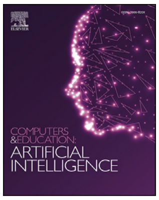
Computers and Education: Artificial Intelligence 4 (2023) 100140

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Can large language models write reflectively

Yuheng Li[a,1], Lele Sha[a,1], Lixiang Yan[a], Jionghao Lin[a], Mladen Raković[a], Kirsten Galbraith[b], Kayley Lyons[c], Dragan Gašević[a], Guanliang Chen[a,*]

[a] *Centre for Learning Analytics, Monash University, Australia*
[b] *Experiential Development and Graduate Education, Faculty of Pharmacy and Pharmaceutical Sciences, Monash University, Australia*
[c] *Centre for Digital Transformation of Health, University of Melbourne, Australia*

ARTICLE INFO

Keywords:
ChatGPT
Generative language model
Reflective writing
Natural language processing

ABSTRACT

Generative Large Language Models (LLMs) demonstrate impressive results in different writing tasks and have already attracted much attention from researchers and practitioners. However, there is limited research to investigate the capability of generative LLMs for reflective writing. To this end, in the present study, we have extensively reviewed the existing literature and selected 9 representative prompting strategies for ChatGPT – the chatbot based on state-of-art generative LLMs to generate a diverse set of reflective responses, which are combined with student-written reflections. Next, those responses were evaluated by experienced teaching staff following a theory-aligned assessment rubric that was designed to evaluate student-generated reflections in several university-level pharmacy courses. Furthermore, we explored the extent to which Deep Learning classification methods can be utilised to automatically differentiate between reflective responses written by students vs. reflective responses generated by ChatGPT. To this end, we harnessed BERT, a state-of-art Deep Learning classifier, and compared the performance of this classifier to the performance of human evaluators and the AI content detector by OpenAI. Following our extensive experimentation, we found that (i) ChatGPT may be capable of generating high-quality reflective responses in writing assignments administered across different pharmacy courses, (ii) the quality of automatically generated reflective responses was higher in all six assessment criteria than the quality of student-written reflections; and (iii) a domain-specific BERT-based classifier could effectively differentiate between student-written and ChatGPT-generated reflections, greatly surpassing (up to 38% higher across four accuracy metrics) the classification performed by experienced teaching staff and general-domain classifier, even in cases where the testing prompts were not known at the time of model training.

1. Introduction

Educators frequently administer reflective writing tasks to elicit students' reflections on their prior learning experiences, within or outside a particular course (Mann et al., 2009). Engagement in this writing task has been shown to promote the development of critical thinking and problem-solving, an important set of skills that can benefit life-long learning (Charon & Hermann, 2012). However, recent advancements in generative language models have raised concerns among educators administering written assignments (Kung et al., 2022, Susnjak, 2022, Yan et al., 2023). For instance, by utilising generative language models to automatically draft their reflective written responses, some students may miss the opportunity to engage in authentic and critical reflections on their own learning experiences. In this way, for example, some students may not be able to evaluate the learning strategies they used in the past, and then modify their learning strategies to ensure more productive learning in the future (Raković et al., 2022). More importantly, the instructors cannot give tailored feedback to improve student learning.

In particular, ChatGPT, a recently released chatbot based on artificial intelligence (AI), has demonstrated the potential to comprehend different requests from users and, per those requests, generate relevant and insightful texts for different purposes and in different domains, e.g., journal articles (Pavlik, 2023), financial reports (Wenzlaff & Spaeth,

* Corresponding author.
E-mail addresses: yuheng.li@monash.edu (Y. Li), lele.sha@monash.edu (L. Sha), lixiang.yan@monash.edu (L. Yan), jionghao.lin1@monash.edu (J. Lin), mladen.rakovic@monash.edu (M. Raković), kirstie.galbraith@monash.edu (K. Galbraith), kayley.lyons@unimelb.edu.au (K. Lyons), dragan.gasevic@monash.edu (D. Gašević), guanliang.chen@monash.edu (G. Chen).

[1] The two authors contributed equally to this paper.

https://doi.org/10.1016/j.caeai.2023.100140
Received 20 April 2023; Received in revised form 6 May 2023; Accepted 9 May 2023
Available online 12 May 2023

2022), and academic literature reviews (Aydın & Karaarslan, 2022). Despite the promises of ChatGPT to make text generation more efficient, many educators are concerned regarding the potentially detrimental effects of using automatic text generation methods to facilitate student writing, including reflective writing (Kung et al., 2022, Stokel-Walker, 2022, Susnjak, 2022). However, those concerns have not been empirically supported as of yet in the context of reflective writing. More research is thus needed to empirically document and understand the capabilities of cutting-edge text generation methods to generate reflective writing, thus providing educators and researchers with new insights regarding the use of these methods in reflective writing. In addition, it remains unknown whether/how AI-generated writing can be accurately differentiated from students' original work. This may be particularly important for educational stakeholders aiming to identify and prevent academic misconduct, e.g., reflective essays generated by ChatGPT, but submitted as students' original work (Stokel-Walker, 2022). To address these challenges, in this study, we set out to (1) empirically examine the quality of reflective responses generated by ChatGPT and (2) empirically investigate the use of state-of-the-art classification approaches to differentiate between the responses generated by the ChatGPT bot and the responses originally generated by students.

Accordingly, we posed the following Research Questions (RQs): **RQ1** – Can ChatGPT generate high-quality reflective writings? **RQ2** – To what extent are reflective responses generated by ChatGPT distinguishable from reflective responses written by university students? To answer the RQs, we have extensively prompted ChatGPT to generate a diverse set of reflective writings. We also involved experienced teaching staff to evaluate the reflective depth presented in the writings. Lastly, we compared the differentiation performance (i.e., whether the ChatGPT-generated writings could be differentiated from student-written ones) among (i) experienced teaching staff; (ii) the state-of-the-art AI text detector released by OpenAI; and (iii) a BERT-based classifier (Devlin et al., 2018) fine-tuned on reflective writings generated by ChatGPT and written by students.

The contribution of this paper is two-fold: 1) we illustrated the capability of the state-of-the-art large language models, specifically ChatGPT, in generating reflective writings and the quality of these ChatGPT-generated content compared to student-written works, and 2) we developed a BERT-based classifier for distinguishing between AI-generated and student-written reflective writings. These timely contributions could inform educational researchers and practitioners regarding ChatGPT and other large language models' potential impacts on reflective writing tasks, such as students might miss out on the opportunity to engage in cognitive reflection if they choose to use ChatGPT to generate reflective writing for the very purpose of completing an assessment.

2. Related work

2.1. Large language models in education

In the domain of education, LLMs have already demonstrated diverse potential in automating and completing various educational tasks. For example, in a recent systematic literature review, Yan et al. (2023) identified a total of 53 different educational tasks that could potentially benefit from the ongoing developments of LLMs. Specifically, LLMs have illustrated proven capability in automatically generating short-answer and multiple-choice questions from educational materials, which can be delivered as formative assessments for teachers to evaluate and identify students' knowledge gaps (Kurdi et al., 2020). Likewise, automatic feedback systems could benefit from the advancement in LLMs as these AI-based systems can potentially become more generalisable and able to deliver more personalised feedback compared to rule-based systems (Cavalcanti et al., 2021). LLMs could also empower educational chatbots to have the ability to engage in more human-like conversations with learners and become capable of answering questions across multiple disciplines (Wollny et al., 2021). Although these systematic reviews illustrated the diverse potential of LLMs in education, the findings were drawn from reviewing studies that used mostly BERT and its variations instead of the state-of-the-art models, such as GPT-3 and later versions.

