Received: 13 December 2024 | Accepted: 19 April 2025

DOI: 10.1111/bjet.13601

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Metacognition meets AI: Empowering reflective writing with large language models

Seyed Parsa Neshaei[1] | **Paola Mejia-Domenzain**[1] | **Richard Lee Davis**[2] | **Tanja Käser**[1]

[1]EPFL, Lausanne, Switzerland

[2]KTH Royal Institute of Technology, Stockholm, Sweden

Correspondence
Seyed Parsa Neshaei, EPFL, Rte Cantonale, 1015 Lausanne, Vaud, Switzerland.
Email: seyed.neshaei@epfl.ch

Funding information
State Secretariat for Education, Research, and Innovation (SERI), Switzerland

Abstract: Reflective writing is known as a useful method in learning sciences to improve the metacognitive skills of students. However, students struggle to structure their reflections properly, limiting the possible learning gains. Previous works in educational technologies literature have explored the paradigms of learning from worked and modelling examples, but (a) their application to the domain of reflective writing is rare, (b) such methods might not scale properly to large-scale classrooms, and (c) they do not necessarily take the learning needs of each student into account. In this work, we suggest two approaches of integrating AI-enabled support in digital systems designed around learning from worked and modelling examples paradigms, to provide personalized learning and feedback to students using large language models (LLMs). We evaluate Reflectium, our reflective writing assistant, show benefits of integrating AI support into the learning from examples modalities and compare the perception of the users and their interaction behaviour when using each version of our tool. Our work sheds light on the applicability of generative LLMs to different types of providing support using the learning from examples paradigm, in the domain of reflective writing.

KEYWORDS
large language models, learning from examples, metacognition, reflective writing, writing assistants

Practitioner notes

What is already known about this topic

- Reflective writing fosters metacognitive skills and improves learning gains and personal growth.
- The learning from *worked* and *modelling* examples paradigms is effective for skill acquisition and applying the acquired knowledge.
- Existing reflective writing assistants usually lack dynamic, AI-driven feedback or interactivity, limiting personalization and adaptability to each user's own needs in the learning process.

What this paper adds

- It introduces Reflectium, an AI-enabled reflective writing assistant, integrating intelligent and interactive writing support for both the learning from *worked* and *modelling* examples paradigms.
- It demonstrates the use of a fine-tuned large language model (LLM) for providing feedback in the learning from *worked* examples version, and an LLM-powered conversational agent simulating instructor interactions for the learning from *modelling* examples version.
- It reports findings from a user study comparing the positive impact of artificial intelligence (AI) support on learners' performance, interaction behaviour and learning experience.

Implications for practice and/or policy

- Digital tutoring systems for teaching reflective writing using the learning from *worked* examples paradigm should incorporate adaptive AI feedback to enhance learning gains.
- Conversational agents simulating peers/instructors and powered by LLMs can provide scalable, interactive support for learning from *modelling* examples, notably in large-scale educational settings.
- Reflective writing tools should be evaluated for their impact on different aspects of the learning process, such as task performance, interaction behaviour and user experience, to guide future improvements.
- Educators and policymakers should consider the integration of AI-driven reflective writing tools into teaching curricula to enhance reflective practices and metacognitive skill development.

INTRODUCTION

Reflective writing is defined as the process of expressing insights, thoughts and experiences on specific events (Williams et al., 2020). It helps learners to revisit past experiences and gain deeper insights into their skills, which in turn can improve learning gains and personal growth (Colomer et al., 2020). Reflective writing is known to improve students' metacognitive skills, which is considered a crucial part of the human learning process (O'Loughlin & Griffith, 2020; Perry et al., 2019). However, forming reflective writings is not necessarily straightforward, especially for new learners. They exhibit issues with using reflective structural models (eg, the Gibbs reflective cycle (Adeani et al., 2020)) to write down their emotions and thoughts (Adeani et al., 2020; Middleton, 2017; Prior et al., 2016).

Learning from examples has the potential to support beginners in reflective writing. Based on the theory of cognitive knowledge acquisition, the *learning from examples* paradigm has been shown to be effective for initial skill learning (Renkl, 2002), enabling learners to borrow knowledge from others (Sweller, 1994) and to apply the knowledge they acquire to problems. Prior works have mainly addressed two types of *learning from examples*: (a) learning from *worked* examples, in which the learners see annotated examples to study and learn from, and (b) learning from *modelling* examples, in which the learners participate in a learning session with a peer or instructor that formulates and solves specific examples, showing the process of solving the example (Hoogerheide et al., 2014). Research on comparing these two paradigms, especially in the domain of reflective writing, is scarce and there is no consensus on which approach leads to higher overall learning gains, with some works claiming similar results obtained from the two approaches (Hoogerheide et al., 2014). Moreover, many current approaches to *learning from worked examples* rely on static predefined examples (Hilbert et al., 2008), failing to integrate intelligent artificial intelligence (AI)-based feedback in the loop. Furthermore, the *learning from modelling examples* approaches either (a) need on-demand access to an instructor, making it difficult to execute in large-scale educational environments, or (b) rely on pre-recorded videos of an instructor (Braaksma et al., 2002; Groenendijk et al., 2013), limiting the interactivity level of the approach and failing to adapt to each student's unique needs. We define *AI support* as the use of AI technologies, such as natural language processing methods, to assist individuals in various aspects of the writing process, including text structure and writing quality, across different stages of writing (eg, planning, drafting and feedback). While AI support has shown to be beneficial in a range of educational writing support tools (Göldi et al., 2024; Lee et al., 2024; Mejia-Domenzain et al., 2024), its integration into different *learning from examples* paradigms has not been explored.

To address the research gaps mentioned above, we suggest two approaches of integrating AI-enabled support in digital systems designed around the *learning from examples* paradigms. For the *learning from worked examples* approach, we integrate a feedback functionality provided by a fine-tuned large language model (LLM). For the *learning from modelling examples* approach, we design and implement a conversational agent (ie, chatbot), powered by an LLM acting as the 'peer/instructor' to simulate the real classroom experience of modelling examples. We design and implement these variations into Reflectium, our intelligent and interactive assistant designed to teach reflective writing.

We evaluate Reflectium in a 2×2 controlled user study with 100 Prolific participants, in which we manipulate (a) the two paradigms of learning from *worked* and *modelling* examples, and (b) the availability of AI support. This design enables us to find and compare the effects of the two *learning from examples* paradigms as well as to analyse the effects of providing intelligent support using *AI models* in each of these paradigms. Particularly, we aim to answer the following three research questions (RQs): What are the effects of embedding AI support into the learning from *worked* and *modelling* examples paradigms on learners' **performance** in reflective writing (RQ1), on their **interaction behaviour** (RQ2) and on their perceived learning **experience** (RQ3)? Answering RQ1 enables us to assess the effect of the combination of AI-support and example-based learning paradigms on learners' reflective writing (see also Huang et al., 2018; Mejia-Domenzain et al., 2024; Wale & Kassahun, 2024). With RQ2, we aim to scrutinize the effect of the AI support on learners' learning behaviour and the impact of these behaviours on their learning gains (see also Cotos et al., 2020; Mejia-Domenzain et al., 2024; Mouchel et al., 2023). Finally, RQ3 assesses learners' perceived learning experience, which has been shown to be an important factor for adoption (Lee et al., 2005; Wambsganss, Kueng, et al., 2021; Wang & Tahir, 2020).

Our results indicate higher learning gains for participants in the *worked* examples and AI support conditions, but we have not found any significant difference in participants'

perception of the tool. Our work sheds light on the applicability of AI support to different *learning from examples* paradigms for the task of reflective writing.

RELATED WORK

Reflective writing

Reflective writing encourages learners to articulate their thoughts and experiences related to specific events (Williams et al., 2020). It has been known for being a source of valuable insights and a guide for future action plans (Boud et al., 2013; Nehyba & Štefánik, 2023) as well as an instrument to support the professional development of learners (Cochran-Smith, 2005). Participating in reflective sessions is considered an important metacognitive skill for students (Colomer et al., 2020; McGuire et al., 2009; Perry et al., 2019). Additionally, prior work underscores the positive impacts of reflective writing on learning outcomes, as exemplified by improved learning in vocational students who participated in reflective practices (Cattaneo & Boldrini, 2016; Cattaneo & Motta, 2021; Hommel et al., 2023; Mejia-Domenzain et al., 2022).

Previous works have explored the usage of structural frameworks to help people recall their experiences, with the aim of improving their learning (Williams et al., 2020). Examples include the models of Boud, Kolb, Schön, Gibbs and Rolfe (Boud et al., 2013; Gibbs, 1988; Kolb, 2014; Rolfe et al., 2001; Schön, 2017). In this work, we focus on the Gibbs reflective cycle (Gibbs, 1988) as the theoretical backbone for reflective writing support. It consists of six main components in a cycle (Gibbs, 1988):

- **Description**: A presentation of the event the learner is reflecting on.
- **Feelings**: Any feelings the learners had at the time of the situation.
- **Evaluation**: The positive and negative aspects of what happened in the situation.
- **Analysis**: The possible reasons for the points mentioned in the Evaluation section.
- **Conclusion**: A summary of what happened and what the learner gained from the event as a learning outcome.
- **Action plan**: Opinions on what the learner would do differently if they were faced with a similar situation in the future.

The Gibbs reflective cycle has been particularly recommended for novice learners and practitioners of reflection (Al-Mutawa et al., 2024), being described as 'one of the simplest and most effective' reflection models (Ahmadpour et al., 2025). As a result, previous research has used Gibbs reflective cycle across a variety of domains to help students structure their reflective writings (Ahmed, 2020; Aneeis Hashim et al., 2023; Markkanen et al., 2020). For example, (Ezezika & Johnston, 2023) have employed the Gibbs reflective cycle in a public health biology course, while (Nurlatifah et al., 2023) implemented it in the context of an English as a Foreign Language (EFL) classroom. Additionally, (Adeani et al., 2020) have compared different reflective models (Kolb, Johnson, Gibbs) and found the Gibbs reflective cycle to be the most appropriate model for use in literature classrooms, as its well-structured approach helped students write better reflections.

Learning tools for reflective writing

The rise of LLMs (eg, the GPT family of models) has contributed to the success of writing assistants across diverse areas (Lee et al., 2024), such as screenplays (Mirowski et al., 2023),

peer reviews (Su et al., 2023; Wambsganss et al., 2023), argumentative texts (Afrin & Litman, 2023) and metaphor creation (Kim et al., 2023). Such writing assistants support users in generating new ideas, delivering feedback and revising writings (Buschek et al., 2021; Peng et al., 2020; Wu et al., 2019).

An emerging area of research on writing assistance investigates reflective writing and in particular its significance and impact in educational settings. Researchers have explored, including guiding questions (Moussa-Inaty, 2015), sentence openers (Kingkaew et al., 2023), reflection manuals (Wong et al., 2016) or conversational agents (Kim et al., 2024; Wolfbauer et al., 2023) in the reflective writing process. However, many of these existing works rely on pre-scripted dialogues or predefined responses, which constrain their adaptivity to the certain unique learning paths of each learner. Only a few works have explored integrating LLMs into reflective writing; however, they have not been evaluated in learning scenarios or connected to theoretical underpinnings of learning. For example, the work of Kim et al. (2024) presents a state machine-based approach of using LLMs in supporting psychiatric patients in documenting their daily experiences by interacting with a conversational agent. However, this work did not aim to teach reflective writing, but rather nudged the users to rely on the conversational agent each time they wanted to write about their experiences. Others (Kumar et al., 2024) have shown that access to an LLM facilitating self-reflection improves students' academic performance. Again, the goal of their work was not to teach reflective writing using AI, but to facilitate reflection. Finally, (Li et al., 2023) have shown limitations of LLMs for reflective writing but have not explored the best ways to embed such models in an educational reflective writing assistant. In this work, we aim to address the gaps in the literature by designing an experimental study comparing the effects of different *learning from examples* paradigms, with and without AI support, on the learning gains of students.

