Journal of Pedagogical Sociology and Psychology
Volume 7, Issue 4, 2025
https://doi.org/10.33902/jpsp.202536789

Research Article

Probing Chain-of-Thought (ProCoT): Stimulating critical thinking and writing of students through engagement with large language models

Tosin Adewumi, Lama Alkhaled, Claudia Buck, Sergio Serrano Hernández, Saga Brilioth, Mkpe Kekung, Yelvin Ragimov and Elisa Barney

Luleå University of Technology, Sweden

Correspondence should be addressed to Tosin Adewumi tosin.adewumi@ltu.se
Received 11 June 2025; Revised 15 October 2025; Accepted 20 October 2025

We introduce a novel writing method called Probing Chain-of-Thought, which potentially prevents students from cheating using a large language model while enhancing their critical thinking. large language models have disrupted education and many other fields. For fear of students cheating, many educationists have resorted to banning their use. We conduct studies in two different courses with 65 students using qualitative research design primarily (i.e. phenomenological) and quantitative methods. The students in each course were asked to prompt a large language model of their choice with one question from a set of four (random) questions and required to affirm or refute statements in the large language model output by using peer-reviewed references as evidence. In addition, the rubric for assessing the students writing included 5 more criteria: focus, logic, content, style and correctness. The average success rate of the writing of students based on the criteria for the two cases is 79.49% (±12.82%). The results of the rubric assessment show two things: (1) Probing Chain-of-Thought stimulates critical thinking and writing of students through engagement with large language models when we compare the large language models-only output to Probing Chain-of-Thought output and (2) Probing Chain-of-Thought may prevent cheating because of clear limitations in the concerned large language models when we compare students' Probing Chain-of-Thought output to large language models' Probing Chain-of-Thought output. In quantitative analysis, we also discover that most students prefer to give answers in fewer words than large language models, which are typically verbose. The average word counts for students in the first course, ChatGPT 3.5, and Phind (v8) are 208, 391 and 383, respectively, while it is 405, 356, and 315 for students, ChatGPT 3.5, and BingAI, respectively, in the second course, where we enforced a minimum word-count of 300 for the students. We provide access to the outputs for possible assessments (available after review).

Keywords: ChatGPT, cheating, education, pedagogy, ProCoT, LLM

1. Introduction

The evidence of critical thinking is important in student writing. A chain-of-thought [CoT] prompting allows a Large Language Model [LLM] that is sufficiently powerful to solve a problem as a clear sequence of steps before a final output and the way a student engages with this output may reveal evidence of critical thinking or the lack of it. An LLM, which is typically a deep neural probabilistic model (Brown et al., 2020), such as ChatGPT (http://www.openai.com), has the ability to generate convincing outputs when given inputs, called prompts. The use of such Artificial Intelligence [AI] tools has gained immense popularity in many fields in the last few years (Allam et al., 2023). They have also gained notoriety for suspected cheating by students in education. This is besides the hallucination (or ungrounded addition) problem LLMs have (Adewumi et al., 2022; Rawte et al., 2023), which occurs partly because they are typically trained on static data, with no updates through time. Humans, on the other hand, can source and use real-world context to verify the veracity of claims (Maslej et al., 2023).

Cheating in writing by students by whatever means has always been a problem in education (Pettersson et al., 2024). One common example of a problem is plagiarism, where someone else's work is presented as one's own, intentionally or otherwise (Barrett & Malcolm, 2006). Such a

problem potentially has become worse with the introduction of LLMs for two main reasons. These are (1) LLMs can generate plausible human-like outputs quickly and (2) the outputs of LLMs do not require critical thinking on the part of the user or student. Topic-based writing has always been a common way of assessing students (Todd et al., 2004). However, the capability of LLMs to completely or partially generate content as desired by a user appears to be putting a threat to students writing on their own. This highlights a gap, especially as there aren't sufficiently enough studies on their effective use in education and some educationists are advocating a ban on their use in education (Kishore et al., 2023). Instead, we believe there should be useful ways of students engaging with LLMs in a beneficial way. This makes an important rationale for doing this study.
We, therefore, introduce Probing Chain-of-Thought [ProCoT], a method aimed at stimulating critical thinking and writing of students through engagement with LLMs and the research question we address in this work is: "How effective is ProCoT with LLMs in stimulating critical thinking and preventing cheating in writing by students?" The generative power of LLMs makes it imperative to ask similar questions regarding coding, mathematics and other similar tasks. We believe ProCoT will empower educators to no longer fear LLMs but see them as technological tools and students will become better critical thinkers. Also, it may be applied to any field that uses writing. Our contributions in this work are stated below.

(1) ProCoT reduces the possibility of cheating through LLMs by students.

(2) ProCoT stimulates critical thinking and writing of students through engagement with LLMs

(3) ProCoT will provide valuable data to further train LLMs to be better models without privacy issues.

2. Literature Review

Mehta and Al-Mahrooqi (2015) showed that critical thinking is enhanced when students engage with or write on areas previously discussed. It follows then that ProCoT can enhance critical thinking in students when they engage with an existing content and write about it. Indeed, Wu (2024) identified some examples of critical thinking design in writing (e.g. thinking independently or thinking comparatively). Usually, in academic or essay writing, students are asked to demonstrate a point of view (Paltridge & Starfield, 2013) and this may be supported by evidence (e.g. through peer-reviewed references). This allows students to show what they have learnt and build their writing skills. It is a means of demonstrating critical thinking, which has been described as the art of evaluating one's thinking with the aim of making it better in terms of clarity and accuracy by removing biases and challenging assumptions (Aveyard & Waite, 2024; Mehta & Al-Mahrooqi, 2015; Paul & Elder, 2013; Wu, 2024).

2.1. Theories of Essay Writing

How students develop academic writing are explained in some theories. Among such theories are (1) process, (2) cognitive process, (3) rhetorical, (4) genre, (5) expressivist, and (6) post-process. The process theory by Murray (1972) is a dominant theory that advocates writing as a process (involving the three stages of prewriting, writing and rewriting instead of a finished product. It has been criticized by others as closing students' minds because it makes a linear progression to a predetermined end instead of opening up other possibilities (Heilker, 1996). The cognitive process theory by Flower and Hayes (1981) focuses mainly on examining what guides the decisions writers make as they write, including decisions about diction, syntactic patterns, and content. Rhetoricists, such as Aristotle, view rhetoric as focusing on the writer-audience- discourse (or writing) relationship. It is designed to achieve some purpose in a social situation by using emotion and logic (Aristotle, 2007).

The genre theory situates writing as different types of discourse (e.g. tragedy, sonnet, etc.) of rhetoric (Freedman & Medway, 2003). The expressivist theory advocated by Elbow (1998) recommends a writer's personal expression and encourages sharing of some of the writing with others without disparaging reviews but with the intention of seeing other perspectives. This

approach may not always be very suitable in academic writing, which emphasizes critical reviews. Finally, the post-process theory challenges process theory and endorses the notion that no codifiable or generalizable writing process exists (Van Cleave, 2024). Three assumptions are made in post-process theory. These are (1) writing is public, (2) writing is situated, and (3) writing is interpretive, thereby suggesting that writing has no one-size-fits-all process and is a complex phenomenon. Post-process theory describes ProCoT better than the other theories, especially as the LLM's writing is situated by context, as well as the students' and writing is interpretive and public, since the LLM's writing is available for students to read.

2.2. Large Language Models (LLMs)

An LLM aims to mimic human language patterns, responses, and general linguistic structures after training. It is a deep probabilistic model that is trained on large amounts of data to generate probabilities over a set of tokens (or words) for predicting the next token in a sequence and there are many types with different sizes (Adewumi et al., 2019, 2023; Brown et al., 2020; Touvron et al., 2023). These models are usually based on the Transformer, which is a deep neural network architecture (Vaswani et al., 2017). They are trained to perform Natural Language Processing [NLP] tasks, like reading comprehension, summarization, and question answering (Radford et al., 2019; Touvron et al., 2023).

