipline similarity, concept classification, cognitive level classification, essay arguments segmentation
Detection	Semantic analyses, detecting off-task messages, confusion detection, urgency detection, conversational intent detection, teachers' behaviour detection
Assessment and Grading	Formative and summative assessment grading, short answer grading, essay grading, subjective question grading, student self-explanation
Teaching Support	Classroom teaching, learning community support, online learning conversation agent, intelligent question-answering, teacher activity recognition
Prediction	Student performance prediction, student dropout prediction, emotional and cognitive engagement detection, growth and development indicators for college students, at-risk student identification
Knowledge Representation	Knowledge graph construction, knowledge entity recognition, knowledge tracing, cause-effect relation extraction
Feedback	Real-time feedback, post-hoc feedback, aggregated feedback, feedback on feedback (peer-review comments)
Content Generation	MCQs generation, open-ended question generation, code generation, reply (natural language) generation
Recommendation	English reference selection and recommendation, resource recommendation, course recommendation

4.2 | Practical Challenges — RQ2

4.2.1 | Technology readiness

According to the Technology Readiness Level scale (Science and Group), the LLMs-based innovations are still in the early development and testing stage. Over three-quarters of the LLMs studies (n=89) are in the applied research stage (TRL-2), which aims to experiment with the capability of LLMs in automating different educational tasks by developing different models and combining LLMs with other machine-learning and deep-learning techniques (e.g., RCNN (Shang et al., 2022)). Thirteen studies have established a proof of concept and demonstrated the feasibility of using LLMs-based innovations to automate certain processes of educational tasks (TRL-3). Nine studies have developed functional prototypes and conducted preliminary validation under controlled laboratory settings (TRL-4), often involv-

ing stakeholders (e.g., students and teachers) to test and evaluate the output of their innovations. Only seven studies have taken a further step and conducted validation studies in authentic learning environments, with most functional components integrated into the educational tasks (TRL-5), such as an intelligent virtual standard patient for medical students training (Song et al., 2022) and an intelligent chatbot for university admission (Nguyen et al., 2021). Yet, none of the existing LLMs-based innovations has been verified through successful operations (TRL-6). Together, these findings suggest although existing LLMs-based innovations can be used to automate certain educational tasks, they have yet to show evidence regarding improvements to teaching, learning, and administrative processes in authentic educational practices. 1

4.2.2 | Performance 2

The performance of LLMs-based innovations varies across different machine-learning and educational tasks. For classification tasks, LLMs-based innovations have shown high performance for simple educational tasks, such as modelling the topics from a list of programming assignments (best F1 = 0.95) (Fonseca et al., 2020), analysing the sentiment of student feedback (best F1 = 0.94) (Truong et al., 2020), constructing subject knowledge graph from teaching materials (best F1 = 0.94) (Su and Zhang, 2020), and classifying educational forum posts (Sha et al., 2022c) (best F1 = 0.92). However, the classification performance of LLMs-based innovations decreases for other educational tasks. For example, the F1 scores for detecting student confusion in the course forum (Geller et al., 2021) and students' off-task messages in game-based collaborative learning (Carpenter et al., 2020) are around 0.77 and 0.67, respectively. Likewise, the F1 score for classifying short-answer responses varies between 0.61 to 0.82, with the lower performance on out-of-sample questions (best F1 = 0.61) (Condor et al., 2021). Similar performances were also observed in classifying students' argumentative essays (best F1 = 0.66) (Ghosh et al., 2020). 3

For prediction tasks, LLMs-based innovations have demonstrated reliable performance compared to ground truth or human raters. For example, LLMs-based innovations have achieved high scores of quadratic weighted kappa (QWK) in essay scoring, specifically for off-topic (QWK = 0.80), gibberish (QWK = 0.80), and paraphrased answers (QWK = 0.94), indicating substantial to almost perfect agreements with human raters (Doewes and Pechenizkiy, 2021). Similar performances on essay scoring have been observed in several other studies (e.g., 0.80 QWK in (Beseiso et al., 2021) and 0.81 QWK in (Sharma et al., 2021)). Likewise, LLMs-based innovations' performances on automatic short-answer grading were also highly correlated with human ratings (Pearson's correlation between 0.75 to 0.82) (Ahmed et al., 2022; Sawatzki et al., 2022). 4

