Socratic Assessment of Student Writing: Beyond Detecting AI-Generated Texts

Anonymous ACL submission

Abstract

The issue of academic integrity, especially in the context of technology-enabled cheating, has become a growing concern in educational settings. A particular challenge involves detecting AI-generated essays, such as those created by advanced language models like ChatGPT. This paper proposes an innovative "Socratic Assessment" system, which integrates the Socratic method with ChatGPT to evaluate a student's understanding of an essay's content through a series of probing questions, verifying that the student has written the essay themselves.

This paper presents the design, implementation, and evaluation of the Socratic Assessment system as a countermeasure against assignment cheating using AI-generated essays. By demonstrating the feasibility and effectiveness of the Socratic Assessment method, this research highlights its potential to promote academic integrity and ensure both an equitable and rigorous evaluation of students' work. Lastly, we further explore the potential benefits, drawbacks, and strategies for addressing concerns in adopting this system in educational settings.

1 Introduction

The emergence of sophisticated artificial intelligence (AI) systems, notably the recent unveiling of GPT-4 and ChatGPT (OpenAI, 2023), has had a profound impact on numerous sectors, including education. One key issue arising from the widespread adoption of large language models (LLM) is the challenge of distinguishing between genuine student essays and those that rely on AI-generated content. This concern has prompted action in some institutions, such as the New York City public schools, which have implemented a ban on ChatGPT to mitigate the risk of academic dishonesty[1].

[1]https://www.cnn.com/2023/01/05/tech/chatgpt-nyc-school-ban

Instances of academic dishonesty, where students rely on AI-generated content, compromise the intended learning outcomes of the assignment. The most threatened task by AI, essay writing is a valuable assignment type to facilitate *creation* of new knowledge, a skill situated at the apex of Bloom's Taxonomy, which classifies objectives of learning (Krathwohl, 2002). By engaging in such practices, students forego the opportunity to benefit from the advantages of writing in learning, such as the ability to readily review and evaluate ideas, establish connections between ideas, and explore unexamined assumptions (Applebee, 1984).

Several types of research have been done to detect AI-generated text to address the harmful misuse of large language models' output. There are attempts to detect the distribution of characteristics in AI-generated texts or to insert watermarks on AI-generated texts (Kirchenbauer et al., 2023; Zhao et al., 2023). Regarding academic essays, DetectGPT has shown remarkable detection performance (Mitchell et al., 2023), as DetectGPT requires inference of the original language models to estimate which model is used to generate the text, such methods cannot be applied in real-world scenarios where the number of models is available to students or a private generative model is available. In addition, Ai-generated texts become harder to detect if students do a little paraphrasing on the model outputs (Sadasivan et al., 2023).

A limitation of such existing efforts is that they have focused only on passively evaluating the student submissions while forgetting the educational purpose of essay writing assessments. By focusing exclusively on detecting AI-generated texts within submitted essays, our approach may become entrenched in a technical perspective, potentially overlooking the broader context and goals of essay writing and assessment. We need to remember what the students are expected to achieve in essay writing assignments: the ability to readily review ideas,

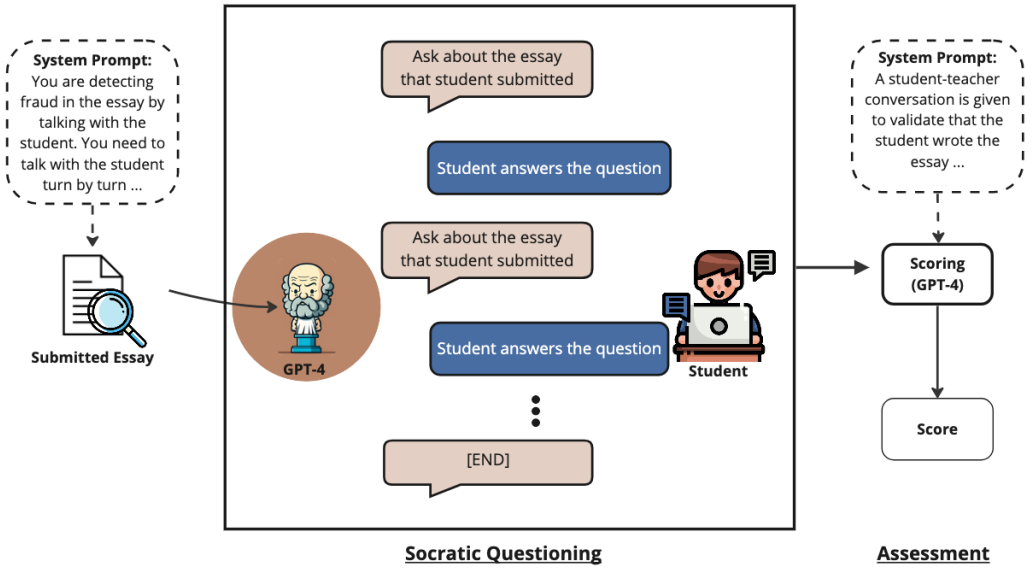


Figure 1: Overview of our method. The entire process is divided into a questioning phase and an evaluation phase. Students are asked questions about their submitted essays over multiple turns. For the entire conversation transcript, including the essay, a final score and confidence are calculated.

establish connections between ideas, and explore unexamined assumptions (Applebee, 1984).

