International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00392-3

ARTICLE

Novelty Evaluation using Sentence Embedding Models in Open-ended Cocreative Problem-solving

Ijaz Ul Haq[1] · Manoli Pifarré[1] · Estibaliz Fraca[2]

Accepted: 4 January 2024

Abstract

Collaborative creativity (cocreativity) is essential to generate original solutions for complex challenges faced in organisations. Effective cocreativity requires the orchestration of cognitive and social processes at a high level. Artificial Intelligence (AI) techniques, specifically deep learning sentence embedding models, have emerged as valuable tools for evaluating creativity and providing feedback to improve the cocreation process. This paper examines the implications of sentence embedding models for evaluating the novelty of open-ended ideas generated within the context of real-life project-based learning. We report a case study research design involving twenty-five secondary students, where a cocreative process was developed to solve a complex, open-ended problem. The novelty of the co-generated ideas was evaluated using eight pre-trained sentence embedding models and compared with experts' evaluations. Correlation and regression analyses were performed to examine the reliability of the sentence embedding models in comparison to the experts' scoring. Our findings disclose that sentence embedding models can solve the challenge of evaluating open-ended ideas generated during the cocreative process. Moreover, the results show that two-sentence embedding models significantly correlate better with experts- Universal Sentence Encoder Transformer (USE-T) and USE Deep Averaging Network (USE-DAN). These findings have a high pedagogical value as they successfully evaluate the novelty generated in a real problem-based environment that uses technology to promote key cocreative processes. Furthermore, the real-time evaluation facilitated by these models can have a strong pedagogical impact because it can provide valuable feedback to teachers and students, thereby optimising collaborative ideation processes and promoting effective cocreative teaching and learning methodologies.

Keywords Novelty · Evaluation · Cocreative process · Project-based learning · Education · Deep learning

Extended author information available on the last page of the article

Published online: 23 January 2024

Introduction

The significance of creativity as a 21st-century skill is now reflected in educational policy initiatives and curricula (e.g., Plucker et al., 2023; Saboorizadeh et al., 2023) as a critical competence for professional and personal skills (e.g., Corbisiero-Drakos et al., 2021). However, creativity is a complex and multifaceted concept (Sawyer, 2021) that emerges in social and situational contexts (Altinay et al., 2022), requiring iterative and improvisational creative processes, particularly in collaborative settings (Sawyer, 2022). In recent years, advancements in interactive technologies have shed light on their potential to foster engagement, facilitate collaborative creative (cocreative) settings (Juusola, 2023), develop high-level cognitive and social processes involved in cocreation (Sun et al., 2022), help both students and teachers in their learning and teaching (Richardson, 2022).

In modern classrooms, Project-Based Learning methodology (henceforth PBL) fosters cocreativity, encourages the generation of novel ideas, and provides a contextualised learning experience. PBL presents students with real and open-ended challenges, placing them at the forefront of the learning process. It also promotes collaborative learning by incorporating technology to enhance and enrich the teaching and learning experience (Haatainen & Aksela, 2021). Furthermore, PBL methodology can embed the four pedagogical features that Sawyer (2022) claims are relevant to developing creativity as a process: a) iteration (not linear; there is a lot of iteration with unpredictable shifts in direction); b) ambiguity (it allows students the opportunity to formulate and solve their problems); c) exploration (discovering directions through experimentation); and d) emergence (ideas emerge from making and doing).

Despite extensive research based on factors related to enhanced creativity in different contexts, the importance of collaboration or teamwork to solve complex problems and generate innovative solutions indicates that groups often perform sub-optimally (e.g., Sawyer, 2021). This is because groups need to orchestrate multi-dimensional variables (such as behavioural, emotional, and cognitive variables) distributed at multiple levels (such as individual, peer, and group levels) and developed over time (Ouyang et al., 2023).

The intricacy of the high-level cognitive and social processes engaged in cocreation has sparked the interest of educational researchers, prompting them to explore methods for evaluating and providing real-time feedback to enhance the complex cocreative process within human groups and in real-class settings (Kenworthy et al., 2023). However, previous research has highlighted that achieving this challenging objective requires further exploration of educational research in three different directions. Firstly, to identify, and automatically evaluate critical creativity dimensions that can support cocreativity processes (Sun et al., 2022). Secondly, researchers should determine the most reliable Natural Language Processing (henceforth NLP) methods for examining group idea generation in open-ended learning contexts (Emara et al., 2021), such as PBL learning scenarios. Lastly, attention should be given to the design and delivery of real-time feedback to optimise the cocreativity process (Algarni, 2022). This paper addresses the first aforementioned research gap by using an AI-driven approach to evaluate the novelty of open-ended ideas co-generated within the context of real-life PBL. Empirical research has started to explore the opportunities of NLP techniques to face the challenge of evaluating the novelty of co-generated ideas during a cocreative process.

However, previous research reveals three potential research gaps that motivated our study. Firstly, existing computational techniques in the literature fail to adequately preserve the semantic and contextual meaning of the words that constitute the sentences. For example, keyword matching (Prasch et al., 2020), LSA (Kenett, 2019; LaVoie et al., 2020), knowledge-based techniques using ontologies (Georgiev & Casakin, 2019) and word embedding models (Olson et al., 2021; Johnson & Hass, 2022) struggle to vectorise the entire sentence in a numerical vector space while maintaining the semantic and contextual meanings of sentence' constituents. Secondly, NLP techniques often require domain-specific data for their computations (Camburn et al., 2019; Simpson et al., 2019), which is typically lacking in real-world classroom environments. Thirdly, automatic creativity evaluations have primarily been applied in closed contexts, such as alternate uses tasks (Beaty & Johnson, 2021; Dumas et al., 2021). This leaves a considerable gap in adapting to open-ended ideation scenarios, particularly in the cocreative process within a PBL environment. To address these gaps, we next explore the AI development of sentence embedding models applicable to PBL and educational contexts.

Considering the aforementioned research gaps, we have chosen sentence embedding models, which present new opportunities for creativity research by addressing the computational challenges of existing NLP techniques (Ul Haq & Pifarré, 2023). Sentence embedding models have recently emerged to represent ideas or thoughts in sentence structures in a numerical vector space while preserving the semantic and contextual meaning of words constituting the sentences. These models provide diverse opportunities to calculate variables associated with the creativity dimensions, such as novelty. Their distinctive feature lies in being pre-trained on large corpora and applicable for downstream applications with general-purpose semantic spaces without the need for training on existing data. Their pre-training on large corpora and the ability to vectorise entire sentences make them suitable for application in PBL, eliminating the need for specific training data or existing data (e.g., in PBL, no pre-existing data is available). Moreover, their adaptability is particularly relevant for PBL, where data is open-ended (no domain-specific data available), and ideas are in sentence structure (more complex than single-word ideas). Despite their potential, a significant research gap exists as sentence embedding models have not been extensively tested in PBL environments, especially in cocreative contexts. A notable study by Kenworthy et al. (2023) used sentence embedding models to assess the ideational dynamics of sports datasets generated by creative groups using a single embedding model, namely USE-DAN. The study computed novelty using a domain dataset that requires a pre-existing domain-specific corpus (i.e., the sports dataset), which is not available in the open-ended PBL cocreative processes considered in this work. Hence, more exploration is needed to apply different sentence embedding models to evaluate the novelty of ideas generated in real-classroom PBL scenarios. Therefore, we expand this exploration by using sentence embedding models to analyse the ideas generated during the different phases of cocreative problem-solving. Our first research question is: How could deep learning sentence embedding models evaluate the novelty of ideas generated in a complex, open-ended, and cocreative ideation process?".

To enhance the pedagogical impact of novelty evaluation, it is crucial to select the most reliable deep-learning sentence embedding model for a real classroom PBL context. However, to our knowledge, a significant research gap exists, as sentence embedding models have not been compared with experts' scores using real-classroom data in PBL to test their reliability against human experts. This lack of empirical evidence underscores the need for our second research question: "Which deep learning sentence embedding model(s) is closer to the experts' evaluation of novelty scores?" By doing so, we can better automatically estimate the teacher's evaluation of the novelty of the students' ideas to solve the proposed project. Furthermore, this information will help increase teachers' trust in AI-driven techniques for creativity assessment because it will help them understand how AI makes decisions compared to human experts (Nazaretsky et al., 2022). Moreover, from an educational standpoint, the scores provided by the sentence embedding model could be valuable information for teachers to design and deliver real-time feedback to help students in the collaborative ideation and subsequent decision processes to creatively solve the challenge. Indeed, evaluating and providing feedback to teachers about how novel ideas are generated in PBL classrooms are key pedagogical issues to promote creativity in real classrooms (Lu et al., 2022). Our study makes strides towards AI-supported orchestration of cocreation processes by examining the possibilities that sentence embedding models offer to evaluate the novelty of ideas generated collaboratively during the process of solving a complex and open-ended challenge.

Simultaneously, our study positions its applications in real-world educational settings, specifically within the digitalised educational landscape. In an era of digitalised education, educational technologies and e-learning platforms have adeptly incorporated creativity and AI into teaching and learning environments. In this regard, the findings of our study could be integrated into these platforms to facilitate AI assistance in real-time novelty evaluation and feedback on students' solutions. The pragmatic implications of our study could contribute to creativity-driven and AI-empowered digital education. This integration substantially benefits students, educators, and e-learning platforms, signalling a promising future for real-world education.

