

Large Language Models for Sustainable Assessment and Feedback in Higher Education: Towards a Pedagogical and Technological Framework

Daniele Agostini^{1,*†}, Federica Picasso^{1,†}

¹University of Trento, Palazzo Fedrigotti, corso Bettini 31, Rovereto (TN), 38068, Italy

Abstract

Nowadays, there is growing attention on enhancing the quality of teaching, learning and assessment processes. As a recent EU Report underlines, the assessment and feedback area remains a problematic issue regarding educational professionals' training and adopting new practices. In fact, traditional summative assessment practices are predominantly used in European countries, against the recommendations of the Bologna Process guidelines that promote the implementation of alternative assessment practices that seem crucial in order to engage and provide lifelong learning skills for students, also with the use of technology. Looking at the literature, a series of sustainability problems arise when these requests meet real-world teaching, particularly when academic instructors face the assessment of extensive classes. With the fast advancement in Large Language Models (LLMs) and their increasing availability, affordability and capability, part of the solution to these problems might be at hand. In fact, LLMs can process large amounts of text, summarise and give feedback about it following predetermined criteria. The insights of that analysis can be used both for giving feedback to the student and helping the instructor assess the text. With the proper pedagogical and technological framework, LLMs can disengage instructors from some of the time-related sustainability issues and so from the only choice of the multiple-choice test and similar. For this reason, as a first step, we are proposing a starting point for such a framework to a panel of experts following the Delphi methodology and reporting the results.

Keywords

Assessment, Evaluation, Higher Education, Educational Technology, Technology-Enhanced Assessment, Artificial Intelligence, Large Language Models

1. AI Teaching, Learning and Assessment in Higher Education: the state of the art

Recent attention has focused on enhancing teaching, learning and assessment quality [1]. However, traditional summative assessments are still dominant in Europe, despite Bologna Process guidelines promoting alternative practices to develop students' lifelong learning skills [2, 3]. With extensive classes, implementing these practices raises sustainability issues for instructors. Large language models (LLMs) may help by processing large text amounts, summarising, and

st International Workshop on High-performance Artificial Intelligence Systems in Education, 2023, Rome, IT

*Corresponding author.


†These authors contributed equally.

✉ daniele.agostini@unitn.it (D. Agostini); federica.picasso@unitn.it (F. Picasso)

🆔 0000-0002-9919-5391 (D. Agostini); 0000-0002-8381-6456 (F. Picasso)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

providing feedback based on specified criteria [4]. Used pedagogically, LLMs could relieve instructors' time pressures, expanding beyond multiple-choice tests.

As a UNICEF definition affirmed:

“AI refers to machine-based systems that can, given a set of human-defined objectives, make predictions, recommendations, or decisions that influence real or virtual environments. AI systems interact with us and act on our environment, either directly or indirectly. Often, they appear to operate autonomously and can adapt their behaviour by learning about the context” [6, p.16].

To connect this topic to the education field, it is possible to affirm that, in terms of technological advancements, theoretical contributions and impact on education, the field of Artificial Intelligence in Education (AIED) has seen success over the past 25 years [7, 8, 9, 10]. Instead of just automating the instruction of students sitting in front of computers, AI could help open up teaching and learning opportunities that would otherwise be difficult to achieve, question conventional pedagogies, or assist teachers in becoming more successful. Other AIED technologies currently track student progress and give tailored feedback to determine if the student has mastered the topic in issue. AIED technologies built to support collaborative learning might gather similar information, and intelligent essay assessment tools can potentially draw conclusions about a student's knowledge. All of this information and more could be gathered during a student's time in formal educational settings (the learning sciences have long recognised the value of students engaging in constructive assessment activities), along with information about the student's participation in non-formal learning (such as learning a musical instrument, a craft, or other skills) and informal learning (such as language learning or enculturation by immersion) [11]. The potential of AI in educational settings, as well as the necessity for AI literacy, places educators at the forefront of these new and exciting breakthroughs that were previously relegated to obscure computer science laboratories. Simultaneously, teachers and administrators are required to have clear perspectives on the potential of AI in education and, eventually, to incorporate this ground-breaking technology into their practice [12]. To deeply focus on the characteristics of AIED concept, for example, Holmes and colleagues [13] created the AIED taxonomy, a system that is helpful to categorise AIED tools and applications into three different but intersecting categories: (1) student-focused, (2) teacher-focused, and (3) institution-focused AIED.