Learning from examples

Learning from examples, also known as *example-based learning*, is a form of learning by observing or imitating what others do (Van Gog & Rummel, 2010) that has been shown to have positive impacts on problem solving and learning across multiple domains (Jackson et al., 2008). Research in the area of *learning from examples* typically focuses on two main approaches to this paradigm (Hoogerheide et al., 2014):

- **Learning from *worked* examples**: *Worked* examples refer to demonstrations of the solution to a problem, typically provided to learners to help them develop problem-solving skills (Renkl, 1997). Worked examples are used in various educational scenarios, including textbooks (eg, problem solutions typically provided in the end of textbooks) (Glasnovic Gracin, 2018), programming tutorials (eg, debugging explanations) (Bofferding et al., 2022) or language learning (eg, essay exercises) (Kyun et al., 2013). When learning how to *write* using the learning from *worked* examples paradigm, text-based worked examples are shown to the learners who are expected to study and learn from them. Research has shown the benefits and effectiveness of learning from worked examples (Sweller & Cooper, 1985), specifically among novice learners (Recker & Pirolli, 1995). This approach is supported by the *cognitive load theory*, which claims that the instruction process should be designed to reduce ineffective memory load and to make more working memory resources available for learning (Hoogerheide et al., 2014; Sweller, 1988). Particularly, in the context of reflective writing, an approach based on the learning from *worked* examples paradigm can consist of a full textual example of a reflective writing, annotated with the components of a reflective framework, for example, the Gibbs reflective cycle. The learners can explore the different annotated text excerpts and learn how to write a similarly structured text by observing the worked example.

- **Learning from *modelling* examples**: The *modelling* examples paradigm refers to instructional demonstrations where a peer student, expert or instructor explicitly shows and explains steps to perform and complete a task, solve a problem or apply a concept. Unlike *worked* examples, the *modelling* examples paradigm emphasizes the thinking *process*. This approach can either happen live (Bjerrum et al., 2013) or in the offline form of pre-recorded videos (Braaksma et al., 2002; Groenendijk et al., 2013), making it more suitable for large-scale classrooms or massive online open courses. Modelling examples are used in various educational scenarios, including programming (eg, live coding by the instructor) (Raj et al., 2020), medical training (eg, modelling diagnostic reasoning in patient cases) (Bjerrum et al., 2013) and soft skills training (eg, role-playing effective communication or negotiation) (Andrew & Meligrana, 2012). The approach is supported by the *social learning theory*, which suggests that individuals acquire new behaviours by observing and modelling the actions of others (Bandura, 1977; Hoogerheide et al., 2014). Specifically, in the context of reflective writing, an approach based on the learning from *modelling* examples paradigm can consist of a peer or instructor providing step-by-step instructions on how to form a reflective writing based on a reflective framework, for example, the Gibbs reflective cycle. The peer or instructor would iterate over the different components one by one, with the aim of forming a full reflective writing in the end, *modelling* how a writer would conduct the process of reflective writing.

There is no consensus on which of the two approaches leads to better results in terms of higher learning gains. For example, (Hoogerheide et al., 2014) found that both approaches were effective at enhancing the test performance of learners and reduced their mental effort. With that said, the current approaches of *learning from examples* suffer from low interactivity and adaptability to each learner's unique needs and learning pace. In particular, most current *learning from worked examples* approaches rely on predefined annotated examples, which naturally limits the adaptability of the system to each user's learning needs while practicing (Erümit & Çetin, 2020). Additionally, the most current *learning from modelling examples* approaches in large-scale educational environments rely on pre-recorded videos (Braaksma et al., 2002; Groenendijk et al., 2013). Similar to a pre-written text, pre-recorded videos naturally limit interactiveness and the possibility of moving towards learning beyond what is mentioned in the video. Moreover, while the learning sciences literature discusses the role of *active inquiry* in learning (Graesser et al., 1993), there is no potential to ask follow-up questions in a pre-recorded video or to make the learning pace suitable to each learner's own needs (Chin, 2006). There have been only a few works in the direction of using AI-based models to enable interactivity and adaptivity in tools using the *learning from examples* paradigm. For instance, (Mejia-Domenzain et al., 2024) have used a retrieval-based approach to adaptively provide students with example texts from peers, focusing on the learning from *worked* examples paradigm. In contrast, in this research, we suggest approaches to integrate LLMs into the process of the two *learning from examples* paradigms to enable adaptiveness and interactivity.

REFLECTIUM—LEARNING FROM WORKED AND MODELLING EXAMPLES WITH AI SUPPORT

To study the effect of embedding LLMs as AI support into the paradigms of learning from *worked* and *modelling* examples on users' learning outcomes, perceptions and behaviours, we designed Reflectium, our reflective writing assistant, around teaching the Gibbs reflective cycle. We implemented Reflectium as a React-based web application in two main versions corresponding to the learning from *worked* and *modelling* examples paradigms.

We then evaluated the initial version of Reflectium in a pilot study with 34 German-speaking students of a nursing and caring vocational school in a Western European country. We used the feedback provided by the participants to adjust the interface and functionalities, as well as to fix issues within the tool.

User interface for worked examples

Figure 1a illustrates the learning from *worked* examples interface of Reflectium. The interface allows the users to hover over sentences from each class of the Gibbs reflective cycle in the text area using their mouse pointer, and see the definition of the highlighted class as well as three other examples on the dashboard on the right side, taking cues from prior works instructing learners on following certain structures in writing (Weber et al., 2024). The users can also ask for more worked examples to learn from, using the 'Another Example' button. A total of four worked examples are present in the system. When they are done with their learning, they can press the 'Done' button, and go to another page where they can write the reflective text of their own as a practice (see Figure 1b). Once done, they can press the 'Feedback' button to obtain personalized feedback. Our design is inspired by prior attempts at adding feedback functionalities and the *learning from errors* principle (Metcalfe, 2017; Wambsganss et al., 2020) to enable interactiveness and adaptivity in the *learning from worked examples* interfaces (Mejia-Domenzain et al., 2024). The feedback relies on a fine-tuned LLM (described in detail in Section "AI model architecture"), classifying each sentence in the user's text into one of the Gibbs reflective cycle classes. We show the results to the user in two modalities: (a) each sentence in the text box on the left is highlighted with the corresponding Gibbs reflective cycle class, emulating an interface similar to the worked examples the users saw earlier, and (b) a dashboard on the right side shows the classes the user included or missed in their writing, arranged in the order of the Gibbs reflective cycle. By hovering their mouse pointer on each class, users can see more *examples* of that class below.

User interface for modelling examples

Figure 2 illustrates the interface of Reflectium using the learning from *modelling* examples paradigm (Figure 2). To enable interactiveness and adaptivity in this version of Reflectium, we implemented a conversational agent module, using an LLM in the role of the *instructor*. The conversational agent provides relevant responses to questions asked by the users regarding the domain of writing with the Gibbs reflective cycle.

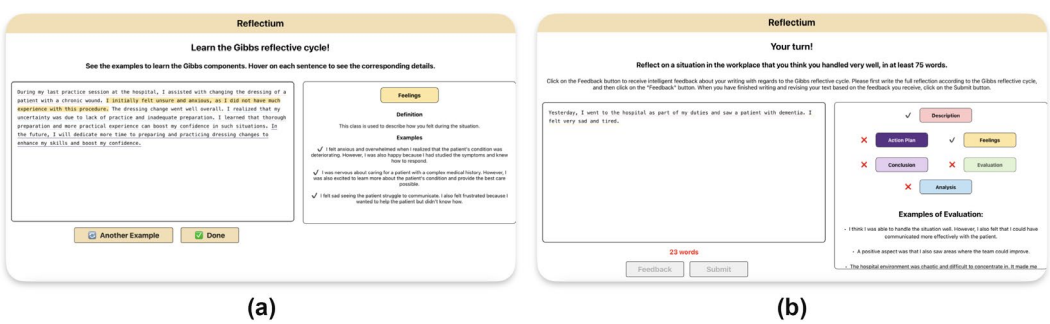

FIGURE 1 Screenshots of the version of Reflectium using the *learning from worked examples* paradigm: (a) observing the worked example and (b) receiving AI-based feedback on the user's writing.

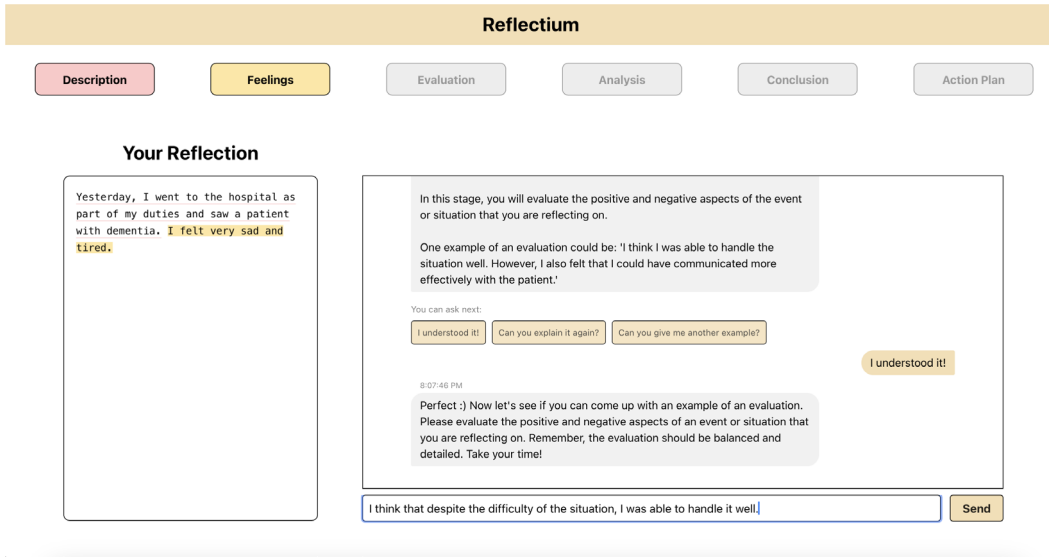

FIGURE 2 Screenshot of the version of Reflectium implementing *learning from modelling examples* using interaction with a conversational agent.

To simulate an instructor in the learning from *modelling* examples paradigm, the conversational agent goes through the components of the Gibbs reflective cycle step by step. For each component, it first begins by explaining the current component, as well as providing suggestions (predefined) for follow-up questions in terms of buttons in the interface (taking cues from Wambsganss et al., 2024). Then, to enable adaptivity based on the learning state of the users at each step, when the users indicate to the conversational agent that they understood the current class, the conversational agent asks the user to write a sentence of that class. It then sends the sentence to the back-end, where it is evaluated and classified by the same fine-tuned LLM (see Section "User interface for worked examples") powering the feedback module in the *worked* examples version. The conversational agent refuses to go to the next step until the user writes a sentence from the expected class (there is a threshold of maximum three tries to account for potential classification errors). The conversational agent then moves to the next class, until all classes of the Gibbs reflective cycle are covered; it concludes the learning session with a congratulating message (taking cues from Kim et al., 2024).

AI model architecture

To provide support to learners in Reflectium, we employed LLMs and generative AI models (as illustrated in Figure 3). In the learning from *worked* examples condition, we used AI models to generate worked examples and to provide adaptive feedback to the texts submitted by the students. In the learning from *modelling* examples condition, we used the models to power the interactive conversational agent. In particular, in the learning from *worked* examples paradigm, we used GPT-4o in an offline stage to generate the full worked examples that were shown to the learner in the user interface of Reflectium. In the learning from *modelling* examples paradigm, GPT-4o was used to power the conversational agent when the users asked follow-up questions, for example, demanding more examples or explanations on certain classes of the Gibbs reflective cycle. In addition, we used fine-tuned BERT models in both the learning from *worked* and *modelling* examples paradigms to classify

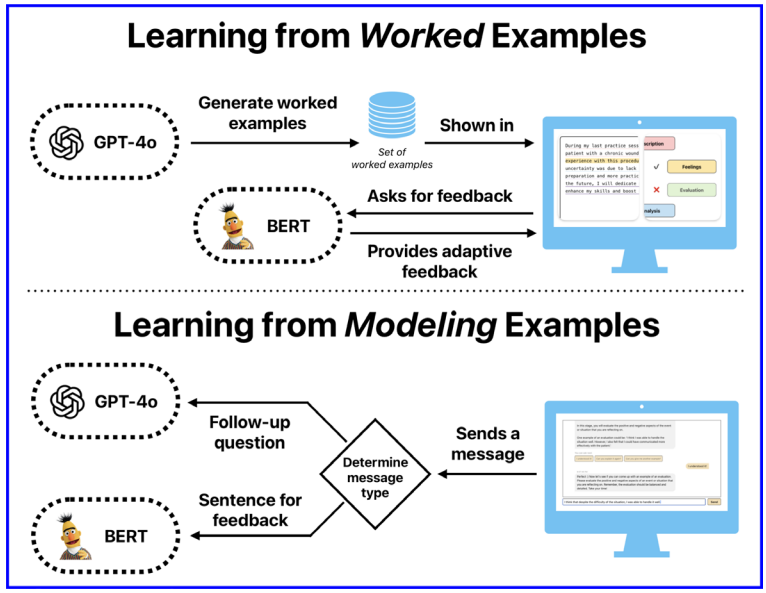

FIGURE 3 AI models used in Reflectium for each of the two interfaces using the paradigms of learning from *worked* and *modelling* examples. GPT-4o is used in both paradigms: Offline generation of full *worked* examples shown to learners and powering conversational agents for follow-up questions for the *modelling* examples. Our fine-tuned BERT model classifies sentences into the Gibbs components, providing in-text highlighting for *worked* examples and conversational agent responses for *modelling* examples.

each sentence of a reflective writing into the classes of the Gibbs reflective cycle, providing feedback in terms of in-text highlighting (for *worked*) or responses from the conversational agent (for *modelling*).

