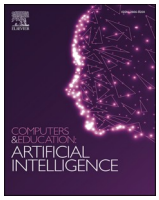
Computers and Education: Artificial Intelligence 6 (2024) 100209

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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



Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays

Johanna Fleckenstein [a,b,*], Jennifer Meyer [b], Thorben Jansen [b], Stefan D. Keller [c], Olaf Köller [b], Jens Möller [d]

[a] University of Hildesheim, Germany
[b] Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[c] Zurich University of Teacher Education, Switzerland
[d] Kiel University, Germany

ARTICLE INFO

Keywords:
Generative AI
Writing assessment
Teachers
Essay writing
ChatGPT

ABSTRACT

The potential application of generative artificial intelligence (AI) in schools and universities poses great challenges, especially for the assessment of students' texts. Previous research has shown that people generally have difficulty distinguishing AI-generated from human-written texts; however, the ability of teachers to identify an AI-generated text among student essays has not yet been investigated. Here we show in two experimental studies that novice ($N = 89$) and experienced teachers ($N = 200$) could not identify texts generated by ChatGPT among student-written texts. However, there are some indications that more experienced teachers made more differentiated and more accurate judgments. Furthermore, both groups were overconfident in their judgments. Effects of real and assumed source on quality assessment were heterogeneous. Our findings demonstrate that with relatively little prompting, current AI can generate texts that are not detectable for teachers, which poses a challenge to schools and universities in grading student essays. Our study provides empirical evidence for the current debate regarding exam strategies in schools and universities in light of the latest technological developments.

1. Introduction

Generative AI is a broad term for any type of artificial intelligence (AI) that can produce new texts, images, videos, or computer code. Based on massive amounts of training data, generative AI can use prompts written in natural language to generate new output. Large language models (LLM), such as the Generative Pre-trained Transformer 3 (GPT-3), can generate human-like text and complete other language-related tasks such as translation, question answering, and coding. The use of LLM like GPT-3 in educational contexts has been a recent matter of discussion (Cotton et al., 2023; Kasneci et al., 2023).

As a variant of the GPT-3 model, the chatbot application ChatGPT is optimized for conversation modeling, which makes it a user-friendly tool. With the growing distribution of ChatGPT, the public has become aware of the impact that automated writing will have on schools and universities. On the one hand, educators see great potential in optimizing teaching and learning processes (Cotton et al., 2023; Sailer et al., 2023). AI is in particular helpful in schools to give feedback to student achievements which is process-oriented and available in real-time to high numbers of students. In their experimental study, Sailer et al. (2023) showed that AI-generated adaptive feedback facilitates pre-service teachers' quality of justifications in written assignments as a part of their diagnostic competence. Jansen et al. (2024) showed that automatic feedback including goal-setting support is in particular helpful when students had to revise their texts. In one of the first experimental studies on the use of LLMs for feedback generation, Meyer et al. (2024) showed that AI-generated feedback increased students' revisions, motivation, and positive emotions. On the other hand, there are certain caveats regarding the role of generative AI. Cotton et al. (2023) discussed the challenges concerning the use of ChatGPT in higher education. These include cheating and deception, causing issues for assessment and learning. Students who use tools like ChatGPT in written assignments have an unfair advantage when it comes to assessment and grading. At the same time, they might miss out on learning

* Corresponding author. University of Hildesheim, Universitätsplatz 1, 31141, Hildesheim, Germany.
E-mail addresses: fleckenstein@uni-hildesheim.de (J. Fleckenstein), jmeyer@leibniz-ipn.de (J. Meyer), tjansen@leibniz-ipn.de (T. Jansen), stefandaniel.keller@phzh.ch (S.D. Keller), koeller@leibniz-ipn.de (O. Köller), jmoeller@ipl.uni-kiel.de (J. Möller).

https://doi.org/10.1016/j.caeai.2024.100209
Received 23 June 2023; Received in revised form 8 January 2024; Accepted 24 January 2024
Available online 25 January 2024

opportunities. Therefore, it is vital for teachers to know if and when AI tools are used, especially for the formative and summative assessment of students' writing performance. If students' writing skills suffer because they use AI tools for their texts very frequently, teachers could counteract, for example, by creating assignment or exam conditions which ensure that students produce their texts without using AI.