Regarding generation tasks, LLMs-based innovations demonstrated high performance across different educational tasks. For example, LLMs-based innovations have achieved an F1 score of 0.92 for generating MCQs with single-word answers (Kumar et al., 2022). Educational technologies developed by fine-tuning Codex also demonstrated the capability of resolving 81% of the advanced mathematics problems (Drori et al., 2022). Text summaries generated using BERT had no significant differences compared with student-generated summaries and can not be differentiated by graduate students (Merine and Purkayastha, 2022). Similarly, BERT-generated doctor-patient dialogues were also found to be indistinguishable from actual doctor-patient dialogues, which can be used to create virtual standard patients for medical students' diagnosis practice training (Song et al., 2022). Additionally, for introductory programming courses, the state-of-the-art LLMs, Codex, could generate sensible and novel exercises for students along with an appropriate sample solution (around three out of four times) and accurate code explanation (67% accuracy) (Sarsa et al., 2022). 5

In sum, although the classification performance of LLMs-based innovations on complex educational tasks is far from suitable for practical adoption, LLMs-based innovations have already shown high performance on several rel- 6

atively simple classification tasks that could potentially be deployed to automatically generate meaningful insights that could be useful to teachers and institutions, such as navigating through numerous student feedback and course review. Likewise, LLMs-based innovations' prediction and generation performance reveals a promising future of potentially automating the generation of educational content and the initial grading of student assessments. However, ethical issues must be considered for such implementations, which we covered in the findings for RQ3.

4.2.3 | Replicability

Most reviewed studies (n=107) have not disclosed sufficient details about their methodologies for other researchers and practitioners to replicate their proposed LLMs-based innovations. Among these studies, 12 studies have open-sourced the original code for developing the innovations but failed to open-source the data they used. In contrast, 20 studies have open-sourced the data they used but failed to release the actual code. Around two-thirds of the reviewed studies (n=75) have failed to release both the original code and the data they used, leaving only 11 studies publicly available for other researchers and practitioners to replicate without needing to contact the original authors. This lack of replicability could become a vital barrier to adoption, as 87 out of the 107 non-replicable studies required fine-tuning the LLMs to achieve the reported performance. This replication issue also limits others from further evaluating the generalisability of the proposed LLMs-based innovations in other datasets, constraining potential practical utilities.

4.3 | Ethical Challenges — RQ3

4.3.1 | Transparency

Based on the transparency index and the three tiers of transparency (Chaudhry et al., 2022), most of the reviewed study reached at-most Tier 1 (n=109), which is merely considered transparent to AI researchers and practitioners. Although these studies reported details regarding their machine learning models (e.g., optimisation and hyperparameters), such information is unlikely to be interpretable and considered transparent for individuals without a strong background in machine learning. For the remaining nine studies, they reached at-most Tier 2 as they often involved some form of human-in-the-loop elements. Specifically, making the LLMs innovations available for student evaluation has been found in three studies (Nguyen et al., 2021; Song et al., 2022; Merine and Purkayastha, 2022). Such evaluations often involved students differentiating AI-generated from human-generated content (Song et al., 2022; Merine and Purkayastha, 2022) and assessing student satisfaction with AI-generated responses (Nguyen et al., 2021). Likewise, two studies have involved experts in evaluating specific features of the content generated by the LLMs-based innovations, such as informativeness (Maheen et al., 2022) and cognitive level (Moore et al., 2022). Surveys have been used to evaluate students' experience with LLMs-based innovations from multiple perspectives, such as the quality and difficulty of AI-generated questions (Drori et al., 2022; Li and Xing, 2021) and potential learning benefits of the systems (Jayaraman and Black, 2022). Finally, semi-structured interviews have been conducted to understand students' perception of the LLM system after using the system in authentic computer-supported collaborative learning activities (Zheng et al., 2022). Although these nine studies had some elements of human-in-the-loop, stakeholders were often involved in a post-hoc evaluation manner instead of throughout the development process, and thus, have limited knowledge regarding the operating principle and potential weakness of the systems. Consequently, none of the existing LLMs-based innovations can be considered as being at Tier 3, which describes an AI system that is considered transparent for educational stakeholders (e.g., students, teachers, and parents).

4.3.2 | Privacy 1

The privacy issues related to LLMs-based innovations were rarely attended to or investigated in the reviewed studies. 2
Specifically, for studies that have fine-tuned LLMs with textual data collected from students, none of these studies has explicitly explained their consenting strategies (e.g., whether students acknowledge the collection and intended usage of their data) and data protection measures (e.g., data anonymisation and sanitisation). This lack of attention to privacy issues is particularly concerning as LLMs-based innovations work with stakeholders' natural languages that may contain personal and sensitive information regarding their private lives and identities (Brown et al., 2022). It is possible that stakeholders might not be aware of their textual data (e.g., forum posts or conversations) on digital platforms (e.g., MOOCs and LMS) being used in LLMs-based innovations for different purposes of automation (e.g., automated reply and training chatbots) as the consenting process is often embedded into the enrollment or signing up of these platforms (Tsai and Gasevic, 2017). This process can hardly be considered informed consent. Consequently, if stakeholders shared their personal information on these platforms in natural language (e.g., sharing phone numbers and addresses with group members via digital forums), such information could be used as training data for fine-tuning LLMs. This usage could potentially expose private information as LLMs are incapable of understanding the context and sensitivity of text, and thus, could return stakeholders' personal information based on semantic relationships (Brown et al., 2022).