In response to address the issue of AI-generated text in student essays by fulfilling the original purpose of the essay assignment, we propose to bring the concept of the Socratic method to assessment, where a human tutor asks a series of questions to assess the solid or weak aspects in the student's understanding of the essay. We coined this task, **Socratic Assessment**, focusing on the aspect of revealing contradictions in an argument.

In Socratic Assessment, we incorporate an instruction-tuned LLM as both a Socratic questioner and an independent confidence estimator. Previously, implementing Socratic questioning required designing a specialized system (Shridhar et al., 2022), and integrating this questioning with advanced student modeling, such as Knowledge Tracing (Corbett and Anderson, 1994), seemed like much more research was needed. However, the instruction-tuned language model allows complex functions to be performed with a simple instruction of describing a persona description as a natural language prompt, making Socratic questioning possible without any further fine-tuning (White et al., 2023).

In a Socratic Assessment, the language models ask questions about the essay and estimate the probability that the student wrote the essay on their own based on a series of question-answer pairs. Asking good questions and evaluating based on those conversations is natural to humans. As humans do, we've tried to have one evaluation model drive the conversation and evaluation in that way, but trying to get more than one kind of structured output from a common prompt has a high template failure rate. To solve this problem, we decompose Socratic Assessment into two phases: questioning and scoring.

To test the viability of the Socratic Assessment, we conducted an experiment in which we asked 24 college-educated employees to submit essays and perform Socratic Assessment. All participants were assigned one of the combinations of three essay types and four essay submission scenarios and could participate in the experiment up to three times in different combinations.

We divide participants into four groups: 1) **Read** group who only read an AI-generated essay, 2) **Rewrite** group who paraphrase AI-generated essay, 3) **Perusal** group who peruse AI-generated essay following instruction, and 4) **Write** group who wrote the essay themselves. After submitting the essay following the scenario, the participants performed Socratic Assessment without seeing their submitted essay. By comparing the results of these

4 groups, we could obtain a useful insight into the discrimination ability of Socratic Assessment in a realistic scenario. According to the automatically measured score, the group that wrote the essay themselves was significantly different, particularly for the narrative essay.

In order to evaluate the automatically measured outcome from Socratic Assessment, we collected labels for the essays and conversations from an expert with a master degree in education and experience as a teacher. The expert's responses lead us to conclude that the questions generated by Socratic Assessment help the expert determine whether the student wrote the essay himself, and that the teacher's judgment can compensate for the failure of Socratic Assessment's detection.

To summarize, our contributions are as follows:

- We proposed a Socratic Assessment task to assess students' high-level understanding of their essays where the existing AI-generated text detection approaches fail.
- We found that Socratic Assessment is more effective than the existing AI-generated text detection in an experimental design that mimics real-world scenarios.
- We analyzed the observations and limitations of the Socratic Assessment and presented directions for future work.

2 Related Work

AI-generated text detection An area of research that addresses similar problems to ours is plagiarism detection of ghostwriting. Our suspected plagiarized author is an AI model, and ghostwriting can only be detected by comparing it to the authentic work of the suspected author (Foltýnek et al., 2019), so its success depends on the success of AI-generated text detection.

In this paper, we categorize the existing literature on detecting AI-generated texts into two main categories, inspired by adversarial machine learning (Nicolae et al., 2018): 1) white-box methods where the detector has complete knowledge of the internal structure and operation of the target language models, and 2) black-box methods where the detector can only observe the inputs and outputs of the target language models.

Firstly, **white-box** methods employ various strategies. For instance, the watermarking approach can force a generative model to produce detectable texts by making certain tokens more likely to appear in the generated samples (Kirchenbauer et al., 2023; Zhao et al., 2023). Another approach involves estimating the log probability of the text itself or using the probability of generating candidates similar to the text to be verified, which shows good performance. DetectGPT, a method that employs this approach, achieves high classification accuracy (Mitchell et al., 2023). However, as such methods can be easily bypassed by simple paraphrasing (Sadasivan et al., 2023), there have been efforts to store all output texts generated by a language model and search for similar results among the stored texts (Krishna et al., 2023).