Fundamentals

Creativity evaluation can be a valuable tool for identifying and supporting students to develop their creative processes, promoting innovation, critical thinking, and improving teaching methods. However, creativity evaluation poses a challenge (Van Hooijdonk et al., 2022). Educational research highlights that the phenomenon of creativity evaluation involves four dimensions: Fluency (number of meaningful ideas), Flexibility (number of different categories), Elaboration (detailed ideas), and novelty (uniqueness of ideas) (Bozkurt Altan & Tan, 2021). Among these four dimensions, novelty holds particular significance, serving as a core dimension (Wang & Deng, 2022) and exhibiting a stronger correlation with creativity compared to other dimensions (Lloyd-Cox et al., 2022).

In automatic creativity evaluation research, novelty is evaluated through various approaches, where we briefly outline the following six: 1) measuring uniqueness or unique solutions, a measure of how unique a concept is relative to others (Doboli et al., 2020); 2) evaluating deviation from existing knowledge or standard solutions (Karampiperis et al., 2014); 3) assessing originality by examining similarity to existing ideas (Prasch et al., 2020; Jimenez-Mavillard & Suarez, 2022); 4) and measuring the extent of differentiation from other ideas (Walter & Back, 2013); 5) considering similarity, where smaller distances indicate similar contexts (LaVoie et al., 2020); and 6) examining the semantic distance between ideas (Dumas et al., 2021).

Likewise, in the aforementioned ways of evaluating novelty, different NLP computational techniques have been applied to evaluate the ideas' novelty: from statistical (Prasch et al., 2020) and knowledge-based techniques (Georgiev & Casakin, 2019) to deep learning models, e.g., the GloVe (Beaty & Johnson, 2021; Johnson & Hass, 2022) or Word2Vec model (Sung et al., 2022). Recently, the emergence of pre-trained sentence embedding models is an important approach for learning contextual representation and can be useful for evaluating open-ended cocreative ideas because these share the following three valuable characteristics. Firstly, these sentence embedding models are trained over a large corpus and then use the potential of transferring the learned knowledge to other NLP tasks (Zheng et al., 2022), such as novelty evaluation of a generic and small dataset generation in PBL during the cocreative process. Secondly, the cocreative ideas generated during the cocreative process are open-ended and have a sentence structure because sentences are the most explicitly specified elements of individual thoughts or ideas. Therefore, sentence embedding models are designed to learn a viable representation of the whole sentence in a vector space, allowing us to access their fine-grained semantics while preserving the semantic structure. Thirdly, novelty is computed as the semantic similarity of an idea with other ideas. Thus, sentence embedding models are evaluated over text similarity tasks, showing significant positive results. All these three important characteristics of sentence embedding models make them suitable for use in a PBL cocreation process.

Sentence embedding models' approach to evaluating novelty involves two computations. As a first step, sentence embedding, or sentence vectorisations, encode sentences into a high-dimensional space. As a second step, cosine similarity among the sentences' vectors is computed. Among the diverse sentence embedding models available in the literature, we have analysed which ones should be used to pursue our goal for the first step. Considering the performance of sentence embedding models on sentence embedding/vectorisation in high dimensional space and text similarity among sentences' vectors as criteria for selecting the eight pre-trained models appropriate to evaluate the novelty of open-ended cocreative ideas. We selected eight sentence embedding models presented in Table 1 in our study that accomplish these two computations, which are: 1) USE-T, 2) USE-DAN (Cer et al., 2018), 3) all-MiniLM-L6 (Wang et al., 2020), 4) all-mpnet-base (Song et al., 2020), 5) SRoBERTa-NLI Large (Liu et al., 2021), 6) ELMo, 7) InferSent with GloVe, and 8)

Table 1 Description of the architecture, training method, data, and key performance information of the 8-sentence embedding models selected for this study

#	Model	Description	Training method	Training data	Performance on sentence embedding	Performance on text similarity tasks
1	USE-T	Uses encoding sub-graph of the transformer architecture utilising attention concept. Attention makes it computationally tractable for a transformer model to consider both the ordering and identity of all the other words, enabling the computation of context-aware representations of words in a sentence	Supervised & unsupervised	Wikipedia, Standard Natural Language Inference (SNLI), web news, and web Questions/Answers	Surpass/outperform word level embedding models on transfer learning using sentence embeddings, achieving high accuracy on sentence embedding tasks	USE Transformer and USE DAN models surpass word-level embeddings on the Semantic Textual Similarity (STS) benchmark
2	USE-DAN	Takes as input embeddings for words and bi grams are first averaged together, and passed through a feedforward Deep Neural Network (DNN) to produce sentence embedding	Supervised & unsupervised	SNLI, Wikipedia, news, web pages	USE-DAN surpasses word level embedding model and is computationally efficient, as sentence length is increased	
3	all-MiniLM-L6	Compresses a large Transformer model into a smaller model using deep self-attention distillation	Supervised & unsupervised	1B + training pairs of sentences	It retains more than 99% accuracy on SQuAD 2.0 and several GLUE benchmark tasks, using 50% of the Transformer parameters and computations	Perform better than BERT-BASE, DistillBERT, and TinyBERT on SQuAD2 and SST-2 benchmarks

Table 1 (continued)

#	Model	Description	Training method	Training data	Performance on sentence embedding	Performance on text similarity tasks
4	all-mpnet-base	Inherits BERT and eXtreme Language Understanding Network (XLNet) leveraging the full position information of the sentence	Supervised & unsupervised	1B + training pairs of sentences	Perform better than BERT and XLNET which suffer from position discrepancy	Perform better than BERT, XLNet, RoBERTa on MNLI, QNLI and STS benchmarks
5	SRoBERTa-NLILarge	Modification of the pre-trained BERT that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings	Supervised	MNLI and XNLI datasets	Perform better than averaging word embeddings such as GloVe and Avg. BERT embedding	Outperform other embedding models such as GloVe and BERT embedding, and InferSent on (STS12-STS16) benchmarks
6	ELMo	Derives vectors from bidirectional Long Short-Term Memory (LSTM) that is trained with a coupled language model, achieves context-dependent embedding	Supervised	One Billion Word Benchmark	Outperform word embeddings such as GloVe and Word2Vec	ELMo perform better than GloVe on Semantic Role Labelling (SRL) and SNLI benchmarks
7	InferSent-GloVe	Bi-directional LSTM architecture with softmax, uses word embeddings based on GloVe	Supervised	SNLI dataset	InferSent embedding models perform better than GloVe, FastText, and SkipThought embedding	Perform better than word2vec, GloVe, TF-IDF, bags of words, and SkipThought on STS benchmark
8	InferSent-FastText	Bi-directional LSTM architecture with softmax, uses word embeddings based on FastText				

InferSent with FastText. These models are summarised in Table 1, which describes the model's introduction, its architecture, how models are trained, and the dataset used for training, along with their performance on sentence embedding and text similarity tasks against other deep learning models.

Method

Participants

Twenty-five secondary education students from an urban secondary school in Lleida, Spain, participated in this study. The students were divided into five small groups, each consisting of five students. The groups were randomly assigned by the teacher to solve an open-ended scientific challenge about identifying the causes of pollution in the Segre River, located in Lleida, Spain. Among the participants, 60% were female, while 40% were male, and their average age ranged between 15–16. The study received approval from the ethical committee of the university.

Study Procedure

This study adopts a case study research design utilising a quantitative approach, allowing for an in-depth investigation of the phenomenon within a real classroom context (Thomas, 2021).

The study procedure is illustrated in Fig. 1. Students were organised in small groups (Fig. 1- a. students' groups) to collaboratively address an open-ended scientific challenge: how to reduce the pollution of the Segre River located in Lleida, Spain. Each small group dedicated 18 h to their cocreative inquiry to solve the scientific challenge. Following the cocreative process (Fig. 1- b. Creative process), embedded in the Viacocrea technology, each group of students had to solve a scientific challenge that was technology-supported by using the Viacocrea application. The Viacocrea prototype offers a multi-user collaborative platform, a graphic representation of cocreative phases, and creative techniques aimed at structuring, supporting, enriching, and orchestrating small group cocreative problem-solving endeavours (Pifarré, 2023) .