2. Theoretical background

2.1. AI and Education: what about sustainable and authentic assessment?

Higher education aims to provide meaningful, relevant courses, where graduates can learn to work and live in an increasingly digital society [14]. In our contemporary education system, our students need to be supported in becoming effective lifelong learners, who must be prepared for the assessment tasks they will encounter in their lives, but they also need to become lifelong assessors, possessing assessment skills acquired through continuous work. It could be possible through the implementation of Assessment for Learning intended as an approach that emphasises the assessment process as an "essential moment of the educational experience,

Table 1

A taxonomy of AIED systems [13, 12 p.550]

	AIED Focus		
	Student focused AIED	Teacher focused AIED	Institution focus AIED
	<ul style="list-style-type: none"> • Intelligent Tutoring Systems (ITS) • AI-assisted Apps (e.g., maths, text-to-speech, language learning) • AI-assisted Simulations (e.g., games-based learning, VR, AR) • Automatic Essay Writing (AEW) • Chatbots • Automatic Formative Assessment (AFA) • Learning Network Orchestrators • Dialogue-based Tutoring Systems (DBTS) • Exploratory Learning Environments (ELE) • AI-assisted Lifelong Learning Assistant 	<ul style="list-style-type: none"> • Plagiarism detection • Smart Curation of Learning Materials • Classroom Monitoring • Automatic Summative Assessment • AI Teaching Assistant (including assessment assistant) • Classroom Orchestration 	<ul style="list-style-type: none"> • Admissions (e.g., student selection) • Course-planning, Scheduling, Timetabling • School Security • Identifying Dropouts and Students at risk • e-Proctoring

characterised by situations in which learners are enabled to analyse and understand the processes in which they are involved and can thus participate in decisions about their learning goals by becoming increasingly aware of their progress" [15, 16, p.56]

Learners as lifelong assessors must be able to:

- Estimate the possession of criteria for evaluating situations or carrying out an assignment;
- Seek and understand contextual feedback to construct new knowledge;
- Interpret and use feedback to achieve daily goals and challenges [17, 16, p.73].

Students need to gain, during their learning process, the assessment expertise, so the competence required for students to effectively understand assessment criteria and to be able to use the feedback received to close the gap and improve their own learning [19]. They will be supported in this process by teachers who, through continuous assessment and making judgements about the learning products of students, will develop more effective standards of judgement to define the expected competence of students themselves [18, 16, p.76]. But how to really support students' development of assessment expertise? As Sadler pointed out [19], it is necessary to involve students in direct and authentic assessment experiences, supporting them in the acquisition of the concept of quality, and training them in order to make complex judgements according to a multiplicity of criteria [19]. What are the principles to apply in order to create sustainable assessment contexts and then scaffold authentic assessment experiences?

Boud [17] drew up nine useful principles for reflecting on and designing sustainable assessment and feedback practices. For the author, it is indeed important that there is a timely sharing of clear assessment criteria for students; it is also crucial that students are seen as individuals who can achieve success and, in terms of assessment processes, these must be useful in making students confident in their success and in this sense it seems useful to consider separating feedback processes from the awarding of grades. The focus on learning during the assessment process must take priority over the focus on performance: it seems important therefore to support the development of self-assessment competencies and encourage the use of reflective peer assessment practices. One of the last fundamental aspects is related to the completion of the feedback loop as a tool for reviewing student work and finally, the importance of introducing a review process of assessment practices with the implementation of formative assessment processes is emphasised.

In terms of assessment and feedback in connection with AI systems, Swiecki and colleagues affirmed that

“AI-based techniques have been developed to fully or partially automate parts of the traditional assessment practice. AI can generate assessment tasks, find appropriate peers to grade work, and automatically score student work. These techniques offload tasks from humans to AI and help to make assessment practices more feasible to maintain” [20, p.2].

The power of AI related to the assessment and feedback processes is connected to the fact that, while traditional assessment practices could provide a partial overview about the students' performance, several AI techniques thanks to their characteristics can promote a wider vision of learning process and progress. In relation to the topic of authenticity and sustainability of assessment, AI systems can help to collect, represent, and assess data in a complex way: authentic assessment processes can be very articulated in terms of task and general design, so AI can be helpful for academics to monitor learning process towards an assessment of student progress [21, 20]. In fact, authentic assessment requires students to

“use the same competencies, or combinations of knowledge, skills, and attitudes, that they need to apply in the criterion situation in professional life” [22, p.69].