Second, **black-box** methods pursue a detector model that explicitly learns to discriminate AI-generated data and real human data (Solaiman et al., 2019; Zellers et al., 2019; Guo et al., 2023). Experiments using a fake news generation model called GROVER have shown that the generation model itself performs well when adapted for discrimination (Zellers et al., 2019). Other studies have shown that simply training another model, such as RoBERTa, as a detector also yields outstanding performance (Solaiman et al., 2019; Guo et al., 2023). OpenAI, the provider of ChatGPT, released a publicly accessible detector model trained on the text generated from 34 different models (AIT).

While several studies have demonstrated that training another model can yield outstanding performance, these methods remain vulnerable to overfitting. A recent finding that a simple linguistic feature-based machine learning method achieved 99% accuracy classification performance between human and ChatGPT-generated text implies that the performance of current classifiers can rely heavily on simple characteristics of generated texts (Desaire et al., 2023). There is also a study that shows that a model that can be used to detect fake text loses its generalizability when trained on small corpora(Bakhtin et al., 2019).

Difficulties in practical application We adopt the premise that it is practically impossible to reliably classify between genuine student essays and AI-generated text. In addition to the theoretical work showing that improvements in generative language models can lead to decreased detection classifier performance, there are too many powerful real-world obstacles (Sadasivan et al., 2023).

One of the most compelling reasons is the potential existence of private models. To the best

of our knowledge, no existing white-box method can reliably detect the output of an independently trained and unpublished model. Language models that behave similarly to ChatGPT, such as Alphaca and Vicuna, have been released to the public for easy use (Taori et al., 2023; Chiang et al., 2023).

This means that a variety of custom LLMs can emerge, not just the ones centrally provided by OpenAI. In order for the detection methodology to be effective, the models must be cooperative with watermarking (Chakraborty et al., 2023), which can be circumvented by uncontrolled private models.

The alternative, which has been reported to be robust to paraphrasing attacks, is for the language model API provider to store all outputs and perform plagiarism checks to identify suspected plagiarized text among the stored outputs. However, it still has too many practical barriers (Krishna et al., 2023). Such a system would still be unable to cope with private models, require significant resources to create and maintain, and offer no guarantee of satisfactory performance when handling massive amounts of data. The black-box methods like the detector, which is trained to detect the generated text, cannot respond because it requires the collection of data generated by a specific model.

Both AI-generated text detection methods are also inherently vulnerable to paraphrasing, personified prompting, and other attacks. The addition of a single exclamation mark ('!') can significantly the detection results for AI-generated essays (Tlili et al., 2023). Methods such as DetectGPT and OpenAI's text classifier all show a sharp drop in performance when the text is paraphrased by another language model. In the case of DetectGPT, paraphrasing reduced its detection accuracy from 70.3% to 4.6% (Krishna et al., 2023).

We thus believe that approaches relying on AI-generated text detection are likely to fail in practice. Therefore, we propose a new methodology to address this issue: Socratic Assessment.

Socratic questioning on education Historically, Socratic questioning is a good method to figure out a deep understanding of the topic. Much research has been conducted to implement Socratic questioning using AI. The Socratic method points out the limits of the partner's knowledge by first getting him (or her) to agree to a hypothesis of understandable difficulty and then exposing contradictions or lack of information (Delić and Bećirović, 2016). It's also often used to develop the critical thinking and reasoning ability of students (Paul and Elder, 2019). A recent study showed that LLMs can create a Socratic subquestion to make math problems better to be solved (Shridhar et al., 2022). Another study showed that LLMs can form Socratic questions like humans, and published a dataset (Ang et al., 2023). While these studies have used Socratic questioning primarily for the learning gain of students, we focus on its aspect of the assessment for more concise problem formulation.

3 Socratic Assessment

In light of the aforementioned limitations of AI-generated text detection, the existing method of interpreting the results of an essay assignment based on student submissions alone does not seem to solve the problem of cheating using AI. Therefore, we propose to design an alternative task that can fulfill this purpose by revisiting the key purpose of performing an essay assignment.

Returning to the purpose, based on the Knowledge-Learning-Instruction (KLI) framework (Koedinger et al., 2012) which categorizes different events in the learning procedure, we categorize the expected role of the essay assignment when it is used in an educational context as follows:

- Role as an *Instructional Event*, where we expect to acquire higher-order knowledge by letting students go through the steps of self-reconstructing and processing specific knowledge during essay writing.

