Prompting and Output Evaluation in ChatGPT for Teaching and Learning - A Review of Empirical Studies Using Machine Learning

Wenting Sun[1], Jiangyue Liu[2], Xiaoling Wang[3]

Humboldt-Universität zu Berlin, Germany[1]
Suzhou University, China[2]
Zhejiang Normal University, China[3]

Abstract

Large variations in output generated by generative Artificial Intelligences (AIs) can be influenced by slight changes in prompting. Understanding prompt usage in education can reduce the trial-and-error efforts for educators and learners using AIs driven by Large Language Models (LLMs). From the human-computer interaction (HCI) researchers' perspective, lack of guidance, representation of tasks and efforts, and generalization of prompts are challenges of interactive use of prompting [1]. Therefore, it is important to glean practice experiences and lessons from existing prompting usage articles. Existing reviews on prompt engineering often contain technical terms or lack synthesis of output evaluation methods.

This review explores how non-AI experts construct prompts in education and the methods used for output content evaluation. Using ChatGPT as an example, this review synthesizes prompt engineering behaviours in education, combining Biggs's Presage-Process-Product (3P) model and the Turn, Expression, Level of Details, Role (TELeR) taxonomy as the analytical framework [2,3]. Data were sourced from the Web of Science and Scopus, following PRISMA guidelines [4], resulting in a dataset of 495 empirical articles on ChatGPT in education. Detailed analysis of 102 articles with prompting details was conducted using thematic analysis, clustering analysis, and Machine Learning (ML) and Natural Language Processing (NLP) techniques to explore the possibility of automatically classifying articles with and without prompt details.

Six groups emerged from the clustering of coding results. The combination of bigrams and the Naïve Bayes (NB) algorithm or TF-IDF and Support Vector Machine (SVM) outperformed in classifying articles with and without prompting details. Findings suggest that domain knowledge can complement the insufficient prompting skills of non-AI experts. The study identifies specific features and patterns in prompt construction across three stages and suggests future directions for analysing ChatGPT usage behaviours.

Keywords: *Prompt engineering, Generated output evaluation, Human-AI interaction, ChatGPT, Machine learning*

1. Introduction

To effectively assist working and learning, using human-like language to drive Large Language Models (LLMs) output ("prompting") has been a potentially significant design technique for non-AI-experts [5]. As a new type of skill needs to be acquired, prompting engineer (or prompt design, prompt programming, prompting) is iterative and interactive, an art of co-creation between humans and AI [6]. As pointed by [7], prompts are instructions steering LLMs such as ChatGPT to generate customized outputs and interaction with LLMs by enforcing rules, automating process, and ensuring specific qualities of outputs. Even for natural language processing (NLP) professionals, developing efficient and generalized prompts is hard since an extension amount of trial and error needs to be taken [6].

From the human-computer interaction (HCI) researchers' perspective, lack of guidance, representation of tasks and efforts, and generalization of prompts are challenges of interactive use of prompting [1]. Therefore, it is important to glean practice experiences and lessons from existing prompting usage articles. Although there is rich contribution of the review analysis of the implementation of ChatGPT in education, prompt engineering behaviours have not yet been studied broadly and systematically. This review would provide insights into prompt engineering by analysing the empirical article with prompt details in the field of ChatGPT in education. This would provide ideas in prompt construction and output evaluation methods in different teaching and learning stages and contribute to the objective measurement option for prompt engineering skills.

2. Related Works

To effectively and efficiently navigate and enjoy these benefits provided by generative AI tools like ChatGPT, users need some efforts and skills, one of which is prompting [8, 9]. By selecting the appropriate prompts, LLMs can be used to generate the desired output to solve the tasks at hand [10]. Although prompting LLMs appears effortless, designing context sensitive prompt strategies, devising prompts to overcome the arisen error from LLMs, and systematically assessing those prompts strategies' effectiveness is a complex interdisciplinary topic [5]. For the field of education, with the increasing popularity of the prompt engineering practice, investigating the skills of prompt engineering is important [6].