Recent studies that investigated the potential of ChatGPT have illustrated the capability of state-of-the-art LLMs in completing complex assessments and writing tasks. For example, ChatGPT has achieved the equivalent of a passing score for a third-year medical student (above 60%) in the United States Medical Licensing Examination Step 1 exam and provided logical justification and informational context across the majority of answers (Gilson et al., 2023). Similar performances (60.8%) were also found when using ChatGPT to complete a parasitology exam (79 questions) for Korean medical students (Huh, 2023). For solving higher-order reasoning questions in pathology, ChatGPT has demonstrated a relational level of accuracy and correctly solved around 80% of the questions (n = 100) (Sinha et al., 2023). ChatGPT was also found capable of answering both first- and second-order knowledge questions in the domain of microbiology with around 80% accuracy (Das et al., 2023). Likewise, ChatGPT's performance on four real exams (containing 95 multiple choice questions and 12 essay questions) at the University of Minnesota Law School was equivalent to C+ students, which would pass all four courses (Choi et al., 2023). ChatGPT has also demonstrated the ability to generate abstracts for research papers and pass plagiarism detectors with 100% originality scores (Macdonald et al., 2023). These diverse performances of ChatGPT on different assessments and writing tasks could be related to the prompting methods and the task contents (e.g., tasks involving figures and images (Huh, 2023)).

2.2. Prompting generative language models

To ensure the results of a language generation model are representative, it is critical to test the model with effective prompts (Gu et al., 2021, Hou et al., 2022, Kung et al., 2022), i.e., short utterances that serve as an input to the model. A prompt provides specific instructions for text generation, e.g., relative to a particular topic (Yu et al., 2022). Previous studies (Liu et al., 2021) have found that adopting effective prompts could greatly leverage the power of a generative language model like GPT-3 (Brown et al., 2020). Therefore, in order to fully exploit the power of the ChatGPT, we first extensively reviewed the literature on existing prompting strategies and, based on our review, designed a set of prompting strategies to empower the generation of student reflective writing in the present study. It is important to note that prior to ChatGPT, generative language models (e.g., GPT-2 (Radford et al., 2019)) were generally limited to performing tasks in a zero-shot setting, where only a natural language description of a task is available to the model (Dong et al., 2022). To remedy this, several prompting strategies have been proposed to improve the quality of generated responses in a few-shot learning setting (Gu et al., 2021, Hou et al., 2022), i.e., using a few representative examples of task-answer pairs as a basis to generate a high-quality response. By contrast, ChatGPT models have greatly improved their zero-shot capability and no longer necessitate the complex few-shot prompting strategies (e.g., model-based (Hou et al., 2022) and embedding-based (Gu et al., 2021)) to understand task requirements (Kung et al., 2022). For this reason, we have omitted few-shot prompting strategies in our review.

In general, two strands of research have emerged from our literature review. First, researchers have mostly explored the use of the template as a basis to formulate a prompt (Srivastava et al., 2022, Tang et al., 2022, Wang et al., 2022) – termed as `Generate by template`. For instance, Kung et al. (2022) utilised a `Query` template which contains just enough information about a request to answer medical questions (e.g., *Can you generate a writing about [a specific topic]?*), which demonstrated around 50% accuracy in a zero-shot setting. Researchers also explored generating more context-specific responses by prefixing the query with contextual information. For example, Wang et al. (2022)

Table 1
Example student-written and ChatGPT-generated reflection snapshots containing the first two sentences and the last sentence.

Author	A snapshot of reflective writing
Student	The second year XXX had a joint workshop with the second-year medicine and nursing cohort. I was very excited and keen to meet the students from the other course and get an idea of what their course was like. ... I will continue to work towards improving my oral communication skills and my leadership skills.
ChatGPT	During my internship at XXX, I had the opportunity to work alongside experienced pharmacists and learn about the various medications and treatments used to manage patients' conditions. ... My internship at XXX Hospital was a valuable learning experience that has helped shape my understanding of the pharmacy profession and the role of compassion in patient care.

added a context prefix to the query (termed as `Context` template) detailing factors such as scenario (e.g., course settings) and task details (e.g., assessment criteria). In a similar vein, researchers also attempted to couple a `Query` directly with a high-quality example (termed as `Example` template), and found that generative LLMs were able to automatically detect important features in the example and generate similarly high-quality responses (Liu et al., 2021). In the broad fields of Artificial Intelligence (AI) and Natural Language Processing (NLP), researchers have explored techniques to further refine/revisit the initial response and regenerate an improved version (termed as `Regenerate to improve`). For instance, Susnjak (2022) asked ChatGPT to evaluate/reflect on its initial response and automatically generate an improved one (termed as `Reflect on Reflections (RoR)`), which they showed that the regenerated responses demonstrated improved quality in terms of clarity, relevance and depth. In a similar vein, researchers sought to trace the intermediate reasoning steps via Chain-of-Thoughts (CoTs) (Zhang et al., 2022), i.e., asking the model to output not only the response but also a series of steps that led to the response (termed as `CoT`). Based on this, researchers were able to give specific instructions and ask ChatGPT to generate an improved version in a legal setting (Yu et al., 2022).

2.3. Evaluation of reflective writing

Although researchers have shown that ChatGPT's capability in generating high-quality content could be greatly improved via the adoption of these prompting strategies, limited research has set out to evaluate the quality of generated responses in an educational setting (Kung et al., 2022, Wang et al., 2022). Particularly, one of the commonly utilised writing tasks by educators is reflective writing in which students' learning progresses can be captured (e.g., the attainment of skills) (Charon & Hermann, 2012). The quality of reflective writing can be assessed from several different perspectives based on prior theories of reflection and experiential learning (Boud et al., 1985, Dennison & Kirk, 1990, Kolb, 1984). Specifically, Boud et al.'s framework (Boud et al., 1985) posits that reflective activity consists of three key components: revisiting past experiences, considering related emotions, and re-evaluating the experience. The re-evaluation process can be further divided into five sub-processes, including connecting past and present emotions and knowledge, integrating the connections for new perspectives, validating new perspectives, relating new perspectives to current or future experiences, and determining actions to take as a result of the reflection. These stages are essential for educators to assess and identify evidence of students' self-awareness and high-quality comprehension during reflective writing (Ryan, 2011). More importantly, such reflective activities can facilitate students' self-regulation, critical thinking, problem solving and etc. (Mann et al., 2009, Dewey, 1933). Nevertheless, since students' tend to lack the capability to reflect critically (Sen, 2010), Ryan (2011) have suggested that exemplifying reflective writings that critically evaluate ones' prior experience, feelings and etc. in terms of the reflective stages can promote students' reflective abilities. To this end, we argue that state-of-the-art generative LLMs and tools implemented on such bases (e.g., ChatGPT) can potentially help with generating examples of high-quality reflective writings. However, whether ChatGPT can generate reflective writings that critically evaluate different reflective stages/processes remains unknown.

Besides, given the demonstrated power of ChatGPT in the broad AI and NLP research (Castelvecchi, 2022, Kung et al., 2022), educators and practitioners have raised concerns regarding ChatGPT's potential negative impacts on educational writing tasks, such as endangering academic integrity and encouraging misconduct via the student's exploitation of the tool for generating contents as their original works (Pavlik, 2023, Sharples, 2022, Susnjak, 2022). More specifically, in the case of reflective writing, students' continuous engagement in such a writing task is warranted to improve their reflective abilities (Lew & Schmidt, 2011). Therefore, it is critical to investigate effective approaches to distinguish between ChatGPT-generated content vs. students' original work, so that any misuse of ChatGPT can be detected and regulated in education (for snapshots of student-written and ChatGPT-generated reflections, see Table 1).