Cotton et al. (2023) stated that it can be difficult for teachers to distinguish between a student's own writing and the responses generated by LLMs. They suggest a few approaches that can be used to distinguish human writing from generative AI. Accordingly, certain patterns and irregularities in the language as well as context unawareness could be signs of AI, whereas grammar and spelling errors typically signal human writing. However, empirical evidence is still missing on the questions of how difficult it actually is for teachers to identify AI-generated texts and how they assess AI-generated texts compared to student texts. This study presents the first experimental data base to address these questions and close this research gap.

2. Related studies

Multiple empirical studies have compared texts generated by AI with texts written by humans in several domains, such as news articles (Graefe et al., 2018), scientific abstracts (Gao et al., 2023), poetry writing (Gunser et al., 2022; Köbis & Mossink, 2021), portrait drawing (Yalçın et al., 2020), and music composition (Zacharakis et al., 2021). For example, in Graefe et al. (2018), participants rated AI-generated news articles as more credible and harder to read and attested them higher journalistic quality than comparable human-written articles. Recently, the study by Gunser et al. (2022) showed that even in the production of literary texts generated by GPT-2 could hardly be distinguished from human-generated texts. The participants read AI-based continuations and either human-written continuations or original continuations of literary texts, generated based on the first lines of texts and poems of classic authors. Participants identified only 60 percent of human-written texts and 58 percent of AI-generated texts correctly. Similarly to a previous study by Köbis and Mossink (2021), Gunser et al. (2022) found that even though the subjects did not succeed in correctly identifying the source (human or AI), they were overconfident in their ability to identify sources. However, their confidence was not related to their classification accuracy. The evaluations of the poems depended on the actual source: Subjects perceived AI-generated poems as less well-written, inspiring, fascinating, interesting, and aesthetic than both human-written and original poems. Gao et al. (2023)found that human reviewers misclassified 32 % of ChatGPT-generated scientific abstracts as being human-written and incorrectly identified 14 % of original abstracts as being AI-generated.

3. Research status and gaps

Despite the prior research on the detectability and assessment of AI-generated texts, there are no studies to our knowledge that have focused on the educational context. This issue, however, has become increasingly relevant: The new developments in the field of text-generative AI have led to the question of how well teachers can identify AI-generated texts among texts written by students. For teachers, on the one hand, it could likewise be the case that they are unable to recognize the source of argumentative texts and thus would be unaware of any attempts at deception. On the other hand, teachers are experts in assessing student texts and could recognize, for example, typical mistakes or inadequate vocabulary use of learners. Such unique characteristics of student texts might help teachers to differentiate them from AI-generated texts. In addition, student essays with their particular shortcomings may be difficult to simulate by AI because student texts may not be part of the AI training data. Furthermore, AI-generated texts might be too perfect in terms of language, structure, and content compared to EFL student texts, making it easy to identify the text source. Thus, the perceived text quality could differ depending on the source, which would have implications for the correctness of their assessment. However, an empirical test of these assumptions has yet to be conducted. Drawing on two samples with presumably different levels of expertise, we aimed to address this research gap.

The main research questions addressed in our two studies are.

(1) Are teachers able to identify AI-generated texts among student-written texts?
(2) How confident are teachers in identifying the source?
(3) How do teachers assess text features like overall quality, language, structure, and content of the texts in relation to their assumed and actual source?

In Study 1, we investigated these research questions in a sample of preservice teachers with little prior teaching experience and English as a foreign language. In Study 2, we aimed to acquire more experienced teachers who were native speakers of English.

4. Study 1

4.1. Methods

4.1.1. Participants

A sample of $N = 89$ pre-service teachers was recruited via an online lecture for pre-service teachers enrolled in a Master of Education program at a German university. The sample consisted of 57 females, 23 males and nine participants, who did not indicate their gender. Participants' age ranged from 21 to 38 years ($M = 25.46$, $SD = 3.35$).

4.1.2. Materials

4.1.2.1. Selection of student-written texts. For our study, we needed essays from students to compare them with AI-written texts. Argumentative essay writing in EFL takes a central place in the curriculum at upper-secondary level in many countries and is a major part of international high-stakes language assessments such as the Test of English as a Foreign Language (TOEFL; Keller et al., 2020). The texts were selected from a large corpus of EFL argumentative essays (Keller et al., 2020) written by upper-secondary level students from the academic track in Germany and Switzerland. The writing prompt came from the essay writing section of the TOEFL iBT® test (for details see Rupp et al., 2019). It consists of argumentative writing on a controversial topic which is formulated in an agree-disagree format: The human-written essays were produced by 11th grade students (academic track) in English as a response to the statement "Do you agree or disagree with the following statement: A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught. Use specific reasons and examples to support your answer.' (30 min writing time, no preparation or resources) ".