Generating worked examples: To generate the worked examples included in the version of Reflectium using the learning from *worked* examples paradigm, we initially prompted GPT-4o iteratively, each time by appending a sentence of the next Gibbs reflective cycle class to the current partially formed text. Two researchers in the domain of learning sciences reviewed the quality of the output texts and checked if the generated sentences were from the intended class of the Gibbs reflective cycle. All texts were approved and thus included in the learning from *worked* examples version of Reflectium. We provide all prompts and texts in the Appendix A.

Providing adaptive feedback: We used a BERT model to build a classifier for labelling each input sentence as one of the classes of the Gibbs reflective cycle, and integrated it in the feedback modules of both learning from examples paradigms. We picked the base models of BERT in English and German and fine-tuned them on a dataset of 96 annotated diaries. The dataset was collected in two experiments conducted in a German-speaking vocational school in a Western European country: the *first* experiment evaluated a conversational agent for providing reflective writing support and teaching students on how to use the Gibbs reflective cycle Neshaei et al., 2025, and in the *second* experiment, we evaluated an early version of Reflectium as a pilot study. In both versions, the final reflective texts written by the students were collected for training our models. All students provided informed consent for their data to be used for research and the studies were approved by the university's ethics committee (Nr. HREC000572 and HREC 013-2021). The texts had an average length of 268.49 words (SD: 163.83). We translated the dataset to English by machine translation, and conducted the data labelling process separately for the two languages. Two researchers annotated five reflections independently, resulting in a Cohen's Kappa of 0.9285, indicating

TABLE 1 Performance metrics (balanced accuracy and F1) of our fine-tuned BERT models, grouped per dataset language.

Language	Balanced accuracy	F1 description	F1 feelings	F1 evaluation	F1 analysis	F1 conclusion	F1 action plan
German	56.81 ± 4.75	86.60 ± 2.07	63.80 ± 4.09	54.60 ± 14.48	13.20 ± 14.48	48.20 ± 10.99	65.00 ± 18.11
English	64.07 ± 1.95	90.60 ± 0.89	80.20 ± 2.17	52.4 ± 5.27	10.20 ± 14.94	64.20 ± 5.12	76.60 ± 5.18

a strong agreement. Then, one of the annotators completed the rest of the annotations on their own. Before using the data to train our models, we processed it by wrapping target sentences to classify around specific `<start>` and `<end>` tokens, with one sentence before and after in the context window, to allow contextual information of a reflective text to be used in the classification process. We then fine-tuned our BERT models on the processed data for three epochs using the `simpletransformers` library. The results of our fivefold cross-validation evaluation on the two models can be seen in Table 1.

Interactive conversational agent: We used the GPT-4o generative AI model in the version of Reflectium using the *learning from modelling examples* approach to simulate the *instructor* by engaging in conversations with the user through the conversational agent interface. Following prior works (Kim et al., 2024), we designed the conversational agent as a state machine. To ensure the relevance of the outputs and reduce the risk of hallucinations, we included a single set of predefined messages shown to the user when entering each step of the Gibbs reflective cycle. However, we forwarded the message history of each user, as well as their current query any time they asked a follow-up question to GPT-4o and displayed the result to the user. Sending the complete message history to GPT-4o enables the model to identify the learning path of the student and the possible mistakes in each class, leading to personalized answers from the model. The predefined messages and the prompts were prepared and finalized by three researchers in a workshop. All prompts and predefined responses can be found in the Appendix A. To ensure the effectiveness of our predefined messages and prompts, we preevaluated the final interaction with the conversational agent in a small-scale pilot study with 12 researchers, who approved the responses as relevant and useful.

Accounting for the low performance of the models: As can be seen in Table 1, we achieved low F1 scores for specific classes (eg, mostly notable for Analysis). To remedy this situation when deploying the models to Reflectium, we implemented two fallback cases:

1. Learning from *worked* examples version: if, after the third feedback attempt, there are still classes missing in the results, we look at the model scores provided for each class and assign the sentence with the maximum score of the missing class to the label of this class. This process allows us to pick the 'next-most-likely' candidate for that class, mitigating model accuracy issues.
2. Learning from *modelling* examples version: After three failed attempts for a class, the interface moves automatically to the next class. Instead of displaying a congratulatory message (starting with 'Perfect') the system instead presents a non-congratulating message (starting with 'OK').

EVALUATION STUDY

We evaluated Reflectium with 100 Prolific participants, with the goal of assessing the effects of AI support on learning gains, interaction and perception of users. In the following, we will describe the study design, procedure, participants and the employed measures in detail.

Study design

As illustrated in Figure 4, we employed a randomized 2×2 study design encompassing two main factors: learning paradigm (*worked* or *modelling* examples) and AI support (without or with). For brevity, we will refer to the groups with AI support as 'Worked w/ AI' and 'Modelling w/ AI' and to the groups without AI support as 'Worked w/o AI' and 'Modelling w/o AI'.

To create the **Worked w/o AI** version, we removed the AI-enabled feedback module from the Worked w/ AI version of Reflectium, only keeping the predefined worked examples. While users were still directed to a writing interface after observing the worked examples, the feedback button was removed.

To build the **Modelling w/o AI** version, we replaced the conversational agent with a pre-recorded instructional video on the Gibbs reflective cycle, as a standard method for learning from *modelling* examples in the literature (Braaksma et al., 2002; Groenendijk et al., 2013). The video was narrated by a learning sciences researcher as the instructor over a set of presentation slides[1] explaining each of the classes of the Gibbs reflective cycle step by step and showing the process of forming a reflective writing iteratively by adding sentences from the classes in order.

Procedure

Our experiment consisted of three main phases: a pre-intervention, a learning intervention and a post-intervention. An overview of the procedure can be seen in Figure 5.

Pre-intervention

The experiment started with a pre-survey, where we tested the effectiveness of the randomization across the four conditions using three different constructs, taking cues from

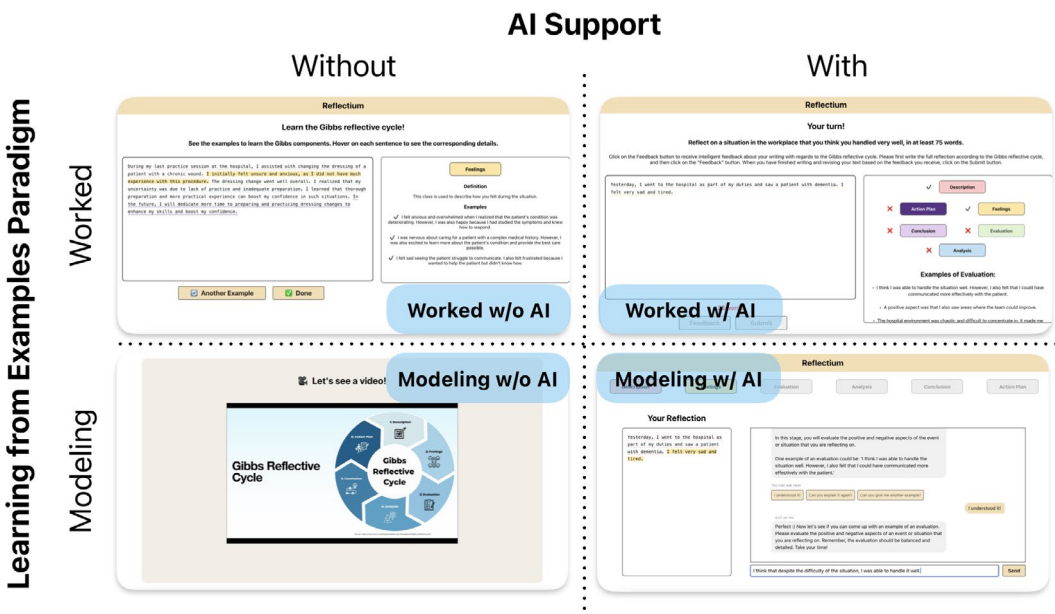

FIGURE 4 Our study setup with four groups, using a randomized 2 (learning from examples paradigm: worked vs. modelling)×2 (AI support: without vs. with) design.

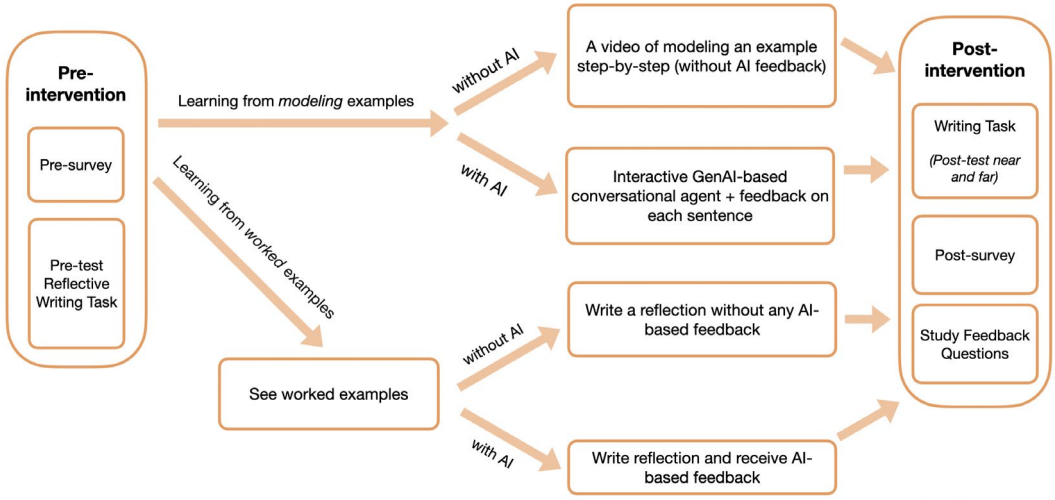

FIGURE 5 Overview of our study procedure for Reflectium, encompassing with- and without-AI versions across the two different approaches to the *learning from examples* paradigm.

TABLE 2 Constructs used in the pre-survey.

Construct	Pretest questions
Information technology usage	1. I like to experiment with new information technology (IT) tools 2. In general, I am willing to try out new IT tools 3. If I heard about a new IT tool, I would look for ways to experiment with it 4. Among my peers, I am usually the first to try out new IT tools
Feedback-seeking	1. It is important for me to receive feedback on my performance 2. I like getting feedback on my behaviour 3. It is important for me to receive feedback on my progress and learning potential 4. I find feedback on my performance useful
Reflective writing knowledge	1. I know what is reflective writing (ie, journaling) 2. I have written reflections before 3. I know what is the Gibbs reflective cycle 4. I have used the Gibbs reflective cycle before

prior works (Neshaei et al., 2024; Wambsganss et al., 2020) (as provided in Table 2). Each pre-survey construct consisted of questions measured on the 1-to-7 Likert scale. We averaged the results for questions inside each construct. Then, to be able to measure improvements in the reflective writings and adhering to the Gibbs reflective cycle from before to after the learning intervention, we asked the users to reflect on a situation at work where they were challenged by an unexpected event, in at least 75 words, as a *pretest* writing task.

Learning intervention

After completing the pre-intervention stage, the users saw a short video as an introduction to reflective writing, without teaching the Gibbs reflective cycle. Then, they got access to one of the four versions of Reflectium as the learning intervention: learning from worked and modelling examples paradigms, each without and with AI-based assistance (see Figure 4). We ensured balance among genders by randomizing the versions of Reflectium each learner

used. All learning interventions consisted of a writing task, in which we asked the users to reflect on a situation in the workplace that they think they handled very well.

Post-intervention

Our post-intervention consisted of two posttest reflective writing tasks and a post-survey questionnaire consisting of a set of constructs. Finally, participants were also asked for feedback on the tool.