4.3.3 | Equality 3

Although most of the studies (n=95) used LLMs that only apply to English content, we also identified application 4
scenarios of LLMs in automating educational tasks in 12 other languages. Specifically, 19 studies used LLMs that can be applied to Chinese content. Ten prior studies used LLMs for Vietnamese (n=3), Spanish (n=3), Italian (n=2), and German (n=2) contents. Additionally, seven studies applied LLMs to Croatian, Indonesian, Japanese, Romanian, Russian, Swedish, and Hindi content. While the dominance of English-based innovations remains a concerning equality issue, the availability of innovations that support a variety of other languages, specifically in non-western, educated, industrialized, rich and democratic (WEIRD) societies (e.g., Indonesia and Vietnam), may indicate a promising sign for LLMs-based innovations to have potential global impacts and levels such equality issues in the future. However, the financial burdens from adopting the state-of-the-art models (e.g., OpenAI's GPT-3 and Codex) could potentially exacerbate the equality issues, making the best-performing innovations only accessible and affordable to WEIRD societies.

4.3.4 | Beneficence 5

A total of seven studies have discussed potential issues related to the violation of the ethical principle of beneficence. 6
For example, one study has discussed the potential risk of adopting underperforming models, which could negatively affect students' learning experiences (Li and Xing, 2021). Such issues could be minimised by deferring decisions made by such models (Schneider et al., 2022) and labelling the AI-generated content with a warning message (e.g., teachers' manual revision is mandatory before determining the actual correctness) (Angelone et al., 2022). Apart from issues with adopting inaccurate models, two studies have suggested that potential bias and discrimination issues may occur if adopting a model that is accurate but unfair (Sha et al., 2021; Merine and Purkayastha, 2022). This issue is particularly concerning as most existing studies focused solely on developing an accurate model. Only nine reviewed studies released information regarding the descriptive data of different sample groups, such as gender and ethicality

(e.g., (Pugh et al., 2021)). Two studies have proposed potential approaches that could address such fairness issues. Specifically, using sampling strategies, such as balancing demographic distribution, has been found as an effective approach to improve both model fairness and accuracy (Sha et al., 2022b,a). These approaches are essential for ensuring that LLMs-based innovations will not perpetuate problematic and systematic biases (e.g., gender biases), especially as the best-performing LLMs are often black-boxed with little interpretability, traceability, and justification of the results (Wu, 2022).

5 | DISCUSSION

5.1 | Main Findings

The current study systematically reviewed 118 peer-reviewed empirical studies that used LLMs to automate educational tasks. For the first research question (RQ1), we illustrated the current state of educational research on LLMs. Specifically, we identified 53 types of application scenarios of LLMs in automating educational tasks, summarised into nine general categories, including profiling and labelling, detection, assessment and grading, teaching support, prediction, knowledge representation, feedback, content generation, and recommendation. While some of these categories resonate with the utilities proposed in prior positioning works (e.g., feedback, content generation, and recommendation) (Kasneci et al., 2023; Rudolph et al., 2023), novel directions such as using LLMs to automate the creation of knowledge graph and entity further indicated the potential of LLMs-based innovations in supporting institutional practices (e.g., creating knowledge-based search engines across multiple disciplines). These identified directions could benefit from the state-of-the-art LLMs (e.g., GPT-3 and Codex) as most of the reviewed studies (92%) focused on using BERT-based models, which often required manual effort for fine-tuning. Whereas, the state-of-the-art LLMs could potentially achieve similar performance with a zero-shot approach (Bang et al., 2023). While the majority of the reviewed studies (63%) focused on using LLMs to automate classification tasks, there could be more future studies that aimed to tackle the automation of prediction and generation tasks with the more capable LLMs (Sallam, 2023). Likewise, although supporting teachers are the primary focus (72%) of the existing LLMs-based innovations, students and institutions could also benefit from such innovations as novel utilities could continue to emerge from the educational technology literature. Together, the findings of the first research question could spark educational researchers with ideas of exploring the potential of state-of-the-art LLMs in augmenting educational practices, specifically, the identified 53 types of application scenarios may all worth to re-explore in the light of ChatGPT and other powerful generative AI models (Kasneci et al., 2023).