3. Method

3.1. Prompting strategies

To empower the ChatGPT model to generate a diverse set of reflective written responses, based on prior research (detailed in Section 2), we identified two representative prompt categories: `Generate by template` and `Regenerate to improve`. For the former, we included three widely-used templates as a basis to formulate a prompt: `Query`, `Example`, and `Context`. For the latter, we included two regeneration strategies: `CoT` (chain of thought) and `RoR` (reflect on reflection), which are detailed below. Importantly, each of the template prompts was coupled with a regeneration strategy and resulted in a total of 9 prompts: 1) `Query`, 2) `Example`, 3) `Context`, 4) $\texttt{Query}_{\texttt{CoT}}$, 5) $\texttt{Query}_{\texttt{RoR}}$, 6) $\texttt{Example}_{\texttt{CoT}}$, 7) $\texttt{Example}_{\texttt{RoR}}$, 8) $\texttt{Context}_{\texttt{CoT}}$, 9) $\texttt{Context}_{\texttt{RoR}}$. We generated 100 responses per prompt, resulting in a total of 900 reflective responses generated by ChatGPT.

- `Generate by template`
  - `Query` (Kung et al., 2022) contains just enough information to generate an appropriate reflective writing. For example, *Can you generate a reflective writing for pharmacy students?*
  - `Example` (Liu et al., 2021) contains a random student-written example and a query, e.g., *"I recently started a position at XXX...". This is an example of reflective writing. Based on this example, can you show me another one?*
  - `Context` (Wang et al., 2022) consists of a specific date, scenario and personality (randomly selected from words that described the Big Five personality traits in (Goldberg, 1990)), e.g., *On XXX, I started my internship at YYY. Based on this context, can you show me an example of reflective writing for a **sensitive** pharmacy student?*
- `Regenerate to improve`
  - `CoT` (Zhang et al., 2022) contains manually identified potential improvement based on CoT reasoning steps, e.g., *Can you provide the chain of thought on the original example you showed me? Can you revise the original example and elaborate more about XXX in the chain of thought?*[2]

[2] The regeneration prompt is manually constructed by the teaching staff based on CoT-reasoning.

Table 2
The descriptive statistics of the reflective writing dataset. **Avg uniq. w /sent** is the average number of unique words per sentence. **Student written** are reflections written by students, while all other columns are ChatGPT-generated by a particular prompt. More statistics are available in Appendix A.

Statistics	Student written	Query	Example	Context	Query		Example		Context	
					CoT	RoR	CoT	RoR	CoT	RoR
No. reflections	900	100	100	100	100	100	100	100	100	100
Avg Words	344.02	253.96	199.78	332.88	300.24	295.78	227.28	284.56	378.34	423.56
Avg unique words	166.04	128.48	109.90	151.02	146.18	145.80	120.70	146.10	167.06	184.16
Avg sentence	13.21 (5.93)	9.92 (1.61)	8.78 (2.56)	13.48 (2.80)	11.48 (2.96)	11.58 (2.72)	9.22 (2.91)	11.48 (2.69)	15.00 (3.19)	16.68 (3.33)
Avg words /sent	26.04 (12.30)	25.60 (6.69)	22.75 (7.36)	24.69 (7.71)	26.15 (7.72)	25.54 (7.23)	24.65 (9.55)	24.79 (8.47)	25.22 (8.00)	25.39 (7.89)
Avg uniq. w /sent	22.80 (9.00)	22.76 (5.08)	20.54 (5.66)	21.90 (5.80)	23.14 (5.82)	22.74 (5.66)	21.93 (6.99)	22.11 (6.44)	22.30 (5.97)	22.48 (5.91)

Table 3
Statistical test for comparing the number of words per reflective writing between student-written and ChatGPT-generated reflections.

Comparison between		KWH p-value	MWU U-value	MWU P-value[*]	MWU effect size
Student Written	*Query*	<0.0001	15504.0	<0.0001	-0.3812
	Example	<0.0001	18329.0	<0.0001	-0.6329
	Context	0.5601	10661.5	0.7197	0.0502
	$Query_{CoT}$	0.0947	12841.0	0.0474	-0.1440
	$Query_{RoR}$	0.0765	12938.0	0.0383	-0.1526
	$Example_{CoT}$	<0.0001	16783.0	<0.0001	-0.4951
	$Example_{RoR}$	0.0162	13550.0	0.0080	-0.2071
	$Context_{CoT}$	0.0020	8238.5	0.9991	0.2661
	$Context_{RoR}$	<0.0001	6348.5	1.0000	0.4344
	ChatGPT-Generated Overall	0.0003	115193.5	0.0001	-0.1402

KWH - Kruskal-Wallis H-test; MWU - Mann-Whitney U test.

[*] The Mann-Whitney U test conducted is one-sided, with the Null Hypothesis being "the distribution underlying the no. words for student-written reflection is stochastically greater than that of the ChatGPT-generated ones".

– `RoR` (Susnjak, 2022) asks ChatGPT to automatically improve based on an evaluation of the generated initial response, e.g., *Can you evaluate the original example you showed and identify areas of improvement? Can you revise the original example based on the areas of improvement you identified?*

3.2. Dataset

We collected 14,908 reflections written by 1,321 Bachelor's and Master's students in the Faculty of Pharmacy and Pharmaceutical Sciences at Monash University between 2017 and 2019[3]. Students were instructed to write reflections about various topics (e.g., exams, internships and placements) in which students were asked to describe their prior learning experiences (e.g., rehearsing a skill for a practical exam) and reflect on feelings related to those learning experiences (Driscoll, 2006). As this was a coaching program where students obtained suggestions for future studying from their skill coaches, students were advised to provide concrete and practical plans when applicable about how they would improve their skills based on what they reflected upon. Then, we utilised ChatGPT to generate a total of 900 reflections using a total of 9 different prompts (as detailed in Section 3.1). To match this number and create a balanced dataset (i.e., ChatGPT-generated vs. student-written), we randomly sampled 900 out of 14,908 student reflections. The dataset statistics are summarised in Table 2. More statistics are included in Appendix A.

Based on the descriptive statistics and the statistical tests in Table 3, we found that ChatGPT-generated reflections were statistically significantly shorter compared to the student-written ones, except for those by `Context`-based prompts. However, on the sentence level (i.e., the last two rows in Table 2), the difference in the number of words and unique words in a sentence was small between the student-written and ChatGPT-generated ones. ChatGPT-generated reflections, in general, had a lower standard deviation, indicating a lower level of variability in terms of the number of sentences, words, and unique words compared to student-written ones.

3.3. Evaluation approach

RQ 1: Assessing the quality of the reflection. To assess the quality of reflective writing, we have adapted a marking rubric widely used in pharmacy education from (Tsingos et al., 2015) which roots in well-established theories (Boud et al., 1985, Mezirow, 1991). Based on the rubric,[4] a score ranging from a minimum of 0 and a maximum of 6 was given to a reflective written response based on the following 6 categories (detailed description in Table 4): **returning to experience**, **attending to feelings**, **association**, **integration**, **validation** and **appropriation**. Since the category **outcome of reflection** (theorised in Boud et al.'s reflection model (Boud et al., 1985)) relates more to students' set goals for their future studying, which deserves to be systematically and comprehensively evaluated in terms of their qualities, we have excluded this category during the assessment and leave such evaluations to our future research.