Reflective writing tasks: Prior works (Hübner et al., 2010) have considered a range of learning deficiencies, particularly mediation (ie, not having the necessary cognitive requirements to improve task performance) and production (ie, being capable but not using strategies spontaneously) deficiencies. To see whether different versions of Reflectium were successful in addressing such learning deficiencies in the domain of reflective writing, we asked the users to write again as a *posttest* (similar to previous works on writing assistants; Mejia-Domenzain et al., 2024). We first posed a generally similar question (*posttest-near*) to those in the pretest and learning intervention to the users, asking them to reflect on a situation in the workplace when things did not go as planned, in at least 75 words without intelligent support from Reflectium, to measure their learning gains by comparing the writings to the pretest (see Section "Measure of learning" for the measurement process). Then, we administered a *transfer* posttest task (*posttest-far*) to measure if the users maintained their mastery of the Gibbs reflective cycle in another domain of writing. In particular, we asked them to reflect on the learning experience they had with Reflectium, in at least 75 words.

Questionnaire: We delivered a questionnaire (see Table 3) to find the users' perception towards the version of Reflectium they used. Each construct, inspired from prior works (Neshaei et al., 2024; Wambsganss et al., 2020), consisted of questions measured on the 1-to-7 Likert scale. We averaged the results for questions inside each construct. We additionally administered the questions from the NASA Task Load Index to the participants (Hart, 2006).

Study feedback questions: We then asked two additional study feedback questions accepting written answers from the users: (1) *In your own words, explain the benefit and importance of the Gibbs reflective cycle, and how do you think it has exactly changed the way you reflect on your experiences?* (in at least 40 words), and (2) *Any more comments on the whole study?* (in at least 30 words).[2]

Participants

We recruited a total of 100 participants using the Prolific experiment crowdsourcing platform. We required participants to be fluent in English, have a degree in health and welfare (eg, medicine or nursing) and have a high school diploma or above. A breakdown of the participants' demographics in each study group can be seen in Table 4. The study was approved by the university's ethics review board (Nr. HREC000572 and HREC 013-2021).

Measure of learning

Following (Ullmann, 2015), we graded each writing in two dimensions:

Breadth of reflection: We used the adherence of the texts submitted by the users in the pre- and posttest writing tasks to the Gibbs reflective cycle as a measure to compare both

TABLE 3 Constructs used in the post-survey.

Construct	Example question
Excitement after interaction	Interacting with Reflectium was exciting
Perceived ease of use	I found Reflectium easy to interact with
Perceived usefulness	I found Reflectium useful for writing reflections
Technology acceptance	Assuming Reflectium is available, the next time I want to write a reflection, I would use it again
Perceived improvement in writing	After using Reflectium, my ability to write reflections has improved
Perceived improvement in writing in the long run	I assume using Reflectium in the long run will help me improve my abilities to write well-structured reflections
Correctness of the suggestions	Adaptive responses, suggestions and feedback from Reflectium were correct

TABLE 4 Demographics of the groups in our experimental evaluation (F refers to those identified as females, M: males, and O: others).

Group	Demographics	Average age	SD age
Worked w/o AI	16 F, 8 M	25.63	3.52
Worked w/ AI	16 F, 8 M	25.08	1.98
Modelling w/o AI	19 F, 7 M	24.88	2.56
Modelling w/ AI	19 F, 6 M, 1 O	24.77	2.31

across time (pretest to posttest near and posttest far) and across study groups. To measure the structure of the reflective texts, we annotated each of the sentences in the writings with one of the classes of the Gibbs reflective cycle. We granted one point for each class found in the text; as a result, each text could obtain a maximum score of 6 if it contained all of the classes of the Gibbs reflective cycle. One of the same two researchers who annotated the reflective writing datasets used for training the classification model (see Section "AI model architecture") labelled the sentences of the texts to identify the relevant Gibbs reflective cycle components.

Depth of reflection: As an additional dimension, we applied the reflective writing criteria defined by (Ullmann, 2015) and used in prior works (Nehyba & Štefánik, 2023) to score each posttest far writing based on a rubric consisting of four criteria: (a) including verbs referring to thinking in combination with experience, (b) emphasizing the thought process, (c) including a first-person perspective and (d) being in the form of a question. To score each text, two researchers in the domain of learning sciences (one being also among the two participating in annotating for the breadth score) participated in a workshop together and then graded five reflections independently across the four items of the rubric. We observed that 19 out of 20 rubric items were graded with the same score by the two researchers, indicating a strong agreement. Then, one of the researchers completed the rest of the scoring on their own.

We additionally computed a **combined** score per writing by multiplying the breadth score (ie, adherence to the Gibbs reflective cycle) with the binarized depth score. We binarized the depth score to split reflections into 'reflective' and 'non-reflective' texts, with the idea that non-reflective texts should obtain an overall score of 0. When analysing the submitted texts, we found that almost all reflections (with the exception of one text) satisfied criterion C of the depth rubric (including a first-person perspective), resulting in a minimum depth score

of 1. This observation is likely due to the nature of our tasks (asking learners to reflect on a situation they experienced), which naturally made them write in first person. We therefore decided to drop criterion C for the binarization and categorized a text as 'reflective', if it contained at least one reflective component (Criterion A, B or D).

Analytic approach

Randomization control: We verified the randomization by checking for differences between the groups in the pre-survey constructs, employing a one-way ANOVA per construct and using Benjamini–Hochberg to correct for multiple comparisons.

RQ1: Measuring learning gains: To answer RQ1 and to reveal if there were any differences in outcomes coming from the different levels of AI support and learning from examples modalities, we conducted two-way ANCOVAs with AI support and Learning Paradigm as the two independent factors and pretest score as a covariate to examine the effects of pretest score, AI support (ie, w/o AI or w/ AI) and Learning Paradigm (ie, Worked or Modelling) on posttest score, as well as their interaction. We ran separate models per posttest (near and far) and score (depth, breadth, combined).

RQ2: Interaction behaviour: To answer RQ2, we tracked the interaction of users with Reflectium during the study and analysed it to find insights explaining the similarities and differences we found in response to RQ1. We performed all analyses in RQ2 separately for the worked and modelling conditions, as the different modalities of the four versions of Reflectium did not allow us to measure a shared relevant interaction variable (eg, the versions without AI support resulted in a shorter time on task as they just required watching one video/reading a worked example and writing one reflective text, with no option to revise).

For the version of Reflectium with the learning from *worked* examples paradigm, we introduced two behavioural features from our collected interaction data: (1) the number of worked examples studied and (2) the average time (in seconds) spent per example. We tracked the number of times the learners demanded to see examples by clicking on the 'Another Example' button. We employed one-way ANOVAs to find differences in the behavioural constructs across groups. To further explore the differential impact of instructional time, learning modalities, AI support and our behavioural variables on writing performance, we utilized a mixed linear model (MLM)[3] approach. We fitted a separate model for each score (depth, breadth, combined) and test point (posttest near, posttest far). We modelled time (pretest, posttest), AI support (without, with), learning paradigm (worked examples, modelling examples), number of studied examples and time per example, as well as their interactions as fixed effects, and used a random intercept to account for variations between individual students. We fitted these new MLMs only for the participants using one of the worked example versions of Reflectium.

To further examine the impact of AI support on learners' revision behaviour, we focused on learners using the version of Reflectium with the learning from *worked* examples paradigm and AI support. We adapted our MLMs by incorporating two additional behavioural variables as fixed effects: (3) the number of revisions and (4) the average time in seconds per revision, by tracking the number of rounds of feedback the users requested from the AI models while writing their text. We also saved the texts written by the learner at each feedback request to enable analysis of the changes in Gibbs reflective cycle adherence performance over time.

For the learning from *modelling* examples version of Reflectium with AI support, we checked the engagement of learners with the system by (1) measuring how many learners asked follow-up questions from the conversational agent, and (2) whether and how they

changed their writing based on the intelligent feedback they received from the models. To enable this analysis, we saved the conversation history of each learner, storing the cases in which learners asked follow-up questions separately for each component of the Gibbs reflective cycle. We also saved the sentences written by the learners for each class before and after receiving each feedback message from the conversational agent. We did not run any MLM analyses for the learning from *modelling* examples version of Reflectium, due to the lack of a representative interaction behaviour variable, as the versions with and without AI support led to very different interaction patterns (watching a video and writing one text vs. interacting with a chatbot) and time on task.

RQ3: User perception: To find the perception score of each construct per user, we calculated the average of the user's responses to the items from that construct in the post-survey. To compare the results across conditions, we conducted a one-way ANOVA per construct, followed by post hoc comparisons in case of significance, employing a Benjamini-Hochberg procedure to correct for multiple comparisons.

RESULTS

In this study, we aimed to assess the effectiveness of AI-based assistance and the different paradigms of learning from *worked* and *modelling* examples on learners' reflective writing (RQ1), their interaction behaviour (RQ2) and their perception and attitude towards each version of our system (RQ3).

In a first preparatory step, we verified the randomization by checking for differences between the four conditions at the beginning of the study. We did not find any difference in the pretest constructs among groups (as can be seen in Table 5). In a second step, we verified the survey reliability of the pre-survey and the post-survey using Cronbach's alpha, obtaining *acceptable* values for the constructs *reflective writing knowledge* and *perceived improvement in writing*, and *good* values for constructs *IT usage*, *feedback-seeking*, *excitement after interaction*, *perceived ease of use*, *perceived usefulness* and *correctness of the suggestions*. The full reliability results can be found in Table A1 in the Appendix A.

TABLE 5 Scores for pre-survey constructs per condition.

Construct	Group	Mean ± SD	Statistical analysis
IT usage	Worked w/o AI	4.20 ± 1.15	$F(3, 96) = 0.81$, $p = 0.4887$, $\eta^2 = 0.02$
	Worked w/ AI	4.26 ± 0.90	
	Modelling w/o AI	3.97 ± 1.01	
	Modelling w/ AI	4.40 ± 0.92	
Feedback-seeking	Worked w/o AI	5.26 ± 0.52	$F(3, 96) = 1.29$, $p = 0.4251$, $\eta^2 = 0.04$
	Worked w/ AI	4.92 ± 1.04	
	Modelling w/o AI	4.87 ± 0.83	
	Modelling w/ AI	4.90 ± 0.65	
Reflective writing knowledge	Worked w/o AI	2.81 ± 1.12	$F(3, 96) = 2.15$, $p = 0.2955$, $\eta^2 = 0.06$
	Worked w/ AI	3.31 ± 1.40	
	Modelling w/o AI	3.38 ± 1.41	
	Modelling w/ AI	3.76 ± 1.23	

Note: All *p*-values are corrected using the Benjamini–Hochberg procedure. There are no significant differences between conditions.

RQ1: Measuring learning gains

To investigate the effect of different learning paradigms and AI supports on learning outcomes, we evaluated each student's writing outputs in the pre- and posttests (near and far) based on the depth, breadth and combined reflective scores (see Section "Measure of learning"). Based on the benefits of AI support and adaptive feedback found in prior works on AI-supported writing assistants (Lee et al., 2024; Mejia-Domenzain et al., 2024), we hypothesized that the versions of Reflectium with AI support would lead to higher learning gains (*H1-1*). We further hypothesized that the paradigm of learning from *modelling* examples would lead to higher learning gains than using *worked* examples, following prior work showing the benefits of interactive and conversational interfaces (Chi & Wylie, 2014; Wambsganss, Guggisberg, & Söllner, 2021), and in particular, in reflective writing (Kim et al., 2024; Wolfbauer et al., 2023) (*H1-2*).