Regarding the second research question (RQ2), we identified several practical challenges that need to be addressed for LLMs-based innovations to have actual educational benefits. The development and educational research on LLMs-based innovations are still in the early stages. Most of the innovations demonstrated a low level of technology readiness, where the innovations have yet to be fully integrated and validated in authentic educational contexts. This finding resonates with previous systematic reviews on related educational technologies, such as reviews on automated question generation (Kurdi et al., 2020), feedback provision (Cavalcanti et al., 2021), essay scoring (Ramesh and Sanampudi, 2022), and chatbot systems (Wollny et al., 2021). There is a pressing need for in-the-wild studies that provide LLMs-based innovations directly to educational stakeholders for supporting actual educational tasks instead of testing on different datasets or in laboratory settings. Such authentic studies could also validate whether the existing innovations can achieve the reported high model performance in real-life scenarios, specifically in prediction and generation tasks, instead of being limited to prior datasets. This validation process is vital for preventing inadequate usage, such as adopting a subject-specific prediction model for unintended subjects. Researchers need to carefully

examine the extent of generalisability of their innovations and inform the limitations to stakeholders (Gašević et al., 2016). However, addressing such needs could be difficult considering the current literature's poor replicability, which increases the barriers for others to adopt LLMs-based innovations in authentic educational contexts or validate with different samples. Similar replication issues have also been identified in other areas of educational technology research (Yan et al., 2022). 1

For the third research question (RQ3), we identified several ethical challenges regarding LLMs-based innovations. In particular, most of the existing LLMs-based innovations (92%) were only transparent to AI researchers and practitioners (Tier 1), with only nine studies that can be considered transparent to educational technology experts and enthusiasts (Tier 2). The primary reason behind this low transparency can be attributed to the lack of human-in-the-loop components in prior studies. This finding resonates with the call for explainable and human-centred AI, which stresses the vital role of stakeholders in developing meaningful and impactful educational technology (Khosravi et al., 2022; Yang et al., 2021). Involving stakeholders during the development and evaluation of LLMs-based innovations is essential for addressing both practical and ethical issues. For example, as the current findings revealed, LLMs-based innovations are subject to data privacy issues but were rarely mentioned or investigated in the literature (Merine and Purkayastha, 2022), which may be due to the little voice that stakeholders had in prior research. The several concerning issues around beneficence also demand the involvement of stakeholders as their perspectives are vital for shaping the future directions of LLMs-based innovations, such as how responsible decisions can be made with these AI systems (Schneider et al., 2022). Likewise, the equality issue regarding the financial burdens that may occur when adopting innovations that leverage commercial LLMs (e.g., GPT-3 and Codex) can also be further studied with institutional stakeholders. 2

5.2 | Implications 3

The current findings have several implications for education research and practice with LLMs, which we have summarised into three recommendations that aim to support future studies to develop practical and ethical innovations that can have actual benefits to educational stakeholders. First, the wide range of application scenarios of LLMs-based innovations can further benefit from the improvements in the capability of LLMs. Updating existing innovations with state-of-the-art LLMs may further reduce the amount of manual effort required for fine-turning and achieve similar performances (Bang et al., 2023). Considering the 53 identified use cases of LLMs in education, there are multiple research trajectories that could foster the development of practical educational technologies. These avenues have the potential to address some of the pressing challenges that plague the global education system. Particularly, the use cases involving teaching support, assessment and grading, feedback, and content generation categories (Table 1) could act as catalysts for the development of educational technologies that could alleviate teachers' workload and mental stress by automating the laborious tasks associated with creating, evaluating, and providing feedback for student assessments (Carroll et al., 2022). Similarly, further exploration of the use cases in profiling and labelling, detection, prediction, and recommendation could lead to the development of educational technologies that can deliver personalised learning support for each student across various disciplines (Wollny et al., 2021). Such improvements could enhance the overall well-being of teachers and increase students' learning opportunities, thereby contributing to the achievement of SDG 4 by 2030 (Boeren, 2019). Nonetheless, researchers should also be mindful of the potential financial and resource burdens that could be imposed on educational stakeholders when innovating with the commercial LLMs (e.g., GPT-3/4 and ChatGPT). 4

The unrivalled natural language generation capabilities exhibited by ChatGPT and other cutting-edge LLMs (e.g., LLaMA and PaLM 2) might also inspire future studies to delve into a broader spectrum of research directions. These 5

include comparisons between the quality of student-generated and ChatGPT-generated writings (Li et al., 2023) and evaluating these LLMs' capability to tackle educational assessments (Gilson et al., 2023). Such explorations would not only unveil the potential of LLMs and generative AI models in educational content generation and evaluation tasks but also expose the possible threats that these models pose to academic integrity, a pervasive issue across the education sector (Kasneci et al., 2023). Intriguingly, leveraging the use cases of LLMs in tasks such as creating knowledge representation (Zheng et al., 2023) and classifying cognitive levels (Liu et al., 2022) could potentially facilitate the transition from outcome-focused to process-focused assessments. Here, LLMs and generative AI models could be employed for learning assessments in a manner similar to learning analytics (Gašević et al., 2022). Consequently, future studies may begin to explore methods of addressing the potential threats of LLMs with LLMs-based solutions. 1