To ensure the reliability of the markings, we involved two teachers with prior course teaching experience at `anonymized` University. Prior to the marking session, the teaching staff went through a 2-hour training session, where the details of the reflection assignment and the marking rubric were introduced for the first half of the session. The rest of the session was used for a mock marking session with 10 reflective writing examples (5 student-written and 5 ChatGPT-generated). After the training session, both teaching staff were assigned the same set of 90 reflective writings, where 45 of them were generated by ChatGPT (5 per prompting strategy) and another 45 were randomly selected from 900 student-written reflections. Lastly, the marking differences were resolved via a one-hour discussion session, and a final score was given to a response.

[3] The collection of students' reflective writings was approved by Monash University under project code 18768.

[4] A complete description of the assessment criteria is provided in Table 12 in Appendix.

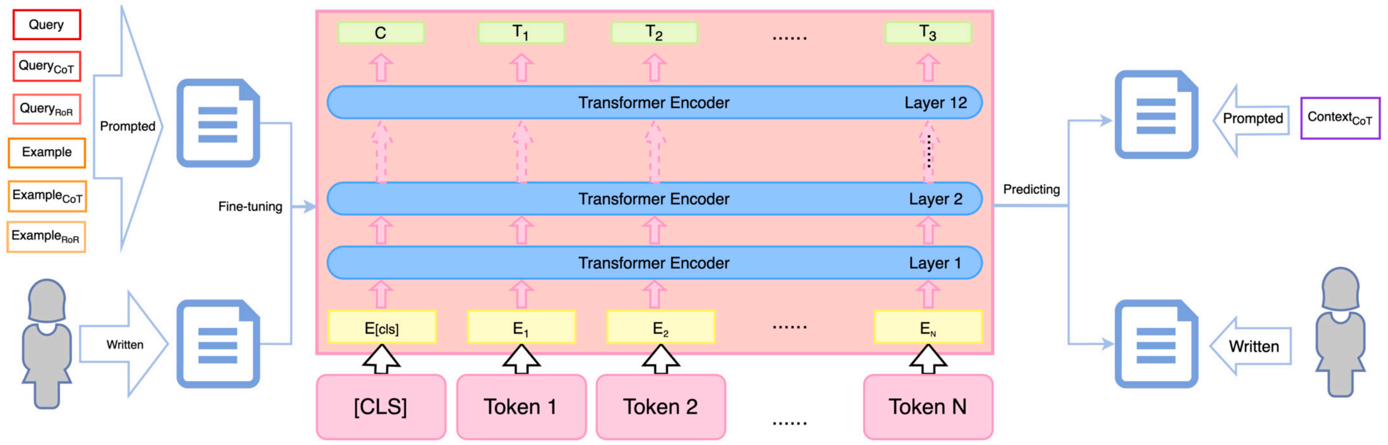
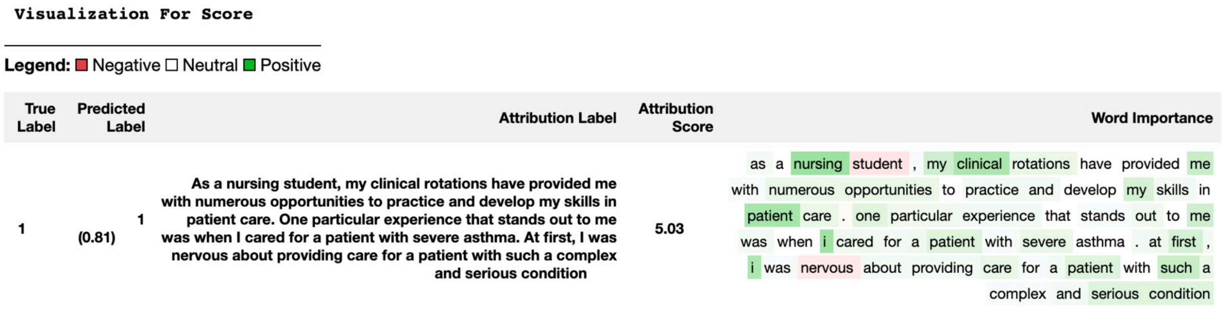


Fig. 1. Example of Model Training without A-priori Known Prompts.



Visualization For Score

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	1 (0.81)	As a nursing student, my clinical rotations have provided me with numerous opportunities to practice and develop my skills in patient care. One particular experience that stands out to me was when I cared for a patient with severe asthma. At first, I was nervous about providing care for a patient with such a complex and serious condition	5.03	as a nursing student , my clinical rotations have provided me with numerous opportunities to practice and develop my skills in patient care . one particular experience that stands out to me was when i cared for a patient with severe asthma . at first , i was nervous about providing care for a patient with such a complex and serious condition

Fig. 2. An example explanation based on Layer Integrated Gradients.

RQ 2: Identifying Human vs. ChatGPT-generated reflections. To understand the extent to which ChatGPT-generated reflections could be identified in an authentic educational assessment scenario, we first asked two teaching staffs, who were not previously involved in the study, to indicate which responses were generated by ChatGPT by assigning a binary label: 1 being ChatGPT-generated and 0 being student-generated, for a test set containing a total of 360 responses (20% of 1,800 responses), where 180 responses were randomly sampled out of 900 ChatGPT-generated responses, and 180 responses were randomly sampled out of 900 student-written responses. Driven by the predictive power of the state-of-the-art deep learning classification methods, which have been documented to outperform humans in recognising distinct patterns in natural language corpora (He et al., 2020, Liu et al., 2019), we built a classifier based on the BERT pre-trained language model – one of the commonly adopted models for classification tasks in educational research (Sha et al., 2021, 2022). In doing so, we aimed to examine the extent to which the reflective responses generated by ChatGPT could be computationally differentiated from the reflective responses written by students. To this end, we treated the differentiation between student and ChatGPT-generated reflections as a binary classification task, where a ChatGPT-generated response was assigned a label 1 and a student-written response was assigned a label 0. We benchedmarked the BERT classifier to the latest OpenAI-released RoBERTa-based classifier (which was purposely built with general-domain datasets to distinguish between AI- and human-generated content OpenAI (2023)). However, the OpenAI classifier does not produce binary results (i.e., the model identified the texts as “very unlikely”, “unlikely”, “unclear if it is”, “possibly”, or “likely” AI-generated). Thus, for ChatGPT-generated reflections, we considered the OpenAI classifier to be correct in the cases of “possibly” and “likely” predictions; for student-written reflections, we regarded the “very unlikely” and “unlikely” predictions as the correct results. To enable the comparison between teachers’ manual classification and the LLM-based classifiers, the performance of LLM classification in the `Overall` scenario was tested using the same set of responses (i.e., 360) that the teachers have previously manually classified. The remaining responses were used for model training (i.e., to fine-tune the BERT classifier).

However, this way of training-testing implied that BERT would have access to the samples generated by the same prompt as the testing samples since the training and testing samples would be randomly selected from a pool that contained the samples generated by all prompts. To take into account a scenario where the BERT model was not aware of the prompt used by a student at the time of training, we separately trained BERT with only the samples generated by the prompts other than the testing prompt, combined with the student-written reflections. For instance, `Context`$_{\texttt{CoT}}$ samples were tested by the BERT model trained using the student-written reflections and samples generated by: `Example`, `Query`, `Example`$_{\texttt{CoT}}$, `Example`$_{\texttt{RoR}}$, `Query`$_{\texttt{CoT}}$ and `Query`$_{\texttt{RoR}}$ (visualised in Fig. 1). Therefore, we could gain insights into the real-world scenario where the prompt used by students to generate reflections (i.e., which is equivalent to the testing reflections) was not known by the instructor during training on the BERT model.