By far, the most popular example of an LLM is OpenAI's ChatGPT, though there are several others, like Gemini (deepmind.google/technologies/gemini) by Google DeepMind and LLAMA-2 (Touvron et al., 2023) by Meta. ChatGPT, now in version 5, in addition to dedicated reasoning versions, is an advanced LLM (OpenAI, 2023). It is the latest in the family of Generative Pre-trained Transformer [GPT] models, which began with GPT-1 while version 2 (Radford et al., 2019) is recognized as possibly the first LLM because of its size of 1.5B parameters (Maslej et al., 2023). ChatGPT (v3.5) is available for free to users through a web interface or for a fee after a certain period through an Application Programming Interface [API].

LLMs offer a myriad of benefits, including assistance in learning, improvement in language skills, access to extensive information, and the provision of customized learning experiences. Particularly, the capabilities of ChatGPT to generate coherent and contextually relevant text based on prompts have made it popular. Similarly to other machine learning [ML] models (Yamamoto et al., 2018), it may be used to assist in evaluating student essays and providing feedback. The use of LLMs in education is not without challenges, however (Meyer et al., 2023; Pasipamire et al., 2025; Trung et al., 2025). Over-reliance on these models can potentially lead to diminished critical thinking abilities in students. Another critical aspect to consider is the inherent biases in LLMs (Adewumi et al., 2022; Meyer et al., 2023). Since these models are trained on existing data, they may unintentionally reflect and reinforce any biases that exist in the data they were trained on, which could affect the quality and fairness of information provided to students. Ensuring responsible AI use is paramount to maintaining a safe and productive learning environment (Asio & Gadia, 2024; Vinuesa et al., 2020).

2.3. LLM Fact-Checking Methods

In recent years, LLM fact-checking methods have emerged for combating misinformation. Retrieval Augmented Generation [RAG] models (Lewis et al., 2020), which combine information retrieval and generation, and special-purpose knowledge bases or datasets are useful in this regard. A couple of datasets that are used for fact-checking in LLMs exist. The first is Fact Extraction and VERification [FEVER] (Thorne et al., 2018) - It is a widely used dataset that consists of claims and evidence, allowing LLMs to learn to validate or debunk statements based on the supporting information. The statements from the dataset are from Wikipedia and claims are classified as Supported, Refuted or NotEnoughInfo. The second is ClaimBuster (Arslan et al., 2020), which is a dataset of 23,533 statements that is designed to identify claims which are worth fact-checking. The statements are extracted from U.S. general election presidential debates. There are three possible categories in the data: non-factual, unimportant factual statement, and check-

worthy factual statement. The availability of these methods have not made LLMs error-free, indicating that more work still needs to be done by the research community in mitigating hallucinations in LLMs.

3. Method

3.1. Research Design

The research design for this work uses the qualitative research design primarily (i.e. phenomenological) and quantitative methods. The phenomenological design is used to describe experiences of participants (Creswell, 2013). Probing the chain-of-thought output of an LLM involves the participants scrutinizing the output by using a reference-based platform to provide time- relevant fact-checked feedback on the output. The method is entrenched in the 'Self-Regulation' method (Zimmerman, 2001), which is the self-directive process by which learners transform their mental abilities into task-related skills and involves the following in an iterative process: *forethought*, *planning*, *monitoring*, *failure detection* and *correction*. We demonstrated the ProCoT method in two university courses (an Erasmus+ blended intensive program [BIP] (European Commission, 2022) and a full semester course) for two sets of participants.

3.2. Participants

The study involved 65 participants who are students in the 2 courses identified earlier. The students are qualified university students and the students in course 2 were awarded credit on successful completion of their task.

3.2.1. Course 1

This case had 39 students who were all Master students. The course is titled 'Text Mining' with the course code *D7058E* and is a regular semester course that is taught every year. The students are a mix of international students from different backgrounds with an average age of about 22 years.

3.2.2. Course 2

This case had 39 students who were all Master students. The course is titled 'Text Mining' with the course code *D7058E* and is a regular semester course that is taught every year. The students are a mix of international students from different backgrounds with an average age of about 22 years.

3.2.3. Instruments

The participants were free to use the LLM of their choice. In the first course 95% of the particpants used ChatGPT 3.5 while 5% used Phind (phind.com - v8). Similarly, most of the students in the second course used ChatGPT (96%) while BingAI had 4% (bing.com/search?q=Bing%20AI&showconv=1&form=MA13FV (limited to 2,000 characters of input)). We gave the same instruction that we provided to the students to the LLMs for generating the LLM ProCoT output. This is to determine if the students can cheat by generating ProCoT feedback successfully using the LLM and pass them off as their own. We designed the instruction to have some of the hallmark questions of critical thinking identified by Aveyard and Waite (2024). These are (1) where did you find it and (2) who has written it?

3.3. Data Collection

3.3.1. Course 1

The students were asked to pose one out of the four questions below to the LLM of their choice. The questions were randomly selected from two blogs (wtamu.edu/cbaird/sq/category/health; magazine.utoronto.ca/research-ideas). The students were under supervision for a 30-minute period while providing answers. The ProCoT instruction was '*Write 1 page to affirm or refute asser- tions/statements made by ChatGPT/LLM in the response by using references from peer-reviewed articles*'.

(1) Did cancer exist before man-made chemicals were around to create it?
(2) Who will benefit from AI in healthcare?
(3) How long do you have to exercise for it to count?
(4) How will we avoid machine bias?

3.3.2. Course 2

Again, each student was asked to pose one out of the four questions below to the LLM of their choice. The questions were randomly selected from a list of 50 exam questions, which are dedicated to the course. The ProCoT instruction for this second case was '*Write, at least, 300 words (excluding the reference section) to affirm or refute statements made by ChatGPT/LLM by using (at least 5) references from peer-reviewed articles or books*'. The students were allowed to go with the task and submit within a week.

(1) What is a large language model?
(2) What is the vanishing gradient problem and how is it tackled?
(3) Describe the two paradigms of using a pre-trained model in transfer learn- ing: feature extraction and fine-tuning.
(4) Name two similarity measurement metrics for vectorized semantic repre- sentation. Which one may be preferred and why?

3.4. Data Analysis

We evaluated the quality of students' writing based on the rubric described in the next subsection, as evaluated by a human subject expert and used statistical analysis to establish the effectiveness of ProCoT. We compared (1) LLM- only outputs, (2) students' ProCoT outputs, and (3) LLM ProCoT outputs. An LLM-only output is the result of asking the model any of the original questions discussed earlier. A student's ProCoT output is the feedback provided (with peer-reviewed references) for the LLM-only output while an LLM ProCoT output is the LLM feedback to their original (LLM-only) output.

3.4.1. Assessment rubric and comparison with other works

Some works, such as those by Mehta and Al-Mahrooqi (2015) and Abdol- reza Gharehbagh et al. (2022), have used or advocated open-ended questions on specific topics for evaluating critical thinking of students. In such works, they ask questions like "What does the writer want to convey?" and "What is the focus of the writer?" Students are also tasked with writing fixed-length essays and evaluated for (1) focus, (2) logic, (3) content, (4) style and (5) correctness based on the rubric by Howard et al. (2012). An important element (or criterion) of research (i.e. evidence) was left out in the work by Mehta and Al-Mahrooqi (2015). This criterion is a key element of our rubric based on the peer-review references we asked for from the students, in addition to the five elements mentioned. For each of the six criteria, we provided three levels of performance: good, fair, and poor. Writings of students that were poor were excluded from the final analysis in the Result (Section 4).

3.5. Ethical Consideration

All the students in each case were given equal test period to complete their task before submission. This is in addition to standardizing other conditions as described in this section. The instructions for carrying out the tasks were made clear and simple. Assessment of the solutions provided by the students were objectively evaluated based on the clear rubrics described earlier. The solutions of the students have been anonymized before publicly making them available in this article.

4. Results

The average success rate (i.e. good and fair outputs) for the two cases is 79.49% with a standard deviation of 12.82%. Table 1 summarizes the word counts. Furthermore, below, we describe the results in the two courses.