For LLMs-based innovations to achieve a high level of technology readiness and performance, the current reporting standards must be improved. Future studies should support the initiative of open-sourcing their models/systems when possible and provide sufficient details about the test datasets, which are essential for others to replicate and validate existing innovations across different contexts, preventing the potential pitfall of another replication crisis (Maxwell et al., 2015). This initiative is particularly vital in the era of generative AI models as most of these models, especially the commercial ones (e.g., ChatGPT and the GPT series), are proprietary. Thus, when using these LLMs for augmenting educational practices, such as scoring student essays (Doewes and Pechenizkiy, 2021), providing real-time feedback (Zheng et al., 2022), or generating questions for learning activities (Sarsa et al., 2022), researchers need to be systematic and transparent about the reporting of the model usage and prompts (Wu, 2022). For example, when using the ChatGPT API for question generation at scale, researchers should at least report the exact models, prompts, and model temperature used in the process, as different models may differ in their ability to generate accurate and reliable content and the prompts are essential for others to replicate the same or similar results (Kasneci et al., 2023). 2

Apart from the aforementioned technical and methodological details, researchers and educational policymakers should also consider the potential wider impacts of LLMs-based solutions on different stakeholders. For example, in terms of detection and academic integrity, some institutions have rapidly adopted AI-detection tools that claim to have high accuracy and a low false positive rate. Yet, as disclosed in a recent report by Turnitin, a company whose AI-detection function has been utilised on more than 38.5 million student submissions, the real-world performance of their solution resulted in a significantly higher occurrence of false positives compared to their laboratory findings (Chechitelli, 2023). Such negligence can be devastating for students who have been falsely accused of academic misconduct, as well as for educators who must handle the repercussions. This example reinforced the importance of conducting rigorous scientific studies with key stakeholders when adopting any LLMs-based solutions that have direct or indirect impacts on students, educators, and other stakeholders. Likewise, the reporting of such studies should also adhere to high standards, incorporating both methodological specifics and detailed data descriptions. These details are especially pertinent when considering the diverse cultural backgrounds of students and the fact that most LLMs are primarily trained on English datasets, which could potentially introduce biases towards non-native English students (Liang et al., 2023). 3

Adopting a human-centred approach when developing and evaluating LLMs-based innovations are essential for ensuring these innovations remain ethical in practice, especially as ethical principles may not guarantee ethical AI due to their top-down manners (e.g., developed by regulatory bodies) (Mittelstadt, 2019). Future studies need to consider the ethical issues that may arise from their specific application scenarios and actively involve stakeholders to identify and address such issues. Specifically, LLM-based innovations should aim to reach at least Tier 3 in the transparency index and TRL-7 in technology readiness. This involves a fully functional system being integrated into authentic learning environments and validated by students and educators in terms of its practicality and ethical considerations. For any decisions made by the LLM-based innovations, the relevant stakeholders should be informed about how 4

the decision was reached, as well as the potential risks and biases involved. For instance, when students receive an assessment that has been automatically graded, these grades should be accompanied by a warning message indicating that they have been graded by LLMs and AI (Angelone et al., 2022). Students should also have the opportunity to consult their teacher regarding any concerns.

The active involvement of stakeholders should also extend beyond the education sector, also involving policymakers and industry companies to establish the guidelines for adopting LLMs-based innovations in learning and teaching practices, as such adoptions could have broader implications on society beyond the education sector. For example, human-AI collaboration might become an essential skill for students to succeed in the job market as AI solutions become an integral component of productivity in the industrial sector (Wang et al., 2020). Therefore, institutions that aim to prohibit AI tools could inadvertently place their students at a disadvantage compared to other institutions that proactively welcome such changes. This could be achieved by consistently refining their policy regarding the use of LLMs and generative AI solutions, based on stakeholder feedback and empirical evidence.