To gain a deeper understanding of the classification results, we have adopted a state-of-art explainer model based on Integrated Gradients, which is widely used to identify which words have a high contribution to a model’s prediction (Sundararajan et al., 2017). An explanation about a `Example` prompt-generated reflection being correctly predicted as ChatGPT-generated (predicted label = 1 with probability = 0.81) is shown in Fig. 2, where the attribute scores indicate the total contribution of input text towards a task label prediction. The words highlighted in green indicate a positive contribution to the ChatGPT-generated task label, and red indicates student-written task labels.

Table 4
The quality score of ChatGPT-generated and student-written reflections across 6 criteria based on the levels of reflection. The number in the bracket indicates the standard deviation of the score.

Criteria (max. mark)	Description	Human	ChatGPT
Overall (6)	Overall mark of all 6 criterias.	2.900 (1.7242)	**4.300 (0.9966)**
Returning to Experience (1)	Experience was clearly described (e.g., chronological information or personal judgements).	0.6500 (0.4810)	**0.9500 (0.2198)**
Attending to Feelings (1)	Personal feelings were described with judgements/reasons provided.	0.6167 (0.4903)	**0.7833 (0.4155)**
Association (1)	Links between prior knowledge, feelings or attitudes, and newly acquired knowledge.	0.6500 (0.4810)	**0.9333 (0.2515)**
Integration (1)	Association between prior and new knowledge, feelings or attitude, and new insights.	0.5500 (0.5017)	**0.8667 (0.3428)**
Validation (1)	Self-assessment of the new insights provided, with reference to prior experience.	0.0167 (0.1291)	**0.0500 (0.2198)**
Appropriation (1)	New insights related to current life and/or future development.	0.4167 (0.4972)	**0.7167 (0.4544)**

3.4. Study setup

Pre-processing. We pre-processed the text contained in reflective writing by performing the following steps: 1) removing invalid characters (e.g., \, *, _, \n, \t), and 2) applying lowercasing. In line with previous studies (Sha et al., 2021), we randomly split the dataset using the 80%:20% train-test ratio, where 20% was used for testing. The training set was further split in the same ratio, with 80% used for model training and 20% for validation during the training. As detailed in Section 3.3, our BERT evaluation is two-fold. First, for the testing of the `Overall` setting, the training and testing data were randomly sampled from all student-written and ChatGPT-generated reflections. Second, for the testing of reflections generated by an individual prompt, the testing samples were restricted to those generated by a particular prompt and student-written ones, while the training samples were reflections generated by other prompts (as detailed in Section 3.3) and student-written ones. We randomly under-sampled the training samples to have a consistent 80%:20% ratio in all cases.

Model implementation. The BERT models were implemented using the hugging-face library.[5] In response to the call for increased reproducibility in educational research (Gardner et al., 2019) and enable practitioners to benefit from our study, we open-source code used in this study.[6]

Evaluation metrics. In line with previous studies (Sha et al., 2021), we adopted the following four metrics to measure the classification accuracy of the correct label (student written vs. ChatGPT generated): Accuracy, Cohen's κ, AUC, and F1 score.

4. Results

4.1. Assessing the quality of ChatGPT-generated reflective writing (RQ1)

To assess the quality of reflective writing, it is critical to examine a reflective response relative to multiple criteria (i.e., reflection levels (King, 2002)). Therefore, we first obtained the comparative difference across the 6 reflection criteria between the ChatGPT-generated and student-written responses, including **returing to Experience**, **attending to Feelings**, **association**, **integration**, **validation**, and **appropriation**. This procedure is detailed in Table 4.

We observed that ChatGPT-generated responses were assigned higher scores across all criteria (6 out of 6) with lower standard deviation values across most criteria (except for **validation**). This indicates that the ChatGPT-generated reflective responses considerably outperformed the responses written by the students in a pharmacy course and that ChatGPT-generated reflective responses were consistently deeper in terms of reflection levels.

Table 5
Classification performance of human evaluators.

Metrics	Is Machine	
	Tutor 1	Tutor 2
Accuracy	0.5489	0.6800
F1	0.5333	0.7241
AUC	0.5491	0.6705

4.2. Differentiating written reflective responses (RQ2)

Differentiation by human evaluators. As shown in Table 5, the manual classification accuracy for the two human evaluators was 0.55 and 0.68, respectively, with a Cohen's κ agreement score of 0.06, indicating none to a slight agreement among the evaluators. This finding indicates that teaching staff were not able to accurately and consistently differentiate between ChatGPT and student-written reflections.

Differentiation by automatic text classifiers.

Given the fact that human evaluators fail to accurately classify such reflections, inspired by the power of state-of-the-art pre-trained LLMs in educational text classification (Yan et al., 2023), we exploited BERT to further explore its capability in differentiating ChatGPT-generated reflections from student-written ones.

We first investigated the performance of LLM-based classifiers when the training data consisted of AI-generated samples randomly sampled from reflections generated by *all the prompts* and human-generated samples randomly drawn from students' reflective writings (as shown in the `Overall` column).

As shown in Table 6, OpenAI's classifier performed poorly in distinguishing ChatGPT-generated reflections from student-written reflections while the BERT model performed significantly better. Although the OpenAI classifier had a high true positive rate, the false positive rate is also high, suggesting that the classifier has the tendency to classify the content as AI-generated on most occasions. This finding is expected as reflective writing in the pharmacy domain is a domain-specific task, which can be difficult for a general-domain classifier (further elaborated in the discussion). Therefore, training domain-specific classifier could potentially achieve better results. Indeed, the BERT model classified reflections with much higher accuracy (up to 0.99 across four metrics as shown in Table 6), indicating that, unlike humans and classifiers trained on general-domain datasets, a BERT model trained with in-domain data was able to satisfactorily differentiate ChatGPT-generated reflections from students-written ones.

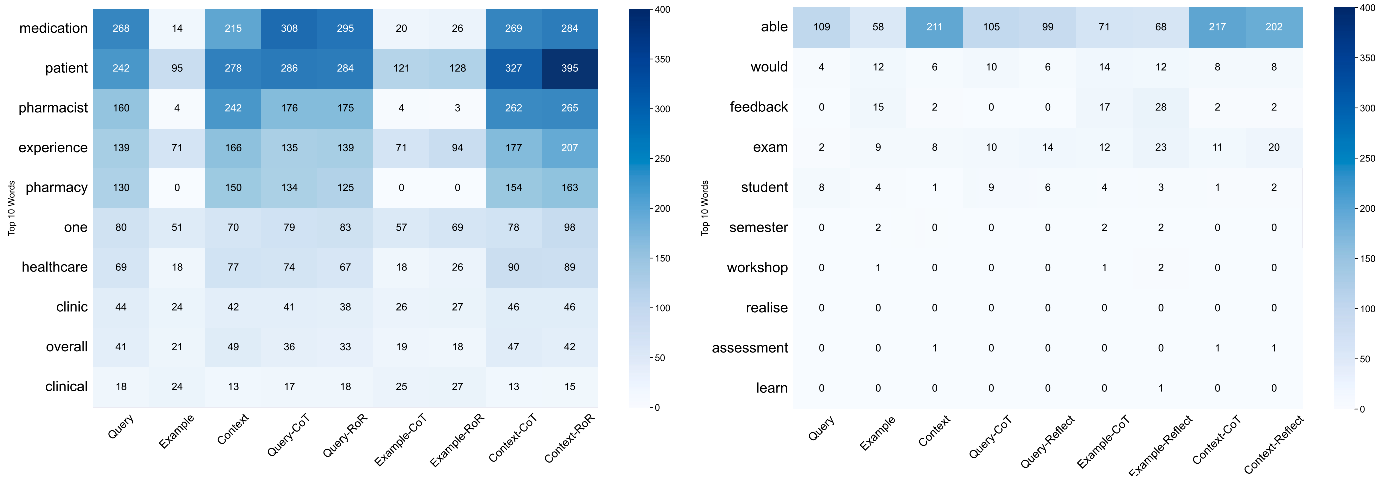
However, in the real world, it may be difficult to know what prompts students would use in advance, so the training samples of BERT may not contain those generated by the prompt applied by students. To take into account such scenarios, the rest of the columns (i.e., other than `Overall`) of the Table 6 are based on the BERT model trained without the use of reflection generated by the same prompt. We found

[5] https://huggingface.co/.