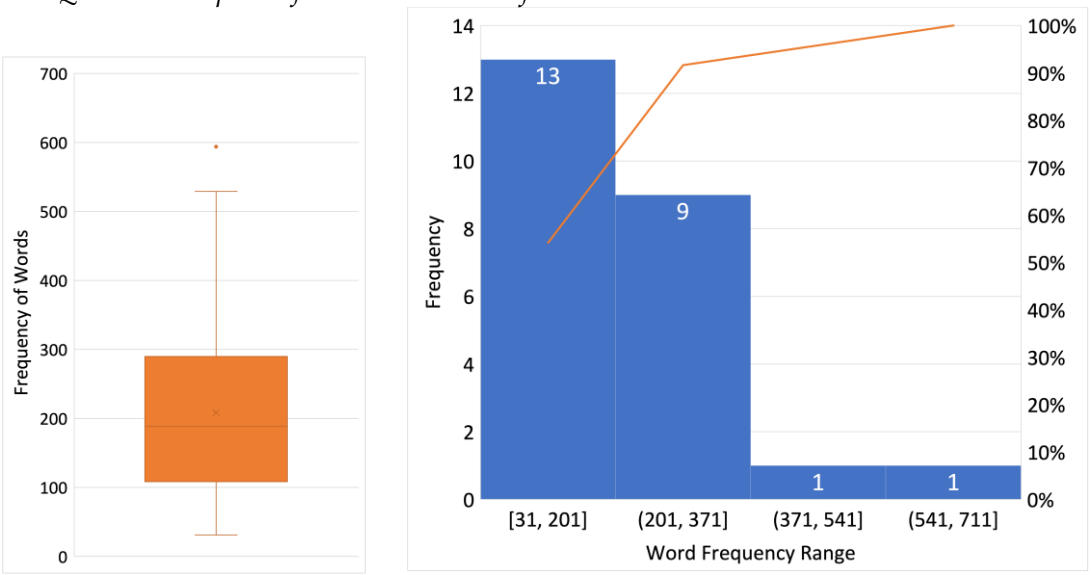
Table 1
Word counts in the two courses

	Word Counts			
Courses	*Minimum*	*Median*	*Maximum*	*Average*
Course 1	31	189	594	208
Course 2	300	389	713	405

4.1. Course 1

Two out of twenty-six submissions from the students were excluded for not meeting the minimum criteria identified earlier (particularly the criterion of evidence, thereby not complying with the ProCoT instruction). This is a success rate of about 92.31%. Also, the minimum, median, maximum, and average word counts for the students' ProCoT answers are 31, 189, 594, and 208, respectively, as shown in the Box plot of Figure 1 (a). Figure 1 (b) shows the Pareto plot for the words in their answers, where most students used word counts in the range 31 and 201. This implies most of the students prefer using fewer words to express their answers. ChatGPT gave an average word count of 391 in its ProCoT answers while Phind gave 383. We observed that the results from the valid students' ProCoT answers show that cheating may be prevented by framing questions as we have done with ProCoT.

Figure 1
Case 1: Quantitative plots of ProCoT number of words in student answers



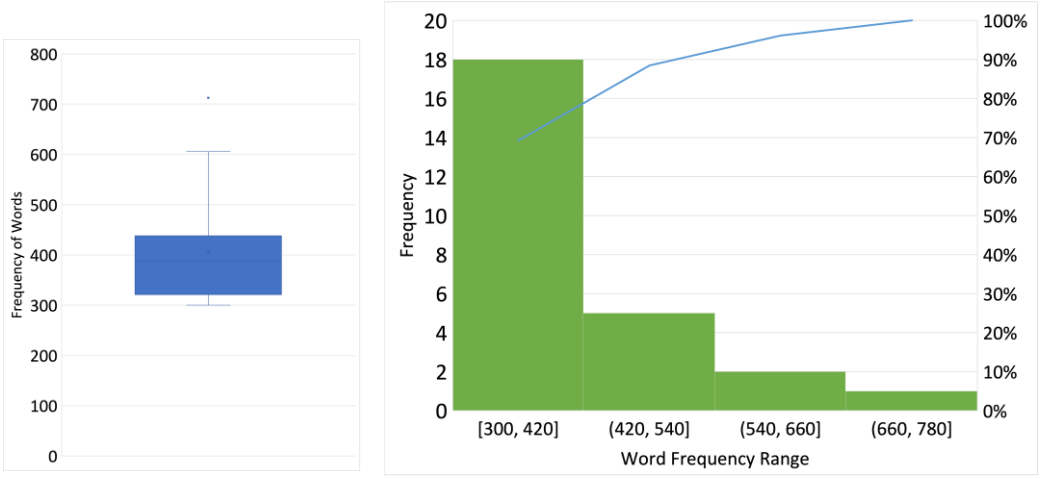
(a) Box and Whisker plot of ProCoT number of words

(a) Pareto plot of ProCoT number of words

4.2. Course 2

Out of the 39 submissions, 26 were valid because they met the criteria outlined earlier, particularly with grounding their arguments with peer-reviewed references. This is a success rate of about 66.67%. This is lower than in the first case because (1) the first case is easier with no demand on the minimum references and (2) no demand on the minimum number of words. Similarly to Course 1, quantitative analysis shows that most students use fewer words, as will be observed from Figure 2, preferring to align with the minimum word count of 300. The median, maximum, and average word counts are 389, 713, and 405, respectively. The average word counts for ChatGPT and BingAI are 356 and 315, respectively. For Course 2, we assume the students' average word count is higher in this case because we required a minimum number of words. It might be helpful to remove this restriction, as done in Course 1, as a possible cue for cheating.

Figure 2
Case 2: Quantitative plots of ProCoT number of words in student answers



(b) Box and Whisker plot of ProCoT number of words

(b) Pareto plot of ProCoT number of words

4.3. Qualitative Examples

4.3.1. Course 1, based on Question 2 (Who will benefit from AI in healthcare?)

In this example (with outputs available in Appendix A), we have (1) the LLM (ChatGPT) output, (2) a student's Pro-CoT answer, and (3) the ProCoT output from ChatGPT and Phind. We can observe that the student affirms the output from ChatGPT with credible references in fewer words meanwhile ChatGPT (3.5) openly declared it "*cannot provide direct references*" and provides a lengthy feedback while Phind also provides a lengthy feedback and two credible references (instead of 3). We provide all the other outputs and writings publicly for download.

Futhermore, comparing the LLM ProCoT outputs (on 12 De- cember, 2023) ChatGPT expressly says '*I cannot provide direct references to peer-reviewed articles*' on the one hand while saying 'numerous studies in peer-reviewed literature support' its answer, on the other hand. In one instance it referenced a *doi*, which directed to a paper that was less than suitable. Phind, meanwhile, appears to be better at providing peer-reviewed references, though some references are blog posts. It also appears to lift the same content supplied to it as if regurgitating (in what may be considered plagiarism) more than paraphrasing or generating original response. Therefore, the quality of students' ProCoT answers (on the basis of grounding by references) is better than those by ChatGPT, though ChatGPT usually gave more comprehensive answers that are typically verbose but not necessarily always factual. They are also better than those by Phind on the basis of originality, though Phind may also give more verbose answers. Section 5.3 presents some qualitative examples. As possible cues for cheating, comparing the average word counts by students and LLMs may be useful. This is besides the LLM artefacts that may result from copying an LLM output unedited. This is because possible cheating through prompt engineering on an LLM to fix the word count is known to work poorly (platform.openai.com/docs/guides/prompt-engineering/strategy-provide-reference-text) It is important to emphasize that word count alone is not enough to conclude that cheating has occurred.

4.3.2. Course 2, based on Question 4 (Describe the two paradigms of using a pre-trained model in transfer learning: feature extraction and fine-tuning.)