Authenticity has been recognised as a fundamental element of assessment design that encourages learning. Authentic assessment tries to reproduce the activities and performance criteria often encountered in the workplace and has been shown to have a favourable influence on student learning, autonomy, motivation, self-regulation, and metacognition; qualities that are significantly associated with employability [23]. Again, international authors even suggest that the authenticity of the assessment tasks is a need for reaching the expert level of problem-solving. Likewise, strengthening the authenticity of an assessment has the potential to have an encouraging effect on student learning and motivation [24, 25, 26, 22].

Finally, UNESCO [26], which has been in the last years amongst the most influential and active institutions that reflect on the implications of AI in society, provided the following guidelines for AI in assessment.

1. Testing and implementing artificial intelligence technologies is crucial for supporting the assessment of various dimensions of competencies and outcomes.

2. Caution is essential when adopting automated assessment with responses to rule-based closed questions.
3. Employing formative assessment leveraged by artificial intelligence as an integrated function of Learning Management Systems (LMS) is key to analysing student learning data with increased accuracy and efficiency and reducing human biases.
4. Progressive assessments based on artificial intelligence are imperative to provide regular updates to teachers, students, and parents.
5. Examining and evaluating the use of facial recognition and other artificial intelligence for user authentication and monitoring in remote online assessments is paramount.

This study moves along those research axes.

2.2. Large Language Models

Over the past few years, Large Language Models (LLMs) have become increasingly prevalent in society and educational settings. These AI-powered models are capable of generating, analysing, and summarising text, as well as engaging in dialogic interactions with humans [27]. One of the most well-known examples of LLMs is OpenAI's ChatGPT, which is based on GPT 3.5 and GPT 4 architectures. Other notable LLMs include Anthropic's Claude (1 and 2), Bing Chat (another GPT4-based model), and Google Bard. While these models are extremely powerful, there are concerns about data privacy and results consistency [28]. However, there are other options available. With the release of open-source and open-access models such as Meta's LLAMA and LLAMA 2, as well as TII's Falcon, and the growth of platforms like HuggingFace, which acts as a repository and framework, there are many possibilities for local LLMs with great capabilities. These models can be customised, fine-tuned, or even trained specifically for one's use case, allowing for greater flexibility and control [29].

To better understand the possibilities related to the use of these models, Tamkin et al. [4] proposed the following crucial points. In fact, LLMs can:

- **Generate:** LLMs can generate human-like text. This can be used to provide detailed explanations, create content, or even generate potential essay or report structures.
- **Summarise:** LLMs can summarise long pieces of text. This can help in providing concise summaries of lengthy student submissions. The summary can take into account different parameters in the text, providing information exactly on the aspects that the teacher wants to assess.
- **Posing and Answering Questions:** LLMs can understand a piece of text and answer questions about it as well as asking questions about it, if required to. This can be used to create interactive feedback and learning experiences.
- **Translate:** LLMs can translate text from one language to another. This can be useful in multilingual educational settings and to adapt content for foreign language student's inclusion. It also adds to the overall sustainability of the teacher's job in such situations.
- **Analyse the sentiment:** LLMs can understand the sentiment expressed in a piece of text. This can be used to gauge student sentiment in feedback, assignments or discussion forums.

- **Classify:** LLMs can classify text into predefined categories. This can be used for assisted grading or categorising student feedback.
- **Detecting plagiarism:** By comparing the similarity between different pieces of text, LLMs can help detect potential cases of plagiarism both between students and between students and the source material.
- **Measure Semantic Similarity:** LLMs can measure the semantic similarity between two pieces of text. This can be used to match student queries with relevant answers or resources and help the teacher in the assessment of the student's work.
- **Generate Feedback:** Based on the assessment of a student's work LLMs can generate personalised feedback. It would work even better if the LLM would have some teacher's notes on the assignment to work with.
- **Assess Knowledge:** LLMs can be used to assess a student's understanding of a topic based on their written submissions, especially if properly trained on correct assignments and having an assessment rubric to refer to.

LLMs are able to analyse massive amounts of text, aggregate it, and then offer feedback based on previously established standards [4]. The outcomes of that analysis can be applied to provide feedback to the student as well as to assist the instructor in evaluating the text. LLMs can remove teachers from some of the time-related sustainability difficulties, and thus from the sole choice of the multiple-choice test and similar, with the correct pedagogical and technical framework. In detail, Kasneci and colleagues [27] define the following opportunities for teachers and students regarding the implementation of AI in teaching and learning university context:

- “For university students, large language models can assist in the research and writing tasks, as well as in the development of critical thinking and problem-solving skills. These models can be used to generate summaries and outlines of texts, which can help students to quickly understand the main points of a text and to organise their thoughts for writing. Additionally, large language models can also assist in the development of research skills by providing students with information and resources on a particular topic and hinting at unexplored aspects and current research topics, which can help them to better understand and analyse the material.
- For personalised learning, teachers can use large language models to create personalised learning experiences for their students. These models can analyse student's writing and responses, and provide tailored feedback and suggest materials that align with the student's specific learning needs. Such support can save teachers' time and effort in creating personalised materials and feedback, and also allow them to focus on other aspects of teaching, such as creating engaging and interactive lessons” [27, pp.2-3].