Posttest near: We observed that all four conditions seemed to improve their reflective writings regarding the combined score (Figure 6, bottom), with the strongest increase observed for the worked examples condition with AI support ($\mu_{pre} = 2.08$, $\mu_{post-near} = 4.21$) and the smallest increase observed for the modelling examples condition without AI support ($\mu_{pre} = 2.04$, $\mu_{post-near} = 2.62$). When investigating the depth and breadth scores (Figure 6, top), we found that the worked examples conditions improved in the depth scores (with AI: $\mu_{pre} = 1.88$, $\mu_{post-near} = 2.67$, without AI: $\mu_{pre} = 1.67$, $\mu_{post-near} = 2.21$) as well as in the breadth score (with AI: $\mu_{pre} = 2.67$, $\mu_{post-near} = 4.25$, without AI: $\mu_{pre} = 2.33$, $\mu_{post-near} = 3.62$) in comparison to the pretest. In contrast, the increased combined scores for learners in the modelling examples conditions mainly stemmed from an improvement in the breadth score (with AI: $\mu_{pre} = 2.35$, $\mu_{post-near} = 3.27$, without AI: $\mu_{pre} = 2.77$, $\mu_{post-near} = 3.35$), while their depth scores (with AI: $\mu_{pre} = 2.5$, $\mu_{post-near} = 2.31$, without AI: $\mu_{pre} = 2.04$, $\mu_{post-near} = 2.15$) did not notably change from the pretest. The two-way

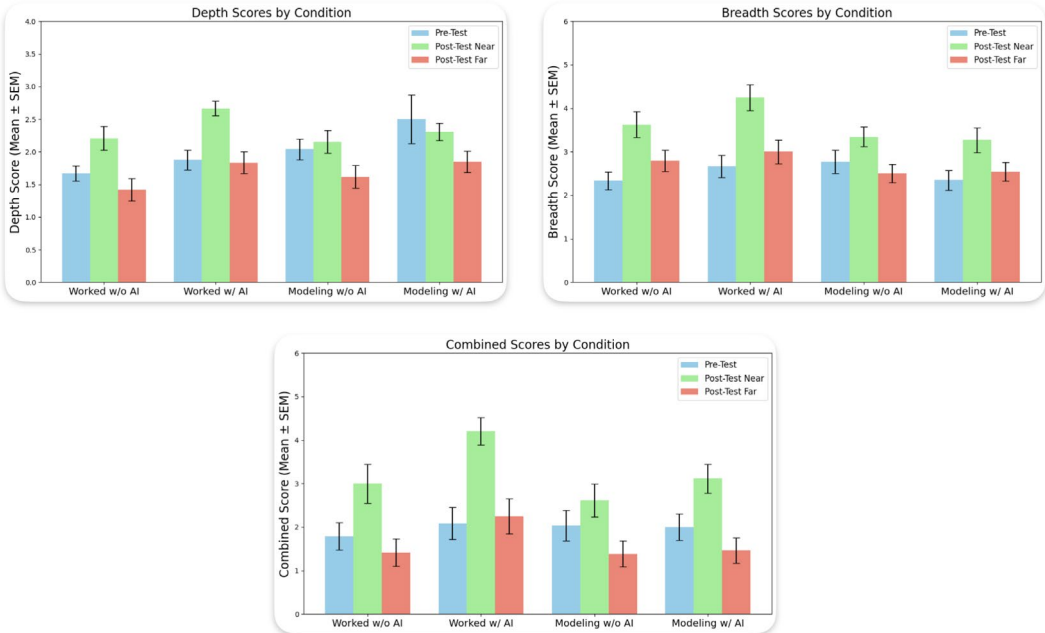

FIGURE 6 Depth, breadth and combined scores obtained in the pretest and posttest writings for users in each condition. SEM = standard error of measurement.

ANCOVA (see Section "Analytic approach") revealed a significant main effect for both AI support ($F(1, 95) = 5.04, p = 0.0271$) and learning paradigm ($F(1, 95) = 4.74, p = 0.0320$) for the combined scores, in favour of w/ AI and Worked conditions respectively. For the breadth score, we found a significant difference between the Worked and Modelling conditions ($F(1, 95) = 5.67, p = 0.0192$), but no significant difference was revealed regarding the AI support ($F(1, 95) = 1.05, p = 0.3089$). The interaction between AI support and learning paradigm was also not significant ($F(1, 95) = 0.81, p = 0.3701$). For the depth scores, we found differences with trends to significance between the different conditions regarding AI support ($F(1, 95) = 3.24, p = 0.0750$), in favour of the condition with AI support. However, no significance was revealed regarding the learning paradigm ($F(1, 95) = 2.48, p = 0.1184$) and there was also no significant interaction between learning paradigms and AI support ($F(1, 95) = 1.10, p = 0.2976$).

Posttest far: In the posttest far, learners were asked to reflect on their experience with the tool (Reflectium), which constituted a far transfer compared with the pretest as well as the task performed with the help of our tool. Indeed, only the worked examples with AI support condition managed to retain their combined score (Figure 6, bottom) with respect to the pretest ($\mu_{post-far} = 2.25$), while all other conditions showed a decrease in performance. When investigating breadth (Figure 6, top right) scores, we found that learners in the worked examples condition with AI ($\mu_{pre} = 2.67$, $\mu_{post-far} = 3.00$) and without AI support ($\mu_{pre} = 2.33$, $\mu_{post-far} = 2.79$) as well as the modelling conditions with AI support ($\mu_{pre} = 2.35$, $\mu_{post-far} = 2.54$) managed to retain or improve their scores in this far transfer task. However, for the depth scores, all conditions showed a decrease in performance. The two-way ANCOVA (see Section "Analytic approach") indicated that there were no significant differences in the combined scores for either AI support ($F(1, 95) = 1.68, p = 0.1985$) or learning paradigm ($F(1, 95) = 1.69, p = 0.1962$). The interaction was also not significant ($F(1, 95) = 1.18, p = 0.2809$ for posttest far). Similarly, for the breadth score, we found no significant differences between learning paradigms ($F(1, 95) = 2.58, p = 0.1112$) or levels of AI support ($F(1, 95) = 0.28, p = 0.5978$) and the interaction between learning paradigm and AI support was again not significant ($F(1, 95) = 0.04, p = 0.8338$). Finally, we found a trend to significance in favour of the conditions with AI support for the depth score ($F(1, 95) = 3.00, p = 0.0867$), while the differences between learning paradigms were not significant ($F(1, 95) = 0.13, p = 0.7241$) and neither was the interaction between learning paradigm and AI support ($F(1, 95) = 0.36, p = 0.5487$).

To conclude, while our analysis confirmed the benefits of providing AI support (*H1-1*), we found higher learning gains when using the learning from *worked* examples paradigm, rejecting *H1-2*.

RQ2: Interaction behaviour

To answer our second research question, we analysed and modelled learners' interaction behaviour separately for each learning paradigm.

Interaction with the learning from worked examples version

In both versions with and without AI, the users in the learning from *worked* examples groups first had the chance to study a set of worked examples. Inspired by the effectiveness of the learning from *worked* examples paradigm (Hoogerheide et al., 2014; Renkl, 1997), we initially hypothesized that both the number of examples that learners

see and the time they spend on each example should positively influence the learning outcomes (*H2-1*). Based on prior works on the helpfulness of revision in writing assistants (Mouchel et al., 2023), we also hypothesized that for learners with AI support, more feedback and revision rounds, as well as higher time spent in each revision round, would lead to higher learning gains (*H2-2*).

Our analyses revealed that users studied an average of 1.71 examples in the Worked w/o AI group (SD = 0.84), and an average of 2.00 in the Worked w/ AI group (SD = 1.32). A one-way ANOVA did not reveal any difference ($F(1, 46) = 0.80, p = 0.3768, \eta^2 = 0.02$). Additionally, the time in seconds they spent per each example was similar across groups (for Worked w/o AI: mean = 90.67 and SD = 67.40; for Worked w/ AI: mean = 82.13 and SD = 82.46; $F(1, 46) = 0.15, p = 0.7023, \eta^2 = 0.00$).

Our MLM analyses, including the two behavioural variables as fixed effects (see Section "Analytic approach") confirmed that neither behavioural feature contributed significantly to the prediction of learning gains measured by depth, breadth and combined scores in posttest near. Similarly, for the MLMs fit for posttest far, the effects of the two behavioural variables (number of studied examples and average time per example) on the depth and breadth scores was not significant. However, we did observe a trend to significance for the average time spent per example for the combined score ($t = 1.798$, $p = 0.079$), indicating that learners spending more time per examples tended to achieve a higher combined score. The detailed results of all MLM analyses can be found in Table A2 in the Appendix A.

Overall, our findings indicate that within the learning from *worked* examples paradigm, individual differences in the number of examples viewed did not significantly account for pretest to posttest learning gains, while time spent on examples seemed to have an effect on performance in posttest far, partially confirming *H2-1*.

To assess *H2-2*, we focused on learners using the version with AI support only. Learners in this condition conducted an average of 3.46 rounds of asking for feedback on their writing ($SD = 3.58$), suggesting that they generally appreciated the feedback functionality. They spent an average time of 299.18 seconds per round ($SD = 209.17$). 10 users (41.7%) used the feedback functionality only once. To measure whether the feedback-seeking had any effects on their learning for this group, we split the users into high- and low-performing groups. All users with a growth of two points or more in the combined score were classified into the high-performing group.[4] We then calculated the same metric separately for each subgroup, but did not find any difference between the two subgroups (for low-performing: a mean of 3.75 rounds of feedback ($SD = 4.65$) and a mean of 313.84 seconds per round ($SD = 221.02$); for high-performing: a mean of 3.20 rounds of feedback ($SD = 3.14$) and a mean of 302.85 seconds per round ($SD = 212.59$)).

Our MLM results for posttest near (incorporating the two revision behaviour features as fixed effects) revealed that there was no significant effect of revision behaviour on the depth, breath or combined scores. Similarly, for the MLMs fit for posttest far, the feedback-related measures failed to show a significant impact. These findings suggest that, contrary to our initial hypothesis (*H2-2*), the frequency of revisions and the time invested per revision did not meaningfully influence learning gains from pretest to posttest in the Worked w/ AI condition. The detailed results of these MLM analyses can be seen in Table A3 in the Appendix A.

To further explore whether the feedback module was beneficial in the learning beyond merely the number of times the users asked for feedback, our trained researcher annotated the first and last texts submitted in the feedback rounds by each user and calculated the breadth score (ie, adherence to the Gibbs reflective cycle). We found nine users (38%) with improvements in at least one class, which can explain the added benefits of the AI support for feedback in this condition. Out of the nine users, six of them also showed

improvement in the same classes from pretest to posttest near, further confirming the role of feedback in their learning process.

Interaction with the learning from modelling examples version

While users in the Modelling w/o AI version merely watched a video and then participated in a writing task without intelligent support, those in the Modelling w/ AI version participated in chatting with a conversational agent, particularly allowing them to ask follow-up questions and receive feedback on their sentences. We analysed the interaction behaviour on each of these two aspects:

Follow-up questions: Eight users (31%) asked the conversational agent to come up with more examples of a certain class, for a total of 12 times (three times for description, once for Feelings, two times for Evaluation, three times for Analysis, once for Conclusion and two times for Action Plan). Three users (12%), one being among the eight students, asked the conversational agent to explain the definition of a class more (once for Description and two times for Analysis).

Receiving feedback: While the conversational agent provided feedback on each sentence written by the user, contrary to our hypothesis that the feedback functionality helps them in learning the Gibbs reflective cycle components, we actually found that the feedback module had a relatively limited effect on their learning; out of the $26 \times 6 = 156$ feedback interactions users had with the tool, the users changed their answer from incorrect to correct after receiving feedback in only five cases. This shows that the initial explanation and examples provided in a conversational manner were sufficient for users to learn how to write a sentence from a specific class.

RQ3: User perception

In addressing our final research question, we evaluated learners' post-surveys (see Section “Analytic approach”). In line with previous works on writing assistants and educational tools showing higher perception metrics when embedding AI and intelligent support (Mejia-Domenzain et al., 2024; Wambsganss, Kueng, et al., 2021; Weber et al., 2024), we initially hypothesized that we would receive higher values for perception constructs (eg, perceived ease of use or excitement after interaction) in the versions of Reflectium with AI support (*H3-1*).

The mean per-construct scores (SD) are provided in Figure 7, while Table 6 reports the results of the statistical analyses. We found no significant difference among conditions in terms of the behavioural measures. In particular:

- Among the versions of Reflectium using the learning from *modelling* examples paradigm, the version with AI support scored higher than the version without AI support in all metrics, but the differences were not significant.
- Among the versions of Reflectium using the learning from *worked* examples paradigm, no consistent trend was found when comparing versions without and with AI support.
- When comparing the two learning from examples modalities, Modelling w/ AI did not show any consistent differences to worked w/ AI. On the contrary, Modelling w/o AI had a lower score than Worked w/o AI in all of the constructs, but the differences were not significant.

Based on these findings, we reject *H3-1*.