[6] https://github.com/anonymised_for_review.

Table 6
OpenAI AI text classifier performance and BERT classification result on Overall and individual prompts across four metrics. The worst-performing 3 results of our models were in **bold**.

Metrics	OpenAI	Overall	Query	Example	Context	Query		Example		Context	
						CoT	RoR	CoT	RoR	CoT	RoR
Accuracy	0.5625	0.9944	0.9900	**0.9100**	0.9800	0.9900	0.9900	**0.9100**	**0.8600**	0.9800	0.9800
F1	0.6847	0.9889	0.9800	**0.8200**	0.9600	0.9800	0.9800	**0.8200**	**0.7200**	0.9600	0.9600
AUC	0.5625	0.9944	0.9900	**0.9100**	0.9800	0.9900	0.9900	**0.9100**	**0.8600**	0.9800	0.9800
Kappa	0.1250	0.9945	0.9910	**0.9011**	0.9804	0.9910	0.9910	**0.9011**	**0.8372**	0.9804	0.9804



(a) Top-words for ChatGPT-generated

(b) Top-words for students-written

Fig. 3. A summary of explanations for correctly-classified ChatGPT-generated reflections. The top 10 most frequently appeared words are on the Y-axis, while the X-axis lists the promoting strategy.

that, even without the samples generated by the same prompt, BERT's classification performance only dropped slightly to 96%–99% measured across the four metrics, with the exception of prompts based on `Example`, which had a performance of 72%–91% across four metrics. Still, about half of the results were above 90% in the reflections generated by `Example` prompts, indicating that even without the use of samples generated by the same prompts (out-of-sample prediction), a domain-specific BERT-based model can still well-surpass the human evaluator and general-domain classifiers in identifying ChatGPT-generated reflections.

Explaining BERT classifications. Given such a high difference in classification performance between the human evaluator and BERT classifier, we further scrutinised BERT's classification results by adopting a widely-used black-box explainer to gain more insights about BERT's classification (detailed in Section 3.3). We summarised the frequency of the top-10 most important words to a "ChatGPT-generated" classification label (in Fig. 3a) and "student-written" classification label (in Fig. 3b) using the `Overall` classifier. We observed that BERT regarded words related to the pharmacy setting (e.g., medication, patient, and pharmacist) as important to a ChatGPT-generated classification label, while for a student-written classification label, themes mostly centred around coursework (e.g., exam, assessment) and action/verbs (able, realise). Besides, there were 16 instances where the top 10 words appeared more than 200 times (the number in the box) for the ChatGPT-generated, while only 3 for the student-written ones. In fact, most of the important words in the classification of the student-written labels were low in frequency, indicating a more sparse word choice by students compared to that of ChatGPT.

Given such a high difference in classification performance between the human evaluators and BERT classifiers, we further scrutinised BERT's classification results by adopting a widely-used black-box explainer to gain more insights about BERT's classification (detailed in Section 3.3). We summarised the top 10 most contributing words for identifying ChatGPT-generated reflective writings and student-written ones, respectively, and counted their occurrence in the reflective writings scaled by their corresponding sample sizes. We observed that BERT deemed words closely related to the pharmacy setting (e.g., medication, patient, and pharmacist) as important to a ChatGPT-generated classification (in Fig. 3a). These words appeared more frequently in ChatGPT-generated texts in contrast to student-written ones, except for those generated by `Example`-based prompts, suggesting its potential in generating more diverse reflective writings and providing explanations for the relatively lower performance in classifying such texts. While for a student-written classification, themes mostly centred around coursework (e.g., exam, assessment, semester, workshop) were considered significant, with a relatively high occurrence in student-written reflections (in Fig. 3b). However, these important words were not as frequent in students' reflective writings compared to the top 10 words for ChatGPT-generated texts, indicating that students' word choices were more sparse compared to that of ChatGPT.

5. Discussion

We investigated ChatGPT's capability in generating high-quality reflective responses in a university pharmacy course and whether the generated reflective responses can be distinguished from those written by students. For the first research question, we found that ChatGPT is capable of generating high-quality reflections, outperforming student-written reflections in all the assessment criteria. Such findings were expected as ChatGPT were trained explicitly for completing complex natural language tasks and have a vast amount of internet knowledge, which could contain numerous examples of reflective writings.

This finding resonates with prior studies that demonstrated ChatGPT's capabilities in completing authentic educational assessments, such as passing medical (Gilson et al., 2023), parasitology (Huh, 2023), and law exams (Choi et al., 2023). However, our findings further suggested that for reflective writing tasks, ChatGPT can outperform students. Such high performance could be related to the nature of the tasks, as reflective writing is entirely textual and content-based, which can be more easily completed by large language models and LLM-based tools like ChatGPT. Whereas other exams may contain a combination of different task modalities, such as graphs and figures, that ChatGPT and language models may struggle with (Huh, 2023).

In terms of the second research question, we identified that human evaluators could not accurately differentiate ChatGPT-generated reflections from students-generated reflections. This finding resonates with the growing concerns among institutions and educators regarding the potential academic misconduct that may occur when students are using ChatGPT to complete assessments, especially for generating written assessments (Kung et al., 2022, Susnjak, 2022, Yan et al., 2023). The poor accuracy and low agreement score also indicate that each evaluator has their own internal assumptions of how to identify ChatGPT-generated text, and unfortunately, these assumptions are often incorrect because ChatGPT-generated and student-written reflections can share similar semantic structures/patterns. On the other hand, the performance of automatic detection tools varies. The poor performance of the OpenAI classifier resonates with prior findings, suggesting that such general-domain classifiers can perform poorly on domain-specific tasks and tend to incorrectly classify human-written texts as AI-generated (Hu, 2023). Whereas the impressive out-of-sample performance of the BERT classifier illustrates the potential effectiveness of using classifiers trained with domain-specific data for differentiating ChatGPT-generated from student-generated reflective writings. Such a significant difference in performance may be owing to general-domain texts being syntactically and semantically more diverse compared to domain-specific texts.