5.3 | Limitations

The current findings should be interpreted with several limitations in mind. First, although we assessed the practicality and ethicality of LLMs-based innovations with seven different items, there could be other aspects of these multi-dimensional concepts that we omitted. Nevertheless, these assessment items were chosen directly from the corresponding definitions and related to the pressing issues in the literature (Adams et al., 2021; Weidinger et al., 2021). Second, we only included English publications, which could have biased our findings regarding the availability of LLMs-based innovations among different countries. Thirdly, as we strictly followed the PRISMA protocol and only included peer-reviewed publications, we may have omitted the emerging works published in different open-sourced archives. These studies may contain interesting findings regarding the latest LLMs (e.g., ChatGPT). Additionally, this review focused on the potential of LLMs-based innovations in automating educational tasks, and thus, other pressing issues, such as the potential threat to academic integrity, were outside of the scope of this systematic scoping review. We briefly touched on these pressing issues in the implications and illustrated the importance of the current findings in supporting future educational studies to address these issues. Moreover, since this study is a systematic scoping review, we did not assess the quality of the included studies, and thus, the findings, particularly, the performance metrics extracted from the reviewed studies, may need further evaluation. The goal of this study is to provide an overview of the different educational tasks that can be augmented by LLMs and generative AI models, which can serve as a reference point for future studies to further develop on using the state-of-the-art models (e.g., ChatGPT and PaLM 2). Furthermore, the transparency index that we adopted for RQ3 did not consider the transparency to students, which could be an important direction for future human-centred AI studies. Finally, we recognise the rapid development in the field of artificial intelligence in education. It is pertinent to mention that a number of recent workshops and preliminary papers, while contributing to this field, were not incorporated in this scoping review due to time constraints (Leiker et al., 2023; Ma et al., 2023; Caines et al., 2023). Their exclusion represents a limitation to the breadth of this study, acknowledging the relentless pace of scholarly advancements in this area.

6 | CONCLUSION

In this study, we systematically reviewed the current state of educational research on LLMs and identified several practical and ethical challenges that need to be addressed in order for LLMs-based innovations to become beneficial and

impactful. Based on the findings, we proposed three recommendations for future studies, including updating existing innovations with state-of-the-art models, embracing the initiative of open-sourcing models/systems, and adopting a human-centred approach throughout the developmental process. These recommendations could potentially support future studies to develop practical and ethical innovations that can be implemented in authentic contexts to automate a wide range of educational tasks.

references

Adams, C., Pente, P., Lemermeyer, G. and Rockwell, G. (2021) Artificial intelligence ethics guidelines for k-12 education: a review of the global landscape. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II*, 24–28. Springer.

Ahmed, A., Joorabchi, A. and Hayes, M. J. (2022) On the application of sentence transformers to automatic short answer grading in blended assessment. In *2022 33rd Irish Signals and Systems Conference (ISSC)*, 1–6. IEEE.

Angelone, A. M., Galassi, A. and Vittorini, P. (2022) Improved automated classification of sentences in data science exercises. In *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference 11*, 12–21. Springer.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W. et al. (2023) A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Becker, H. J. (2000) Findings from the teaching, learning, and computing survey. *Education policy analysis archives*, **8**, 51–51.

Beseiso, M., Alzubi, O. A. and Rashaideh, H. (2021) A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, **33**, 727–746.

Boeren, E. (2019) Understanding sustainable development goal (sdg) 4 on "quality education" from micro, meso and macro perspectives. *International review of education*, **65**, 277–294.

Brown, H., Lee, K., Mireshghallah, F., Shokri, R. and Tramèr, F. (2022) What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2280–2292.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020) Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

Bulut, O. and Yildirim-Erbasli, S. N. (2022) Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education*, **9**, 72–87.

Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., Gao, Y., Andersen, O., Yuan, Z., Elliott, M., Moore, R., Bryant, C., Rei, M., Mullooly, A., Nicholls, D. and Buttery, P. (2023) On the application of large language models for language teaching and assessment technology. In *AIED Workshops*, in press.

Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E. and Lester, J. C. (2020) Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*, 55–66. Springer.

Carroll, A., Forrest, K., Sanders-O'Connor, E., Flynn, L., Bower, J. M., Fynes-Clinton, S., York, A. and Ziaei, M. (2022) Teacher stress and burnout in australia: examining the role of intrapersonal and environmental factors. *Social Psychology of Education*, **25**, 441–469.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D. and Mello, R. F. (2021) Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, **2**, 100027.

Chaudhry, M. A., Cukurova, M. and Luckin, R. (2022) A transparency index framework for ai in education. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II*, 195–198. Springer.

Chechitelli, A. (2023) Ai writing detection update from turnitin's chief product officer. https://www.turnitin.com/blog/ai-writing-detection-update-from-turnitins-chief-product-officer. Accessed: 2023-06-12.

Condor, A., Litster, M. and Pardos, Z. (2021) Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Doewes, A. and Pechenizkiy, M. (2021) On the limitations of human-computer agreement in automated essay scoring. *International Educational Data Mining Society*.

Doyle, W. and Ponder, G. A. (1977) The practicality ethic in teacher decision-making. *Interchange*, **8**, 1–12.

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., Liu, K., Chen, L., Tran, S., Cheng, N. et al. (2022) A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, **119**, e2123433119.

Ertmer, P. A. (1999) Addressing first-and second-order barriers to change: Strategies for technology integration. *Educational technology research and development*, **47**, 47–61.

