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

Presenter: Wenting Sun

Wenting Sun, Jiangyue Liu, Xiaoling Wang
Humboldt-Universität zu Berlin, Germany
Suzhou University, China
Zhejiang Normal University, China

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 (Zamfirescu-Pereira et al., 2023).
- 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 (Oppenlaender et al., 2023)
- 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 (Dang et al., 2022). Therefore, it is important to glean practice experiences and lessons from existing prompting usage articles.

Research questions

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.

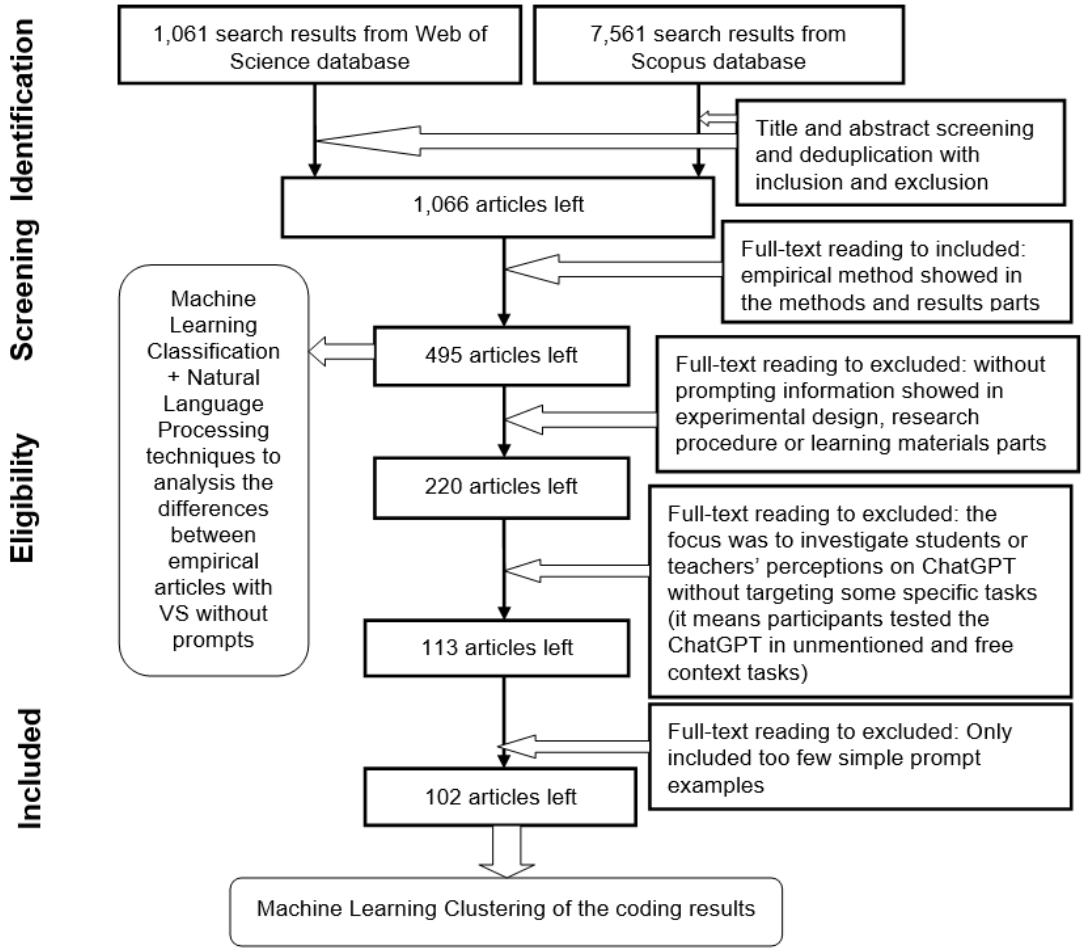
Therefore, it would be valuable to have a review of the prompt strategies and their usage scenarios to present examples to non-AI-experts and beginners about whether there are existing prompt cases similar to their problem to help them develop their own prompts and evaluate the quality of ChatGPT outputs.

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?

Research methods

- PRISMA guidance in research review
- Database: Web of Science, Scopus
- Search string: “chatgpt* OR gpt* OR chatbot* OR Bing OR Bard OR Copilot” AND “learn* OR educat* OR train* OR teach*”.



Research methods

Prompt features clustering:

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.

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.

Research methods

Categorisation of ChatGPT in teaching and learning stages

Biggs's Presage-Process Product (3P) model of teaching and learning was used to explain the phases of ChatGPT usage scenarios in education.

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 (Biggs et al., 2001).

Categorisation of the prompting features

After comparing multiple prompting features or strategies analysis frameworks (as we mentioned in part 2), we chose TELeR by (Santu & Feng, 2023), a general taxonomy of LLM prompts, as the frameworks to analyse prompt types and details.

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.

Research methods

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 Braun and Clarke (2006), 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 tables.

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]

Results

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.

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

Results

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 left, feedback, prompt, LLM, response, assessment, quality, and performance appeared more frequently in empirical studies with prompt details. As shown in Figure 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.

Results

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.

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)

Discussions and conclusions

- 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.
- 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 Santu and Feng (2023) 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.

Implications

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 (Yan et al., 2024). 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.

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.

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.

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. 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.
5. Oppenlaender J., Linder R., Silvennoinen J., “Prompting ai art: An investigation into the creative skill of prompt engineering”, arXiv preprint arXiv:2303.13534, 2023.
6. Braun V., Clarke V., “Using thematic analysis in psychology”, Qualitative Research in Psychology, 3(2), 77–101, 2006.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.