The students' essays were rated on the official TOEFL iBT® independent writing rubric on a scale from 0 to 5 (ETS, 2022). The expert raters reached a high exact agreement of 62.5 % of texts and showed a quadratic weighted kappa of $\kappa = 0.67$. The machine rating score was created for each essay within the MEWS study using the e-Rater® of the ETS (see Rupp et al., 2019 for a detailed description of the process).

We used the corpus to randomly select two human-written essays under certain restrictions: One low-quality text should score 2 on the writing rubric from 0 to 5, one high-quality text should score 4. The machine scores and both human scores should be 2 and 4, respectively, to ensure reliability of the rating. Both texts should be free of German words.

4.1.2.2. Prompting of AI-generated texts. Using ChatGPT allowed us to create two texts of similar quality for the same prompt as our human-

written texts. We prompted ChatGPT with the respective descriptors of official TOEFL iBT® writing rubric level 2 and level 4 (see Supplement for the detailed prompts and selected texts). For validation of the text quality manipulation, the AI-generated texts at the two levels were scored by a supervised machine learning algorithm (Zesch & Horbach, 2018) developed used in previous studies (Jansen et al., 2024). Thus, we obtained two AI-generated texts of similar quality as the student texts. This way, we could analyze the combined effects of text source (AI- vs. human-written) and text quality (low vs. high) while controlling for text quality in both experimental conditions.

4.1.3. Procedure

The assessment of student texts took place using a digital tool, the *Student Inventory*, which was developed based on the free online survey application Limesurvey to assess student texts in an experimental computer-based setting. The *Student Inventory* has already been used for studies on text assessments in other contexts (Jansen et al., 2021; Möller et al., 2022). It allows scoring through a screen split, presenting student texts on the left part of the screen, and assessment items on the right.

In our experimental study, each subject was informed about the research purpose and the procedure: „In the following, you are asked to decide for 4 texts whether they were written by an artificial intelligence or a student. Afterwards, we will ask you to evaluate the texts on the dimensions of "language", "structure" and "content", and finally to provide an overall assessment of the respective writing performance. On a separate page, pre-service teachers were informed, that half of the texts were not written by a student, but by a computer program (artificial intelligence): „Please tick for each text whether you think it was written by an artificial intelligence or not.“ The next page presented the first text to the pre-service teachers and they were asked to decide: „This text was written by an artificial intelligence.“ (yes/no). Furthermore, they had to judge their confidence in the decision (“How confident are you in your decision?” Six-point-scale ranging from “50 % - I guessed” to “100 % - very sure”) thus following the procedure by Gunser et al. (2022). After the first evaluation phase all texts were presented again and participants were asked to rate the three analytic criteria *language* (word usage, mechanics of language, grammar, style), *structure* (paragraphs, introduction, conclusion), *content* (topic, argumentation, conclusion) and overall quality on a scale from “1 – very low quality” to “7 – very high quality”. Participants could read more detailed information on the assessment criteria. The participants were debriefed immediately after their evaluations at the end of the study and were able to see the correct expert judgments.

4.2. Results

4.2.1. Source identification

Table 1 shows percentages of correct and incorrect classifications. Pre-service teachers identified only 45.1 percent of AI-generated texts correctly and53.7 percent of student-written texts. They correctly classified 46.4 percent of low-quality texts and 51.2 percent of high-quality texts. Descriptively, the highest number of misclassifications (59.8 %) appeared for low-quality texts that were AI-generated (see Table 1).

Firstly, a one-sample *t*-test against 1 was performed to test whether the percentage of correctly identified source significantly differed from a perfect identification of 100 percent. Results showed that the correct source identification was far from perfect for AI-generated texts ($t(163) = -14.08$, $p < .001$, $d = -1.10$) and student-generated texts ($t(163) = -11.870$, $p < .001$, $d = -0.936$). Secondly, a 2 (actual source: human vs. artificial) x 2 (assumed source: human vs. artificial) x 2 (quality: high vs. low) Chi-Square-test revealed independency of real source and assumed source ($\chi^2(1) = 0.05$, $p = .825$, $\varphi = -0.01$), meaning that preservice teachers cannot correctly identify the source of the texts being AI-generated or student-generated. It holds true for different levels of text quality as well as for the different actual sources.