- Role as an *Assessment Event*, which tests students' critical thinking skills, ability to organize logic and understanding of the topic in the process of writing an essay.

Thus, the essay assignment performs both instructional and assessment roles, triggering learning and measuring abilities. This means that an alternative approach should be capable of facilitating both the acquisition of knowledge and the verification of understanding, as obtained through traditional essay assignments.

We propose Socratic Assessment as a strong candidate for this replacement. The Socratic method consists of several stages, the core of which is to reveal gaps in knowledge by asking questions that point out holes in logic. Socratic Assessment relies on these stages to fulfill its function as an assessment. Its role as an assessment is based on the

hypothesis that "*someone who can answer questions about an essay at a level similar to that of someone who actually wrote the essay has already fulfilled the purpose of the essay assignment*".

Specifically, as shown in Figure 1, Socratic Assessment is composed of two parts. One is to collect conversations between students and LLMs, and the other is to verify how effective the conversations are in diagnosing students. In this study, both parts were implemented by injecting natural language instruction prompts into GPT-4 API.

In the Socratic Questioning phase, GPT-4 is used to ask a question to the student turn-by-turn and is instructed to print **[END]** when enough dialog has taken place. Instruction prompts are presented as system roles in the GPT-4 API.

In the Assessment phase, we included the dialog and asked GPT-4 to output the "probability for whether the student wrote the essay himself". To calculate AUROC directly, we asked it to output 'probability' and performed greedy sampling by setting the API parameter top_p = 1 to exclude randomness. We did not perform a majority-voting or apply the chain-of-thought method (Wei et al., 2022) and used the values from a single prompt for our analysis. To eliminate any concerns about potentially hacking the metrics, we didn't make any adjustments to the prompt once we found the output was in a parsable form.

4 Experiments

In this section, we will validate the research question, " *Can a probability obtained from an automated Socratic Assessment classify whether a student has written an essay?*" To analyze this in a real-world application, we designed a human study where the participants submit an essay and experience Socratic Assessment in person. Based on the results, we compare the classification ability of the Socratic Assessment.

4.1 Experiment setting

We experimented with three different essay types and four different essay writing scenarios. For each scenario, we collected student-AI dialogue about essays and analyzed them in several aspects. Figure 2 shows an overview of our experiment.

Types of essay and constraints There are three types of essays, *Narrative*, *Source Dependent*, *Argumentative* as presented in Table 1. We followed the classification of essay types from the ASAP++ benchmark, the widely used benchmark for essay scoring (Mathias and Bhattacharyya, 2018). They differ in type and in how they use outside information. All essays are required to be between 150 and 250 words.

For the *Narrative*, participants were asked to rely solely on their experience. Participants are requested to describe the challenging experience in their life and how it led to personal growth.

For the *Source Dependent* essay, a specific article[2] of appropriate length and specialized content is selected and participants are asked to analyze it. The target article is a knowledgeable reflection on an environmental issue and requires an explanation and discussion of the topic.

The *Argumentative* essay type is an essay that asks participants to freely find sources of information and construct an argument. It asks to write about whether they agree or disagree with the social issue of basic income.

Types of experiment scenario To ensure that we could meaningfully categorize cases where students wrote their own essays, we controlled participants' understanding of the essays into four levels in order to check feasibility in real-world scenarios as follows:

- **Read**: Participants read an AI-generated essay for five minutes and then submit it.
- **Rewrite**: Participants receive an AI-generated essay and rewrite it in their own style within five minutes.
- **Perusal**: Participants carefully read an AI-generated essay, following a provided checklist. The checklist consists of three open-ended questions that allow participants to think again about the main content of the essay, and there is no time limit.
- **Write**: Participants write an essay on their own, within a 30-minute time limit. The essay should be no less than 150 words and no more than 250 words, and participants have up to 30 minutes to complete it.

All groups were fully informed about the Socratic Assessment and were instructed to try to get the best possible score. For all timed experiment scenarios, the Source-dependent essay type gives