In specific relation to assessment and feedback practice, the correlation with LLM can be summarised in four different points:

1. **Automatisation:** Assessment and Feedback can be totally or partially automated, although, at the moment, only supervised, human-mediated, assessment is advised [20, 30], there are some cases in which even a complete automation worked very well [29].

2. Sustainability: relative to the time variable, these models make scalable types of assessment that previously were not. This allows the teacher to always apply the most suitable method of assessment for verifying learning objectives [31, 20].
3. Objectivity: if trained correctly, LLMs should not have bias, they tend to be more objective and consistent than a human being and adhere to the established criteria.
4. AI in the loop: LLM, teachers and students can be part of the same process in which the IA is assigned only those tasks in which it is super-human [32, 33].

3. Research methods

3.1. Objectives and research question

In light of the evidence already produced by the international literature on the topic of AI and education, this research study aims to create and validate a model for the use of AI in Educational Assessment in Higher Education. The work is based on one main research question:

- Could university teachers use AI tools to adopt approaches that support more effective, sustainable and authentic assessment?

3.2. The Model

The designed model takes into account the existing literature connected to the topic of Assessment for Learning, Authentic Assessment and Sustainable Assessment [15, 22, 17]. Starting from these literature pieces of evidence, we are working on the development of a model useful to adopt AI in the assessment processes in the Higher Education context. The model considers the role that AI plays in the assessment and feedback practices connected to academics and students in the virtuous cycle of the learning spiral. The model itself will be assessed following four different levels proposed by Kaptelinin and colleagues [34] through the Activity Theory checklists for the design, the evaluation and the use of technology:

- Design: we will introduce this checklist in order to evaluate the design process itself.
- Evaluation Phase 1: in the first phase of the evaluation process, thanks to the introduction of the Delphi study approach and then, thanks to the collaboration of the experts, we will use the checklist connected to the Activity Theory to assess the structure of the model itself and collect prompts and suggestions.
- Evaluation Phase 2: in the second phase of the evaluation process, we aim to introduce an evaluation of how we are going to propose the use of the model itself, again based on the validated checklists.
- Use: in this last phase, we will introduce a specific checklist to evaluate the model and its related impacts on the teaching, learning and assessment processes.

Every checklist is developed following four different areas:

1. Means and ends: the extent to which the technology facilitates and constrains the attainment of users' goals and the impact of the technology on provoking or resolving conflicts between different goals.

2. Social and physical aspects of the environment: integration of target technology with requirements, tools, resources, and social rules of the environment.
3. Learning, cognition, and articulation: internal versus external components of activity and support of their mutual transformations with target technology.
4. Development: developmental transformation of the foregoing components as a whole [34].

The model will be composed of two levels of adoption:

- AI-Mediated Summative Assessment: level focused on assessment processes connected to Technology Enhanced Assessment practices, so the power of AI in connection to the possibility of introducing assessment and feedback timely, customised and informed by AI data [30].
- AI-Mediated Formative Assessment: level focused on the power of the AI implementation in assessment and feedback in order to monitor the whole learning process and to guide formative design actions and students' self and peer assessment processes [35, 36].

3.3. The Delphi technique

To validate the proposed model, we planned to introduce in our research the Delphi Study technique, intended as

“a scientific method to organise and manage structured group communication processes with the aim of generating insights on either current or prospective challenges [...] the Delphi technique builds on the anonymity of participating experts who are invited to assess and comment on different statements or questions related to a specific research topic” [37, p.2].

In a Delphi survey, the opinions, generated by the individuated group of experts across the multiple discussion rounds organised on a specific topic, are collected. The multi-round structure can be introduced sequentially, or immediately thanks to specific software. The structured group communication process should create a convergence or a divergence of opinions, producing a more dynamic and accurate collection of data in comparison to traditional opinion-polling techniques. This method allows researchers to focus on the sharing process, reducing risks related to group dynamics that may emerge during in-person collaborative processes [38, 39, 40, 41, 42, 37].