Table 1
Percentage of correct and incorrect classifications.

Text quality			Source assumed by preservice teachers	
			Student	AI
Low	Real Source	Student	53.7 %	46.3 %
		AI	59.8 %	40.2 %
	Total		56.7 %	43.3 %
High	Real Source	Student	53.7 %	46.3 %
		AI	50.0 %	50.0 %
	Total		51.8 %	48.2 %
Overall	Real Source	Student	53.7 %	46.3 %
		AI	54.9 %	45.1 %
	Total		54.3 %	45.7 %

4.2.2. Confidence of source identification

On average, pre-service teachers estimated the confidence in their source identification with 77.3 percent for AI-generated texts and 76.9 percent for student-written texts. We performed a multiple regression analysis to predict their confidence (see Table 2). The real source (student = 0; AI = 1), the assumed source (student = 0; AI = 1), and the text quality (low level = 0; high level = 1) were included as predictors as well as the first- and second-order interaction terms between the three variables. Results showed that only the assumed source significantly predicted confidence: Pre-service teachers were less confident in their judgment when they assumed the text to be AI-generated. The real source and the interaction term of real and assumed source, however, did not predict confidence. Results did not support the assumption that pre-service teachers are more confident when their source identification is correct. Text quality did not affect the confidence in source identification.

4.2.3. Text assessment

In four additional multiple regression analyses, we investigated the effects of real and assumed source and text quality on the holistic (overall score) and analytic (language score, structure score, content score) assessments of the texts (see Table 3). As to be expected, text quality significantly predicted all assessment scores, indicating the successful manipulation of the text quality. The main effects of real and assumed source were not significant. However, the significant interactions between the two (for all aspects but language) showed that preservice teachers assigned lowest scores to student texts they correctly assumed to be written by students. AI-generated texts were assessed more positively independently whether they were identified correctly or not. (see Fig. 1). Further interaction terms were included and did not change the results.

4.3. Discussion

The central goal of the study was to investigate to what extent pre-service teachers are capable to distinguish between AI-generated texts and student-written texts and how their identification of the source relates to their text assessment. Our results show that the participants were not able to detect AI-generated argumentative essays. The percentages

Table 2
Regression of confidence estimation on real source, assumed source, and text quality.

Predictors	*B*	SE_B	*p*
(Intercept)	82.05	2.35	<.001
Real source	−2.25	3.24	.488
Assumed source	−9.68	3.45	<.01
Real source * Assumed source	4.43	4.92	.369
Text quality	−2.05	3.32	.539
Real source * Text quality	3.23	4.68	.491
Assumed source * Text quality	1.52	4.88	.756
Real source * Assumed source * Text quality	−4.32	6.93	.534

Table 3
Regression of assessment scores on real source, assumed source, and text quality.

Predictors	*B*	SE_B	*p*
Overall score			
(Intercept)	3.21	0.19	<.001
Real source	0.38	0.27	.155
Assumed source	0.37	0.28	.197
Real source * Assumed source	−0.99	0.40	<.05
Text quality	1.38	0.28	<.001
Real source * Text quality	0.30	0.38	.441
Assumed source * Text quality	−0.12	0.40	.768
Real source * Assumed source * Text quality	0.95	0.57	.095
Language score			
(Intercept)	2.88	0.23	<.001
Real source	0.32	0.31	.297
Assumed source	0.46	0.33	.162
Real source * Assumed source	−0.79	0.47	.093
Text quality	1.74	0.32	<.001
Real source * Text quality	0.18	0.45	.688
Assumed source * Text quality	−0.37	0.47	.428
Real source * Assumed source * Text quality	0.83	0.66	.210
Structure score			
(Intercept)	3.17	0.22	<.001
Real source	0.43	0.30	.164
Assumed source	0.39	0.32	.235
Real source * Assumed source	−1.01	0.46	<.05
Text quality	0.98	0.32	<.01
Real source * Text quality	0.77	0.44	.080
Assumed source * Text quality	0.23	0.46	.610
Real source * Assumed source * Text quality	0.27	0.65	.675
Content score			
(Intercept)	3.71	0.22	<.001
Real source	0.18	0.29	.532
Assumed source	0.02	0.31	.943
Real source * Assumed source	−0.92	0.45	<.05
Text quality	0.95	0.31	<.01
Real source * Text quality	0.39	0.42	.354
Assumed source * Text quality	0.13	0.44	.775
Real source * Assumed source * Text quality	0.85	0.63	.175