[2]https://www.nature.com/articles/d41586-023-00953-x

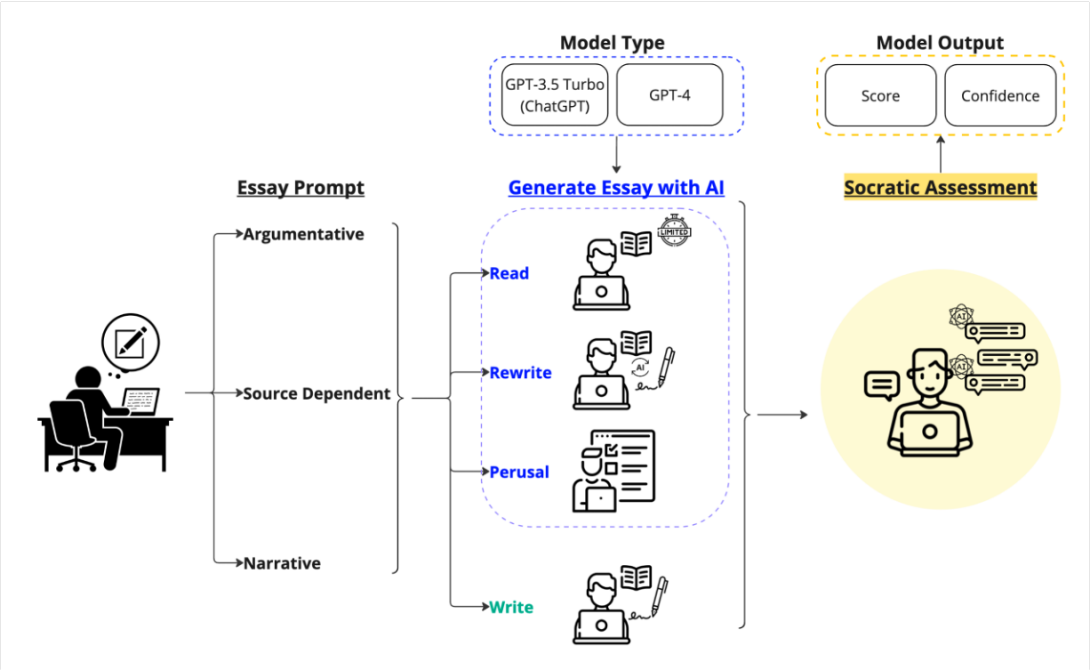


Figure 2: Overview of our experiment. We group participants into four experiment settings, *Read*, *Rewrite*, *Perusal*, *Write* randomly. The type of essay and the generation model for the essay was also determined randomly. After the participant submits their essay, their understanding of the essay is evaluated by Socratic Assessment.

participants an additional 7 minutes to read and understand the article.

Participants detail Since it is essential to guarantee that participants genuinely write their own essays without using ChatGPT, we need to recruit individuals who we could trust to be truthful. So we thus recruited participants for the experiment from within our organization. Twenty-four insiders responded to a total of 41 pairs of essay-dialogue data. Each person could participate in the experiment up to three times.

No one was assigned to the same essay topic or the same scenario. In total, 13 people responded once, 5 people responded twice, and 6 people responded three times. All participants had a bachelor's degree and were instructed to never use ChatGPT except as provided within the experiment.

Expert evaluation We asked an expert, who has a master's degree in education and has worked as a math teacher in a US high school, to label the essay and dialogue data. For each essay-dialogue pair, the expert was asked two questions: "How likely is it that the student wrote this essay on their own, regarding the dialogue about the essay?" and "Did the AI-generated questions help you determine if the essay was written by a student?" They can be answered on a Likert-7 scale, where 1/7 represents 'strongly disagree' and 'strongly agree' respectively.

Prompt setting for GPT-4 To build our experiments, we utilized GPT-4 in three ways. 1) generating essays to help students cheat on their essays; 2) generating questions to reveal whether a student wrote an essay on their own; and 3) calculating the probability that a student wrote an essay on their own by watching the AI's dialog with the student about the essay.

For the first use case, half of the users received an essay generated by gpt-3.5-turbo and the other half received an essay generated by gpt-4 to see the difference in the essay writing model. In the second use case, to test how well the AI text classifier performs in Socratic questioning, half of the users were prompted with a guide to Socratic questioning (Paul and Elder, 2019), while the other half was simply asked a question to determine whether the student had written an essay.

4.2 Result

Data description For experiment scenario and essay type settings, we applied random grouping to participants, so obtained 9, 11, 11, and 10 samples for *Write*, *Read*, *Rewrite*, *Perusal* respectively, and 14, 13, 14 samples for *Narrative*, *Source-Dependent*, and *Argumentative* respectively.

Essay Type	Source	Essay Prompt
Narrative	One's experience	Describe a challenging experience in your life and explain how it has shaped your personal growth.
Source Dependent	Specific Article	Briefly explain the time-machine analogy presented in the article and discuss its implications for the effectiveness of current carbon dioxide removal (CDR) methods as a climate solution.
Argumentative	Allow to search	Argue for or against the implementation of universal basic income as a solution to poverty and income inequality.

Table 1: Types of essay and its prompt

For each experiment setting, we got the total probability of whether the student is likely to be a real author of the essay by prompting GPT-4. As shown in Figure 5, the final probabilities were distributed differently across the experimental scenarios, with the highest in the group that wrote their own essays. The distribution was also different for different types of essays.