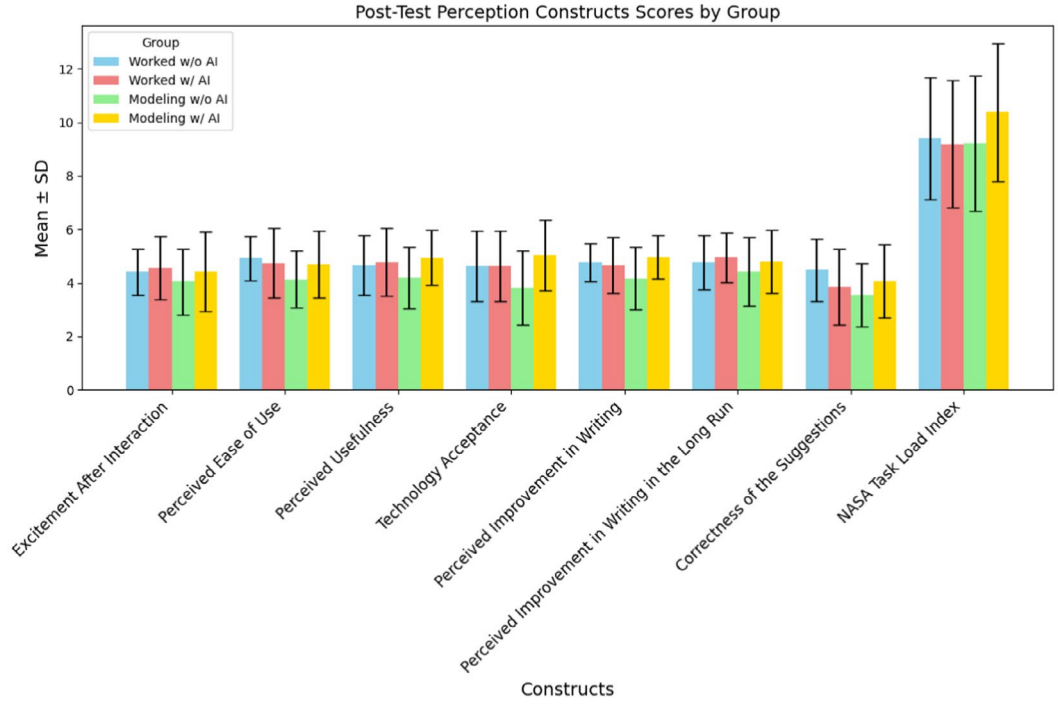

FIGURE 7 Scores for each posttest perception construct in each condition.

DISCUSSION

In this work, we set out to investigate how integrating AI-based methods and LLMs into the paradigm of *learning from examples* would impact learning outcomes, user interaction and perceived experiences. Particularly, to measure whether AI-based methods can further decrease the possibility of mediation and production deficiencies (Hübner et al., 2010), we designed a 2×2 study (inspired by previous studies in the domain of educational writing assistants (Mejia-Domenzain et al., 2024)), comparing the two AI-enabled versions of Reflectium using different *learning from examples modalities* versus their no-AI equivalent. Contrary to our hypothesis, we observed higher learning gains for the learning from *worked* examples version of Reflectium. However, confirming another of our hypotheses, we found a significant positive impact of including AI support on learning outcomes. The usefulness of AI support confirms our initial hypothesis and aligns with previous research indicating that AI-driven tools can enhance writing skills by providing personalized feedback and support (Lee et al., 2024).

While including AI support shows a positive impact, the investigations from RQ2 for the *worked* examples conditions do not reveal any significant effect of the feedback behaviour, which was the main differentiating factor in the AI-enabled versions. Thus, we hypothesize that the positive impact in the version with AI support might come from the principle of *positive reinforcement*, in which getting a confirmatory feedback can improve learning (Grünke et al., 2017). This hypothesis is consistent with findings employing positive reinforcement strategies, where students' motivation and learning outcomes improved through praising rewards (Hurlock, 1925; Simonsen et al., 2008). A follow-up study can measure this effect independently by, for example, tracking fine-grained student behaviour before and after they receive feedback on each sentence.

TABLE 6 Scores obtained for the postsurvey constructs in each condition.

Construct	Group	Mean ± SD	Statistical analysis
Excitement after interaction	Worked w/o AI	4.42 ± 0.86	$F(3, 96) = 0.83$, $p = 0.4801$, $\eta^2 = 0.03$
	Worked w/ AI	4.56 ± 1.18	
	Modelling w/o AI	4.04 ± 1.22	
	Modelling w/ AI	4.42 ± 1.49	
Perceived ease of use	Worked w/o AI	4.92 ± 0.83	$F(3, 96) = 2.22$, $p = 0.1808$, $\eta^2 = 0.06$
	Worked w/ AI	4.74 ± 1.31	
	Modelling w/o AI	4.13 ± 1.06	
	Modelling w/ AI	4.69 ± 1.25	
Perceived usefulness	Worked w/o AI	4.67 ± 1.11	$F(3, 96) = 1.98$, $p = 0.1953$, $\eta^2 = 0.06$
	Worked w/ AI	4.77 ± 1.26	
	Modelling w/o AI	4.19 ± 1.14	
	Modelling w/ AI	4.94 ± 1.03	
Technology acceptance	Worked w/o AI	4.62 ± 1.32	$F(3, 96) = 3.72$, $p = 0.1125$, $\eta^2 = 0.10$
	Worked w/ AI	4.62 ± 1.32	
	Modelling w/o AI	3.81 ± 1.39	
	Modelling w/ AI	5.04 ± 1.32	
Perceived improvement in writing	Worked w/o AI	4.75 ± 0.71	$F(3, 96) = 3.05$, $p = 0.1294$, $\eta^2 = 0.09$
	Worked w/ AI	4.65 ± 1.04	
	Modelling w/o AI	4.17 ± 1.18	
	Modelling w/ AI	4.96 ± 0.80	
Perceived improvement in writing in the long run	Worked w/o AI	4.75 ± 1.01	$F(3, 96) = 1.00$, $p = 0.4534$, $\eta^2 = 0.03$
	Worked w/ AI	4.96 ± 0.93	
	Modelling w/o AI	4.42 ± 1.28	
	Modelling w/ AI	4.81 ± 1.18	
Correctness of the suggestions	Worked w/o AI	4.48 ± 1.16	$F(3, 96) = 2.23$, $p = 0.1808$, $\eta^2 = 0.07$
	Worked w/ AI	3.85 ± 1.42	
	Modelling w/o AI	3.54 ± 1.19	
	Modelling w/ AI	4.06 ± 1.37	
NASA task load index	Worked w/o AI	9.41 ± 2.28	$F(3, 96) = 1.34$, $p = 0.3550$, $\eta^2 = 0.04$
	Worked w/ AI	9.19 ± 2.37	
	Modelling w/o AI	9.21 ± 2.54	
	Modelling w/ AI	10.39 ± 2.58	

On the contrary, contrary to our initial hypothesis, we found no differences between conditions regarding a set of postsurvey metrics spanning across the areas of usability, perceived learning, technology acceptance and cognitive load. This indicates that the improved learning gains of the AI-supported version of our tool did not come at the cost of a diminished learning experience. With that said, this observation aligns with studies suggesting that integrating AI into educational tools and writing assistants can enhance learning without increasing cognitive load or negatively impacting user experience (McLaren et al., 2015; Weber et al., 2024). When digging deeper into the reasons behind similarities and differences

among groups, we only found *limited* usage of a set of features unique to the versions of Reflectium with AI support. We hypothesize that this was the case because the users of our Prolific study did not have the motivation to explore all available functionalities in the tool, when not explicitly instructed to by a goal-setting strategy (Abbas & Gadiraju, 2022). In that case, our results can explain the relative indifference in the posttest constructs across conditions and necessitate measures to persuade or reward learners to use such features more often as part of a future follow-up study.

Together, these results support prior learning sciences literature (Hoogerheide et al., 2014) by indicating that incorporating LLMs into tutoring systems designed around the *learning from examples* paradigm can provide meaningful support for the development of students' reflective writing skills, in terms of the depth of reflections as well as nudging students to follow the theory model of the Gibbs reflective cycle. As a result, our work extends upon prior usages of LLMs for reflective writing (Kim et al., 2024; Li et al., 2023) by contributing to the theoretical aspect of embedding AI models in tutoring systems using the *learning from examples* paradigm and showing the effectiveness of AI support to mitigate the risks of mediation and production deficiencies (Hübner et al., 2010). We additionally uncover the differences in terms of learning gains between the two *learning from examples* paradigms and show generally higher scores obtained by the learning from *worked* examples paradigm, which stays as a well-known paradigm for teaching writing skills (Kyun et al., 2013; Mejia-Domenzain et al., 2024). Finally, we contribute to the practice by designing and implementing Reflectium, our reflective writing assistant, to embed the two learning modalities and help learners in writing their reflective texts. Our work extends upon prior reflection assistants (Kim et al., 2024; Kocielnik et al., 2018; Wolfbauer et al., 2022) by providing insights into the interplay of the *learning from examples* paradigms and AI support, and how it affects the learners in terms of learning how to improve the structure of their reflections.

There are several limitations of this study. First, we did not evaluate the final version of Reflectium in an in-person classroom and only relied on a Prolific online experiment for our results. We call for future work to evaluate how the different paradigms used in Reflectium can be transferred to other learning environments. We also call for future work to integrate more metrics to evaluate the quality of the reflective writings (eg, automated methods) or indirect learning gains (eg, scores on a quiz or test) in the study design. Additionally, the fine-tuned models that we used in Reflectium come with low F1 scores for certain classes, limiting their robustness to different sets of inputs from students. We call for future work to improve the accuracy of the models, e.g., by conducting data augmentation processes. Finally, Reflectium, specifically the Modelling w/ AI group, can benefit significantly from more natural modalities of interacting with learners (eg, a voice-based conversational agent), which we leave to future works to explore.

CONCLUSION

In this work, we explored the integration of AI-enabled assistants into reflective writing education through the paradigms of learning from *worked* and *modelling* examples. We developed Reflectium, our intelligent assistant for reflective writing, in four versions: two versions with and without AI models for each of the two learning from examples paradigms. In an online study on Prolific, we demonstrated promising results regarding the positive effect of integrating AI models into Reflectium. Our findings highlight the potential of integrating adaptive AI support into the paradigm of *learning from examples* to help learners in writing reflectively and improving their metacognitive skills.

FUNDING INFORMATION
This project is supported by the State Secretariat for Education, Research and Innovation (SERI), Switzerland.

CONFLICT OF INTEREST STATEMENT
The authors do not have any conflict of interest.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT
This study was approved by the university's ethics review board.

ORCID
Seyed Parsa Neshaei https://orcid.org/0000-0002-4794-395X
Paola Mejia-Domenzain https://orcid.org/0000-0003-1242-3134
Richard Lee Davis https://orcid.org/0000-0002-6175-9200

Endnotes
[1] The video and the presentation slides will be published.

[2] Analysis of the answers provided to the open questions is not included in this paper and is left for future follow-up work.

[3] Using the *lme4* and *afex* R packages.

[4] We excluded users with a low combined score growth but a pretest combined score of 5 or more from the low-performing groups, as this indicated sufficient prior knowledge and thus limited possibility of learning further.

REFERENCES
Abbas, T., & Gadiraju, U. (2022). Goal-setting behavior of workers on crowdsourcing platforms: An exploratory study on MTurk and prolific. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. *10*, pp. 2–13). https://ojs.aaai.org/index.php/HCOMP/issue/view/530.

Adeani, I. S., Febriani, R. B., & Syafryadin, S. (2020). Using GIBBS' reflective cycle in making reflections of literary analysis. *Indonesian EFL Journal*, *6*(2), 139–148.

Afrin, T., & Litman, D. (2023). Predicting desirable revisions of evidence and reasoning in argumentative writing. *arXiv preprint arXiv:2302.05039*.

Ahmadpour, N., Shariati, A., & Moghadam, M. P. (2025). Effect of narrative writing based on Gibbs' reflective model on the empathy and communication skills of nursing students. *BMC Medical Education*, *25*(1), 10.

Ahmed, A. M. (2020). From reluctance to addiction: The impact of reflective journals on Qatari undergraduate students' learning. *Reflective Practice*, *21*(2), 251–270.

Al-Mutawa, N. A. A., Mahmoud, M. H., Baisa, K. G. G., Daher-Nashif, S., & Al Wahedi, Z. (2024). Reflective writing among healthcare practitioners in primary care: A qualitative study from Qatar. *Cogent Education*, *11*(1), 2373555.

Andrew, J., & Meligrana, J. (2012). Evaluating the use of role playing simulations in teaching negotation skills to university students. *Creative Education*, *3*(6), 696–707.

Aneeis Hashim, S. N., Yaacob, A., Suryani, I., Mohd Asraf, R., Bahador, Z., & Supian, N. (2023). Exploring the use of Gibbs' reflective model in enhancing in-service ESL teachers' reflective writing. *Arab World English Journal (AWEJ)*, *14*, 236–253.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs.

Bjerrum, A. S., Hilberg, O., van Gog, T., Charles, P., & Eika, B. (2013). Effects of modelling examples in complex procedural skills training: A randomised study. *Medical Education*, *47*(9), 888–898.

Bofferding, L., Kocabas, S., Aqazade, M., Haiduc, A.-M., & Chen, L. (2022). The effect of play and worked examples on first and third graders' creating and debugging of programming algorithms. In *Computational thinking in PreK-5: Empirical evidence for integration and future directions* (pp. 19–29). ACM. https://dl.acm.org/doi/abs/10.1145/3507951.3519284.