Some studies contribute to prompting LLMs to solve complex tasks. For instance, to solve complicated reasoning tasks, the chain-of-thought prompting (CoT) proposed by [11] allows LLMs to decompose multi-step problems into intermediate steps and provide chances to debug when the reasoning process met errors. By including examples of CoT sequences as exemplars of few-shot prompting, CoT has the potential to facilitate reasoning capabilities of off-the-shelf LLMs. Targeting at understanding LLMs' potential for performing complex tasks involving multiple steps and subtasks, [3] developed a general prompt taxonomy, TELeR (Turn, Expression, Level of details, Role), to serve as a standard for comparing and benchmarking LLMs' performances and designing prompts.

As of now, some research on prompting engineering is more from the technology-oriented perspective [10]. For non-expert users who tend to adopt experience from human-to-human interaction, the behaviour of prompt initiated might be unsystematic and opportunistic [5]. It is time-consuming to construct promoting because ChatGPT need to sense the context of the inquiry to generate actional outputs. More knowledge about how to prompt is a skill needed to combine domain knowledge and LLMs knowledge. For non-AI-experts, there are some prompt features, components or strategies to be referred to (as mentioned above). However, the question is the proposed prompting techniques are too abstract, and it is hard to guide non-AI-experts or beginners to implement these strategies in their actual practice. Therefore, it would be valuable to have a review of the prompt strategies and their usage scenarios to present examples to non-AI-experts and beginners about whether there are existing prompt cases similar to their problem to help them develop their own prompts and evaluate the quality of ChatGPT outputs. In this article, choosing ChatGPT as one LLM from the generative AI field, we want to introduce a skill-based approach to prompt engineering as an important factor for enabling educators and students to manage ChatGPT effectively.

To make contributions to the prompting construction of ChatGPT in education, two research questions (RQs) led this review:

RQ1: What the performance of text classification techniques to identify empirical studies with and without detailed prompts?

RQ2: What prompting features can be found in the teaching and learning context?

3. Methods

3.1 Data Collection

The data were searched in two big comprehensive academic databases, Web of science (WoS) and Scopus. Using topic (title, abstract, keywords) for the search place and English and Chinese for the language, two groups of keywords were used: "chatgpt* OR gpt* OR chatbot* OR Bing OR Bard OR Copilot" AND "learn* OR educat* OR train* OR teach*".

The search procedure followed the Preferred reporting items for systematic reviews and meta-analysis (PRISMA) guidance in the search and selection flow [4]. The research area (education, educational research) was chosen in the WoS database otherwise the number of search results was more than 10,000. Peer-reviewed journal papers or long conference papers, or chapters from books were chosen as paper types. The search was conducted in June and July of 2024. The selection workflow and results can be seen in Figure 1.