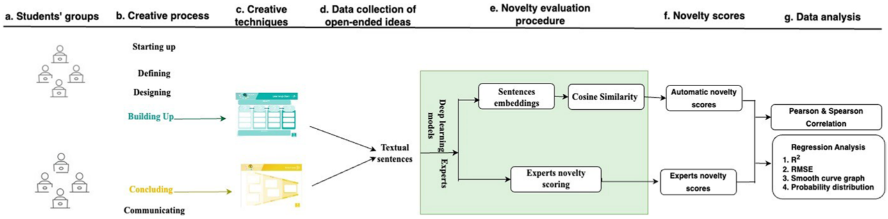
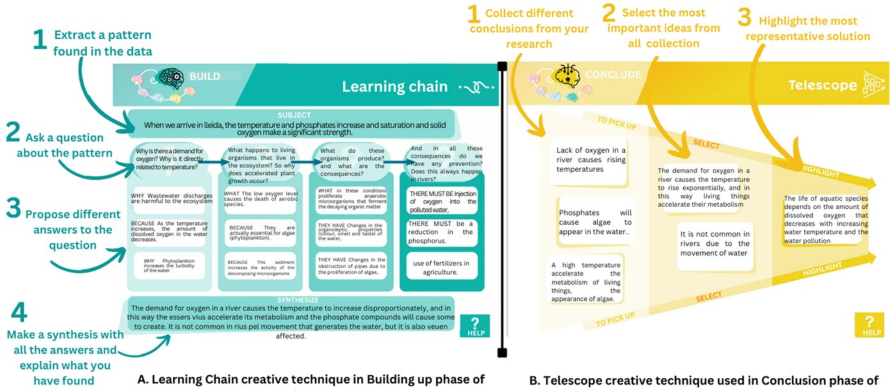


Fig. 1 The procedure followed in this study. Evaluating the novelty of cocreative ideas in the two Viacocrea techniques: Learning Chain and Telescope

In our study, each small group followed the six Viacocrea phases (Fig. 1- b. Creative process): 1) Starting up (creating a group and gaining inspiration from various resources); 2) Defining (defining and understanding the problem); 3) Designing (designing an action plan to solve the problem); 4) Building Up (building up new knowledge, its organisation, and analysis); 5) Concluding (highlighting the relevant solutions); 6) Communicating (sharing their creative solutions with others). In each phase, the small groups performed two digital creative techniques (Fig. 1 – c. creative techniques) selected by the teacher to solve the open-ended challenge, resulting in a total of 12 techniques. These creative techniques were implemented within a shared multi-user collaborative digital space in which all small-group students were engaged in face-to-face interactions and small-group technology-based interactions to cocreatively generate ideas (all small-group members can annotate in Viacocrea). Notably, there was no interaction between groups, only within each group. Figure 2(A, B) displays two of these techniques. The dataset for this study comprises the ideas annotated by all participants within these two creative techniques.

Figure 2 illustrates the design of each creative technique, visually guiding small group thinking towards a specific creative objective. Each technique has different sections with specific objectives. The first creative technique used in this study called the Learning Chain (Fig. 2A), is a creative technique used in the Building-up phase of the creative process. This technique aims to find a novel and comprehensive explanation for a problem and effectively communicate it through a series of interconnected questions and answers. Figure A highlights four sections within this technique: 1) Subject (the idea that needs to be developed and elaborated); 2) Questions asked (Participants shared different queries about the subject to think and go deeper); 3) Novel ideas to answer the specific question; and 4) Conclusion.

The second creative technique, the Telescope (Fig. 2B), is used in the Conclusion phase of the creative process because it aims to draw conclusions from the creative problem-solving process. The Telescope is a technique to be applied when they have different ideas about a subject and need to select the most relevant to narrow down


1 Extract a pattern found in the data
2 Ask a question about the pattern
3 Propose different answers to the question
4 Make a synthesis with all the answers and explain what you have found
BUILD
Learning chain
SUBJECT
SYNTHESIZE
? HELP
1 Collect different conclusions from your research
2 Select the most important ideas from all collection
3 Highlight the most representative solution
CONCLUDE
Telescope
TO PICK UP
SELECT
HIGHLIGHT
HIGHLIGHT
SELECT
TO PICK UP
? HELP

When we arrive in lleida, the temperature and phosphates increase and saturation and solid oxygen make a significant strength.

Why is there a demand for oxygen? Why is it directly related to temperature?

WHY Wastewater discharges are harmful to the ecosystem

BECAUSE As the temperature increases, the amount of dissolved oxygen in the water decreases.

WHY Phytoplankton increases the turbidity of the water

What happens to living organisms that live in the ecosystem? So why does accelerated plant growth occur?

WHAT The low oxygen level causes the death of aerobic species.

BECAUSE They are actually essential for algae (phytoplankton).

BECAUSE This sediment increases the activity of the decomposing microorganisms

What do these organisms produce? and what are the consequences?

WHAT in these conditions proliferate anaerobic microorganisms that ferment the decaying organic matter

THEY HAVE Changes in the organoleptic properties (colour, smell and taste) of the water,

THEY HAVE Changes in the obstruction of pipes due to the proliferation of algae,

And in all these consequences do we have any prevention? Does this always happen in rivers?

THERE MUST BE Injection of oxygen into the polluted water,

THERE MUST be a reduction in the phosphorus.

use of fertilizers in agriculture.

The demand for oxygen in a river causes the temperature to increase disproportionately, and in this way the essers vius accelerate its metabolism and the phosphate compounds will cause some to create. It is not common in rius pel movement that generates the water, but it is also veuen affected.

Lack of oxygen in a river causes rising temperatures

Phosphates will cause algae to appear in the water..

A high temperature accelerate the metabolism of living things, the appearance of algae.

The demand for oxygen in a river causes the temperature to rise exponentially, and in this way living things accelerate their metabolism

It is not common in rivers due to the movement of water

The life of aquatic species depends on the amount of dissolved oxygen that decreases with increasing water temperature and the water pollution

A. Learning Chain creative technique in Building up phase of creative process

B. Telescope creative technique used in Conclusion phase of creative process

Fig. 2 Example of the Learning Chain and Telescope techniques used in the Viacocrea application and for novelty evaluation were already completed by one small group

the number of options. Telescope supports the small-group process of synthesising and selecting the best ideas to solve a problem as it helps to elaborate, justify and communicate the idea' selection. The Telescope design contains three objectives or sections (as highlighted in Fig. 2B): pick up, select, and highlight.

In this paper, we analyse all the ideas generated by the small groups in these two techniques. Table 2 presents the number of ideas introduced in each section of the creative techniques and the number of similarity comparisons made within each section (Fig. 1 – d. Data collection of open-ended ideas). To illustrate, we consider the technique "Learning Chain—section subjects". In this section, the participants introduced a total of 8 ideas. By comparing each idea with the remaining seven, we obtained 36 similarity scores, excluding duplicate scores, as indicated in the first row of Table 2. The same process was repeated for the different sections (specified in the second column of Table 2) of the two techniques (mentioned in the first column of Table 2). The final dataset for this study consisted of a total of 62 ideas, resulting in 588 similar comparisons, as shown in the last row of Table 2. These similarity comparisons were used to calculate the novelty score (Fig. 1- e.Novelty evaluation procedure), representing ideas' semantic similarity, as evaluated by experts and sentence embedding models.

Regarding the experts' novelty scoring, a team of three postgraduate experts with experience in creative projects and evaluating creative work using the Torrance Test evaluated the novelty of ideas in the 588 comparisons. The evaluation process followed two key criteria: 1) assessing the general meaning of the two ideas compared in each similarity comparison and 2) considering the use of key concepts, key topics, and details of the ideas (as details can influence the meaning). Considering these two characteristics, firstly, each expert individually assigned a score to each idea according to a similarity scale ranging from 0 (completely dissimilar) to 1 (completely similar). Secondly, the three experts convened for a meeting where they reviewed the scores, discussed them, and reached a consensus on the experts' scores for each idea. Lastly, we divided the novelty scoring scale from 0 to 1 into three categories: *high* similarity scores (0.7, 0.8, 0.9), *medium similarity* scores (0.4, 0.5, 0.6), and *low similarity* scores (0, 0.2, 0.2,

Table 2 The number of ideas examined in various sections of the creative techniques used in our study, along with the comparisons between these ideas

Creative techniques	Section of creative technique	No. of ideas written by the 5 small-groups	Number of the different similarity comparisons
Learning chain	Subject	8	36
	Question asked	20	210
	Conclusion	11	66
Telescopi	Ideas in collect and select sections	23	276
Total		62	588

The comparison encompassed different aspects, such as comparing the subject section to other subjects from other groups, comparing questions asked with other questions asked, and so forth

0.3). This categorisation aligns with previous research that has characterised novelty into low, medium, and high categories (Birkey & Hausserman, 2019; Chowdhury et al., 2022; Jagtap, 2019).

Regarding the automatic novelty evaluation of ideas, we employed eight off-the-shelf pre-trained sentence embedding models in our open-ended cocreation context using the Viacocrea application. These models were implemented in the Python programming language, utilising the computational resources and environment provided by Google Colab (Bisong & Bisong, 2019). To compute novelty, the sentence embedding models were implemented using deep-learning libraries, e.g., Hugging Face (Jain, 2022). Our Python program implementation takes each cocreative idea as input, resulting from the creative techniques described in section "Study Procedure" *Study procedure*. Furthermore, the Python program performs two primary computations, as illustrated in Fig. 3. First, it conducts sentence-level vectorisation in a high-dimensional mathematical vector space using in-built pooling techniques, as highlighted in Fig. 2. Second, it calculates the distance between sentence vectors by employing cosine similarity functions. The cosine measure similarity scores range from 0 (completely dissimilar) to 1 (completely similar). Like expert scoring, the cosine similarity scores were categorised into three categories, e.g., *high* similarity scores (0.7, 0.8, 0.9, 1), *medium* similarity scores (0.4, 0.5, 0.6), and *low* similarity scores (0, 0.2, 0.2, 0.3).

To answer the research objectives of this study, we analysed and compared the novelty evaluation scores (illustrated in Fig. 1 – f. Novelty scores) obtained from experts and sentence embedding models using statistical methods described in the following section "Data Analysis".