To figure out the vulnerability of the black-box AI text detection method in a practical essay assignment setting, we analyzed each submitted essay text with the OpenAI AI text classifier. Figure 4 shows the distribution of the detection result.

Discrimination of user scenarios In the experimental scenario, the *Write* group wrote the essay by themselves and all other groups used AI texts, so the grouping on scenario plays the role of the gold label for whether the essay is AI-generated. Based on the label, we measured the AUROC of AI-generated text detection for GPT-4, Human Expert, and the OpenAI AI text classifier, respectively. Figure 3 shows the results, which show that the performance of each methodology varies significantly depending on the type of essay.

Evaluation on the question usability Since we didn't perform specific prompt tuning, we used two different types of prompts that asked students questions to see if the Socratic Assessment could be enhanced, and in each case, we analyzed the difference in how much the questions helped the expert make a judgment in Likert 7 scale. We found a trend toward higher scores for the prompt that simply asked students to discern authenticity, with a mean of 5.6(1.7), and for the prompt that included the methodology of Socratic Questioning (Paul and Elder, 2019), with a mean of 6.1(0.9). The range of the score is 1~7 and () is the standard deviation. This result shows that there is room for improvement with better questioning by adjusting prompts. Furthermore, the overall positive response for question usability, 5.88/7 implies that Socratic Assessment can provide helpful information about a student's essay fraud detection just by providing the teacher with a conversation about the essay, without providing the score directly.

Robustness to paraphrasing By comparing the AUROC of detection with and without the *Rewrite* group, we can assess how AIT and Socratic Assessment are impacted by paraphrasing. The AUROC for Socratic Assessment increases slightly from 62.5 for the full population to 63.2 without the Rewrite group, while AIT experiences a significant increase from 59.0 to 66.4. For the narrative essay type, AIT increases from 81.3 to 90.0 for the full population, while Socratic Assessment decreases marginally from 81.8 to 79.2. This suggests that AIT is vulnerable to paraphrasing, while Socratic Assessment remains relatively unaffected.

5 Conclusion & Future Work

5.1 Conclusion

In this paper, we argue that AI-text detection methods cannot prevent cheating in essay assignments and propose a similar practical methodology, Socratic Assessment. We find that Socratic Assessment is significantly effective with a simple prompt design, which demonstrates the field applicability of this work.

Our analysis of 41 essay-dialog data shows that Socratic Assessment can discriminate whether a student is AI-generated or not for certain types of essays, namely Narrative and Source-dependent essays, with AUROCs of 81.8 and 75.0, respectively. We also found that a teacher expert could identify student cheating with high accuracy by watching the student-AI dialog about the essay, suggesting that even if the accuracy of Socratic Assessment's automated scores is somewhat lower, it can serve as a supplementary tool to help teachers diagnose student cheating through intervention.

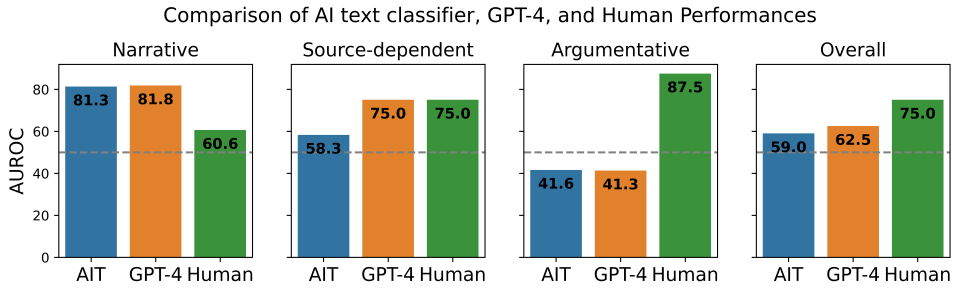


Figure 3: AUROC of various detection methods: OpenAI AI text classifier (AIT), GPT-4, and the expert.

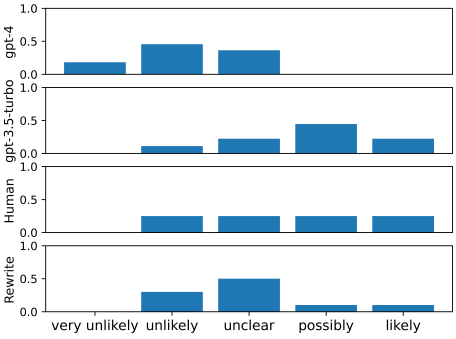


Figure 4: Prediction of the AI text classifier (AIT). Both high type-I errors on the text generated by GPT-4 and high type-II errors on the human-written text contributed to low AUROC of AI text detection.