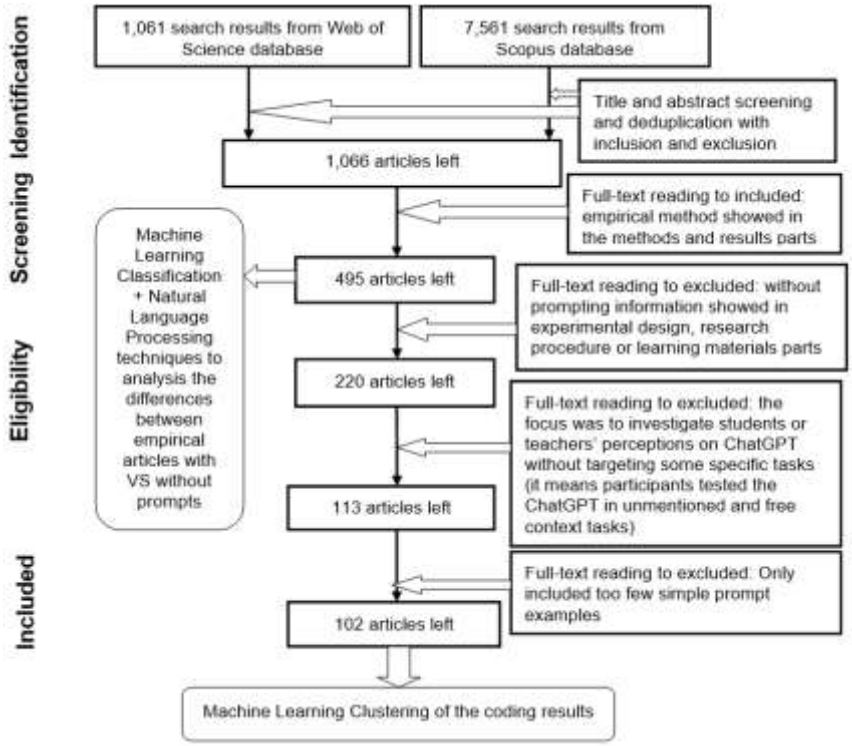

Figure 1. PRISMA flowchart guidelines and general description of data analysis.

3.2 Data Analysis

Extraction method: The extracted data were aggregated in excel for further analysis.
Prompt features clustering: To better explore the prompting construction and output content evaluation, we adopted a clustering algorithm to automatically identify similar prompting characteristics. Clustering is an unsupervised Machine learning (ML) algorithm that parts similar samples into the same cluster. Considering our dataset is mixed (including both categorical and numerical variables), we used the K-Prototype clustering algorithm (like K-Means clustering but K-Means is more suitable for numerical variables). To find optimal k (the number of clusters), we used elbow method.
ML and Natural language processing (NLP) model training and evaluation: After vectorization of the raw text by TF-IDF and bigram, this study adopted six classifiers including Naïve Bayes (NB), Random Forest (RF), K Nearest Neighbours (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost. These are the commonly used classifiers in educational research. These six classifiers were implemented using the Python package Scikit-learn. The text of titles and abstracts from all empirical studies were randomly split into 80% training dataset and 20% testing dataset after data balancing considering the number of included articles (with prompts details) is less than excluded articles (without prompts details). This study evaluated the performance of each classifier on the reserved 20% test dataset. For ML evaluation metrics, accuracy, precision, recall, and F1 score were included to compare the performance of these models.
Visualizing the empirical studies with prompt details: To better visualize the highly frequent terms of empirical studies with prompt details in education, word cloud was employed. All titles and abstracts from included articles and excluded articles were used as the input text data whereas two word clouds were generated.
Different code schemes were used. Both top-down and bottom-up methods were used to extract more information. Two researchers independently coded the articles, with one coding 20% of the data and the other coding all of them. The agreement achieved at 80%. More details can be found below:
Categorisation of ChatGPT in teaching and learning stages
In this review, Biggs's Presage-Process Product (3P) model of teaching and learning was used to explain the phases of ChatGPT usage scenarios in education. We think it is reasonable to divide the usage scenarios of prompting into these three phases considering the user roles, time limitation, output requirement, and task context would impact prompt creating and human-AI interaction. Biggs's 3P model divides educational events into three stages, namely presage, process, and product. These

stages contain different but strongly mutually interactive learning and teaching activities, forming a complex and dynamic context [2]. In the presage stage, students' factors and teaching context initially set the course climate, then in the process stages, the dynamic balance of selective memorizing and seeking meaning promotes the optimal time and space management for the learning-focused ongoing activities. To reflect the effects of the prior two stages, diverse and hierarchy levels of learning outcomes are selected and evaluated in the product stage. This model has been used in a review of ChatGPT in education which demonstrates a promising landscape to explain the engagement of teachers and students using ChatGPT in education [12].