To sum up, the Delphi process can be divided in the following different steps:

1. Defining and recruiting experts: experts could be professionals with specific knowledge and relevant experience in a particular discipline and research area. The panel size is calibrated depending on the study's purpose.
2. Developing Delphi questionnaire: a Delphi questionnaire can be structured from primary data (interviews) or literature analysis to enhance validity. Experts can be involved with Paper-based or e-Delphi.
3. Round 1: this phase can be qualitative, really powerful to generate ideas (e.g. open-ended questions) or quantitative (e.g. rating scale). To certify the rigour of each round, Kilroy and Driscoll (2006) suggest that the response rate should not fall below 70%.

4. Analysis and design of Round 2: phase characterised by the results analysis and in connection with non-consensus issues, another questionnaire containing non-consensus issues and the Round 1 results are sent out to the experts (Round 2). The feedback sustains the experts' comparison of their initial opinions with the group result. Additional rounds are organised until the achievement of an acceptable level of consensus free of issues or controversies [43].

4. Results and discussions

Starting from JISC's [30] and UNESCO's [26] guidelines, we developed our model called AI-MAAS (AI-Mediated Assessment Academics and Students), composed of two different levels of application and interpretation, with a focus on the implementation of AI in Technology Enhanced Assessment and Formative Assessment processes. The model revolves around three focal points, i.e. with respect to the cyclic and balanced intersection of AI, teachers and students, following Vygotskij and Leont'ev's model of artefact mediation [44]. AI Mediated Summative Assessment level of implementation describes the three elements (Academics, AI and Students) and the connection between them as follows.

4.1. AI Mediated Summative Assessment Level

This first level (Fig. 1) focuses on general, usual, assessment. In this kind of process, AI can give the assessment a formative twist and significance, adding to the interaction with the student, but keeping the whole process sustainable for the academic. In this approach, most of formative exchanges would be between AI and Student, and mostly one-way (i.e. AI to Student), with early feedback on the product, and a final report on the assessment and future actions being the most important. It is important to notice that the AI's final feedback to the student must be, at this stage, moderated by the academic. This approach is supported by the capability of AI in connection to the possibility of introducing assessment and feedback timely, customised and informed by AI data [31].

Elements of this process are:

- AI: -Constructive role: AI can help teachers with the construction and delivery of early feedback and assessment. In connection, academics can define and share rubrics and assessment criteria to scaffold the assessment process. -Feedback mechanism: academics play a key role as actors who can give reinforcement feedback to the AI system itself, always to improve jointly developed evaluation processes. -Evaluation and Reporting: the relationship between AI and students is characterised by the exchange of the students' products to be assessed and then AI as the producer of specific reports that contain suggestions for learning improvement. AI with the role of tutor that shares early and timely feedback supported by the academics' expertise.
- Academics: -Experts provision: academics as experts able to build and share tailored information to sustain AI actions. -Feedback management: academics as professionals who are able to manage timely, personalised and AI-informed feedback.

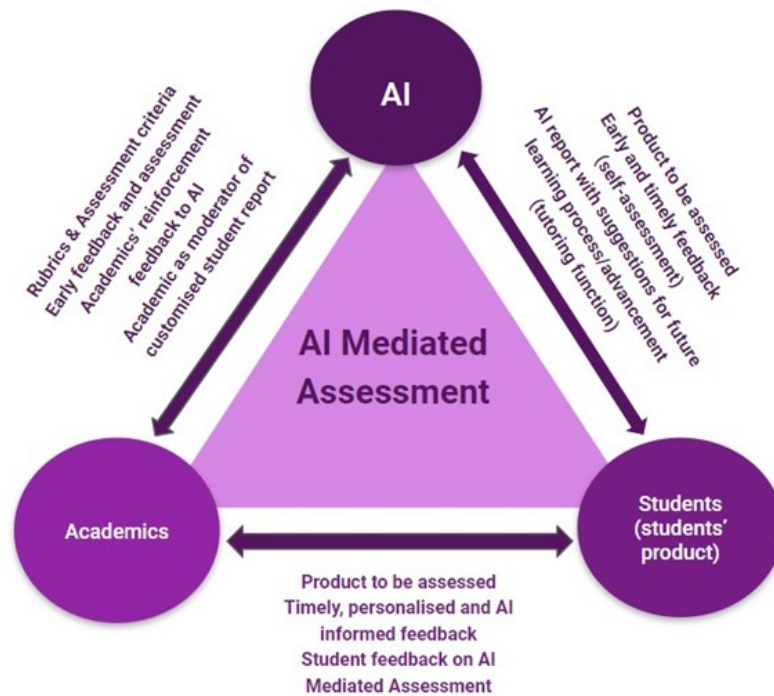


Figure 1: AI-Mediated Summative Assessment level

- Students (students' product): -Product creation: students as crucial actors are able to build specific products to be assessed thanks to the collaboration between academics and AI. -Guidance role: students as important elements to guide the AI Mediated Assessment processes with focused feedback.