5.1. Implications

The ChatGPT's capability in generating high-quality reflections (outperforming students in all criteria as shown in Section 4.1) while being indistinguishable by experienced teaching staff (around 0.61 accuracies when manually classifying ChatGPT-generated vs. student written reflections) may provide concerning motivations for students to exploit ChatGPT to achieve a high grade with minimum effort. But this may deprive them of developing critical thinking and idea-synthesising skills that were purposefully designed for the reflective writing task, leading to fewer learning opportunities. Besides, given the ChatGPT's capability to generate high-quality reflective writing, broader written assessment tasks in higher education may be at risk, which may present challenges for educators and practitioners to prevent the undesirable use of ChatGPT across all written tasks. One potential approach to prevent such academic misconduct is using classifiers trained with domain-specific data. We showed that even the experienced teaching staff could not remotely match the classification accuracy of a BERT-based classifier. This classifier was able to effectively detect the ChatGPT-generated reflections with high accuracy – even in testing out-of-sample scenarios where a different prompt was used to generate training reflection samples. Given such a stark performance difference between human and machine classification, we posit that computational classifiers (such as the BERT classifier evaluated in the present study) are needed to address the challenges presented by ChatGPT potentially beyond the task of reflective writing, so as to facilitate the detection and regulation of the use of ChatGPT in education. However, such classifiers as the BERT-based one used in the current study are not without flaws. For instance, reflections generated by `Example` prompts are harder to identify correctly. Besides, ChatGPT's generative capability may further advance. This highlighted the need for the development of more advanced classifiers to counter ChatGPT in education. Lastly, to our surprise, the state-of-the-art AI-text detector released by OpenAI performed poorly compared to the BERT classifiers. This may be due to the fact that the OpenAI classifier was trained on the general domain corpus, and thus, is not capable of handling domain-specific datasets (e.g., reflective writings in pharmaceutical curricula) in terms of differentiating between AI-generated and human-written texts, which further highlights the need for building domain-specific classifiers.

5.2. Limitations

We acknowledged the following limitations of our study. First, we generated 100 reflective writings for each of the 9 prompts, resulting in a total of 900 ChatGPT-generated reflective writings. To validate the findings for future work, a larger scale evaluation may be needed using a larger sample size and prompts. Secondly, we focused on reflective writing in a pharmacy course setting in higher education. We acknowledge that other course settings within the same domain and across different domains, and types of written assessment (e.g., essays) should also warrant a similar evaluation. For future work, we plan to replicate the present study in other course settings and written assessments to validate our findings.

6. Conclusion

In this study, we investigated the effectiveness of ChatGPT in generating reflective writing and the potential challenges it poses to academic integrity in education. The results showed that ChatGPT can generate high-quality reflective writing that outperforms student-written reflections in all assessment criteria. However, human evaluators could not accurately differentiate ChatGPT-generated reflections from students' original work, highlighting the need for effective approaches to distinguish between the two. We also demonstrated the potential of using computational classifiers trained with domain-specific data to detect and regulate the use of ChatGPT in education. While ChatGPT's capabilities in generating high-quality content are promising, educators and practitioners must be cautious about its potential negative impacts on educational writing tasks and academic integrity. Further research is needed to explore the implications of using ChatGPT in education and develop effective strategies to regulate its use.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Descriptive statistics of the reflective writing dataset

Table 7
Descriptive statistic w.r.t. the number of words per reflective writing for student-written and ChatGPT-generated reflections.

	Number of Words per Reflective Writing						
	Average	Standard Deviation	Max.	Min.	Median	25th Percentile	75th Percentile
Student Written	344.02	159.32	1,109.00	38.00	321.00	238.00	418.00
Query	253.96	38.37	356.00	174.00	247.50	227.00	273.75
Example	199.78	51.66	343.00	102.00	189.50	166.50	236.00
Context	332.88	66.07	494.00	179.00	337.00	286.25	382.00
Query-CoT	300.24	78.90	552.00	193.00	277.00	251.50	352.00
Query-RoR	295.78	66.19	462.00	162.00	283.50	249.75	347.25
Example-CoT	227.28	65.45	413.00	96.00	223.50	184.25	272.25
Example-RoR	284.56	54.89	400.00	166.00	296.00	242.25	322.25
Context-CoT	378.34	74.10	523.00	227.00	378.50	329.75	428.00
Context-RoR	423.56	84.98	677.00	260.00	425.50	364.00	481.50
ChatGPT-Generated Overall	299.60	93.88	677.00	96.00	283.50	231.25	362.50

Table 8
Descriptive statistics w.r.t. the number of unique words per reflective writing for student-written and ChatGPT-generated reflections.

	Number of Unique Words per Reflective Writing						
	Average	Standard Deviation	Max.	Min.	Median	25th Percentile	75th Percentile
Student Written	166.04	58.78	441.00	32.00	157.00	125.00	197.00
Query	128.48	14.95	174.00	94.00	125.50	119.50	137.00
Example	109.80	20.82	160.00	67.00	109.00	96.50	122.00
Context	151.02	22.50	206.00	94.00	149.00	137.00	167.00
Query-CoT	146.18	26.30	218.00	103.00	140.50	129.00	160.50
Query-RoR	145.80	24.72	203.00	96.00	145.00	128.25	161.00
Example-CoT	120.70	25.68	183.00	62.00	120.50	103.25	140.50
Example-RoR	146.10	20.87	185.00	101.00	146.50	132.00	163.50
Context-CoT	167.06	23.88	216.00	124.00	166.00	155.25	182.00
Context-RoR	184.16	24.39	253.00	133.00	184.00	168.75	199.75
ChatGPT-Generated Overall	144.37	31.42	253.00	62.00	141.00	122.25	166.00

Table 9
Descriptive statistics w.r.t. the number of sentences per reflective writing for student-written and ChatGPT-generated reflections.

	Number of Sentences per Reflective Writing						
	Average	Standard Deviation	Max.	Min.	Median	25th Percentile	75th Percentile
Student Written	13.21	5.93	40.00	1.00	12.00	9.00	16.00
Query	9.92	1.61	14.00	7.00	10.00	9.00	11.00
Example	8.78	2.56	15.00	4.00	9.00	7.00	11.00
Context	13.48	2.80	20.00	8.00	14.00	11.25	15.00
Query-CoT	11.48	2.96	19.00	7.00	11.00	9.00	13.00
Query-RoR	11.58	2.72	19.00	7.00	11.00	9.25	13.00
Example-CoT	9.22	2.91	18.00	4.00	9.00	7.25	11.00
Example-RoR	11.48	2.69	18.00	7.00	12.00	10.00	13.00
Context-CoT	15.00	3.19	21.00	9.00	15.00	12.25	17.00
Context-RoR	16.68	3.33	24.00	11.00	17.00	14.25	19.00
ChatGPT-Generated Overall	11.96	3.75	24.00	4.00	11.00	9.00	14.00

Table 10
Descriptive statistics w.r.t. the number of words per sentence for student-written and ChatGPT-generated reflections.

	Number of Words per Sentence						
	Average	Standard Deviation	Max.	Min.	Median	25th Percentile	75th Percentile
Student Written	26.04	12.30	192.00	2.00	24.00	18.00	32.00
Query	25.60	6.69	49.00	9.00	25.00	21.00	30.00
Example	22.75	7.36	72.00	6.00	22.00	18.00	27.00
Context	24.69	7.71	58.00	8.00	24.00	19.00	30.00
Query-CoT	26.15	7.72	60.00	8.00	25.00	21.00	31.00
Query-RoR	25.54	7.23	52.00	8.00	25.00	20.00	30.00
Example-CoT	24.65	9.55	106.00	9.00	23.00	18.00	29.00
Example-RoR	24.79	8.47	63.00	6.00	24.00	19.00	30.00
Context-CoT	25.22	8.00	60.00	8.00	25.00	19.00	30.75
Context-RoR	25.39	7.89	63.00	8.00	25.00	20.00	30.00
ChatGPT-Generated Overall	25.05	7.92	106.00	6.00	24.00	20.00	30.00

Table 11
Descriptive statistics w.r.t. the number of unique words per sentence for student-written and ChatGPT-generated reflections.