Ferguson, R., Hoel, T., Scheffel, M. and Drachsler, H. (2016) Guest editorial: Ethics and privacy in learning analytics. *Journal of Learning Analytics*, **3**, 5–15.

Fonseca, S. C., Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Carvalho, L. S. and Cristea, A. I. (2020) Automatic subject-based contextualisation of programming assignment lists. *International Educational Data Mining Society*.

Gašević, D., Dawson, S., Rogers, T. and Gasevic, D. (2016) Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, **28**, 68–84.

Gašević, D., Greiff, S. and Shaffer, D. W. (2022) Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond. *Computers in Human Behavior*, **134**, 107304. URL: https://www.sciencedirect.com/science/article/pii/S0747563222001261.

Geller, S. A., Gal, K., Segal, A., Sripathi, K., Kim, H. G., Facciotti, M. T., Igo, M., Hoernle, N. and Karger, D. (2021) New methods for confusion detection in course forums: Student, teacher, and machine. *IEEE Transactions on Learning Technologies*, **14**, 665–679.

Ghosh, D., Klebanov, B. B. and Song, Y. (2020) An exploratory study of argumentative writing by young students: A transformer-based approach. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 145–150.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., Chartash, D. et al. (2023) How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, **9**, e45312.

Holmes, W. and Porayska-Pomsta, K. (2022) *The Ethics of Artificial Intelligence in education: Practices, challenges, and debates*. Taylor & Francis.

Jayaraman, J. and Black, J. (2022) Effectiveness of an intelligent question answering system for teaching financial literacy: A pilot study. In *Innovations in Learning and Technology for the Workplace and Higher Education: Proceedings of 'The Learning Ideas Conference'2021*, 133–140. Springer.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. et al. (2023) Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, **103**, 102274.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. and Gašević, D. (2022) Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, **3**, 100074.

Kumar, N., Mali, R., Ratnam, A., Kurpad, V. and Magapu, H. (2022) Identification and addressal of knowledge gaps in students. In *2022 3rd International Conference for Emerging Technology (INCET)*, 1–6. IEEE.

Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S. (2020) A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, **30**, 121–204.

Leiker, D., Finnigan, S., Gyllen, A. R. and Cukurova, M. (2023) Prototyping the use of large language models (llms) for adult learning content creation at scale. In *AIED Workshops*, in press.

Li, C. and Xing, W. (2021) Natural language generation using deep learning to support mooc learners. *International Journal of Artificial Intelligence in Education*, **31**, 186–214.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D. and Chen, G. (2023) Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 100140.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. and Zou, J. (2023) Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*.

Liu, S., Liu, S., Liu, Z., Peng, X. and Yang, Z. (2022) Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement. *Computers & Education*, **181**, 104461.

Liu, Z., He, X., Liu, L., Liu, T. and Zhai, X. (2023) Context matters: A strategy to pre-train language model for science education. *arXiv preprint arXiv:2301.12031*.

Ma, Q., Wu, S. and Koedinger, K. (2023) Is llm the better programming partner? In *AIED Workshops*, in press.

Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O. and Ghadi, Y. Y. (2022) Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Computer Science*, **8**, e1010.

Maxwell, S. E., Lau, M. Y. and Howard, G. S. (2015) Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, **70**, 487.

Merine, R. and Purkayastha, S. (2022) Risks and benefits of ai-generated text summarization for expert level content in graduate health informatics. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 567–574. IEEE.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I. and Roth, D. (2021) Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Mittelstadt, B. (2019) Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, **1**, 501–507.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T. and Stamper, J. (2022) Assessing the quality of student-generated short answer questions using gpt-3. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings*, 243–257. Springer.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A. and Aromataris, E. (2018) Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, **18**, 1–7.

Nguyen, T. T., Le, A. D., Hoang, H. T. and Nguyen, T. (2021) Neu-chatbot: Chatbot for admission of national economics university. *Computers and Education: Artificial Intelligence*, **2**, 100036.

Nye, B., Mee, D. and Core, M. G. (2023) Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *AIED Workshops*, in press.

Oleny, A. (2023) Generating multiple choice questions from a textbook: Llms match human performance on most metrics. In *AIED Workshops*, in press.

OpenAI (2023) Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2023-02-25.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E. et al. (2021) The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, **88**, 105906.

Pardo, A. and Siemens, G. (2014) Ethical and privacy principles for learning analytics. *Br J Educ Technol*, **45**, 438–450.

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J. and D'Mello, S. K. (2021) Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. *International Educational Data Mining Society*.

Ramesh, D. and Sanampudi, S. K. (2022) An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, **55**, 2495–2527.

Rudolph, J., Tan, S. and Tan, S. (2023) Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, **6**.

Sallam, M. (2023) The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, 2023–02.