Categorisation of the prompting features

After comparing multiple prompting features or strategies analysis frameworks (as we mentioned in part 2), we chose TELeR by [3], a general taxonomy of LLM prompts, as the frameworks to analyse prompt types and details. This taxonomy targets ill-defined complex tasks which means ill-defined, abstract goal-oriented, and highly dependent on subjective interpretation. TELeR categorizes LLM prompts from four dimensions, namely turn (single or multi-turn), expression (question style or instruction style), role (system role defined or undefined), and levels of details. The levels of details in task specification are divided into seven levels (levels 0-6) according to clear goals, associated data, distinct sub-tasks, evaluation criteria/few-shot examples, additional information fetched via information retrieval techniques, and explanation/justification seeking. This multi-level structured analysis framework could help to analysis both simple and complex prompting. For promptings with multiple levels of details co-occur, we tagged them using the highest levels.

Categorisation of the ChatGPT output evaluation

We used thematic analysis to extract information about the evaluation method of ChatGPT output. We followed the six-step proposed by [13], consisting of familiarizing with the data, generating initial codes, searching for sub-themes and themes, reviewing sub-themes and themes, defining and naming sub-themes and themes, and reporting. Using this method, we developed a code scheme about ChatGPT output quality evaluation method, details in Table 1.

Table 1. ChatGPT output evaluation method.

Measurement	Description	Example
Bottom-up analysis	Interpretive analysis, thematic analysis, other qualitative methods without explicitly mention data analysis method but organize the data into increasingly more abstract units of information without using analysis framework in advance	Writing skills development process by Punar Özçelik and [14]
General-based evaluation rubric	Correctness, Explanation Sophistication levels, execute the coding solution, validity, accuracy, clarity, adaptation, alignment, verification, suitability, readability, consistency, other rubrics that can be used in general area	Formative feedback guidelines by [15]
Domain-based evaluation rubric	Explicitly mentioned the domain or course-based evaluation	Field course design evaluation by [16]
ML evaluation rubrics	Machine learning evaluation metrics	Detect incoherent math answers by [17]
User-perceptions	User perceptions about the domain knowledge generated by ChatGPT	ChatGPT as writing assistant by [18]
Learning-performance	Learning performance impacted by outputs generated by ChatGPT	Embedded systems course quiz by [19]

4. Results and Discussion More information about the existing reviews of ChatGPT in education, the inclusion criteria, coding results, code solutions, and selected empirical studies in this study can be found at the OSF link: https://osf.io/qfrmt/?view_only=e7282e863bca4b64b34fbd990b591bbb

4.1 RQ1: Automatically Classify Articles with Prompts

The abstract text from 495 empirical studies about ChatGPT in education after the 1st round of reading have been developed into a corpus. These data were categorized into two types, one is with prompt details, and one is without. It means the dataset included two columns, one is the raw text from the abstract, and one is the category.
As shown in Table 2, we chose both Term Frequency-Inverse Document Frequency (TF-IDF) and bigrams to do feature embedding to transform text into number vector (vectorization) and then the results can be fed into Naïve Bayes (NB), Random Forest (RF), K Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), XGBoost algorithms function as classifiers. Based on traditional ML evaluation metrics, it was found that the combination of bigrams and the Naïve Bayes (NB) algorithm or TF-IDF and Support Vector Machine (SVM) outperformed.

Table 2. Summary of the performance of different algorithms.

Vectorization method	Algorithm	Category	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Bigrams	LR	Without prompt	0.68	0.67	0.7	0.68
		With prompt		0.70	0.67	0.68
	NB	Without prompt	**0.73**	**0.74**	**0.70**	**0.72**
		With prompt		**0.73**	**0.76**	**0.74**
	RF	Without prompt	0.61	0.58	0.70	0.64
		With prompt		0.65	0.52	0.58
	KNN	Without prompt	0.71	0.72	0.65	0.68
		With prompt		0.70	0.76	0.73
	SVM	Without prompt	0.61	0.60	0.60	0.60
		With prompt		0.62	0.62	0.62
	XGBoost	Without prompt	0.66	0.64	0.70	0.67
		With prompt		0.68	0.62	0.65
TF-IDF	LR	Without prompt	0.71	0.70	0.70	0.70
		With prompt		0.71	0.71	0.71
	NB	Without prompt	0.71	0.75	0.60	0.67
		With prompt		0.68	0.81	0.74
	RF	Without prompt	0.61	0.60	0.60	0.60
		With prompt		0.62	0.62	0.62
	KNN	Without prompt	0.63	0.67	0.50	0.57
		With prompt		0.62	0.76	0.68
	SVM	Without prompt	**0.73**	**0.74**	**0.70**	**0.72**
		With prompt		**0.73**	**0.76**	**0.74**
	XGBoost	Without prompt	0.61	0.60	0.60	0.60
		With prompt		0.62	0.62	0.62