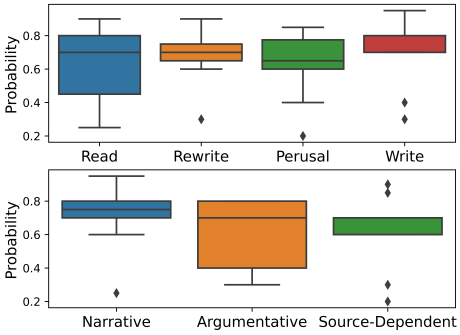


Figure 5: Boxplot of the probability by prompted GPT-4 that the student wrote the essay on their own, given the essay and the dialogue about the essay of the student. Above is the distribution by experiment scenario; Below is the distribution by essay type.

Therefore, Socratic Assessment has the potential to be a countermeasure against essay cheating using AI-generated texts by verifying that students possess the core knowledge that they are expected to achieve through essay assignments. With further complementary research, we are anticipating it to grow into a tool that allows students to practice their creative knowledge in a more dynamic form than before.

5.2 Future work

(Nelson and Dawson, 2014) argues that assessment should be an educational experience, not just a diagnostic. Socratic assessment should go beyond simply checking that students have completed the assignment, and consider how it can work as a device for students to think more deeply about the essay topic and learn from it. In order for Socratic Assessment to work as meaningful feedback, it should aim to enable the gap between the actual level of performance and the desired learning goal to be bridged (Evans, 2013).

The focus of this study is not on generating pointed questions for students, but rather on predicting the confidence level of our judgment based on the resulting conversations. Due to the limitations of our methodology in analyzing the educational effects of the question-and-answer cycle on students, we focus on the effects of the assessment in this study. Therefore, in this study, we focus less on prompt engineering and analysis of the language model that performs the role of asking questions.

Based on the qualitative feedback we received from participants and the labeler after the experiment, their prior knowledge or usual interest in the topic seems to have a significant impact on the outcome of the Socratic Assessment. In order for this system to work, more analysis of the competencies needed to write an essay needs to be done to take into account students who have not written an essay themselves, but already possess the competencies needed to write one.

In addition, a more detailed analysis of the potential variables that could affect the scores on the Socratic Assessment is needed. This study is subject to the limitations outlined in the following sections, and more follow-up research is needed to address

the potential risks of using these systems with real students.

6 Limitations

Availability of cheating in Socratic Assessment One limitation of the proposed Socratic assessment is that even the Socratic questions can be answered by LLM itself. Therefore, when Socratic assessment is applied in real-world scenarios, controlling and supervising the test environment must be taken into account. For example, one possible approach could be to enforce video proctoring during the exam, which may help to encourage students to act conscientiously. Alternatively, requiring the completion of essay writing within a short time frame could be considered, as this would limit the time available for querying LLMs for cheating.

Human experiment setting Since the purpose of this study is to explore the potential of Socratic Assessment as a means of countering the violations of the right to learn by generative AI, we did not aim for strict statistical precision. There were two practical limitations in our human experiments.

Firstly, our experiments were conducted on the employees of an education company, not actual students in educational organizations, due to the difficulty of designing and controlling the experiments. The experimenters were not informed in advance about the type of essay or the scenario of the experiment, but because they are part of an education company, they may have inferred the intent of the experiment and been influenced by their responses. Due to such differences, there will be some differences between our experimental results and the real-world scenario for students.

The second problem is our limited resources for experiments. In order to gain as much insight as possible within a limited number of parameters, we introduced many variables into the experiment, so the distribution of scores for each group did not follow a normal distribution, and we were unable to test whether each variable caused a significant change in Socratic Assessment scores. In addition, to achieve a high-quality Expert label, we requested highly qualified experts and suffer difficulty to find, so we couldn't conduct cross-validation by multiple labelers.

In order to apply the findings to the practical education purpose, more rigorous experimental validation is necessary.

Implementation Cost Using Large Language Model API for Socratic Assessment can be expensive. We used GPT-4 API, but using the current GPT-4 API is slower and 15~30 times more expensive than gpt-3.5-turbo. To provide this as an infrastructure for educational purposes, collective investment from a larger scale is necessary.

Assessment incompleteness and low reliability As a preliminary experimental validation, this methodology has many limitations. As stated by Turnitin's AI writing detection service[3], *"it should not be used as the sole basis for adverse actions against a student."* Teachers or evaluators should check the alerts for high-risk cases of plagiarism and use their subjective assessment to detect plagiarism by directly viewing the student's verification conversation history.