4.2. AI Mediated Formative Assessment Level

AI Mediated Formative Assessment level of implementation describes the three elements and the connection between them as follows:

The second level (Fig. 2) focuses on proper formative assessment processes. In this approach the lecturer has designed the teaching to follow this approach, and the assessment is continuous, not relegated to the final stages of the course. Most interactions are bi-directional and occur between AI and Students, and Students and the lecturer. In this model, AI is directed by the lecturer and impersonates various roles, always in the form of collaboration with the students as a mentor, a tutor, or a peer. At the same time, AI capabilities to monitor the whole learning process and to inform formative design actions will be employed to support the academic [36, 37].

Elements of this process are:

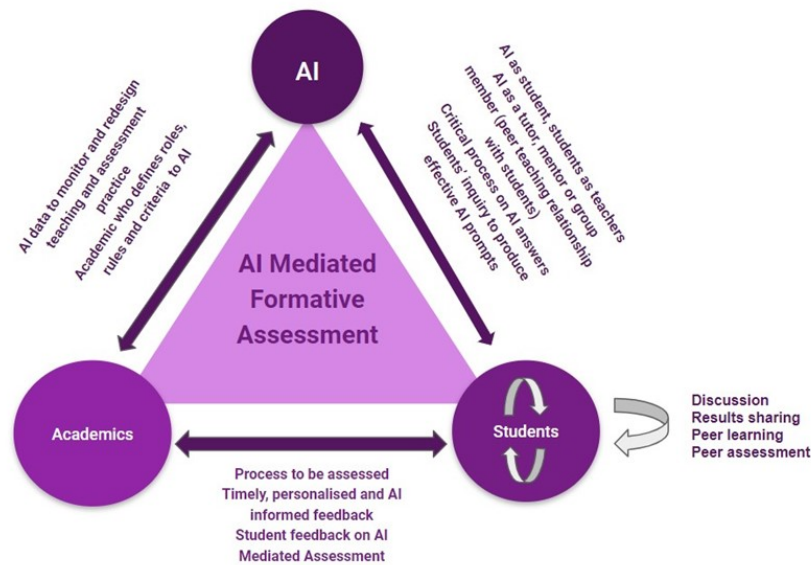


Figure 2: AI-Mediated Formative Assessment level

- **AI:** -Constructive role: the relationship between AI and academics is set up with a dynamic process of exchange in terms of expertise, resources and tasks. -Feedback mechanism: the data produced by AI can be fundamental for monitoring and redesigning academics' teaching and assessment practice. -Evaluation and Reporting: AI can play the role of student and the students can act as teachers in order to support and give prompts to AI that, at the same time, can play the role of tutor, mentor or group member (peer teaching relationship with students).
- **Academics:** -Expertise Provision: in connection with AI, academics define roles, rules and criteria for AI itself. -Feedback management: in terms of relationships with students, academics pay attention to the assessment of the whole learning process, giving timely, personalised and AI-informed feedback.
- **Students (students' product):** -Constructive role: students can activate critical thinking actions on AI answers, in order to stimulate deep and complex reflective processes, through specific students' inquiry to produce effective insights for AI. At the same time, students can discuss, share results produced by academics and AI, and also activate peer learning and assessment processes. -Guidance role: students can generate feedback on AI Mediated Assessment itself.

As previously mentioned, the model will be assessed and validated using the Delphi method [5] and following the Activity Theory checklist [34] during the design, validation and experimentation processes.

5. Conclusions and future research actions

Starting from the opportunities connected to the use of AI in education from both perspectives of students and teachers [27], it is important to understand how to better include these new opportunities to enhance teaching, learning and assessment processes in Higher Education contexts. For this purpose, our research is contextualised in an academic environment that has to cope with constant renewal in terms of approaches and strategies to deal with a major change at design, organisational and conceptual level. In connection with the topic of assessment and feedback from a perspective of assessment for learning, sustainability and authenticity [15, 17, 22], it is important to reflect and design specific formative and practical actions to sustain students and teachers in the implementation of AI systems as powerful agents to support the progress of the educational system. In terms of future research perspective, the designed actions include, after the validation of the AI-MAAS model through the Delphi study, experimentation using the model with academics, with a following phase which will comprehend the impact analysis and the assessment of the efficacy.