Data Analysis

The data analysis (Fig. 1 – g. Data Analysis) conducted in this study compares the novelty scores obtained from the eight-sentence embedding models with the agreed-upon experts' scores of cocreative ideas generated using the Viacocrea prototype in an open-ended science project. For this comparison, we used three types of correlation:

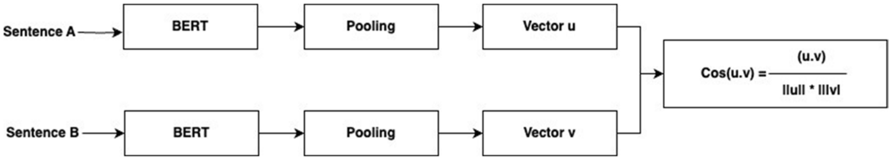


Fig. 3 The process of sentence vectorisation and the mathematical expression of cosine similarity used to determine the distance between sentence vectors are illustrated

1. Firstly, we calculated Pearson and Spearman correlation analyses among the novelty scores provided by the three experts.
2. Secondly, we adopted Pearson and Spearman correlation analyses using SPSS, which is applied in similar studies to measure the correlation between experts and automatic scores in creativity research (LaVoie et al., 2020).
3. Thirdly, we also conducted regression analysis using JMP software (Beaty & Johnson, 2021) to evaluate the correlation between the experts' scores and the scores generated by the sentence embedding models. This analysis provided four measures, namely: 1) R-squared (R^2) correlation; 2) Root Mean Squared Error (RMSE); 3) scatter graph with a smooth curve; and 4) probabilities of values in three categories. These measures provided a more detailed correlation analysis of the sentence embedding models in relation to the agreed-upon experts' scores.

Findings

The findings of the study are presented according to our two research objectives.

Novelty Results from Sentence Embedding Models

The complex nature of PBL through the cocreative process poses a challenge for NLP techniques. Recently, sentence embedding models have emerged as a powerful tool for solving open-ended problems. We applied eight-sentence embedding models without requiring feature extraction, pattern identification, or fine-tuning. This was possible because the ideas in our study were very few, open-ended, unstructured, and varied in length, making traditional approaches impractical. These models allowed us to obtain a list of ideas as input and produce output as a fixed-length vector representation for the entire idea.

In order to evaluate the novelty of cocreative ideas, pre-trained sentence embedding models performed two computations. Firstly, they transformed the open-ended textual ideas of different lengths into numerical vector space. Secondly, they computed the cosine similarity between the embedding vectors of an idea and all other ideas to determine their similarity. Showing all the results of cosine similarity scores from sentence embedding models would be difficult and overwhelming because of the high number of similarity scores (five hundred eighty-eight). Therefore, we have chosen an example of the results obtained; specifically, we present the results of a subset of eight ideas which students shared in the Subject section (1) of the Learning Chain creativity technique (depicted in Fig. 2A). Table 3 presents these eight ideas written by the different small groups in the technique section. Next, Fig. 4 displays the 36 similarity scores obtained for these eight ideas using each of the eight-sentence embedding models.

The subset of eight ideas mentioned above resulted in 36 similarity comparisons, so their corresponding scores are presented in Fig. 4, along with heatmaps with each of the eight embedding models. Taking a closer look at Fig. 4A, we observe the heatmap and cosine similarity scores of the 36 comparisons among the subset ideas,

Table 3 A subset of 8 cocreative ideas was written by the different small groups in the Subject section (1) of the Learning Chain creativity technique, as highlighted in Fig. 2A

Idea #	Idea examples
1	When we arrive in Lleida, the temperature and phosphates increase, and saturation and solid oxygen make a significant strength
2	In both rivers the hydrological quality is not lower than a score of 11/15, it is very high in both cases
3	As the water temperature rises, less oxygen dissolves
4	The level of water pollution affects the species that live there
5	Phosphates are found in fertilisers and detergents and can reach the water with agricultural runoff, industrial waste, and sewage discharges
6	Nitrate is one of the most common groundwater contaminants in rural areas. It should be controlled in drinking water-primarily because excessive levels can cause methaemoglobinaemia, or "blue baby disease”
7	Temperature, conductivity, pH, nitrates
8	The concentration of nitrates increases when the river reaches an urban area, in this case Lleida

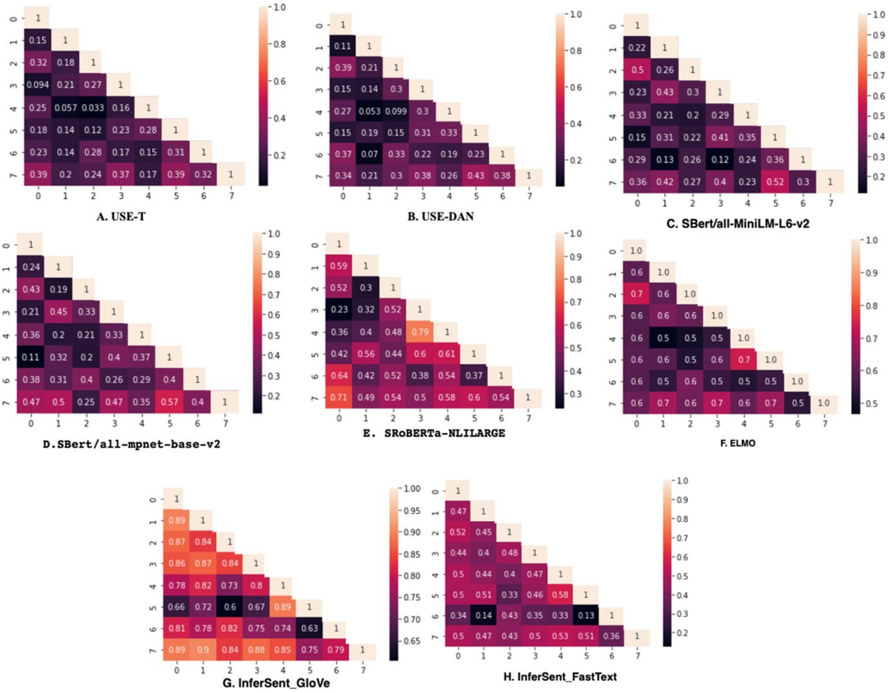


Fig. 4 Heatmap and cosine similarity scores from comparisons of a subset of eight ideas computed by sentence embedding models

which provide similarity scores ranging from 0 to 1. Specifically, we could focus on comparing ideas 0 and 1, which yielded a similarity score of 0.15. This indicates a low level of similarity between these two ideas. Considering the other values in the first column of Fig. 4 (corresponding to model A), we find that the comparison of idea 0 with the remaining ideas exhibits low similarity because all the values in column 1 are less than 0.4. However, we can observe discrepancies in the results of the eight embedding models when comparing any two ideas. This can be seen, for example, in the comparison between ideas 0 and 1, where inconsistencies arise. In Fig. 4, model A scores the similarity between ideas 0 and 1 as low (0.15), as do models B (0.11), C (0.22), and D (0.24). On the other hand, three models, E (0.59), F (0.6), and H (0.47), classify the similarity between these ideas as a medium. Contrastingly, model G yields a high similarity score of 0.89. These results show that eight-sentence embedding models display different similarity scores. Therefore, there is a need to validate with experts' novelty scores to reach the second research objective.

Evaluation of Sentence Embedding Models to Assess the Novelty of Open-ended Ideas

Previous research has assessed the reliability of automatic novelty scores by applying correlation analyses to the experts' scores (LaVoie et al., 2020). In line with this, in our study, the three experts individually assigned novelty scores to each idea. We found significantly high Pearson and Spearman correlations among the experts' scores, as displayed in Table 4.

In order to reach an agreement, the experts revisited the ideas and based on discussion, gave an agreed-upon score to the ideas. Subsequently, we computed Pearson and Spearman's correlations between each sentence embedding model with individual and agreed experts' scores. The results, presented in Table 5, show that the correlation of the sentence embedding model is higher with the "agreed experts' score" (displayed in the first row in Table 5) than with the individual expert scores. Specifically, when considering the correlation of the sentence embedding models with the agreed experts' scores, we found that USE-T displayed a high correlation with the expert's scores ($r=0.860$, Spearman = 0.784), followed by USE-DAN ($r=0.827$, Spearman, 0.728), all-MiniLM-L6 ($r=0.839$, Spearman = 0.753) and all-mpnet-base ($r=0.804$, Spearman = 0.713).

Table 4 Correlation among the experts' novelty scores

	Categories scores	
Comparison	Pearson	Spearman
Expert 1-Expert 2	0.879**	0.833**
Expert 1- Expert 3	0.880**	0.796**
Expert 2- Expert 3	0.856**	0.784**

Correlation is significant at the 0.01 level (2-tailed)$_{**}$

Table 5 Pearson and Spearman's correlations of deep learning models with experts' scores

Comparison	Correlation	USE-T	USE-DAN	All-MiniLM-L6	all-mpnet-base	SRoBERTa-NLI Large	ELMo	InferSent_Glove	InferSent-FastText
Agreed experts scores	Pearson (r)	0.860**	0.827**	0.839**	0.804**	0.606**	0.565**	0.106**	0.627**
	Spearman	0.784**	0.728**	0.753**	0.713**	0.608**	0.565**	0.108**	0.570**
Expert1	Pearson	0.851**	0.810**	0.827**	0.789**	0.576*	0.538**	0.079**	0.614**
	Spearman	0.769**	0.715**	0.737**	0.691**	0.576**	0.521**	0.101**	0.554**
Expert 2	Pearson	0.851**	0.825**	0.817**	0.791**	0.592**	0.538**	0.062**	0.633**
	Spearman	0.769**	0.720**	0.718**	0.703**	0.581**	0.541**	0.075**	0.575**
Expert 3	Pearson	0.839**	0.795**	0.813**	0.782**	0.577**	0.528**	0.055**	0.616**
	Spearman	0.717**	0.667**	0.709**	0.669**	0.588**	0.539**	0.73**	0.569**