(Note: with/without prompt in this table means whether prompting information existed that can extract features. Both categories are empirical studies about ChatGPT in education.)

Two word clouds were also generated to provide a broader view of the high-frequency words used in the articles on ChatGPT usage in education with detailed prompts (n=102) compared with articles mentioned prompts (including the prior n=102 with detailed prompts) (n=495). The larger and bolder the term is, the more often the term appears in the corpus. Except for the high frequency of ChatGPT, student, study, potential, and education, can be found differences between two figures. Except for the high frequency of ChatGPT, student, study, potential, and education, some differences can be found between these two figures. As shown in Figure 2 (left), feedback, prompt, LLM, response, assessment, quality, and performance appeared more frequently in empirical studies with prompt details. As shown in Figure 2 (right), use, tool, learning, research, excluded, impact, artificial intelligence, and AI, appeared frequently in the corpus which included empirical studies with and without prompt details.

Figure 2. Word cloud generated from empirical articles (left: ChatGPT in education with prompt details (n=102); right: ChatGPT in education (n=495))

4.2 RQ2: Clustering Results of Prompt Features

We chose six clusters to synthesize the prompting features and related output evaluation methods. That is because the elbow method demonstrated that six clusters could be the optimized clusters. For further examination, it was found that when choosing six clusters, each teaching and learning stages would have two clusters. Six clusters would largely reduce the information loss from our dataset. These prompt features were clustered into the stages of the 3P model of teaching and learning, as shown in Table 3. As the results are shown in Table 3, we have drawn the following observations:

Table 3. Summary of clustered groups using K-Prototype clustering.

Stages	Quality	Turn	Expression	Role	Detail levels	Cluster#(n)
Presage	General-based evaluation rubric	Multiturn	Instruction	Undefined	L3.73	5(26)
	Bottom-up analysis	Multiturn	Question	Undefined	L2	2(13)
Process	Bottom-up analysis	Multiturn	Mixed	Defined	L2.22	3(9)
	Domain-based evaluation rubric	Multiturn	Question	Undefined	L1.04	0(24)
Product	Domain-based evaluation rubric	Single	Instruction	Undefined	L2.85	1(13)
	General-based evaluation rubric	Single	Instruction	Defined	L3.94	4(17)

(Note: to generate groups around centroids during clustering, some articles were included in other stages by machine, which might be different from human manually coded results. L=detail levels.)

In general, it can be found that articles in the product stage preferred richly detailed prompts and conducted a single turn of prompting. Articles in presage and process stages demonstrated multi-turn prompting while presage showed a comparatively high level of prompt details. Compared to using instruction as expression style, when articles used questions as the expression style, they provided fewer details and more likely undefined ChatGPT a role to generate outputs.

In the presage stage, it can be found that most studies did not define a role in ChatGPT before providing the actual prompt. To perform a complex task, multiple turns of prompts were normally conducted. The cluster#2 using question as expression style demonstrated an average of L2 details

prompting which means these studies more often enter prompts with multi-sentence expressing the high-level goal and the sub-tasks that need to be performed to achieve the goal. The articles in this cluster preferred to employ bottom-up analysis as the ChatGPT output quality evaluation method. Cluster #5 using instruction as expression style showed an average of L3.73 detailed prompts which means prompts in these studies not only include high-level goals and sub-tasks but also used a bulleted list of sub-tasks and a guideline on how ChatGPT output will be evaluated. 66.7% (26 out of 39) articles in this stage chose this kind of highly detailed prompting. Articles in this cluster more likely chose a general-based evaluation rubric to evaluate the ChatGPT output quality.