	Number of Unique Words per Sentence						
	Average	Standard Deviation	**Max.**	Min.	Median	25th Percentile	75th Percentile
Student Written	22.80	9.00	104.00	2.00	22.00	17.00	28.00
Query	22.76	5.08	42.00	9.00	23.00	20.00	26.00
Example	20.54	5.66	54.00	6.00	20.00	17.00	24.00
Context	21.90	5.80	44.00	8.00	22.00	18.00	26.00
Query-CoT	23.14	5.82	48.00	8.00	23.00	19.00	27.00
Query-RoR	22.74	5.66	41.00	8.00	23.00	19.00	26.00
Example-CoT	21.93	6.99	74.00	9.00	21.00	17.00	26.00
Example-RoR	22.11	6.44	43.00	6.00	22.00	18.00	27.00
Context-CoT	22.30	5.97	49.00	8.00	22.00	18.00	26.00
Context-RoR	22.48	5.91	48.00	8.00	22.00	18.00	26.00
ChatGPT-Generated Overall	22.26	5.97	74.00	6.00	22.00	18.00	26.00

Appendix B. Assessment criteria for reflective writings

Table 12
The descriptions to the assessment criteria used to assess the levels of reflection for student-written and ChatGPT-generated reflections.

Criteria	Non-Reflective (0)	Reflective (1)
Returning to Experience	Experience was not clearly described	Experience was clearly described (the description may include chronological information or personal judgements)
Attending to Feelings	Personal feelings were not described or were described without evaluation	Personal feelings were described with judgements/reasons provided
Association	No links between prior knowledge, feelings or attitudes, and newly acquired knowledge, feelings or attitudes	Links between prior knowledge, feelings or attitudes, and newly acquired knowledge, feelings or attitudes
Integration	Association between prior and new knowledge, feelings or attitude provided, but new insights not provided	Association between prior and new knowledge, feelings or attitude provided, and new insights provided
Validation	No self-assessment of the new insights or no reference to prior experience provided	Self-assessment of the new insights provided, with reference to prior experience
Approximation	New insights not related to current life and/or future development	New insights related to current life and/or future development (may specifically infer the impact of the connections)

References

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. Available at SSRN 4308687.

Boud, D., Keogh, R., & Walker, D. (1985). Promoting reflection in learning: A model. In *Boundaries of adult learning 1* (pp. 32–56).

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

Castelvecchi, D. (2022). Are ChatGPT and AlphaCode going to replace programmers? *Nature*.

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

Charon, R., & Hermann, M. N. (2012). A sense of story, or why teach reflective writing? Academic medicine. *Journal of the Association of American Medical Colleges*, *87*, 5.

Choi, J.H., Hickman, K.E., Monahan, A., & Schwarcz, D. (2023). Chatgpt goes to law school. Available at SSRN.

Das, D., Kumar, N., Longjam, L. A., Sinha, R., Roy, A. D., Mondal, H., & Gupta, P. (2023). Assessing the capability of chatgpt in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*, *15*.

Dennison, W.F., & Kirk, R. (1990). Do, review, learn, apply: A simple guide to experiential learning. Blackwell Education.

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

Dewey, J. (1933). How we think: A restatement of the relation of reflective thinking to the educative process. *American Journal of Psychology*, *46*, 528.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A survey of natural language generation. *ACM Computing Surveys*, *55*, 1–38.

Driscoll, J. (2006). *Practising clinical supervision: A reflective approach for healthcare professionals*. Elsevier Health Sciences.

Gardner, J., Yang, Y., Baker, R. S., & Brooks, C. (2019). *Modeling and experimental design for mooc dropout prediction: A replication perspective*In *EDM*. Springer.

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

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of personality and social psychology*, *59*, 1216.

Gu, X., Yoo, K. M., & Lee, S. W. (2021). Response generation with context-aware prompt learning. arXiv preprint, arXiv:2111.02643.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint, arXiv:2006.03654.

Hou, Y., Dong, H., Wang, X., Li, B., & Che, W. (2022). Metaprompting: Learning to learn better prompts. arXiv preprint, arXiv:2209.11486.

Hu, G. (2023). Challenges for enforcing editorial policies on ai-generated papers. *Accountability in Research*, 1–3.

Huh, S. (2023). Are chatgpt's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *Journal of Educational Evaluation for Health Professions*, *20*, 1.

King, T. (2002). Development of student skills in reflective writing. In *4th ICED*.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development, Vol. 1*.

Kung, T.H., Cheatham, M., Medinilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. et al. (2022). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. medRxiv.

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

Lew, M. D., & Schmidt, H. G. (2011). Self-reflection and academic performance: Is there a relationship? *Advances in Health Sciences Education*, *16*, 529–545.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What makes good in-context examples for gpt-3? arXiv preprint, arXiv:2101.06804.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint, arXiv:1907.11692.

Macdonald, C., Adeloye, D., Sheikh, A., & Rudan, I. (2023). Can chatgpt draft a research article? An example of population-level vaccine effectiveness analysis. *Journal of Global Health*, *13*, Article 01003.

Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in health sciences education*, *14*, 595–621.

Mezirow, J. (1991). *Transformative dimensions of adult learning*. London, England: ERIC.

OpenAI. https://beta.openai.com/ai-text-classifier (2023).

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*. 10776958221149577.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 9.

Raković, M., Bernacki, M. L., Greene, J. A., Plumley, R. D., Hogan, K. A., Gates, K. M., & Panter, A. T. (2022). Examining the critical role of evaluation and adaptation in self-regulated learning. *Contemporary Educational Psychology*, *68*, Article 102027.

Ryan, M. (2011). Improving reflective writing in higher education: A social semiotic perspective. *Teaching in Higher Education*, *16*, 99–111.

Sen, B. A. (2010). *Reflective writing: A management skill*. *Library Management*.

Sha, L., Rakovic, M., Li, Y., Whitelock-Wainwright, A., Carroll, D., Gaševic, D., & Chen, G. (2021). Which hammer should I use? A systematic evaluation of approaches for classifying educational forum posts. *International Educational Data Mining Society*.

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

Sharples, M. (2022). Automated essay writing: An aied opinion. *International Journal of Artificial Intelligence in Education*, *32*, 1119–1126.

Sinha, R. K., Roy, A. D., Kumar, N., Mondal, H., & Sinha, R. (2023). Applicability of chatgpt in assisting to solve higher order problems in pathology. *Cureus*, *15*.

Srivastava, P., Ganu, T., & Guha, S. (2022). Towards zero-shot and few-shot table question answering using gpt-3. arXiv preprint, arXiv:2210.17284.

Stokel-Walker, C. (2022). AI bot chatgpt writes smart essays-should academics worry? *Nature*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR (pp. 3319–3328).

Susnjak, T. (2022). Chatgpt: The end of online exam integrity? arXiv preprint, arXiv: 2212.09292.

Tang, T., Li, J., Zhao, W. X., & Wen, J. R. (2022). Context-tuning: Learning contextualized prompts for natural language generation. arXiv preprint, arXiv:2201.08670.

Tsingos, C., Bosnic-Anticevich, S., Lonie, J. M., & Smith, L. (2015). A model for assessing reflective practices in pharmacy education. *American journal of pharmaceutical education*, *79*.

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education* (pp. 153–166). Springer.

Wenzlaff, K., & Spaeth, S. (2022). Smarter than humans? Validating how openai's chatgpt model explains crowdfunding, alternative finance and community finance. In *Validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance*.

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

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic literature review. arXiv preprint, arXiv:2303.13379.

Yu, F., Quartey, L., & Schilder, F. (2022). Legal prompting: Teaching a language model to think like a lawyer. arXiv preprint, arXiv:2212.01326.

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic chain of thought prompting in large language models. arXiv preprint, arXiv:2210.03493.