Correlation is significant at the 0.01 level (2-tailed)$_{**}$

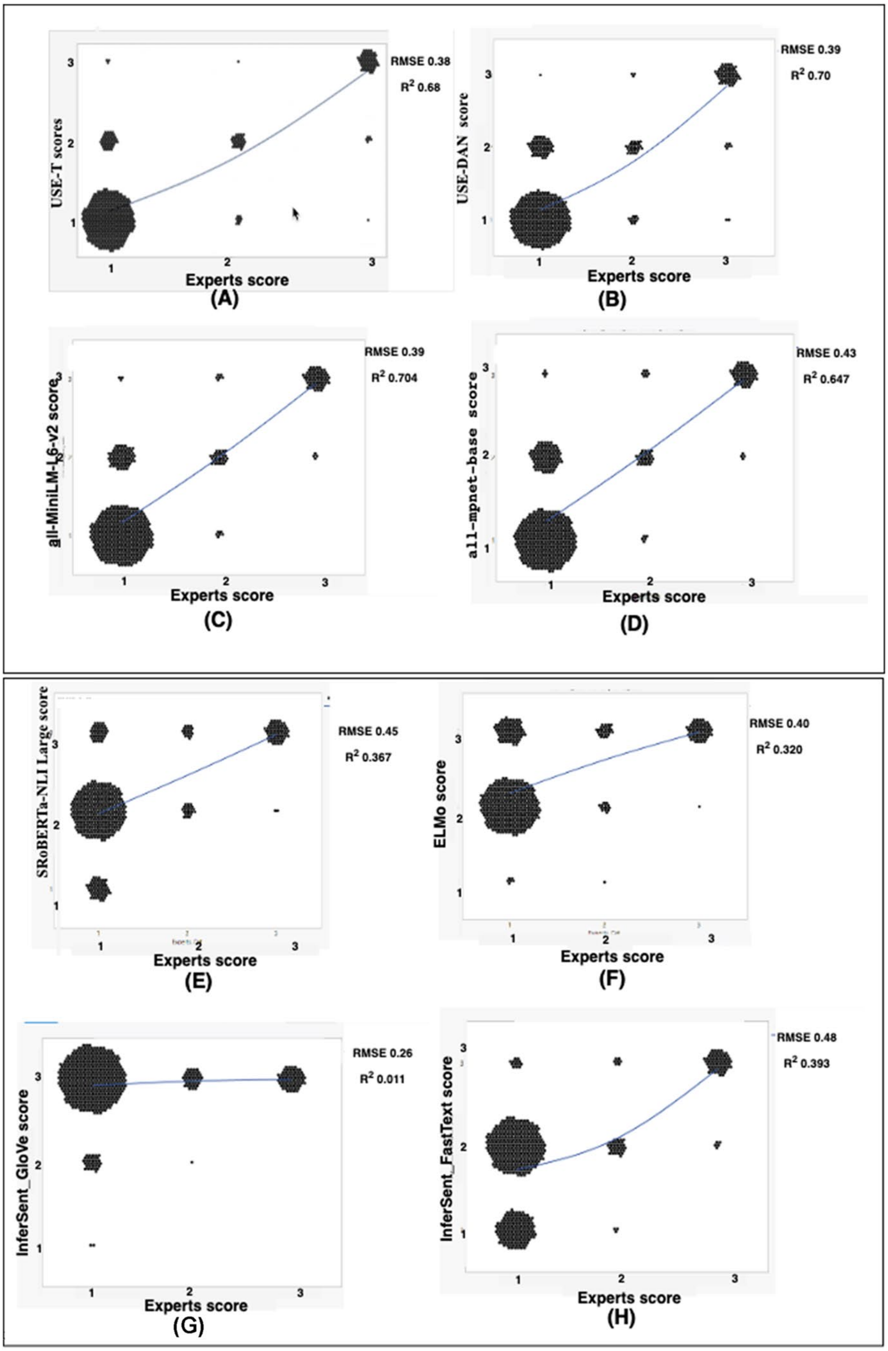


Fig. 5 Scatter graph with RMSE and R2 values between the agreed experts' scores on the x-axis and each sentence embedding model score on the y-axis. The values on the x-axis and y-axis (1, 2, and 3) show the three categories of novelty

Furthermore, regression analysis was conducted correlation (R^2 and RMSE analysis) and a scatter graph with a smooth curve to visualise the correlation between the agreed expert scores and the sentence embedding models. The results of the R^2 and RMSE analyses confirmed that USE-T and USE-DAN resulted in the highest correlation, as evidenced in the curve graph in Fig. 5 (A, B), with R^2 values of 0.684 and 0.704 and RMSE values of 0.38 and 0.39, respectively. Following these, all-MiniLM-L6 showed a correlation of RMSE 0.39 and R^2 0.704, while all-mpnet-base exhibited an RMSE of 0.43 and an R^2 of 0.643, as shown in Fig. 5 (C, D). However, the remaining four models in Fig. 5 (E, F, G, and H) had low R^2 and RMSE values.

In Fig. 5, curve graphs for the eight embedding models are presented, where the X-axis represents the expert scores and the Y-axis represents the scores assigned by the sentence embedding models, while 1, 2, and 3 show the three categories (low, medium, and high novelty). As displayed in Fig. 5, all the models positively correlate with the agreed experts' scores because all the curved lines are in an upward direction. However, the decree of correlation varies among the models, with some demonstrating high correlation and others displaying low correlation. For models with high correlation, as depicted in Fig. 5 (A, B, C, D), the curve line intersects or bends towards category 1 in case of category 1, the curve line intersects or bends towards values in category 2 in case of category 2, and similarly, in the case of category 3, curves intersect with values in category three or bend towards category three. On the contrary, the curve line corresponding to Fig. 5 (E, F, G, H) shows a low correlation between the experts' scores and the sentence embedding models.

Finally, we performed a probability distribution analysis of values to assess the agreement between the sentence embedding models and the experts' scores across the three categories: low, medium, and high. The results are presented in the first column of Table 6. Regarding ***category 1*** (first row in Table 6), the probability ratio for the experts' scores is 80%. The models that have a probability distribution score closer to the experts are USE-T (74%), USE-DAN (71%), all-MiniLM-L6 (69%), and all-mpnet-base (64%). On the contrary, the models with different probability distribution scores from the experts are: SRoBERTa-NLILarge (10%), ELMo (1%), InferSent_Glove (25%), and FastText (0.3%). Regarding ***category 2*** (second row in Table 6), the probability ratio for the experts' scores is 7%. The models that show a closer alignment with the experts are USE-T (14%), USE-DAN (16%), all-MiniLM-L6 (18%), all-mpnet-base (22%), and InferSent_FastText (5%). Although InferSent_FastText performs better in category 2, its low performance can be attributed to the majority of values belonging to category 1. Regarding ***category 3*** (third row in Table 6), the probability ratio for the experts' scores is 11.9%. The models that have a probability distribution closest to that of the experts in this category are USE-T (11.4%)), USE-DAN (11.3%), all-MiniLM-L6 (12.8%), and allMPNETt-base (13%).

In conclusion, the probability distribution analysis across the three categories presented in Table 6 confirms previous results. It confirms that the four embedding models, namely USE-T, USE-DAN, all-MiniLM-L6, and all-mpnet-base, highlighted in bold italics, outperform the experts' values on our dataset.

Table 6 The values in the table show the probability distribution across the three categories

Categories	Experts	USE-T	USE-DAN	all-MiniLM-L6	all-mpnet-base	SRoBERTa-NLI Large	ELMo	InferSent_Glove	InferSent-FastText
1 (low)	***0.80822***	***0.74315***	***0.717***	***0.69007***	***0.64212***	0.10445	0.0119	0.25171	0.00342
2 (medium)	***0.07192***	***0.14212***	***0.169***	***0.18151***	***0.22603***	0.67295	0.6695	0.60274	***0.05822***
3 (high)	***0.11986***	***0.11473***	***0.113***	***0.12842***	***0.13185***	0.22260	0.3184	0.14555	0.93836

Discussion

Our first research goal is the application of deep learning sentence embedding models for the assessment of the novelty of ideas generated by a group of students during an open-ended project activity. The information provided by AI-driven techniques for creativity assessment could be valuable for teachers and students to improve and regulate the cocreation actions during the resolution of the project. However, evaluating novelty in PBL is challenging because the ideas generated are limited in number and non-domain-specific, and their presentations have meaningful sentence structures. Here, we discuss using sentence embedding models to address these challenges.

Previous studies in automatic creativity evaluation have employed statistical (LaVoie et al., 2020) and word embedding techniques (Johnson & Hass, 2022) to evaluate novelty. These computational techniques, such as the LSA and GloVe model, are useful for single-word creativity tasks (Beaty & Johnson, 2021). However, these techniques have limitations when evaluating the novelty of open-ended cocreative ideas because they fail to capture the syntactic, semantic, and contextual similarities among sentence constituents. Our study shows that these limitations can be overcome by using sentence embedding models.

The open-ended ideas obtained in our PBL context have the following four characteristics: (a) ideas are presented in sentence structures; (b) ideas are generic and derived from a small dataset; (c) sentence embedding outperforms word embedding in textual similarity tasks; and (d) the objectives of creativity techniques are integral to the cocreative process. Next, we discuss our first research question by showing how sentence embedding models can be used to evaluate the novelty of ideas with the aforementioned four characteristics.