In this example (with outputs available in Appendix B), we have (1) the LLM (BingAI) output, (2) a

student's ProCoT output, and (3) the ProCoT outputs from ChatGPT and BingAI. We can observe that the student affirms much of the LLM output though the student's writing has some grammatical errors. ChatGPT ProCoT output provides (credible) references this time. However, BingAI ProCoT output is strikingly very similar to its original output to the initial question, in terms of content, style and word count (about 252 words), in what may be considered self-plagiarism. Also, it uses only the same one reference of a code notebook throughout.

In all the ChatGPT ProCoT outputs (on 20 December, 2023) only in-text citation is provided, when it does use references. With regards to question 1 about LLM (What is a large language model?), ChatGPT's ProCoT output does not give any reference and merely regurgitates what was stated earlier rather than affirm or refute what was given. With regards to question 2 about the vanishing gradient, ChatGPT's ProCoT output gives 7 relevant references. With regards to question 3 on transfer learning, ChatGPT's ProCoT output gives 5 relevant references. With regards to ChatGPT's answer to question 4 on similarity metrics, its ProCoT output fails to follow the instruction, as it did not provide the minimum references, which is one of the reasons for rejecting some students' submissions. BingAI, on the other hand, in all cases regurgitated the input and provided 4-6 blogs as references (when it provides references) instead of peer-reviewed articles. It then adds a short paragraph as conclusion in 2 out of the 4 instances.

5. Discussion

As in Lin et al. (2025) and Mehta and Al-Mahrooqi (2015), we observed that critical thinking is enhanced by reflections brought about by engagement with the topic the students wrote about. Beyond that, finding validation for the views expressed in the content through peer-reviewed articles as evidence added additional layer of critical thinking. Menlah and Boateng (2025) had also demonstrated in their work involving Maths that AI had the capacity to raise students' cognitive engagement. The outcome of our work is consistent with such previous studies. As discussed about post-process theory (Van Cleave, 2024), ProCoT offers an engaging approach to writing though it may be a complex phenomenon. Hence, post-process theory describes ProCoT better and also fits the constructivist viewpoint.

The constructivist viewpoint points out that successful learning occurs when students actively engage with knowledge, possibly through interactions with supportive tools (Adewumi et al., 2025; Dong & Gan, 2025). In addition, as earlier pointed out, the results, particularly from the students' ProCoT answers revealed that cheating may be prevented by framing questions as we have done with ProCoT. This is because of the need for references or reliable sources. For example, Calma and Davies (2025) identified the importance of clarification questions in their work involving the evaluation of essays.

Given the concerns in some quarters about AI-generated outputs being plagiarism (Wu, 2024), the pedagogy of essay writing evaluation requires a comprehensive approach that focuses on student learning and development. Central to this is the use of rubrics, which are detailed guides laying out the expected criteria for an essay (Lipnevich et al., 2023). They may cover aspects like the strength of an argument, the use of evidence (or references), as demonstrated with ProCoT, and the organization of ideas (Paltridge & Starfield, 2013). A formative approach to evaluation, based on quality feedback, helps students identify their strengths and weaknesses and fosters a growth mindset (Porter, 2022; Spector et al., 2016). These methods also provide transparency and ensure alignment with intended learning outcomes [ILOs} (Le et al., 2023; Ramesh & Sanampudi, 2022).

Peer review and self-assessment are also important in pedagogy. By engaging in peer review, students develop critical thinking skills (Cheong et al., 2023) while self-assessment encourages students to reflect on their own writing, fostering a sense of ownership and responsibility (Culver, 2023). ProCoT challenges the reviewing skills of students and encourages reflection, which is another vital component in the pedagogical evaluation of any writing. It helps students to be more independent and confident writers (Pang, 2022). The pedagogical evaluation of writing is a comprehensive yet subjective process. It is a crucial component of the educational journey that

plays a significant role in the overall development of students, both as critical thinkers and skilled writers.

6. Conclusion

We presented Probing Chain-of-Thought for stimulating critical thinking and reducing possible cheating in the writing of students. ProCoT shows that educators need not fear LLMs but embrace them as powerful tools to enrich the experiences of students in learning. We demonstrated that weaknesses exist in the LLMs that were evaluated in this work. This makes it important for teaching students to be better writers by grounding their writing in peer-reviewed references. We observed that the average success rate of the writing of students based on the 6 rubric criteria was 79.49% (±12.82%). Also, ProCoT stimulates critical thinking and writing of students through engagement with LLMs and it may prevent cheating because of clear limitations in the concerned LLMs. We observed improvements in Course 2 over Course 1, which had one week difference between them, though it should be noted that the writing topics are different.

Hence, new methods may be required to evaluate students' writing performance and learning in the future since LLMs are likely to improve in capabilities. In the future, multimodal evaluation may be the solution or methods that are grounded in interview/oral assessments. Furthermore, pair-wise comparison of ProCoT and non-ProCoT students' writing will be another valuable method for assessing performance of the ProCoT method.

While ProCoT stimulates critical thinking in students and may prevent or minimize the possibility of cheating in writing, it has its limitations, especially because the AI field is developing at a rapid pace that within a few months more capable models are being released, hence the versions of LLMs evaluated in this work may have been surpassed in capabilities by newer models. For example, the newer ChatGPT models (e.g. GPT-5) are capable of providing credible references in many instances. For example, in the appendix, we provide the ProCoT output for the new GPT-5 for question 2 in Course 1. This study suggests students can engage with LLMs in creative ways and this should be encouraged.

Acknowledgment: The authors wish to thank the anonymous reviewers and the editor for their valuable feedback during the review process.

Author contributions: Tosin Adewumi: Conceptualization, Writing, Review; Lama Alkhaled, Claudia Buck, Sergio Serrano Hernandez, Saga Brilioth, Mkpe O. Kekung, and Yelvin Ragimov: Writing; Elisa Barney: Supervision.

Declaration of interest: The authors declare that no competing interests exist.

Data availability: Data from the studies are publicly available at the stated link in the article.

Ethical declaration: Ethical consideration was made during this project and data anonymization performed for the data made available.

Funding: The authors acknowledge and thank the Department of Computer Science Electrical and Space Engineering of Luleå University of Technology for the strategic funding for pedagogical publication.

References

Abdolreza Gharehbagh, Z., Mansourzadeh, A., Montazeri Khadem, A., & Saeidi, M. (2022). Reflections on using open-ended questions. *Medical Education Bulletin*, *3*(2), 475–482.

Adewumi, T., Adeyemi, M., Anuoluwapo, A., Peters, B., Buzaaba, H., Samuel, O., Rufai, A. M., Ajibade, B., Gwadabe, T., Koulibaly Traore, M. M., Ajayi, T. O., Muhammad, S., Baruwa, A., Owoicho, P., Ogunremi, T., Ngigi, P., Ahia, O., Nasir, R., Liwicki, F., & Liwicki, M. (2023). Afriwoz: Corpus for exploiting cross-lingual transfer for dialogue generation in low-resource, african languages. In B. Verma & N. K. Kasabov (Eds.), *2023 International Joint Conference on Neural*

Networks (IJCNN) (pp 1–8). IEEE. https://doi.org/10.1109/IJCNN54540.2023.10191208

Adewumi, T., Liwicki, F., & Liwicki, M. (2022). State-of-the-art in open- domain conversational AI: A survey. *Information*, *13*(6). https://doi.org/10.3390/info13060298

Adewumi, T. P., Liwicki, F., & Liwicki, M. (2019). Conversational systems in machine learning from the point of view of the philosophy of science—using alime chat and related studies. *Philosophies*, *4*(3), 41. https://doi.org/10.3390/philosophies4030041

Adewumi, T., Liwicki, F. S., Liwicki, M., Gardelli, V., Alkhaled, L., & Mokayed, H. (2025). *Findings of mega: Maths explanation with LLMS using the Socratic method for active learning.* arXiv. https://doi.org/10.48550/arXiv.2507.12079

Allam, O., Williams, M., Almeida, M., Alper, D., Craver, A., Persing, J., & Alperovich, M. (2023). Generative pre-trained transformers (gpt) artificial intelligence – assessing the accuracy of ChatGPT as an adjunct for peri-operative care. *Plastic and Reconstructive Surgery - Global Open*, *11*, 132–133. https://doi.org/10.1097/01.GOX.0000992588.09873.87

Aristotle. (2007). *On rhetoric: A theory of civic discourse* (G. A. Kennedy, Trans.). Oxford University Press. (Original work published ca. 4th century B.C.E.)