In the process stage, it can be found that the prompt details were less rich than in the other two stages and more likely used multi-turn prompting. It makes sense considering users in this teaching and learning stage need to have a quick question and answer (Q&A) session with ChatGPT and adjust next time prompts orienting on a certain topic or task within limited time. Articles in cluster#0 preferred question as expression style and did not define ChatGPT a role. Prompts in these articles are in L1.04 details which means most articles include high-level goals. 72.7% (24 out of 33) articles in this stage chose this kind of prompt strategy. These articles chose domain-based evaluation rubrics to evaluate the ChatGPT output quality. Articles in cluster#3 preferred a mixed expression style combining instruction and questions in prompts. And most of them define ChatGPT a role to communicate with. Prompts in these articles are in L2.22 details which means most articles include high-level goals and sub-tasks while some organize the sub-tasks in a bulleted list. These articles chose bottom-up analysis to evaluate the ChatGPT output quality.

In the product stage, prompts using instruction as the expression style in single turn is commonly used. Cluster#1 preferred L2.85 detailed prompts and did not define ChatGPT a role. Normally, they use domain-based evaluation rubric as ChatGPT output measurement. Cluster#4 had the highest level of prompt details with L3.94. Articles in this cluster define ChatGPT a role to support the generation or analysis of the product in teaching and learning. Normally, they used general-based evaluation rubric to measure the ChatGPT output quality.

5. Conclusion and Implications

In this study, we emphasized prompting construction in three stages of teaching and learning. This review contributes to prompting construction for non-AI-experts, ChatGPT output content evaluation, ML, and NLP techniques used in review articles. Based on the analysis framework combining Biggs's Presage-Process Product (3P) model of teaching and learning and a general taxonomy of LLM prompts TELeR by [3] and thematic results of ChatGPT outputs evaluation methods, six groups were generated using a clustering algorithm. It was found that bottom-up, general evaluation rubrics, domain evaluation rubrics, ML evaluation metrics, user perceptions, and learning performance are the commonly used ChatGPT outputs evaluation methods. To further explore the dataset, ML and NLP techniques were used to automatically identify empirical studies with prompt details and without in ChatGPT in education. Word clouds were drawn to explore the high-frequency terms.

For researchers, low transparency of AI tools might relate to practical and ethical issues of their implementation in real society, for which explainable and human-centered AI is called for meaningful and impactful educational technology [20]. Including human-in-the-loop components in prior stage studies might be one potential solution, which would be hard at the beginning but the continues new datasets from real life will be collected. AI tools would be not limited to test verification in laboratory settings and on different datasets. Included voices from the related stakeholders in the education system would create a virtuous cycle for the innovative integration of AI into traditional educational systems in the long term. In this way, ethics and integration into the classroom would be deeply explored and discussed with the close stakeholders from real educational scenarios whether in micro (classroom/course), meso (school/university/institution), or macro (education system/policy) levels.

For research methods in review writing, benefit from the development of NLP and LLMs techniques, more articles can be included in the review process to help understand the big picture of a field. When a lot of articles are included and many coding results are collected, it is hard to synthesise regularity from these large amount of data with less information cost. In this situation, ML and NLP techniques could assist the analysis of these data. Using clustering, several clustered groups can be generated, and based on this, researchers can comparatively be easier to find the similarities and differences among the included articles. Using NLP, the text data can be developed into a corpus and explored further from the frequency of terms and semantic similarities of the sentence aspects. Even for the excluded articles, the data from them is also a kind of complementary information for the topic one review explored. These data can be collected to be putted into ML classifier to automatically

categorize included articles and excluded articles when the search results from academic databases are huge.