Firstly, during cocreative ideation, the ideas are expressed in sentences of different lengths (examples are presented in Table 3). While for single-word tasks, statistical (e.g., LSA) and word embedding (e.g., GloVe) can be useful in PBL cocreation, most of the ideas or thoughts are in a sentence structure. Also, the meaning of text only becomes clear at the sentence level because it represents the semantic and contextual relationships among the words in the sentence. Therefore, sentence embedding models allow access to fine-grained semantics while preserving the meaning and context of sentence constituents.

Secondly, ideas generated for solving open-ended problems are diverse because they can span different domains (e.g., science, linguistics, education, etc.). Given the limited dataset available (i.e., 588 comparisons of ideas in this study), training deep learning models is unfeasible with this dataset. The sentence embedding models are pre-trained over a large corpus of data and then transfer the learned knowledge to other downstream NLP tasks, in our case, evaluating a small generic dataset of cocreative ideas. We applied eight embedding models to compute similarity scores for open-ended ideas from small-size datasets generated during cocreation in a real classroom setting. Hence, our approach is generalisable and applicable to other PBL cocreation scenarios requiring novelty evaluation of open-ended ideas.

Thirdly, we applied sentence embedding models for novelty evaluation, which include sentence vectorisation and then sentence similarity, with the data in a real classroom in a more open-ended PBL cocreation process. Sentence embedding models offer dependable and high-quality outcomes in various NLP tasks. They can help vectorise entire sentences into a vector space, effectively capturing semantic and contextual meanings. This natural mechanism facilitates the measurement of sentence similarity (Lamsiyah et al., 2021). Furthermore, these sentence embedding models perform highly on semantic textual similarity benchmarks (Cer et al., 2018; Reimers & Gurevych, 2019). Therefore, semantic embedding models provide a semantic space that forms a cognitive map, enabling one to distinguish between novel and non-novel solutions by mapping semantically similar ideas closer to each other, while ideas with different meanings are mapped farther apart.

Lastly, our study confirms the applicability of sentence embedding models in evaluating open-ended ideas' novelty generated during the different iterative stages of a cocreative process. The design of educational interventions that help students' iterative creative process over different phases (six phases in the Viacocrea application) and creative technique demands that sentence embedding models be enabled to measure the novelty of cocreative ideas in different phases or different sections of creative techniques. In our view, this information could be used by teachers and students to reflect and explore different directions to solve the project creatively (Kenworthy et al., 2023). Therefore, we applied sentence embedding at different phases of the creative process, for example, Building and Conclusion phases. Additionally, the design of different creative techniques demands comparing different sections of creative techniques, such as evaluating novelty in the Subject section of the creative technique Learning (Fig. 2A). To sum up, our case study confirms that pre-trained sentence embedding models can effectively evaluate the novelty of open-ended, generic, small datasets of cocreative ideas in PBL through the cocreative process described in Research context 2.1.

The second research goal of this study was to provide some data concerning the selection of a reliable sentence embedding model for novelty evaluation in a PBL context. To reach this objective, we compared eight embedding models with experts' evaluations of novelty in open-ended cocreative ideas. Our results show that most of these pre-trained sentence models have already been validated on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019) and have been applied to several NLP tasks. However, to the best of our knowledge, this study is the first to empirically analyse eight-sentence embedding models on the novelty evaluation of open-ended cocreative ideas using classroom data.

Our analysis shows that four of the eight embedding models had a significantly high correlation with experts' evaluations. These models include USE-T, USE-DAN, all-MiniLM-L6, and all-mpnet-base. The significant correlation of these four models can be due to three reasons: 1) they are trained by using a combination of unsupervised training (training on the unlabelled dataset) and fine-tuning on supervisor datasets (labelled dataset) such as the SNLI dataset. The inclusion of SNLI training is crucial for obtaining higher quality sentences; 2) the nature of the dataset on which these models are trained is useful for our context. These four models are trained over a large amount of data from a variety of sources such as Wikipedia, SNLI, web news, web Questions/Answers and book corpus that might be beneficial for our dataset; 3) the variation in results comes from these models' different abilities and architectures. The different abilities and architectures of pre-trained embedding models are necessary for learning contextual representations of words and the semantic vectorisation of whole sentences. The transformer, deep averaging network and fine-tuned BERT on SNLI with softmax architectures have shown better results.

Furthermore, we examine the three factors mentioned above that influence the performance of the remaining four models: SRoBERTa-NLI Large, ELMo, InferSent_Glove, and InferSent-FastText. 1) These four models use supervised training in contrast to models trained in supervised and unsupervised ways, contributing to performance differences. 2) The performance of these models is influenced by the datasets on which they were trained. For example, SRoBERTa-NLI Large was trained on datasets like MNLI and XNLI, ELMo on the One Billion Word Benchmark, and InferSent-GloVe and InferSent-FastText were trained on datasets categorising sentences as "entailment," "contradiction," or "neutral." Although these models are used for semantic similarity in sentences, variations in performance may be due to the nature of the datasets, which differ from our dataset. 3) The variation in results may stem from the diverse abilities and architectures of these models, being not appropriate for novelty evaluation. For instance, SRoBERTa-NLI Large, originally designed to discern entailment or contradiction between sentences, and ELMo, focused on predicting the next word in a sentence, have primary objectives different to novelty evaluation.

Finally, in comparing these four highly correlated models with experts' evaluations, we found that USE-T and USE-DAN models outperformed all-MiniLM-L6 and all-mpnet-base. One of the reasons is that USE-T and USE-DAN models are trained to embed sentences from scratch, whereas SBERT fine-tunes a trained BERT model to optimise and predict masked words and the next sentences. Also, USE-T performs better than USE-DAN because the transformer architecture targets high accuracy, while DAN targets efficiency in time and memory usage at a slightly reduced accuracy (Cer et al., 2018). Hence, we confirmed with real classroom data that USE-T, followed by USE-DAN, can be efficient tools to evaluate ideas automatically. This evaluation opens the possibility of implementing real-time scoring, which can be used to design and deliver feedback with an eye on supporting teaching and learning. We can envisage a future where AI applications in creativity optimise the generation of valuable and novel solutions to shared challenges in a cocreative process.

The findings of our study affirm the technological robustness and reliability of specific AI-sentence embedding models, demonstrating their effectiveness in automatic novelty evaluation. Moreover, the novelty scores generated by these AI techniques accurately portray the level of novelty and the evolution of novel idea generation during iterative cocreative processes. A well-designed educational setting has the potential to integrate these AI techniques offering valuable feedback to teachers and, more importantly, to student groups engaged in creative problem-solving. This feedback acts as a mediator, facilitating communication among students about their problem resolution and serving as the catalyst for fruitful group creative-thinking mechanisms (Pifarré, 2019) .

Additionally, the novelty scores can prompt actions, providing real-time feedback for low, medium, and high creative ideas (Birkey & Hausserman, 2019; Hassan et al., 2019). For instance, translating novelty scores from a pool of ideas co-generated by student groups into actionable insights can offer tailored guidance across three distinct categories: high novelty, medium novelty, and low novelty ideas. In each category, students can receive specific feedback to enhance their communication and orchestrate their cocreation processes, thereby increasing the likelihood of co-generating better and more novel ideas. The provision of real-time novelty feedback addresses a crucial pedagogical concern in promoting creativity, creative thinking (Ndolo, 2021), problem-solving skills (Chevalier et al., 2022), and social skills (Sun et al., 2022) in real classrooms.

Conclusion

This study primarily aims to contribute to the automatic novelty evaluation of open-ended ideas generated within real-life project-based classrooms. PBL environments are among the most used methodologies to promote cocreativity in real-life settings. For teachers to effectively facilitate students' creative processes, they need real-time information on how students are generating novel ideas, be able to provide feedback and help group creative processes. Educational research highlights the importance of providing instructional guidance during creative teaching (Sawyer, 2022). To meet this objective, we examined and compared with experts the evaluation capabilities of eight pre-trained sentence embedding models in evaluating the novelty of ideas generated by groups of students during cocreative problem-solving.

As a first contribution, our study shows that deep learning sentence embedding models can be used for the novelty evaluation of ideas in a cocreative ideation process in the real classroom context. We applied eight pre-trained sentence embedding models, which showed that they have the potential to evaluate generic, open-ended and small datasets during the cocreative process. We implemented the sentence embedding models to our real-classroom data and found cosine similarity among the sentence vectors as novelty scores. This contribution adds to the existing literature, as sentence embedding models provide a robust measure in line with the semantic theory in creativity research (Li et al., 2023), offering consistent ways to understand how people are involved in thinking and learning. Moreover, the utilisation of sentence embedding models represents progress in strengthening the statistical (Acar et al., 2021), word embeddings (Buczak et al., 2023; Organisciak et al., 2023), and standard subjective scoring methods for assessing the creative process and its output (Kenett, 2019). Furthermore, our study revealed that sentence embedding models effectively evaluate the novelty of generic, open-ended, and small datasets in PBL during the cocreative process.

As a second contribution, we conducted experiments to evaluate the reliability of sentence embedding models for automatic novelty evaluation in the context of open-ended cocreative ideas. While previous literature has applied sentence embedding models to evaluate novelty (Kenworthy et al., 2023), to the best of our knowledge, our work is among the first studies to test the reliability of sentence embedding models in evaluating novelty in open-ending ideas generated during real-classroom cocreative problem-solving. Our results confirmed that sentence embedding models yielded significant correlations with experts' novelty scores. USE-T exhibited the highest correlation among the models tested, followed by USE-DAN, demonstrating their efficacy in evaluating open-ended, short ideas and small datasets even without further fine-tuning. This contribution adds to the literature that the evaluation evidence supports the reliability of these two deep learning models, which can be applied in PBL and creative thinking tasks in a real-time classroom context.