Arslan, F., Hassan, N., Li, C., & Tremayne, M. (2020). A benchmark dataset of check-worthy factual claims. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 821–829. https://doi.org/10.1609/icwsm.v14i1.7346

Asio, J. M. R., & Gadia, E. D. (2024). Predictors of student attitudes towards artificial intelligence: Implications and relevance to the higher education institutions. *International Journal of Didactical Studies, 5*(2), 27763. https://doi.org/10.33902/ijods.202427763

Aveyard, H. & Waite, M. (2024). *A beginner's guide to critical thinking and writing in health and social care, 3E*. McGraw-Hill Education.

Barrett, R. & Malcolm, J. (2006). Embedding plagiarism education in the assessment process. *International Journal for Educational Integrity*, *2*(1). https://doi.org/10.21913/IJEI.v2i1.23

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …, & Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Calma, A., & Davies, M. (2025). Assessing students' critical thinking abilities via a systematic evaluation of essays. *Studies in Higher Education*. Advance online publication. https://doi.org/10.1080/03075079.2025.2470969

Cheong, C. M., Luo, N., Zhu, X., Lu, Q., & Wei, W. (2023). Self-assessment complements peer assessment for undergraduate students in an academic writ- ing task. *Assessment & Evaluation in Higher Education*, *48*(1), 135–148. https://doi.org/10.1080/02602938.2022.2069225

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Sage Publications.

Culver, C. (2023). Learning as a peer assessor: evaluating peer-assessment strate- gies. *Assessment & Evaluation in Higher Education*, *48*(5), 581–597. https://doi.org/10.1080/02602938.2022.2107167

Dong, S., Zhi, R., & Gan, F. (2025). Integrating robotics in art education: leveraging constructivist and experiential learning frameworks to enhance cognitive development, problem-solving skills and collaboration among diverse learners. *European Journal of Education, 60*(2), e70114. https://doi.org/10.1111/ejed.70114

Elbow, P. (1998). *Writing without teachers*. Oxford University Press.

European Commission. (2022). *Blended mobility implementation guide for Erasmus+ higher education mobility ka131*. https://doi.org/10.1093/oso/9780195120165.001.0001

Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition & Communication*, *32*(4), 365–387. https://doi.org/10.58680/ccc198115885

Freedman, A. & Medway, P. (2003). *Genre in the new rhetoric*. Routledge. https://doi.org/10.4324/9780203393277

Heilker, P. (1996). *The essay: theory and pedagogy for an active form.* Hampton Press.

Howard, R. D., McLaughlin, G. W., and Knight, W. E. (2012). *The handbook of institutional research*. John Wiley & Sons.

Kishore, S., Hong, Y., Nguyen, A., & Qutab, S. (2023). Should ChatGPT be banned at schools? organizing visions for generative artificial intelligence (AI) in education. In S. Paul, S. Sarker & v. K. Tuunainen (Eds.), *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors*

with Digital Technologies (p. 5). Association for Information Systems. https://aisel.aisnet.org/icis2023/learnandiscurricula/learnandiscurricula/5/

Le, X.-M., Phuong, H.-Y., Phan, Q.-T., & Le, T.-T. (2023). Impact of using analytic rubrics for peer assessment on efl students' writing performance: An experimental study. *Multicultural Education*, *9*(3), 41-53.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lin, C.-J., Lee, H.-Y., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2025). Enhancing reflective thinking in stem education through experiential learning: The role of generative ai as a learning aid. *Education and Information Technologies*, *30*(5), 6315–6337. https://doi.org/10.1007/s10639-024-13072-5

Lipnevich, A. A., Panadero, E., & Calistro, T. (2023). Unraveling the effects of rubrics and exemplars on student writing performance. *Journal of Experimental Psychology: Applied*, *29*(1), 136. https://doi.org/10.1037/xap0000434

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, W., Clark, J., & Perrault, R. (2023). *Artificial intelligence index report 2023.* arXiv. https://doi.org/10.48550/arXiv.2310.03715

Mehta, S. R. & Al-Mahrooqi, R. (2015). Can thinking be taught? Linking critical thinking and writing in an EFL context. *RELC Journal*, *46*(1), 23–36. https://doi.org/10.1177/0033688214555356

Menlah, C. K. A., & Boateng, F. O. (2025). Examining the effect of AI-based tutoring systems on students' mathematical problem-solving skills: The moderating role of mathematics anxiety. *Journal of Pedagogical Sociology and Psychology, 7*(3), 5-17. https://doi.org/10.33902/jpsp.202536137

Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., & Gonzalez-Hernandez, G. (2023). ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, *16*(1), 20.

Murray, D. (1972). Teach writing as a process not product. *The Leaflet*, *71*(3), 11– 14.

OpenAI. (2023). *ChatGPT release notes.* OpenAI Help Center. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Paltridge, B., & Starfield, S. (2013). *The handbook of English for specific purposes* (Vol. 592). Wiley-blackwell.

Pang, N. S.-K. (2022). Teachers' reflective practices in implementing assessment for learning skills in classroom teaching. *ECNU Review of Education*, *5*(3), 470– 490. https://doi.org/10.1177/2096531120936290

Pasipamire, N., Chigwada, J., & Maturure, R. (2025). Exploring the use and impact of artificial intelligence in higher education in Africa. *Journal of Pedagogical Sociology and Psychology, 7*(3), 108-131. https://doi.org/10.33902/jpsp.202532046

Paul, R. & Elder, L. (2013). *Critical thinking: Tools for taking charge of your professional and personal life*. Pearson Education.

Pettersson, J., Hult, E., Eriksson, T., & Adewumi, T. (2024). Generative AI and teachers–for us or against us? a case study. In F. Westphal, E. Peretz-Andersson, M. Riveiro, K. Bach, & F. Heintz (Eds.), 14th Scandinavian Conference on Artificial Intelligence SCAI 2024 (pp. 1-7). https://doi.org/10.3384/ecp208005

Porter, G. (2022). Collaborative annotation: Links to formative assessment and issues of scale for pedagogy. In R. F. Kizilcefc (Ed.), *Proceedings of the Ninth ACM Conference on Learning@ Scale* (pp. 313–316). Association for Computing Machinery. https://doi.org/10.1145/3491140.3528322

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ramesh, D. and Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rawte, V., Sheth, A., & Das, A. (2023). *A survey of hallucination in large foundation models.* arXiv. https://doi.org/10.48550/arXiv.2309.05922

Spector, J. M., Ifenthaler, D., Samspon, D., Yang, L., Mukama, E., Warusavitarana, A., Lokuge Dona, K., Eichhorn, K., Fluck, A., Huang, R., Bridges, S., Lu, J., Ren, Y., Gui, X., Deneen, C. C., San Diego, J., & Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of Educational Technology & Society, 19*(3), 58–71. http://www.jstor.org/stable/jeductechsoci.19.3.58

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *FEVER: a large-scale dataset for fact extraction and VERification.* NAACL-HLT. https://doi.org/10.18653/v1/N18-1074

Todd, R. W., Thienpermpool, P., & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing Writing*, *9*(2), 85– 104. https://doi.org/10.1016/j.asw.2004.06.002

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bash- lykov, N., Batra, S., Bhargava, P., & Bhosale, S. (2023). *Llama 2: Open foundation and fine-tuned chat models.* arXiv. https://doi.org/10.48550/arXiv.2307.09288

Trung, L. T. B. T., Thanh Trung, T., & Dung, T. M. (2025). ChatGPT in Vietnamese math classrooms: What are the influencing factors behind teachers' adoption?. *Journal of Pedagogical Research, 9*(2), 72-88. https://doi.org/10.33902/JPR.202531924

Van Cleave, J. (2024). Writing toward theoretical resonance. In J. R. Wolgemuth, K. W. Guyotte, & S. A. Shelton (Eds.), *Expanding approaches to thematic analysis* (pp. 106–122). Routledge. https://doi.org/10.4324/9781003389149-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł ., & Polosukhin, I. (2017). *Attention is all you need.* arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-019-14108-y

Wu, Y. (2024). Critical thinking pedagogics design in an era of ChatGPT and other AI tools—shifting from teaching "what" to teaching "why" and "how". *Journal of Education and Development*, *8*(1), 1. https://doi.org/10.20849/jed.v8i1.1404

Yamamoto, M., Umemura, N., & Kawano, H. (2018). Automated essay scoring system based on rubric. In R. Lee (Ed.), *Applied computing & information technology* (pp. 177–190). Springer https://doi.org/10.1007/978-3-319-64051-8_11

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement*: *Theoretical perspectives* (2nd ed., pp. 1–37). Lawrence Erlbaum Associates.