REFERENCES

[1] Dang H., Mecke L., Lehmann F., Goller S., Buschek D., “How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models”, arXiv preprint arXiv:2209.01390, 2022.
[2] Biggs J., Kember D., Leung D. Y., “The revised two-factor study process questionnaire: R-SPQ-2F”, British Journal of Educational Psychology, 71(1), 133-149, 2001.
[3] Santu S. K. K., Feng D., “TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks”, arXiv preprint arXiv:2305.11430, 2023.
[4] Page M. J., McKenzie J. E., Bossuyt P. M., Boutron I., Hoffmann T. C., Mulrow C. D., Moher D., “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”, International Journal of Surgery, 88, 105906, 2021.
[5] Zamfirescu-Pereira J. D., Wong R. Y., Hartmann B., Yang Q., “Why Johnny can’t prompt: how non-AI experts try (and fail) to design LLM prompts”, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1-21, 2023.
[6] Oppenlaender J., Linder R., Silvennoinen J., “Prompting ai art: An investigation into the creative skill of prompt engineering”, arXiv preprint arXiv:2303.13534, 2023.
[7] White J., Fu Q., Hays S., Sandborn M., Olea C., Gilbert H., Schmidt D. C., “A prompt pattern catalog to enhance prompt engineering with chatgpt”, arXiv preprint arXiv:2302.11382, 2023.
[8] Ansari A. N., Ahmad S., Bhutta S. M., “Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review”, Education and Information Technologies, 1-41, 2023.
[9] Cronjé J., “Exploring the Role of ChatGPT as a Peer Coach for Developing Research Proposals: Feedback Quality, Prompts, and Student Reflection”, Electronic Journal of e-Learning, 22(2), 1-15, 2023.
[10] Liu P., Yuan W., Fu J., Jiang Z., Hayashi H., Neubig G., “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”, ACM Computing Surveys, 55(9), 1-35, 2023.
[11] Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Zhou D., “Chain-of-thought prompting elicits reasoning in large language models”, Advances in neural information processing systems, 35, 24824-24837, 2022.
[12] Mai D. T. T., Da C. V., Hanh N. V., “The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach”, Frontiers in Education, 9, 1328769, 2024.
[13] Braun V., Clarke V., “Using thematic analysis in psychology”, Qualitative Research in Psychology, 3(2), 77–101, 2006.
[14] Punar Özçelik N., Yangın Ekşi G., “Cultivating writing skills: the role of ChatGPT as a learning assistant—a case study”, Smart Learning Environments, 11(1), 10, 2024.
[15] Zhang Z., Dong Z., Shi Y., Price T., Matsuda N., Xu D., “Students’ perceptions and preferences of generative artificial intelligence feedback for programming”, Proceedings of the AAAI Conference on Artificial Intelligence, 38(21), 23250-23258, 2024.
[16] Tupper M., Hendy I. W., Shipway J. R., “Field courses for dummies: To what extent can ChatGPT design a higher education field course?”, Innovations in Education and Teaching International, 1-15, 2024.
[17] Urrutia F., Araya R., “Who's the Best Detective? Large Language Models vs. Traditional Machine Learning in Detecting Incoherent Fourth Grade Math Answers”, Journal of Educational Computing Research, 61(8), 187-218, 2024.
[18] Barrett A., Pack A., “Not quite eye to AI: student and teacher perspectives on the use of generative artificial intelligence in the writing process”, International Journal of Educational Technology in Higher Education, 20(1), 59, 2023.
[19] Shoufan A., “Can students without prior knowledge use ChatGPT to answer test questions? An empirical study”, ACM Transactions on Computing Education, 23(4), 1-29, 2023.
[20] Yan L., Sha L., Zhao L., Li Y., Martinez-Maldonado R., Chen G., Gašević D., “Practical and ethical challenges of large language models in education: A systematic scoping review”, British Journal of Educational Technology, 55(1), 90-112, 2024.