The results of our study hold significant ecological and pedagogical value because it effectively evaluates novelty generated in a real problem-based environment that uses technology to promote key cocreative processes. The real-time evaluation conducted in this context can have a strong pedagogical impact because it can support giving feedback to teachers and students during the cocreation. This feedback can, in turn, promote teaching and learning methodologies that optimise the collaborative ideation process, thereby enhancing the overall educational experience. Moreover, the real-time novelty scores of co-generated ideas could be embedded in a more ambitious technology that could provide tailored feedback to improve, orchestrate and regulate the group cocreation actions.

Therefore, our study has provided experimental evidence regarding using specific sentence embedding models during the cocreative process to evaluate the novelty of open-ended ideas generated within project-based learning contexts. We see novelty evaluation as the first step that could allow researchers and educators to design and deliver adjusted and contextualised feedback to teachers and students, capable of shaping and enhancing the whole cocreative process. This approach aligns with previous studies supporting the notion that feedback improves students' engagement (Karaoglan Yilmaz & Yilmaz, 2022; Hobscheid & Kerbavaz, 2022), fosters higher levels of intrinsic motivation (Su et al., 2022; Wu, 2023), enhances meta-cognitive awareness, academic achievement, stimulates idea generation, and improves the quality of ideas (Karaoglan Yilmaz & Yilmaz, 2022) in the cocreative process.

Our study carries significant practical implications for real classroom settings. Creative problem solving, especially through the cocreative process in PBL, is gaining prominence in educational settings. The effective utilisation of an AI-driven approach to evaluate novel solutions and offer tailored feedback for enhancing students' creative and problem-solving skills can serve as a valuable tool. This approach helps equip future citizens with the essential competencies to generate innovative solutions to the world's complex economic, environmental, and social challenges.

Limitations and Future Work

From our perspective, this study encompasses three limitations that should be considered for future research directions.

First and foremost, this study focused exclusively on the dimension of novelty. To gain a more comprehensive understanding of creativity, future research should incorporate real-time assessments of the other three core dimensions, namely flexibility, elaboration, and fluency, throughout the creative process. Moreover, incorporating a combination of various creativity dimensions can prove valuable. For instance, Kenworthy et al. (2023) conducted research in which they designed a computational model that combines novelty, elaboration, and relevance scores to determine the overall quality of ideas. This approach highlights the potential benefits of integrating multiple dimensions in assessing creativity.

Secondly, our study's dataset is limited in size because the Viacocrea application generates data from a real classroom setting aimed at solving a specific science project related to local issues concerning river pollution. Notwithstanding this limitation, we believe that the small-sized dataset generated within a real context holds value in addressing our research objectives and allowing us to draw pedagogical conclusions. Besides, the utilisation of correlation analysis, which employs statistical measures such as the correlation coefficient, enables quantifying the relationship between the two variables, even when working with a relatively small dataset (Chok, 2010; Temizhan et al., 2021). This approach remains valuable in achieving our research objectives.

Thirdly, our dataset is imbalanced, with 80% of the data falling into the low category, 11% in the high category, and the remaining percentage in the medium category. It is important to note that real-world, open-ended problem-solving processes are often skewed, where certain categories may be more prevalent than others. Despite this imbalance, our data holds the linearity assumption in the scatter graph and exhibits meaningful differences among the categories, thereby yielding valid results. Consequently, our study provides preliminary but significant contributions and paves the way for new areas in evaluating novelty within open-ended cocreation.

Fourthly, we conducted correlation and regression analyses, which have been used in previous research to assess the reliability of computational techniques compared to experts' scores. Nevertheless, in addition to these standard analyses, there is potential for further mathematical validation methods that can confirm findings, thereby paving the way for future studies.

Finally, in future research, we aim to evaluate the real-time novelty of cocreative ideas. We intend to use this evaluation data to provide creative novelty feedback that can serve as guidance for students and teachers, producing novel solutions and promoting cocreativity. This will be done in the context of the Viacocrea application, as described in *Research Context 2.1*, thereby introducing new features to enhance this promising cocreation platform.

Acknowledgements This research has been funded by the Ministry of Science and Innovation of the Government of Spain under Grant PID2022-139060OB-I00.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The datasets used and analysed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics Approval All authors declare that ethical standards have complied with the approval of the study by the university's Institutional Review Board.

Informed Consent All the participants have given their written informed consent.

Disclosure Statement The authors have no conflicts of interest to disclose.

References

Acar, S., Berthiaume, K., Grajzel, K., Dumas, D., Flemister, C., & Organisciak, P. (2021). Applying automated originality scoring to the verbal form of torrance tests of creative thinking. *Gifted Child Quart, 67*, 3–17. https://doi.org/10.1177/00169862211061874

Algarni, A. (2022). Evaluating co-creation in collaborative drawing using creative thinking modes *(Doctoral dissertation, The University of North Carolina at Charlotte)*.

Altinay, L., Kromidha, E., Nurmagambetova, A., Alrawadieh, Z., & Madanoglu, G. K. (2022). A social cognition perspective on entrepreneurial personality traits and intentions to start a business: Does creativity matter? *Management Decision, 60*(6), 1606–1625. https://doi.org/10.1108/MD-12-2020-1592

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Birkey, R., & Hausserman, C. (2019). Inducing creativity in accountants' task performance: The effects of background, environment, and feedback. In *Advances in accounting education: Teaching and curriculum innovations* (Vol. 22, pp. 109–133). Emerald Publishing Limited. https://doi.org/10.1108/S1085-462220190000022006

Bisong, E., & Bisong, E. (2019). Google colaboratory. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners* (59–64). https://doi.org/10.1007/978-1-4842-4470-8_7

Bozkurt Altan, E., & Tan, S. (2021). Concepts of creativity in design based learning in STEM education. *International Journal of Technology and Design Education, 31*(3), 503–529. https://doi.org/10.1007/s10798-020-09569-y

Buczak, P., Huang, H., Forthmann, B., & Doebler, P. (2023). The machines take over: A comparison of various supervised learning approaches for automated scoring of divergent thinking tasks. *The Journal of Creative Behavior, 57*(1), 17–36. https://doi.org/10.1002/jocb.559

Camburn, B, He, Y, Raviselvam, S, Luo, J, & Wood, K. (2019). *Evaluating crowdsourced design concepts with machine learning. In Proceedings of the ASME 2019 international design engineering technical conferences and computers and information in engineering conference. Volume 7: 31st International conference on design theory and methodology*. Anaheim, California, USA: ASME. https://doi.org/10.1115/DETC2019-97285

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*. https://doi.org/10.48550/ARXIV.1803.11175

Chevalier, M., Giang, C., El-Hamamsy, L., Bonnet, E., Papaspyros, V., Pellet, J. P., ... & Mondada, F. (2022). The role of feedback and guidance as intervention methods to foster computational thinking in educational robotics learning activities for primary school. *Computers & Education*, *180*, 104431. https://doi.org/10.1016/j.compedu.2022.104431

Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. University of Pittsburgh: Master dissertation.

Chowdhury, J. R., Zhuang, Y., & Wang, S. (2022). Novelty controlled paraphrase generation with retrieval augmented conditional prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10535–10544).

Corbisiero-Drakos, L., Reeder, L. K., Ricciardi, L., Zacharia, J., & Harnett, S. (2021). Arts integration and 21st century skills: A study of learners and teachers. *International Journal of Education & the Arts*, *22*(2). https://doi.org/10.26209/ijea22n2

Doboli, S., Kenworthy, J., Paulus, P., Minai, A., & Doboli, A. (2020). A cognitive inspired method for assessing novelty of short-text ideas. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206788

Dumas, D., Organisciak, P., & Doherty, M. (2021). Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts, 15*(4), 645–663. https://doi.org/10.1037/aca0000319

Emara, M., Hutchins, N. M., Grover, S., Snyder, C., & Biswas, G. (2021). Examining student regulation of collaborative, computational, problem-solving processes in open-ended learning environments. *Journal of Learning Analytics, 8*(1), 49–74. https://doi.org/10.18608/jla.2021.7230

Georgiev, G. V., & Casakin, H. (2019). Semantic measures for enhancing creativity in design education. In *Proceedings of the design society: International conference on engineering design* (Vol. 1, No. 1, pp. 369–378). Cambridge University Press. https://doi.org/10.1017/dsi.2019.40

Haatainen, O., & Aksela, M. (2021). Project-based learning in integrated science education: Active teachers' perceptions and practices. *LUMAT: International Journal on Math, Science and Technology Education, 9*(1), 149–173. https://doi.org/10.31129/LUMAT.9.1.1392

Hassan, M. A., Habiba, U., Khalid, H., Shoaib, M., & Arshad, S. (2019). An adaptive feedback system to improve student performance based on collaborative behavior. *In IEEE Access*, *7*, 107171–107178. https://doi.org/10.1109/ACCESS.2019.2931565

Hobscheid, M., & Kerbavaz, K. (2022). Flexibility is key: Co-creating a rubric for programmatic instructional assessment. *Communications in Information Literacy, 16*(1), 3. https://doi.org/10.15760/comminfolit.2022.16.1.3