Boud, D., Keogh, R., & Walker, D. (2013). *Reflection: Turning experience into learning*. Routledge.

Braaksma, M. A., Rijlaarsdam, G., & Van Den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology*, *94*(2), 405.
Buschek, D., Zürn, M., & Eiband, M. (2021). The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native English writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).
Cattaneo, A., & Boldrini, E. (2016). Individual and collaborative writing-to-learn activities in vocational education: An overview of different instructional strategies. In *Writing for professional development* (pp. 188–208). Brill.
Cattaneo, A. A., & Motta, E. (2021). "I reflect, therefore I am… a good professional". On the relationship between reflection-on-action, reflection-in-action and professional performance in vocational education. *Vocations and Learning*, *14*(2), 185–204.
Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.
Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. *International Journal of Science Education*, *28*(11), 1315–1346.
Cochran-Smith, M. (2005). *Studying teacher education: What we know and need to know* (Vol. *56*, pp. 301–306). Sage.
Colomer, J., Serra, T., Cañabate, D., & Bubnys, R. (2020). Reflective learning in higher education: Active methodologies for transformative practices. *Sustainability*, *12*(9), 3827.
Cotos, E., Huffman, S., & Link, S. (2020). Understanding graduate writers' interaction with and impact of the research writing tutor during revision. *Journal of Writing Research*, *12*(1), 187–232.
Erümit, A. K., & Çetin, İ. (2020). Design framework of adaptive intelligent tutoring systems. *Education and Information Technologies*, *25*(5), 4477–4500.
Ezezika, O., & Johnston, N. (2023). Development and implementation of a reflective writing assignment for undergraduate students in a large public health biology course. *Pedagogy in Health Promotion*, *9*(2), 101–115.
Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods. Further education unit*. Oxford Polytechnic.
Glasnovic Gracin, D. (2018). Requirements in mathematics textbooks: A five-dimensional analysis of textbook exercises and examples. *International Journal of Mathematical Education in Science and Technology*, *49*(7), 1003–1024.
Göldi, A., Wambsganss, T., Neshaei, S. P., & Rietsche, R. (2024). Intelligent support engages writers through relevant cognitive processes. In Proceedings of the CHI conference on human factors in computing systems. (pp. 1–12). Hawaii, USA.
Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In *Psychology of learning and motivation* (Vol. *29*, pp. 411–436). Elsevier.
Groenendijk, T., Janssen, T., Rijlaarsdam, G., & van den Bergh, H. (2013). Learning to be creative. The effects of observational learning on students' design products and processes. *Learning and Instruction*, *28*, 35–47.
Grünke, M., Sperling, M., & Burke, M. D. (2017). The impact of explicit timing, immediate feedback, and positive reinforcement on the writing outcomes of academically and behaviorally struggling fifth-grade students. *Insights into Learning Disabilities*, *14*(2), 135–153.
Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. *50*, pp. 904–908). Sage.
Hilbert, T. S., Renkl, A., Schworm, S., Kessler, S., & Reiss, K. (2008). Learning to teach with worked-out examples: A computer-based learning environment for teachers. *Journal of Computer Assisted Learning*, *24*(4), 316–332.
Hommel, M., Fürstenau, B., & Mulder, R. H. (2023). Reflection at work—A conceptual model and the meaning of its components in the domain of VET teachers. *Frontiers in Psychology*, *13*, 923888.
Hoogerheide, V., Loyens, S. M., & Van Gog, T. (2014). Comparing the effects of worked examples and modeling examples on learning. *Computers in Human Behavior*, *41*, 80–91.
Huang, Y.-T., Chen, M. C., & Sun, Y. S. (2018). Development and evaluation of a personalized computer-aided question generation for English learners to improve proficiency and correct mistakes. *arXiv preprint arXiv:1808.09732*.
Hübner, S., Nückles, M., & Renkl, A. (2010). Writing learning journals: Instructional support to overcome learning-strategy deficits. *Learning and Instruction*, *20*(1), 18–29.
Hurlock, E. B. (1925). An evaluation of certain incentives used in school work. *Journal of Educational Psychology*, *16*(3), 145.
Jackson, A., Gaudet, L., McDaniel, L., & Wright, O. (2008). Instructional design: Benefits and advantages of worked examples throughout the e-learning experience. In *E-learn: World conference on e-learning in corporate, government, healthcare, and higher education* (pp. 1677–1680). Association for the Advancement of Computing in Education (AACE).

Kim, J., Suh, S., Chilton, L. B., & Xia, H. (2023). Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (pp. 115–135). Pittsburgh, Pennsylvania, USA.

Kim, T., Bae, S., Kim, H. A., Lee, S.-W., Hong, H., Yang, C., & Kim, Y.-H. (2024). MindfulDiary: Harnessing large language model to support psychiatric patients' journaling. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–20). Hawaii, USA.

Kingkaew, C., Theeramunkong, T., Supnithi, T., Chatpreecha, P., Morita, K., Tanaka, K., & Ikeda, M. (2023). A learning environment to promote awareness of the experiential learning processes with reflective writing support. *Education Sciences*, *13*(1), 64.

Kocielnik, R., Xiao, L., Avrahami, D., & Hsieh, G. (2018). Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 1–26.

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT Press.

Kumar, H., Xiao, R., Lawson, B., Musabirov, I., Shi, J., Wang, X., Luo, H., Williams, J. J., Rafferty, A., Stamper, J., & Liut, M. (2024). Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In Proceedings of the Eleventh ACM Conference on Learning@ Scale (pp. 86–97). Atlanta, Georgia, USA.

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The effect of worked examples when learning to write essays in English literature. *The Journal of Experimental Education*, *81*(3), 385–408.

Lee, M., Gero, K. I., Chung, J. J. Y., Shum, S. B., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E. A., August, T., Bhat, A., Choksi, M. Z., Dutta, S., Guo, J. L. C., Hoque, M. N., Kim, Y., Knight, S., Neshaei, S. P., … Siangliulue, P. (2024). A design space for intelligent and interactive writing assistants. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–35). Hawaii, USA.

Lee, M. K., Cheung, C. M., & Chen, Z. (2005). Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information & Management*, *42*(8), 1095–1104.

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, *4*, 100140.

Markkanen, P., Välimäki, M., Anttila, M., & Kuuskorpi, M. (2020). A reflective cycle: Understanding challenging situations in a school setting. *Educational Research*, *62*(1), 46–62.

McGuire, L., Lay, K., & Peters, J. (2009). Pedagogy of reflective writing in professional education. *Journal of the Scholarship of Teaching and Learning*, *9*, 93–107.

McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education*, *25*, 520–542.

Mejia-Domenzain, P., Frej, J., Neshaei, S. P., Mouchel, L., Nazaretsky, T., Wambsganss, T., Bosselut, A., & Käser, T. (2024). Enhancing procedural writing through personalized example retrieval: A case study on cooking recipes. *International Journal of Artificial Intelligence in Education*, *35*, 1–37.

Mejia-Domenzain, P., Marras, M., Giang, C., Cattaneo, A., & Käser, T. (2022). Evolutionary clustering of apprentices' self-regulated learning behavior in learning journals. *IEEE Transactions on Learning Technologies*, *15*(5), 579–593.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*(1), 465–489.

Middleton, R. (2017). Critical reflection: The struggle of a practice developer. *International Practice Development Journal*, *7*, 1–6.

Mirowski, P., Mathewson, K. W., Pittman, J., & Evans, R. (2023). Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–34). Hamburg, Germany.

Mouchel, L., Wambsganss, T., Mejia-Domenzain, P., & Käser, T. (2023). Understanding revision behavior in adaptive writing support systems for education. *arXiv preprint arXiv:2306.10304*.

Moussa-Inaty, J. (2015). Reflective writing through the use of guiding questions. *International Journal of Teaching and Learning in Higher Education*, *27*(1), 104–113.

Nehyba, J., & Štefánik, M. (2023). Applications of deep language models for reflective writings. *Education and Information Technologies*, *28*(3), 2961–2999.

Neshaei, S. P., Wambsganss, T., El Bouchrifi, H., & Käser, T. (2025). MindMate: Exploring the Effect of Conversational Agents on Reflective Writing. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*(pp. 1–9. https://doi.org/10.1145/3706599.3720029

Neshaei, S. P., Rietsche, R., Su, X., & Wambsganss, T. (2024). Enhancing peer review with AI-powered suggestion generation assistance: Investigating the design dynamics. In Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 88–102). Greenville, South Carolina, USA.

Nurlatifah, L., Purnawarman, P., & Sukyadi, D. (2023). The implementation of reflective assessment using Gibbs' reflective cycle in assessing students' writing skill. In *AIP Conference Proceedings* (Vol. *2621*). AIP Publishing.

O'Loughlin, V. D., & Griffith, L. M. (2020). Developing student metacognition through reflective writing in an upper level undergraduate anatomy course. *Anatomical Sciences Education*, *13*(6), 680–693.

Peng, Z., Guo, Q., Tsang, K. W., & Ma, X. (2020). Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–15).

Perry, J., Lundie, D., & Golder, G. (2019). Metacognition in schools: What does the literature suggest about the effectiveness of teaching metacognition in schools? *Educational Review*, *71*(4), 483–500.

Prior, J., Ferguson, S., & Leaney, J. (2016). Reflection is hard: Teaching and learning reflective practice in a software studio. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1–8). Canberra, Australia.

Raj, A. G. S., Gu, P., Zhang, E., Williams, J., Halverson, R., & Patel, J. M. (2020). Live-coding vs static code examples: Which is better with respect to student learning and cognitive load? In Proceedings of the Twenty-Second Australasian Computing Education Conference (pp. 152–159). Melbourne VIC Australia.

Recker, M. M., & Pirolli, P. (1995). Modeling individual differences in students' learning strategies. *The Journal of the Learning Sciences*, *4*(1), 1–38.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*(1), 1–29.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction*, *12*(5), 529–556.

Rolfe, G., Freshwater, D., & Jasper, M. (2001). *Critical reflection for nursing and the helping professions: A user's guide*. Palgrave MacMillan

Schön, D. A. (2017). *The reflective practitioner: How professionals think in action*. Routledge.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, *31*(3), 351–380.

Su, X., Wambsganss, T., Rietsche, R., Neshaei, S. P., & Käser, T. (2023). Reviewriter: AI-Generated instructions for peer review writing. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 57–71). Toronto, Canada.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89.

Ullmann, T. D. (2015). *Automated detection of reflection in texts: A machine learning based approach*. Open University.

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*, 155–174.

Wale, B. D., & Kassahun, Y. F. (2024). The transformative power of AI writing technologies: Enhancing EFL writing instruction through the integrative use of Writerly and Google Docs. *Human Behavior and Emerging Technologies*, *2024*(1), 9221377.

Wambsganss, T., Benke, I., Maedche, A., Koedinger, K., & Käser, T. (2024). Evaluating the impact of learner control and interactivity in conversational tutoring systems for persuasive writing. *International Journal of Artificial Intelligence in Education*, 1–32. https://doi.org/10.1007/s40593-024-00409-x.

Wambsganss, T., Guggisberg, S., & Söllner, M. (2021). Arguebot: A conversational agent for adaptive argumentation feedback. In *Innovation through information systems: Volume II: A collection of latest research on technology issues* (pp. 267–282). Springer.

Wambsganss, T., Kueng, T., Soellner, M., & Leimeister, J. M. (2021). ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13).

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14).

Wambsganss, T., Su, X., Swamy, V., Neshaei, S. P., Rietsche, R., & Käser, T. (2023). Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. *arXiv preprint arXiv:2311.03311*.

Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! For learning—A literature review. *Computers & Education*, *149*, 103818.

Weber, F., Wambsganss, T., Neshaei, S. P., & Soellner, M. (2024). Legal-writer: An intelligent writing support system for structured and persuasive legal case writing for novice law students. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–23). Hawaii, USA.
Williams, K., Woolliams, M., & Spiro, J. (2020). *Reflective writing*. Bloomsbury Publishing.
Wolfbauer, I., Bangerl, M. M., Maitz, K., & Pammer-Schindler, V. (2023). Rebo at work: Reflecting on working, learning, and learning goals with the reflection guidance chatbot for apprentices. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. (pp. 1–7). Hamburg, Germany
Wolfbauer, I., Pammer-Schindler, V., Maitz, K., & Rosé, C. P. (2022). A script for conversational reflection guidance: A field study on developing reflection competence with apprentices. *IEEE Transactions on Learning Technologies*, *15*(5), 554–566.
Wong, Y. M., Mansor, R., & Samsudin, S. (2016). The use of critical reflection manual in writing reflective journal: A case study of Malaysian student teachers' perceptions. *Geografia*, *12*(1), 8–18.
Wu, S., Reynolds, L., Li, X., & Guzmán, F. (2019). Design and evaluation of a social media writing support tool for people with dyslexia. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–14). Glasgow, UK.