7 Ethical Considerations

Developing and implementing the Socratic Assessment's AI system for use in the educational sector will prioritize ethical considerations. Feedback from teachers, students, and educational experts will be considered. This section discusses the main ethical considerations associated with the research and proposes strategies to mitigate potential ethical risks. By evaluating the Socratic Assessment using the Human AI Experience Guidelines created by Microsoft, we will consistently improve the assessment based on these guidelines (Amershi et al., 2019). The guidelines are best practices for designing human interaction with AI-based products and feature responsibly.

Preventing Harm and Misuse from Users We will take a diligent approach to explain its capabilities and limitations to prevent any harm or misuse of Socratic Assessment's AI system. We will establish clear expectations about what the AI system is designed for. Unclear user expectations regarding supported tasks or domains can lead to misuse, product abandonment, and harm. By providing detailed information about the AI system, users can make informed decisions about its use. Additionally, we will provide user-friendly documentation so that users can easily understand the system's features and limitations. This will help prevent misunderstandings and ensure the AI system is used properly and effectively.

[3] https://www.turnitin.com/products/features/ai-writing-detection

Mitigating Social Biases of Assessment We are committed to identifying, testing, and mitigating social biases to ensure that the AI system's language and behaviors do not reinforce undesirable or unfair stereotypes and biases. If the Socratic Assessment questions unintentionally produce unethical or biased questions and writing, we will allow the user to report it. During regular interactions with the AI system, we will ask the user to provide feedback indicating their preferences. Additionally, we will empower the user to influence how the AI evolves by enabling them to provide explicit feedback on AI system outputs and behaviors. To minimize these risks, the research will also use diverse essays by receiving feedback from teachers.

Improving Trust of Assessment We will explain why a certain essay received a particular score to ensure transparency and provide teachers with a more in-depth understanding of the grading process. This additional information will not only increase user trust, but it will also allow teachers to better communicate with their students and provide constructive feedback. Moreover, we believe this approach will enable students to learn from their mistakes and improve their writing skills. We understand that the grading process is imperfect, so we will receive feedback from teachers on the scoring. This feedback will be carefully analyzed and used to update the Socratic Assessment, ensuring our grading system remains fair, accurate, and consistent. By continuously improving our assessment process, we aim to provide teachers and students with the best possible experience and help them achieve their goals.

Protecting Data and Privacy The Socratic Assessment system collects students' names, essays, and records of their interactions with the AI. However, we have anonymized students' names and deleted them. We limited data collection to the minimum amount required for the research study. We did not ask any unnecessary questions unless they were directly relevant to the project's goals. During the pilot test, we followed responsible data management practices, emphasizing maintaining confidentiality throughout the process. We will implement strict protocols for handling and storing data to minimize the risks of unauthorized disclosures or breaches. We will regularly review and update security measures to ensure compliance with applicable regulations and industry standards.

Ultimately, the goal is to create a system that effectively evaluates students' writings and promotes responsible Socratic assessment by considering students' diverse cultures and cognitive learning abilities. This will benefit teachers and students, providing them with a fair, accurate, and consistent evaluation tool.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Beng Heng Ang, Sujatha Das Gollapalli, and See-kiong Ng. 2023. Socratic question generation: A novel dataset, models, and evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 147–165, Dubrovnik, Croatia. Association for Computational Linguistics.

Arthur N Applebee. 1984. Writing and reasoning. *Review of educational research*, 54(4):577–596.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4:253–278.

Haris Delić and Senad Bećirović. 2016. Socratic method as an approach to teaching. *European Researcher. Series A*, (10):511–517.

Heather Desaire, Aleesa E Chua, Madeline Isom, Romana Jarosova, and David Hua. 2023. Chatgpt or academic scientist? distinguishing authorship with

over 99% accuracy using off-the-shelf machine learning tools. *arXiv preprint arXiv:2303.16352*.

Carol Evans. 2013. Making sense of assessment feedback in higher education. *Review of educational research*, 83(1):70–120.

Tomáš Foltỳnek, Norman Meuschke, and Bela Gipp. 2019. Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *ArXiv*, abs/2301.10226.

Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Robert Nelson and Phillip Dawson. 2014. A contribution to the history of assessment: how a conversation simulator redeems socratic method. *Assessment & Evaluation in Higher Education*, 39(2):195–204.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. 2018. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*.

OpenAI. 2023. Gpt-4 technical report.

Richard Paul and Linda Elder. 2019. *The thinker's guide to Socratic questioning*. Rowman & Littlefield.

Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Xuandong Zhao, Yu xiang Wang, and Lei Li. 2023. Protecting language generation models via invisible watermarking. *ArXiv*, abs/2302.03162.