Jagtap, S. (2019). Design creativity: Refined method for novelty assessment. *International Journal of Design Creativity and Innovation, 7*(1–2), 99–115. https://doi.org/10.1080/21650349.2018.1463176

Jain, S. M. (2022). Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems* (pp. 51–67). Apress. https://doi.org/10.1007/978-1-4842-8844-3_4

Jimenez-Mavillard, A., & Suarez, J. L. (2022). A computational approach for creativity assessment of culinary products: The case of elBulli. *AI & SOCIETY, 37*(1), 331–353. https://doi.org/10.1007/s00146-021-01183-3

Johnson, D. R., & Hass, R. W. (2022). Semantic context search in creative idea generation. *The Journal of Creative Behavior, 56*(3), 362–381. https://doi.org/10.1002/jocb.534

Juusola, K. (2023). Enhancing teaching and learning through the co-creative learning community approach. *Educational Action Research, 31*(1), 102–117. https://doi.org/10.1080/09650792.2023.2166090

Karampiperis, P., Koukourikos, A., & Koliopoulou, E. (2014). Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 508–512). https://doi.org/10.1109/ICALT.2014.150

Karaoglan Yilmaz, F. G., & Yilmaz, R. (2022). Learning analytics intervention improves students' engagement in online learning. *Technology, Knowledge and Learning, 27*(2), 449–460. https://doi.org/10.1007/s10758-021-09547-w

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences, 27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Kenworthy, J. B., Doboli, S., Alsayed, O., Choudhary, R., Jaed, A., Minai, A. A., & Paulus, P. B. (2023). Toward the development of a computer-assisted, real-time assessment of ideational dynamics in collaborative creative groups. *Creativity Research Journal*, 1–16. https://doi.org/10.1080/10400419.2022.2157589

Lamsiyah, S., El Mahdaouy, A., Espinasse, B., & Ouatik, S. E. A. (2021). An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications, 167*, 114152. https://doi.org/10.1016/j.eswa.2020.114152

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement, 80*(2), 399–414. https://doi.org/10.1177/0013164419860575

Li, Y., Du, Y., Xie, C., Liu, C., Yang, Y., Li, Y., & Qiu, J. (2023). A meta-analysis of the relationship between semantic distance and creative thinking. *Advances in Psychological Science, 31*(4), 519. https://doi.org/10.3724/SP.J.1042.2023.00519

Liu, Z., Lin, W., Shi, Y., & Zhao, J. (2021). A robustly optimized BERT pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics* (pp. 471–484). Springer International Publishing.

Lloyd-Cox, J., Pickering, A., & Bhattacharya, J. (2022). Evaluating creativity: How idea context and rater personality affect considerations of novelty and usefulness. *Creativity Research Journal, 34*(4), 373–390. https://doi.org/10.1080/10400419.2022.2125721

Lu, S. Y., Wu, C. L., & Huang, Y. M. (2022). Evaluation of disabled STEAM-students' education learning outcomes and creativity under the UN sustainable development goal: Project-based learning oriented STEAM curriculum with micro: Bit. *Sustainability, 14*(2), 679. https://doi.org/10.3390/su14020679

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology, 53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Ndolo, S. (2021). Effective feedback strategies that promote critical thinking skills in online learning environments: an online assessment learning perspective. *Expanding global horizons through technology enhanced language learning*, 179–190. https://doi.org/10.1007/978-981-15-7579-2_10

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(25), e2022340118. https://doi.org/10.1073/pnas.2022340118

Organisciak, P., Newman, M., Eby, D., Acar, S., & Dumas, D. (2023). How do the kids speak? Improving educational use of text mining with child-directed language models. *Information and Learning Sciences, 124*(1/2), 25–47. https://doi.org/10.1108/ILS-06-2022-0082

Ouyang, F., Wu, M., Zhang, L., Xu, W., Zheng, L., & Cukurova, M. (2023). Making strides towards AI-supported regulation of learning in collaborative knowledge construction. *Computers in Human Behavior, 142*, 107650. https://doi.org/10.1016/j.chb.2023.107650

Pifarré, M. (2023). Designing, implementing and evaluating a co-creative support technology. In EDULEARN 23 Proceedings (pp. 4364–4367). IATED.

Pifarré, M. (2019). Using interactive technologies to promote a dialogic space for creating collaboratively: A study in secondary education. *Thinking Skills and Creativity, 32*, 1–16. https://doi.org/10.1016/j.tsc.2019.01.004

Plucker, J. A., Meyer, M. S., Karami, S., & Ghahremani, M. (2023). Room to run: Using technology to move creativity into the classroom. In *Creative provocations: Speculations on the future of creativity, technology & learning* (pp. 65–80). Springer International Publishing. https://doi.org/10.1007/978-3-031-14549-05

Prasch, L., Maruhn, P., Brünn, M., & Bengler, K. (2020). Creativity assessment via novelty and usefulness (CANU)–Approach to an easy to use objective test tool. In *Proceedings of the Sixth International Conference on Design Creativity (ICDC 2020)* (pp. 019–026). https://doi.org/10.35199/ICDC.2020.03

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

Richardson, C. (2022). Supporting collaborative creativity in education with the i5 Framework. *Educational Action Research, 30*(2), 297–312. https://doi.org/10.1080/09650792.2020.1810731

Saboorizadeh, J., He, H., Burgoyne, S., Pfeiffer, F., Hunt, H., & Strobel, J. (2023). Theatre-based creativity activities for the development of entrepreneurial mindsets in engineering. In: S. Kaya-Capocci & E. Peters-Burton (Eds.), *Enhancing entrepreneurial mindsets through STEM education. Integrated science* (Vol. 15). Springer. https://doi.org/10.1007/978-3-031-17816-0_16

Sawyer, R. K. (2021). The iterative and improvisational nature of the creative process. *Journal of Creativity, 31*, 100002. https://doi.org/10.1016/j.yjoc.2021.100002

Sawyer, R. K. (2022). The dialogue of creativity: Teaching the creative process by animating student work as a collaborating creative agent. *Cognition and Instruction, 40*(4), 459–487. https://doi.org/10.1080/07370008.2021.1958219

Simpson, E., Do Dinh, E. L., Miller, T., & Gurevych, I. (2019). Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5716–5728). https://doi.org/10.18653/v1/P19-1572

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems, 33*, 16857–16867.

Su, W., Qi, Q., & Yuan, S. (2022). A moderated mediation model of academic supervisor developmental feedback and postgraduate student creativity: Evidence from China. *Behavioral Sciences, 12*(12), 484. https://doi.org/10.3390/bs12120484

Sun, M., Wang, M., Wegerif, R., & Peng, J. (2022). How do students generate ideas together in scientific creativity tasks through computer-based mind mapping? *Computers & Education, 176*, 104359. https://doi.org/10.1016/j.compedu.2021.104359

Sung, Y. T., Cheng, H. H., Tseng, H. C., Chang, K. E., & Lin, S. Y. (2022). Construction and validation of a computerized creativity assessment tool with automated scoring based on deep-learning techniques. *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/aca0000450

Temizhan, E., Mirtagioglu, H., & Mendes, M. (2021). Which correlation coefficient should be used for investigating relations between quantitative variables. *American Scientific Research Journal for Engineering, Technology, and Sciences*, *85*, 265-277. https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/7326

Thomas G. (2021). *How to do your case study* (3rd ed., pp. 1–320). SAGE publications.

Ul Haq, I., & Pifarré, M. (2023). Dynamics of automatized measures of creativity: Mapping the landscape to quantify creative ideation. In *Frontiers in education* (Vol. 8, pp. 1240962). Frontiers Media SA. https://doi.org/10.3389/feduc.2023.1240962

Van Hooijdonk, M., Mainhard, T., Kroesbergen, E. H., & Van Tartwijk, J. (2022). Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks☆. *Thinking Skills and Creativity, 43*, 100994. https://doi.org/10.1016/j.tsc.2021.100994

Walter, T. P., & Back, A. (2013). A text mining approach to evaluate submissions to crowdsourcing contests. In *2013 46th Hawaii International Conference on System Sciences* (pp. 3109–3118). IEEE. https://doi.org/10.1109/HICSS.2013.64

Wang, K., Dong, B., & Ma, J. (2019). Towards computational assessment of idea novelty. In *Proceedings of the 52nd Hawaii international conference on system sciences*. https://ssrn.com/abstract=3393611

Wang, H. H., & Deng, X. (2022). The bridging role of goals between affective traits and positive creativity. *Education Sciences, 12*(2), 144. https://doi.org/10.3390/educsci12020144

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems, 33*, 5776–5788.

Wu, M. (2023). Effects of feedback on individual creativity in social learning: An experimental study. *Kybernetes, 52*(5), 1795–1815. https://doi.org/10.1108/K-07-2021-0602

Zheng, Z., Lu, X. Z., Chen, K. Y., Zhou, Y. C., & Lin, J. R. (2022). Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Computers in Industry, 142*, 103733. https://doi.org/10.1016/j.compind.2022.103733

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ijaz Ul Haq[1] · **Manoli Pifarré**[1] · **Estibaliz Fraca**[2]

✉ Manoli Pifarré
manoli.pifarre@udl.cat

Ijaz Ul Haq
ijaz.ul-haq@udl.cat

Estibaliz Fraca
e.fraca@ucl.ac.uk

[1] Department of Education, Psychology and Social Work, Avinguda Estudi Genera 4, University of Lleida, 25001 Lleida, Spain

[2] Department of Computer Science, University College of London, London, UK