References

- [1] EHEA. (2015). Yerevan Communiqué. <https://www.ehea.info/page-ministerial-conference-yerevan-2015>
- [2] European Commission/EACEA/Eurydice, 2018 <https://eurydice.eacea.ec.europa.eu/publications/2018-eurydice-publications>
- [3] Y. Punie, C. Redecker, editor(s) (2017). European Framework for the Digital Competence of Educators. DigCompEdu, EUR 28775 EN, Publications Office of the European Union, Luxembourg, 10.2760/159770 (online), JRC107466.
- [4] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," arXiv preprint arXiv:2102.02503, 2021.
- [5] N. Dalkey, "An experimental study of group opinion: the Delphi method," *Futures*, vol. 1, no. 5, pp. 408-426, 1969.
- [6] UNICEF, "Policy guidance on AI for children," Author, 2021. [Online]. Available: <https://www.unicef.org/globalinsight/media/2356/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf.pdf>
- [7] K. VanLehn, "The behavior of tutoring systems," *International journal of artificial intelligence in education*, vol. 16, no. 3, pp. 227-265, 2006.
- [8] K. R. Koedinger and A. Corbett, "Cognitive tutors: Technology bringing learning sciences to the classroom," 2006.
- [9] N. T. Heffernan and C. L. Heffernan, "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching," *International Journal of Artificial Intelligence in Education*, vol. 24, pp. 470-497, 2014.
- [10] I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," *International Journal of Artificial Intelligence in Education*, vol. 26, pp. 582-599, 2016.

- [11] W. Holmes, M. Bialik, and C. Fadel, "Artificial intelligence in education," Globethics Publications, 2023.
- [12] W. Holmes and I. Tuomi, "State of the art and practice in AI in education," *European Journal of Education*, vol. 57, no. 4, pp. 542-570, 2022.
- [13] W. Holmes, M. Bialik, and C. Fadel, "Artificial intelligence in Education: Promises and implications for teaching & learning," The Center for Curriculum Redesign, 2019.
- [14] J. Nieminen, M. Bearman, and R. Ajjawi, "Designing the digital in authentic assessment: is it fit for purpose?," *Assessment & Evaluation in Higher Education*, vol. 48, no. 4, pp. 529-543, 2023.
- [15] K. Sambell, L. McDowell, and C. Montgomery, "Assessment for learning in higher education," Routledge, 2013.
- [16] V. Grion and A. Serbati, "Valutazione sostenibile e feedback nei contesti universitari. Prospettive emergenti, ricerche e pratiche," *PensaMultimedia*, 2019.
- [17] D. Boud, "Sustainable assessment: rethinking assessment for the learning society," *Studies in continuing education*, vol. 22, no. 2, pp. 151-167, 2000.
- [18] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," *Studies in higher education*, vol. 31, no. 2, pp. 199-218, 2006.
- [19] D. R. Sadler, "Formative assessment: Revisiting the territory," *Assessment in Education*, vol. 5, no. 1, pp. 77-84, 1989.
- [20] Z. Swiecki, H. Khosravi, G. Chen, R. Martinez-Maldonado, J. M. Lodge, S. Milligan, N. Selwyn, and D. Gašević, "Assessment in the age of artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 3, 100075, 2022.
- [21] V. Murphy, J. Fox, S. Freeman, and N. Hughes, "'Keeping it Real': A review of the benefits, challenges and steps towards implementing authentic assessment," *All Ireland Journal of Higher Education*, vol. 9, no. 3, 2017.
- [22] J. T. Gulikers, T. J. Bastiaens, and P. A. Kirschner, "A five-dimensional framework for authentic assessment," *Educational technology research and development*, vol. 52, no. 3, pp. 67-86, 2004.
- [23] V. Villarroel, S. Bloxham, D. Bruna, C. Bruna, and C. Herrera-Seda, "Authentic assessment: creating a blueprint for course design," *Assessment & Evaluation in Higher Education*, vol. 43, no. 5, pp. 840-854, 2018.
- [24] J. Herrington and A. Herrington, "Authentic assessment and multimedia: How university students respond to a model of authentic assessment," *Higher Educational Research & Development*, vol. 77, no. 3, pp. 305-322, 1998.
- [25] K. Sambell, L. McDowell, and S. Brown, "But is it fair?: An exploratory study of student perceptions of the consequential validity of assessments," *Studies in Educational Evaluation*, vol. 23, no. 4, pp. 349-371, 1997.
- [26] S. Gielen, Y. Dochy, and S. Dierick, "The influence of assessment on learning," in M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of quality and standards*, pp. 37-54, Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003.
- [27] F. Miao, W. Holmes, R. Huang, and H. Zhang, "AI and education: A guidance for policymakers," UNESCO Publishing, 2021. [Online]. Available: <https://doi.org/10.54675/PCSP7350>
- [28] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G.