Appendix A. Outputs for Course 1

Table 1
ChatGPT (Original) Output

AI has the potential to benefit various stakeholders within the healthcare industry, including:

(1) Patients: AI can lead to better and more personalized care, improved diagnostics, quicker and more accurate treatment recommendations, remote monitoring, and more efficient healthcare services.

(2) Healthcare Providers: Doctors, nurses, and other healthcare professionals can benefit from AI by gaining access to advanced diagnostic tools, more accurate predictions, assistance in decision-making, reducing administrative burdens, and improving patient outcomes.

(3) Hospitals and Healthcare Facilities: AI can streamline operations, optimize workflows, man- age resources more efficiently, and enhance patient experiences, ultimately reducing costs and improving the quality of care.

(4) Researchers and Scientists: AI can accelerate the pace of medical research by analyzing massive amounts of data, identifying patterns in diseases, aiding in drug discovery, and helping in the development of new treatment methods.

(5) Insurance Companies: AI can assist in risk assessment, fraud detection, and personalized premium setting by analyzing patient data and predicting health outcomes more accurately.

(6) Pharmaceutical Companies: AI can help in drug discovery, clinical trials, and optimizing manufacturing processes, potentially reducing the time and cost involved in bringing new drugs to market.

(7) Government and Policy Makers: AI can provide insights and data that can help in shaping healthcare policies, resource allocation, and improving public health initiatives.

While AI offers numerous benefits, it's important to address concerns such as data privacy, security, and ethical use of AI in healthcare to ensure that these advancements are implemented responsibly and for the betterment of all involved.

Table 2
Student ProCoT Answer

1 V¨aa¨na¨nen, A., Haataja, K., Vehvil¨ainen-Julkunen, K., & Toivanen, P. (2021). AI in healthcare: A narrative review. F1000Research, 10, 6.

AI can have remarkable possibilities in reducing healthcare costs, providing preventive healthcare, ease the work burden of healthcare professionals, and providing more accurate diagnosis faster and easier. The above paper concluded that the utilization of AI in healthcare can improve access to advanced healthcare services, especially for underserved populations in developing countries. AI solutions offer wide-ranging applications across healthcare sectors, including diagnosis, preventive healthcare, patient monitoring, and virtual nurse assistants, showing potential to significantly impact various healthcare services. Projections suggest that AI implementations in the healthcare industry could potentially generate substantial savings of around $150 billion by 2026, indicating that investing in AI can be financially beneficial for the healthcare sector.

2 Shaheen, M. Y. (2021). Applications of Artificial Intelligence (AI) in healthcare: A review. ScienceOpen Preprints. Artificial intelligence (AI) integration in healthcare is progressively increasing, offering various benefits for both patient care and administrative tasks. While some AI innovations claim to perform as well as or better than humans in tasks like diagnosing illnesses, the full replacement of human roles in healthcare by AI is expected to take considerable time.

3 Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. NPJ digital medicine, 2(1), 77. Health systems are presented with a critical choice: whether to temper enthusiasm regarding AI's potential due to challenges such as data ownership and trust, or to invest in resolving these issues and enhancing data infrastructure to realize the potential benefits of AI in healthcare.

Table 3
ChatGPT ProCoT Output

The statements are generally affirmed. Numerous studies and reviews in peer-reviewed literature support the potential benefits of AI in healthcare for various stakeholders. While I cannot provide direct references to specific articles, I can highlight the broad consensus from the literature.

Patients:

- Improved Diagnostics: Studies have explored the use of AI in improving diagnostic accuracy across various medical domains. For example, AI models have been developed for early detection of diseases such as cancer and diabetic retinopathy.
- Personalized Medicine: AI-driven approaches, including machine learning algorithms, have been investigated for tailoring treatment plans based on individual patient characteristics. Personalized medicine is an active area of research in oncology and other fields.
- Remote Monitoring: AI-powered devices and applications have been studied for remote monitoring of health conditions. This includes wearable devices and smartphone apps that enable continuous monitoring of vital signs and chronic disease management.
- Enhanced Access: Telemedicine, powered by AI applications, has demonstrated potential in improving access to healthcare services, especially in underserved or remote areas.

Healthcare Providers:

- Clinical Decision Support: The literature supports the use of AI for clinical decision sup- port, providing insights and recommendations to healthcare professionals in diagnostics and treatment planning.
- Administrative Efficiency: Administrative tasks, such as appointment scheduling and medical coding, are areas where AI applications have shown promise in improving efficiency.
- Predictive Analytics: AI-driven predictive analytics has been explored for predicting disease outbreaks and optimizing resource allocation in healthcare settings.
- Reduced Burnout: Studies have investigated the impact of automation and AI in reducing the burden of repetitive tasks, contributing to the potential reduction in burnout among healthcare workers.

Researchers and Drug Development:

- Drug Discovery: AI has been increasingly utilized in drug discovery, offering computational approaches for identifying potential drug candidates and predicting their efficacy.
- Data Analysis: The application of AI in analyzing large-scale medical data for research purposes has been well-documented in the literature.
- Clinical Trials: AI has been explored for optimizing patient recruitment in clinical trials and assisting in the monitoring of trial data.

Healthcare Payers and Insurance Companies:

- Improved Risk Assessment: AI applications have been studied for more accurate risk assessment, contributing to the development of personalized insurance plans.
- Fraud Detection: AI has been employed for fraud detection in healthcare, aiding insurance companies in saving costs.
- Population Health Management: AI-driven approaches are investigated for population health management, aiming to enhance the effectiveness of healthcare services and reduce overall expenses.

Governments and Public Health Organizations:

- Disease Surveillance: AI's role in disease surveillance and monitoring outbreaks has been acknowledged in the literature.
- Health Policy and Planning: The potential of AI to inform health policy and planning through data-driven insights has been discussed in the literature.

Caregivers and Family Members:

- Home Healthcare Devices: AI-driven home healthcare devices and monitoring systems have been explored to provide better care and support for caregivers and family members.

Ethical Considerations: The statement emphasizes the importance of addressing ethical, privacy, and security concerns associated with AI in healthcare, which is a common consideration discussed in the literature. Additionally, equitable access and addressing biases are recognized as important aspects to ensure the fair distribution of AI benefits.

In conclusion, while specific references are not provided, the statements align with the broader consensus in the peer-reviewed literature on the potential benefits of AI in healthcare across various stakeholders. Researchers interested in specific studies on these topics may refer to reputable journals in healthcare informatics, medical imaging, bioinformatics, and related fields.

Table 4
Phind ProCoT Output

The statements provided are generally affirmative, as they align with the current understanding of how AI can benefit various stakeholders in the field of healthcare. However, it's important to note that while AI has the potential to provide numerous benefits, it's crucial to address ethical, privacy, and security concerns to ensure that these technologies are used responsibly and for the benefit of all stakeholders in the healthcare ecosystem.