Sarsa, S., Denny, P., Hellas, A. and Leinonen, J. (2022) Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, 27–43.

Sawatzki, J., Schlippe, T. and Benner-Wickner, M. (2022) Deep learning techniques for automatic short answer grading: Predicting scores for english and german answers. In *Artificial Intelligence in Education: Emerging Technologies, Models and Applications: Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology*, 65–75. Springer.

Schneider, J., Richner, R. and Riser, M. (2022) Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 1–31.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. and Kersting, K. (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, **4**, 258–268.

Science, D. and Group, T. () Technology readiness levels definitions and descriptions. `https://www.dst.defence.gov.au/sites/default/files/basic_pages/documents/TRL%20Explanations_1.pdf`. Accessed: 2023-01-20.

Selwyn, N. (2019) What's the problem with learning analytics? *JLA*, **6**, 11–19.

Sha, L., Li, Y., Gasevic, D. and Chen, G. (2022a) Bigger data or fairer data? augmenting bert via active sampling for educational text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1275–1285.

Sha, L., Raković, M., Das, A., Gašević, D. and Chen, G. (2022b) Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, **15**, 481–492.

Sha, L., Raković, M., Lin, J., Guan, Q., Whitelock-Wainwright, A., Gašević, D. and Chen, G. (2022c) Is the latest the greatest? a comparative study of automatic approaches for classifying educational forum posts. *IEEE Transactions on Learning Technologies*.

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D. and Chen, G. (2021) Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22*, 381–394. Springer.

Shang, J., Huang, J., Zeng, S., Zhang, J. and Wang, H. (2022) Representation and extraction of physics knowledge based on knowledge graph and embedding-combined text classification for cooperative learning. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1053–1058. IEEE.

Sharma, A., Kabra, A. and Kapoor, R. (2021) Feature enhanced capsule networks for robust automatic essay scoring. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part V 21*, 365–380. Springer.

Song, W., Hou, X., Li, S., Chen, C., Gao, D., Sun, Y., Hou, J., Hao, A. et al. (2022) An intelligent virtual standard patient for medical students training based on oral knowledge graph. *IEEE Transactions on Multimedia*.

Sridhar, P., Doyle, A., Agarwal, A., Bogart, C., Savelka, J. and Sakr, M. (2023) Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives. In *AIED Workshops*, in press.

Su, Y. and Zhang, Y. (2020) Automatic construction of subject knowledge graph based on educational big data. In *Proceedings of the 2020 The 3rd International Conference on Big Data and Education*, 30–36.

Truong, T.-L., Le, H.-L. and Le-Dang, T.-P. (2020) Sentiment analysis implementing bert-based pre-trained language model for vietnamese. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 362–367. IEEE.

Tsai, Y.-S. and Gasevic, D. (2017) Learning analytics in higher education—challenges and policies: a review of eight learning analytics policies. In *Proceedings of the seventh international learning analytics & knowledge conference*, 233–242.

Tsai, Y.-S., Whitelock-Wainwright, A. and Gašević, D. (2020) The privacy paradox and its implications for learning analytics. In *Proceedings of the tenth international conference on learning analytics & knowledge*, 230–239.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems*, **30**.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y. and Wang, Q. (2020) From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 1–6.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. et al. (2021) Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M. and Drachsler, H. (2021) Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, **4**, 654924.

Wu, J. (2022) Analysis and evaluation of the impact of integrating mental health education into the teaching of university civics courses in the context of artificial intelligence. *Wireless Communications and Mobile Computing*, **2022**.

Wu, X., He, X., Li, T., Liu, N. and Zhai, X. (2023) Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. *arXiv preprint arXiv:2301.08771*.

Yan, L., Zhao, L., Gasevic, D. and Martinez-Maldonado, R. (2022) Scalability, sustainability, and ethicality of multimodal learning analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, 13–23.

Yang, S. J., Ogata, H., Matsui, T. and Chen, N.-S. (2021) Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, **2**, 100008.

Zawacki-Richter, O., Marín, V. I., Bond, M. and Gouverneur, F. (2019) Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, **16**, 1–27. 1

Zeng, Z., Gašević, D. and Chen, G. (2023) On the effectiveness of curriculum learning in educational text scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2

Zheng, L., Niu, J., Long, M. and Fan, Y. (2023) An automatic knowledge graph construction approach to promoting collaborative knowledge building, group performance, social interaction and socially shared regulation in cscl. *British Journal of Educational Technology*, **54**, 686–711. 3

Zheng, L., Niu, J. and Zhong, L. (2022) Effects of a learning analytics-based real-time feedback approach on knowledge elaboration, knowledge convergence, interactive relationships and group performance in cscl. *British Journal of Educational Technology*, **53**, 130–149. 4