- Groh, S. Günemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, 102274, ISSN 1041-6080, 2023. [Online]. Available: <https://doi.org/10.1016/j.lindif.2023.102274>
- [29] L. Chen, M. Zaharia, and J. Zou, "How is ChatGPT's behavior changing over time?," arXiv preprint arXiv:2307.09009, 2023. [Online]. Available: <https://arxiv.org/pdf/2307.09009.pdf>
- [30] P. P. Martin, D. Kranz, P. Wulff, and N. Graulich, "Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry," *Journal of Research in Science Teaching*, pp. 1–36, 2023. [Online]. Available: <https://doi.org/10.1002/tea.21903>
- [31] M. Webb, "A Generative AI Primer," JISC, 2023. [Online]. Available: <https://nationalcentreforai.jiscinvolve.org/wp/2023/05/11/generative-ai-primer/#3-1>
- [32] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: A systematic review," *Applied Sciences*, vol. 11, no. 12, 5467, 2021.
- [33] D.C. Englebart, "Augmenting human intellect: A conceptual framework," SRI Summary Report AFOSR-3223, October 1962. [Online]. Available: <https://www.doungelbart.org/pubs/augment-3906.html>
- [34] T. W. Malone, "Superminds: The surprising power of people and computers thinking together," Little, Brown Spark, 2018.
- [35] V. Kaptelinin and B. Nardi, "Acting with Technology: Activity Theory and interaction Design," MIT Press, 2006. [Online]. Available: <https://doi.org/10.5210/fm.v12i4.1772>
- [36] E. R. Mollick and L. Mollick, "Assigning AI: Seven Approaches for Students, with Prompts," June 12, 2023. [Online]. Available at SSRN: <https://ssrn.com/abstract=4475995> or <http://dx.doi.org/10.2139/ssrn.4475995>
- [37] OpenAI, "Teaching with AI," 2023. [Online]. Available: <https://openai.com/blog/teaching-with-ai>
- [38] D. Beiderbeck, N. Frevel, H. von der Gracht, S. L. Schmidt, and V. M. Schweitzer, "Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements," *MethodsX*, vol. 8, 101401, 2021.
- [39] S. Aengenheyster, K. Cuhls, L. Gerhold, M. Heiskanen-Schüttler, J. Huck, and M. Muszynska, "Real-Time Delphi in practice - A comparative analysis of existing software-based tools," *Technological Forecasting and Social Change*, vol. 118, pp. 15-27, 2017.
- [40] T. Gnatzy, J. Warth, H. von der Gracht, and I. L. Darkow, "Validating an innovative real-time Delphi approach—A methodological comparison between real-time and conventional Delphi studies," *Technological Forecasting and Social Change*, vol. 78, no. 9, pp. 1681-1694, 2011.
- [41] T. Gordon and A. Pease, "RT Delphi: An efficient, 'round-less' almost real time Delphi method," *Technological Forecasting and Social Change*, vol. 73, no. 4, pp. 321-333, 2006.
- [42] H. P. McKenna, "The Delphi technique: a worthwhile research approach for nursing?," *Journal of advanced nursing*, vol. 19, no. 6, pp. 1221-1225, 1994.
- [43] P. L. Williams and C. Webb, "The Delphi technique: A methodological discussion," *Journal of advanced nursing*, vol. 19, no. 1, pp. 180-186, 1994.
- [44] S. Chuenjitwongsa, "How to conduct a Delphi study," *Medical Education*, 2017. [Online].

Available: <https://meded.walesdeanery.org/how-to-guides>.

- [45] N. V. Cong-Lem, "Vygotsky's, Leontiev's and Engeström's cultural-historical (activity) theories: Overview, clarifications and implications," *Integrative Psychological and Behavioral Science*, vol. 56, no. 4, pp. 1091-1112, 2022.