(1) Patients: AI can indeed assist in faster and more accurate disease diagnosis, leading to earlier and more effective treatments. AI can also help in person- alized medicine by tailoring treatment plans to individual patient profiles, increasing the likelihood of successful outcomes. AI-powered devices and apps can enable remote monitoring of health conditions, allowing patients to better manage chronic illnesses and receive timely interventions. AI can also im- prove access to healthcare services, especially in underserved or remote areas, through telemedicine and virtual health assistants.[1]

(2) Healthcare Providers: AI can provide healthcare professionals with valuable insights and recommendations, aiding in diagnosis and treatment decisions. AI can streamline administrative tasks, such as appointment scheduling and medical coding, reducing the administrative burden on healthcare staff. AI can predict disease outbreaks and patient needs, helping hospitals allocate re- sources more efficiently. Automation of repetitive tasks can reduce the burnout experienced by healthcare workers.[2]

(3) Researchers and Drug Development: AI accelerates drug discovery by identi- fying potential drug candidates and predicting their efficacy. AI can analyze vast amounts of medical data to uncover patterns and insights that may lead to breakthroughs in medical research. AI can optimize patient recruitment for clinical trials and assist in monitoring trial data.[11]

(4) Healthcare Payers and Insurance Companies: AI can help insurers assess risk more accurately, leading to more personalized and competitive insurance plans. AI can detect and prevent healthcare fraud, saving costs for insurance com- panies. AI can assist in managing the health of covered populations more effectively, reducing overall healthcare expenses.[11]

(5) Governments and Public Health Organizations: AI can help in monitoring dis- ease outbreaks, facilitating rapid responses to public health emergencies. AI can provide insights for policymakers to make data-driven decisions on health- care infrastructure and resource allocation. AI-driven home healthcare devices and monitoring systems can help caregivers and family members provide bet- ter care and peace of mind for their loved ones.[11]

(6) Caregivers and Family Members: AI can provide numerous benefits in health- care, but it's important to address ethical, privacy, and security concerns to ensure that these technologies are used responsibly and for the benefit of all stakeholders in the healthcare ecosystem. Additionally, equitable access to AI-driven healthcare solutions and addressing potential biases are important considerations to ensure that the benefits of AI are distributed fairly.[10]

Appendix B. Outputs for Course 2

Table 5
BingAI (Original) Output

Certainly! In the context of transfer learning, there are two primary paradigms for utilizing a pre-trained model:

Feature Extraction: In this approach, we leverage the representations learned by a previously trained network to extract meaningful features from new data. The process involves adding a new classifier (usually a fully connected layer) on top of the pre- trained model. The pre-trained model's convolutional base (which contains generic feature maps) remains frozen during this step. By doing so, we can repurpose the feature maps learned previously for a different dataset without retraining the entire model. However, the final classification layer of the pre-trained model is specific to its original task and class labels. Example: Suppose we use a pre-trained model trained on a large-scale image classification task and add a new classifier for a specific set of classes (e.g., cats vs. dogs).

Fine-Tuning: Fine-tuning involves unfreezing a few of the top layers of the pre-trained model's base. We then jointly train both the newly-added classifier layers and the last layers of the base model. The goal is to adapt the higher-order feature representations in the base model to make them more relevant for the specific task. Fine-tuning is particularly useful when we have sufficient labeled examples for the target task. Example: After feature extraction, we fine-tune the base model's top layers to improve its performance on the specific classification task.

In summary, feature extraction allows us to reuse learned features, while fine-tuning refines the model for a specific task by adjusting its higher-level representations.[3]

Table 6
Student ProCoT Output

I use LTU's library - Articles to evaluate the same meaning as I run in ChatGPT and BingAI: "Describe the two paradigms of using a pre-trained model in transfer learning: feature extraction and fine-tuning."

ChatGPT provides the explanation for both feature extraction and fine-tunning in a descriptive and well detailed manner at the same time simple manner. Moreover, ChatGPT provides process description and advantages though in the initial statement I did not ask for that. ChatGPT does not provide references for the answers as per default. On the other hand, BingAI provides the more condensed answer with more technical terms, highlighting the important according to BingAI words that describe the topic and providing the direct link as reference to the answer.

In the scientific literature at LTU's digital libraries, there are many papers that I got match on but I had to modify my sentence and search for the different words. Most of the authors mention using feature extracting and fine-tuning for improving the images qualities in different areas of industries. According to one of the authors Malladi et. al (2023), fine-tunning becomes standard for NLP tasks solving and fine-tuning as a model with 108 or more parameters. Furthermore, Malladi et. al (2023), mention that fine-tuning starts from a parameter initialization from pre-training which is correlated to the ChatGPT description of fine-tuning as the pre-trained model using multiple parameters. Malladi et. al (2023) also mention that fine-tunning models adapt the pre-trained model and use new distribution SFT and C-way downstream classification task. This comprehensive description of fine-tuning and the mathematical formulation of it is not representative in the answer by ChatGPT nor BingAI.

Bhade & Bamnote (2022) describe feature extraction as extracting information from raw data, it's performed after the pre-processing phase and can be divided in two feature selection and classifica- tion. Bhade & Bamnote (2022) explain also aims of feature extraction as to improve the recognition rate, build a similar feature set in the field of same image. The process of feature extraction and the difference between the feature extraction and transformation is explained more generously and more precisely at Bhade & Bamnote (2022) than in ChatGPT and Bing AI. Another authors like Razavian et. al (2014) describe the feature extraction using freeze parameters of the pre-trained model which is consistent to ChatGPT and BingAI.

The inconsistency for ChatGPT nor BingAI mention that fine-tuning all parameters of a network may cause overfitting which several authors like Ying (2019) and Jeddi, Shafiee & Wong(2020) confirm. Liu, Sha & Peng (2023) used fine-tuning on the small source datasets and described the process of fine-tuning as feeding input to a pre-trained model and learning embedding. Furthermore, they mention the risk of the overcapacity using large pre-trained model on small datasets. ChatGPT and BingAI mention which datasets its better to use for fine-tuning and for feature extraction but the explanation is not extensive and detailed and do not mention instability and overcapacity risks as Liu, Sha & Peng (2023).

In summary, ChatGPT and BingAI present a relatively consistent description of the topic in this Lab and give the medium depth explanations. However, in the scientific literature the same topic is covered more in depth, with more accuracy and technical grounds. Thus to get more comprehensive understanding in the topic ChatGPT can eventually get more precise questions and the user of ChatGPT need to complete the knowledge using the scientific sources.

References

- Bhade, A.W. & Bamnote, Dr. G.R. (2022). Comparison of pre-processing and feature extraction techniques for Alzheimer detection using MRI images. 2022 Fifth International Con- ference on Computational Intelligence and Communication Technologies (CCICT). Viewed 2023-12-18. DOI: 10.1109/CCiCT56684.2022.00020
- Jeddi, A., Shafiee, M. & Wong A. (2020). A Simple Fine-tuning Is All You Need: To- wards Robust Deep Learning Via Adversarial Fine-tuning. Viewed 2023-12-18.Available at: https://arxiv.org/abs/2012.13628v1
- Liu J., Sha , & P eng , (2023)). Improving Fine-tuning Pre-trained Models on Small Source Code Datasets via Variational Information Bottleneck. IEEE. 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER). Viewed: 2023-12-18. Available at: DOI 10.1109/SANER56733.2023.00039
- Malladi S.S., W Wettig, A ., Yu, D ., Chen, D., Arora, (2023). A Kernel- Based View of Language Model Fine-Tuning. Viewed 2023-12-18. Available at: https://doi.org/10.48550/arXiv.2210.05643
- Razavian, S., H. Azizpour, J. Sullivan, S. Carlsson (2014). CNN features off-the-shelf: an astounding baseline for recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014 pp. 806–813. 1, 5 . Viewed 2023-12-18
- Ying, X. (2019). An Overview of Overfitting and its Solutions. J. Phys.: Conf. Ser. 1168 022022. Viewed 2023-12-18. DOI 10.1088/1742-6596/1168/2/022022.