How to cite this article: Neshaei, S. P., Mejia-Domenzain, P., Davis, R. L., & Käser, T. (2025). Metacognition meets AI: Empowering reflective writing with large language models. *British Journal of Educational Technology*, *00*, 1–33. https://doi.org/10.1111/bjet.13601

APPENDIX A

TABLE A1 Individual Cronbach's alpha values for each pre-survey and post-survey construct.

Stage	Construct	Cronbach's alpha
Pre-survey	IT usage	0.8453
	Feedback-seeking	0.8026
	Reflective writing knowledge	0.7678
Post-survey	Excitement after interaction	0.8026
	Perceived ease of use	0.8809
	Perceived usefulness	0.8080
	Technology acceptance	N/A
	Perceived improvement in writing	0.7378
	Perceived improvement in writing in the long run	N/A
	Correctness of the suggestions	0.8382

Note: N/A indicates constructs with only one question. We did not include NASA TLX in the analysis, as it is a complete survey that has been validated in prior works (Hart, 2006).

TABLE A2 Mixed linear model (MLM) results for depth, breadth and combined scores of RQ2 (interaction with the worked examples).

	Posttest near			Posttest far		
	Depth	**Breadth**	**Combined**	**Depth**	**Breadth**	**Combined**
(Intercept)	2.86*** (0.23)	4.66*** (0.43)	4.58*** (0.60)	1.93*** (0.24)	2.78*** (0.40)	1.97*** (0.56)
Time (pre)	−0.79*** (0.18)	−1.58*** (0.33)	−2.13*** (0.43)	0.04 (0.19)	−0.33 (0.31)	−0.17 (0.42)
AIS (Static)	−0.47* (0.20)	−0.69 (0.38)	−1.28* (0.52)	−0.45* (0.22)	−0.21 (0.35)	−0.89 (0.49)
Num. of Examples	−0.07 (0.07)	−0.21 (0.14)	−0.22 (0.19)	−0.09 (0.08)	0.04 (0.12)	−0.05 (0.18)
Time per Example	−0.0006 (0.0011)	0.00001 (0.002)	0.0007 (0.0028)	0.0010 (0.0011)	0.0017 (0.0018)	0.0047 (0.0026)
Time (pre)×AIS (Static)	0.25 (0.25)	0.29 (0.46)	0.92 (0.60)	0.21 (0.27)	−0.13 (0.44)	0.54 (0.59)
N (learner)	48	48	48	48	48	48
AIC	235.59	346.17	402.12	245.93	335.38	393.66
BIC	256.10	366.68	422.63	266.45	355.89	414.17
R^2 (fixed)	0.23	0.27	0.23	0.08	0.05	0.08
R^2 (total)	0.43	0.46	0.48	0.29	0.24	0.33

Abbreviations: AIS, AI support; pre, pretest.

*$p < 0.05$; ***$p < 0.001$.

TABLE A3 Mixed linear model (MLM) results for depth, breadth and combined scores of RQ2 (revision behaviour).

	Posttest near			Posttest far		
	Depth	**Breadth**	**Combined**	**Depth**	**Breadth**	**Combined**
(Intercept)	2.74*** (0.35)	4.96*** (0.77)	4.89*** (0.91)	1.71** (0.47)	2.74** (0.78)	1.62 (1.19)
Time (pre)	−0.79*** (0.18)	−1.58*** (0.33)	−2.13*** (0.46)	0.04 (0.18)	−0.33 (0.32)	−0.17 (0.39)
Num. of Examples	−0.12 (0.08)	−0.25 (0.18)	−0.25 (0.21)	−0.10 (0.11)	0.04 (0.18)	−0.07 (0.28)
Time per Example	−0.0011 (0.0012)	−0.0002 (0.0028)	0.0006 (0.0033)	0.0011 (0.0017)	0.0014 (0.0028)	0.0044 (0.0044)
Num. of Revisions	0.03 (0.03)	−0.04 (0.08)	−0.02 (0.09)	0.02 (0.05)	−0.02 (0.08)	0.06 (0.12)
Time per Revision	0.0005 (0.0006)	−0.0002 (0.0014)	−0.0005 (0.0016)	0.0006 (0.0009)	0.0005 (0.0014)	0.0007 (0.0021)
N (learner)	24	24	24	24	24	24
AIC	143.79	203.17	223.73	154.96	200.86	226.58
BIC	158.76	218.13	238.70	169.93	215.83	241.55
R^2 (fixed)	0.31	0.28	0.30	0.05	0.03	0.04
R^2 (total)	0.39	0.50	0.41	0.47	0.38	0.55

Abbreviation: pre, pretest.

**$p < 0.01$; ***$p < 0.001$.

A.1. | Prompts and textual responses in Reflectium

The following are textual responses used in the version of Reflectium using the learning from *modelling* examples paradigm with AI support:

System prompt
All your responses should be relevant to a nursing vocational student who is learning how to reflect on the events happened in their theory or practice sessions using the Gibbs reflective cycle. Respond in the same language that the user uses. Please respond very briefly and avoid long responses.
Welcome message
Welcome to Reflectium! In this session, you are expected to reflect on a situation in the workplace that you think you handled very well. In the process, I will help you by teaching you about the Gibbs reflective cycle, which can help you better reflect on the situation happened in the workplace for you. Let me know if you are ready to start! *Next message suggestions*: • Yes, I'm ready!
Description: Explain
Great! Let's start with the first stage of the Gibbs reflective cycle: Description. In this stage, you will describe the event or situation that you are reflecting on. One example of a description could be: 'Yesterday, I was assigned to a new patient with dementia. I had to take care of him, as a part of my duties'. *Next message suggestions*: • I understood it! • Can you explain it again? • Can you give me another example?
Description: Example to Evaluate
Perfect :) Now let's see if you can come up with an example of a description. Please describe an event or situation that you are reflecting on. Remember, the description should be clear and detailed. Take your time! *Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown*: Sorry, I found that your response does not indicate a description. Please try again. Common issues in writing a description include: • Not providing enough details • Not being clear about the event or situation • Not mentioning the key points
Feelings: Explain
Great! Now let's move on to the second stage of the Gibbs reflective cycle: Feelings. In this stage, you will describe how you felt during the event or situation. One example of a feeling could be: 'I felt anxious and overwhelmed when I realized that the patient's condition was deteriorating. However, I was also happy because I had studied the symptoms and knew how to respond'. *Next message suggestions*: • I understood it! • Can you explain it again? • Can you give me another example?

Feelings: Example to Evaluate

Perfect :) Now let's see if you can come up with an example of a feeling. Please describe how you felt during an event or situation that you are reflecting on. Remember, the feeling should be honest and detailed. Take your time!
Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown:
Sorry, I found that your response does not indicate a feeling. Please try again. Common issues in writing a feeling include:
- Mentioning on factual information instead of feelings
- Failing to express feelings using clear adjectives
- Evaluating the situation critically instead of merely focusing on personal feelings at this stage

Evaluation: Explain

Great! Now let's move on to the third stage of the Gibbs reflective cycle: Evaluation.
In this stage, you will evaluate the positive and negative aspects of the event or situation that you are reflecting on.
One example of an evaluation could be: 'I think I was able to handle the situation well. However, I also felt that I could have communicated more effectively with the patient'.
Next message suggestions:
- I understood it!
- Can you explain it again?
- Can you give me another example?

Evaluation: Example to evaluate

Perfect :) Now let's see if you can come up with an example of an evaluation. Please evaluate the positive and negative aspects of an event or situation that you are reflecting on. Remember, the evaluation should be balanced and detailed. Take your time!
Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown:
Sorry, I found that your response does not indicate an evaluation. Please try again. Common issues in writing an evaluation include:
- Not providing a balanced view of the situation
- Not mentioning both positive and negative aspects
- Focusing on personal feelings rather than a holistic or detailed evaluation

Analysis: Explain

Great! Now let's move on to the fourth stage of the Gibbs reflective cycle: Analysis.
In this stage, you will analyze the event or situation that you are reflecting on, providing the reasons behind the points you mentioned in the Evaluation part.
One example of an analysis could be: 'I think the reason I was able to follow the correct procedures was because I had practiced the textbook given by our teacher during the past week'.
Next message suggestions:
- I understood it!
- Can you explain it again?
- Can you give me another example?

Analysis: Example to evaluate

Perfect :) Now let's see if you can come up with an example of an analysis. Please analyze the event or situation that you are reflecting on, providing the reasons behind the points you mentioned in the Evaluation part. Remember, the analysis should be logical and detailed. Take your time!
Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown:
Sorry, I found that your response does not indicate an analysis. Please try again. Common issues in writing an analysis include:
- Not providing logical reasons behind the evaluation (eg, not using words and expressions, such as 'because')
- Not explaining the connection between the evaluation and the analysis
- Merely indicating the positive and negative aspects without providing reasons

(Continues)

Conclusion: Explain

Great! Now let's move on to the fifth stage of the Gibbs reflective cycle: Conclusion.
In this stage, you will draw conclusions from the event or situation that you are reflecting on and think about what you have learned from it.
One example of a conclusion could be: 'From my experience with the patient, I learned that it is important to remain calm and follow the correct procedures in emergency situations'.
Next message suggestions:
- I understood it!
- Can you explain it again?
- Can you give me another example?

Conclusion: Example to evaluate

Perfect :) Now let's see if you can come up with an example of a conclusion. Please draw conclusions from the event or situation that you are reflecting on and think about what you have learned from it. Remember, the conclusion should be insightful and detailed. Take your time!
Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown:
Sorry, I found that your response does not indicate a conclusion. Please try again. Common issues in writing a conclusion include:
- Not reflecting on what has been learned from the situation
- Not providing insights or lessons learned
- Failing to connect the conclusion to the event or situation

Action plan: Explain

Great! Now let's move on to the sixth stage of the Gibbs reflective cycle: Action Plan.
In this stage, you will think about what you would do differently in a similar situation in the future, identify areas for future improvement and plan how to address them.
One example of an action plan could be: 'In future situations, I will make sure to practice the correct procedures, even if I feel anxious or overwhelmed'.
Next message suggestions:
- I understood it!
- Can you explain it again?
- Can you give me another example?

Action plan: Example to evaluate

Perfect :) Now let's see if you can come up with an example of an action plan. Please think about what you would do differently in a similar situation in the future. Remember, the action plan should be specific and actionable. Take your time!
Goes to the next Gibbs reflective cycle component, in case the classifier labels the input of the user from the same class. Otherwise, the following error message is shown:
Sorry, I found that your response does not indicate an action plan. Please try again. Common issues in writing an action plan include:
- Not being specific or actionable
- Failing to identify areas for future improvement
- Not planning how to address the areas for improvement

Congratulation

Congratulations! *[party popper emoji]* You have successfully completed the Gibbs reflective cycle. You now know how to describe an event or situation, share your feelings, evaluate the positive and negative aspects, analyze the event, draw conclusions and think about what you would do differently in the future. Well done! :)

The following are the prompts used by the authors to pre-generate the worked examples that are embedded in Reflectium:

Generating a reflective text for inclusion as a worked example in Reflectium: starting from Description

You are a nursing vocational student who is an expert in the Gibbs reflective cycle. You are asked to provide a worked example of a reflection (from your experience in a real-world nursing and caring scenario) that follows the Gibbs reflective cycle. The classes are: Description (describing what happened), Feelings (thoughts and feelings of the situation), Evaluation (good or bad aspects of the event), Analysis (the reasons behind good and bad aspects of the event), Conclusion (concluding what has been learned from the event) and Action Plan (outlining suggestions to do differently in the future). First, start with the Description class, describing a recent event happened in your nursing practice sessions at the hospital. Provide only one sentence of class Description and nothing else.

Generating a reflective text for inclusion as a worked example in Reflectium: Extending with other Gibbs reflective cycle components

You are an expert in the Gibbs reflective cycle. You are asked to provide a worked example of a reflection (from your experience in a real-world nursing and caring scenario) that follows the Gibbs reflective cycle. The classes are: Description (describing what happened), Feelings (thoughts and feelings of the situation), Evaluation (good or bad aspects of the event), Analysis (the reasons behind good and bad aspects of the event), Conclusion (concluding what has been learned from the event) and Action Plan (outlining suggestions to do differently in the future). This your current text until now:
“‘

```
{current partial text}
```

’”

Now, please extent the text with only one sentence of class `{Gibbs reflective cycle component}`. Only output the new sentence and nothing else.