Table 7
ChatGPT ProCoT Output

The provided information accurately describes two common paradigms in transfer learning: feature extraction and fine-tuning. These approaches are widely employed to leverage pre-trained models for new tasks. The assertions made are supported by multiple references from peer-reviewed articles and books.

Feature extraction involves using a pre-trained model's convolutional base to extract meaningful features from new data. This process is well-documented in the literature. For instance, in the book "Deep Learning" by Ian Goodfellow et al. (2016), the authors discuss the concept of transfer learning and feature extraction in the context of convolutional neural networks (CNNs). The idea of freezing the convolutional base during feature extraction is consistent with the work of Razavian et al. (2014) in the paper "CNN Features Off-the-shelf: an Astounding Baseline for Recognition." The study explores using pre-trained CNNs for image classification tasks by extracting features from the convolutional layers. The example provided, using a pre-trained model for a large-scale image classification task and adding a new classifier for specific classes (e.g., cats vs. dogs), aligns with the practical application of feature extraction, as discussed by Yosinski et al. (2014) in "How transferable are features in deep neural networks?"

Fine-tuning, as described, involves unfreezing a few of the top layers of the pre-trained model and jointly training both the new classifier layers and the last layers of the base model. This process is extensively discussed in the literature. For instance, in the paper "Visualizing and Understanding Convolutional Networks" by Zeiler and Fergus (2014), the authors explore fine-tuning strategies to improve the performance of pre-trained CNNs on specific tasks. The concept that fine-tuning is particularly useful when there are sufficient labeled examples for the target task is supported by the work of Donahue et al. (2014) in "Decaf: A Deep Convolutional Activation Feature for Generic Visual Recognition." The study demonstrates the effectiveness of fine-tuning with ample labeled examples for image classification.

In summary, the information provided accurately describes the feature extraction and fine-tuning paradigms in transfer learning, and the statements align with findings from various peer-reviewed sources. These paradigms offer practical strategies for utilizing pre-trained models and adapting them to specific tasks.

Table 8
BingAI ProCoT Output

In the context of transfer learning, there are two primary paradigms for utilizing a pre-trained model:

Feature Extraction and Fine-Tuning. Feature Extraction involves leveraging the representations learned by a previously trained network to extract meaningful features from new data. The process involves adding a new classifier (usually a fully connected layer) on top of the pre-trained model. The pre-trained model's convolutional base (which contains generic feature maps) remains frozen during this step. By doing so, we can repurpose the feature maps learned previously for a different dataset without retraining the entire model. However, the final classification layer of the pre-trained model is specific to its original task and class labels. For example, suppose we use a pre-trained model trained on a large-scale image classification task and add a new classifier for a specific set of classes (e.g., cats vs. dogs).

Fine-Tuning involves unfreezing a few of the top layers of the pre-trained model's base. We then jointly train both the newly-added classifier layers and the last layers of the base model. The goal is to adapt the higher-order feature representations in the base model to make them more relevant for the specific task. Fine-tuning is particularly useful when we have sufficient labeled examples for the target task. For example, after feature extraction, we fine-tune the base model's top layers to improve its performance on the specific classification task.

In summary, feature extraction allows us to reuse learned features, while fine-tuning refines the model for a specific task by adjusting its higher-level representations.

Appendix C. Example of ProCoT with GPT-5

Table 9
GPT-5 ProCoT Output

Here's a refined, evidence-based evaluation of your outlined statements about AI's potential impacts in healthcare—affirming or critiquing each claim—supported by peer-reviewed research:

1. Patients

Claim: AI brings better and personalized care, improved diagnostics, faster and accurate treatment recommendations, remote monitoring, and more efficient services.

Evidence:

- AI-driven remote patient monitoring has been shown to reduce hospitalizations and hospitalization rates—with one study noting a drop from 13% to 2.8% in cancer patients, along with improved pain control and reduced emergency visits.
- AI in mental health**, such as chatbot-based CBT apps like Woebot, has delivered symptom reduction similar to brief human interventions in mild-to-moderate depression.
- Clinical decision support systems (CDSS)** have improved clinician performance in 64% of studies, and patient outcomes in 13%—notably for glucose, transfusion, and pressure ulcer management.

Verdict: Affirmed, with strong support for remote monitoring, mental health support, and decision tools—though effectiveness varies by context and disease area.

2. Healthcare Providers

Claim: Providers benefit from advanced diagnostics, predictions, decision-making help, reduced admin workload, and improved outcomes.

Evidence:

- AI-based **ambient scribes** reduce documentation burden: in a UK rollout, 4 in 5 GPs reported time savings and enhanced patient rapport thanks to AI scribes 7.
- Congress.gov briefings** (peer-reviewed but governmental) note AI's potential to reduce medical errors, improve diagnostics, and ease administrative functions.
- CDSS evidence**, as above, supports better performance and safety.

Verdict: Affirmed, with credible evidence for administrative support and diagnostic assistance, though impacts on outcomes can be condition-specific.

3. Hospitals & Healthcare Facilities**

Claim: AI streamlines operations, workflows, resource use, enhances patient experience, reduces costs, and improves quality.

Evidence:

- A systematic review found AI demonstrably improves **operational efficiency**, resource management, decision-making quality assurance, and emergency response effectiveness 9.
- Oracle's overview reports AI aids billing, scheduling, image diagnostics, logistics, and drug delivery automation in hospitals 10.
- HealthEvoke** notes AI reduces administrative costs and enhances patient satisfaction by streamlining operations 11. Verdict: Affirmed, especially regarding efficiency and operational enhancements—broadly supported across reviews.

4. Researchers & Scientists**

Claim: AI accelerates research via large-scale data analysis, pattern discovery, drug discovery, and new treatment development.

Evidence:

- Cleveland Clinic highlights AI's role across diagnostics, treatment development, and research innovation.
- Reviews confirm AI's influence in drug discovery, particularly in uncovering molecular patterns and expediting leads—though data quality and ethical challenges must be managed.
- Systematic reviews underscore AI's value in healthcare innovation and research directions.

Verdict:

- Affirmed, with substantial evidence for AI's contributions to accelerated and data-driven research.

5. Insurance Companies Claim: AI aids risk assessment, fraud detection, and personalized premiums. Evidence:

- Direct peer-reviewed evidence is sparse. However, reports on AI operations (e.g. within San Antonio entities) show predictive analytics enhancing customer service—in analogous sectors like finance 15. Verdict: Partially supported—the claim is plausible, but lacks direct peer-reviewed validation within insurance contexts.

6. Pharmaceutical Companies Claim: AI supports drug discovery, trials, optimizes manufacturing—reducing time and cost.

Evidence:

- Confirmed roles in drug discovery acceleration, informed by AI pattern identification, while noting challenges around data quality and explainability.
- Oracle also indicates AI is integral in clinical trials and drug development pipelines.

Verdict:

- Affirmed, well-backed by peer-reviewed reviews and industry analysis.

7. Government & Policy Makers

Claim: AI provides insight for policy, resource allocation, and public health strategy.

Evidence:

- A report on the UK's NHS indicates that poor policy alignment and regulation slow AI adoption in diagnostics and genomics—suggesting AI could inform innovation, but still faces bureaucratic hurdles 16.
- Financial Times reporting highlights that better data governance and AI applications could reduce NHS costs by 50% and improve outcomes by 40%—though public trust and regulation remain critical 17. Verdict: Affirmed, with caveats—the potential is recognized, though effective policy translation is nascent
  and dependent on governance frameworks.

Ethical Concerns: Privacy, Security, Responsible Use Claim: AI raises concerns around data privacy, security, ethics, requiring careful, responsible deployment.

Evidence:

- Narrative reviews focus on risks involving bias, transparency, privacy, and safety in AI healthcare use 18.
- Ethical frameworks stress justice, fairness, transparency, accountability, and consent for equitable and trustworthy AI.
- Research on data privacy highlights legal and ethical challenges—citing case studies in oncology and retinopathy—calling for balance between innovation and patient trust. Verdict: Affirmed, with robust peer-reviewed and archival literature underscoring critical ethical issues requiring governance and transparency.