Real source: student = 0; AI = 1; assumed source: student = 0; AI = 1; text quality: low = 0; high = 1

of correctly identified sources indicate that they merely guessed whether the text was generated by AI or written by a student. This result is even more noteworthy because pre-service teachers were overly confident in their judgments regarding the real source of the text. They felt rather confident in their classification between AI-generated and human-written texts when, in fact, the misclassification rates were high. These findings can be considered the first empirical evidence for the widely discussed assumption that AI-generated texts might not be recognized by teachers. Moreover, they indicate that teachers may not be aware of their inability to identify AI-generated texts among student essays. Students therefore have the opportunity to successfully trick the preservice teachers into believing that the texts generated by the AI originate from themselves. Furthermore, the findings show that the preservice teachers' assessment of the texts depended not only on the actual quality of the texts but also on an interaction of real and assumed source. When they assumed a text to be written by a student, they assessed it more positively when it was actually generated by AI. From an educational perspective, this is problematic as it implies that using AI to generate texts in school contexts may even be beneficial for students in terms of grades.

Based on Study 1, the chance that teachers will be able to identify texts that students have not written themselves, but generated with the help of AI, seems rather slight. However, the limitations of this study need to be addressed before drawing far-reaching conclusions: First, the sample consisted of preservice teachers who may not have had the training, the experience, and the linguistic proficiency to fulfill the task successfully. One might expect a more experienced and more proficient sample of teachers to outperform the novices in source identification and assessment of text quality. Second, the participants were told that half of the texts they saw were AI-generated. Even though their classifications did not necessarily reflect these instructions, this limits the ecological validity of the experiment. A teacher in a real-life educational setting would not have any information on how many texts turned in by students had actually been AI-generated. Third, we only used two student texts that were selected from a larger corpus. These two texts may possess certain characteristics that influence the outcome of our experiment. Using a larger variety of student texts could also enhance the ecological validity of the study. In order to remedy these limitations, we conducted a second study, using a sample of experienced teachers, a larger number of texts, and an adjusted procedure in order to enhance the validity of the findings.

5. Study 2

5.1. Methods

5.1.1. Participants

A sample of $N = 200$ teachers was recruited via the online research platform Prolific. Participants were prescreened according to their occupation (teacher) and their first language (English)). Additionally, they were asked to indicate whether they currently worked as a teacher (93.0 %) and whether they were teaching English as a subject (61.5 %). The sample consisted of 140 (70.0 %) females and 60 (30.0 %) males. Participants' age ranged from 22 to 77 years ($M = 42.83$, $SD = 11.51$). The vast majority of participants were from the UK and North America.

5.1.2. Materials

Unlike in Study 1, we did not use a small sample of preselected student texts but used an algorithm that randomly picked texts from the large corpus of EFL argumentative essays (Keller et al., 2020; see Study 1 for a detailed description). The only restrictions pertained to the text quality score: Each participant was provided with one student-written low-quality text (score 2 on the TOEFL iBT® writing rubric) and one student-written high-quality text (score 4). For the AI-generated texts, we reused the essays generated by ChatGPT from Study 1 (see the respective methods section for a description of the prompting).

5.1.3. Procedure

In Study 2, the participants were only informed that some of the texts were AI-generated (not that it was half of the texts) in order to avoid a strong dependence of the decisions. Besides this, the procedure was identical to the one described in Study 1. Like in Study 1, participants were asked to identify the source, indicate their confidence in the decision, and assign a holistic score to the texts as well as subscores for language, structure, and content.

5.2. Results

5.2.1. Source identification

Table 4 shows percentages of correct and incorrect classifications of the text source (student vs. AI). In total, teachers decided in one third of the cases that the texts were generated by AI. They identified only 37.8 percent of AI-generated texts correctly, but 73.0 percent of student-written texts. They correctly classified more high-quality texts (47.0) than low-quality texts (63.8 %). Descriptively, the highest number of misclassifications appeared for low-quality texts that were AI-generated (see Table 4). The frequencies show that – irrespective of the actual source – teachers tended to classify low-quality texts as student-written texts.

Firstly, one-sample *t*-tests against 1 were performed to test whether the percentage of correctly identified source significantly differed from a perfect identification of 100 percent. Results showed that the correct source identification was far from perfect for both AI-generated texts ($t(399) = -25.65$, $p < .001$, $d = -1.28$) and student-generated texts (*t*

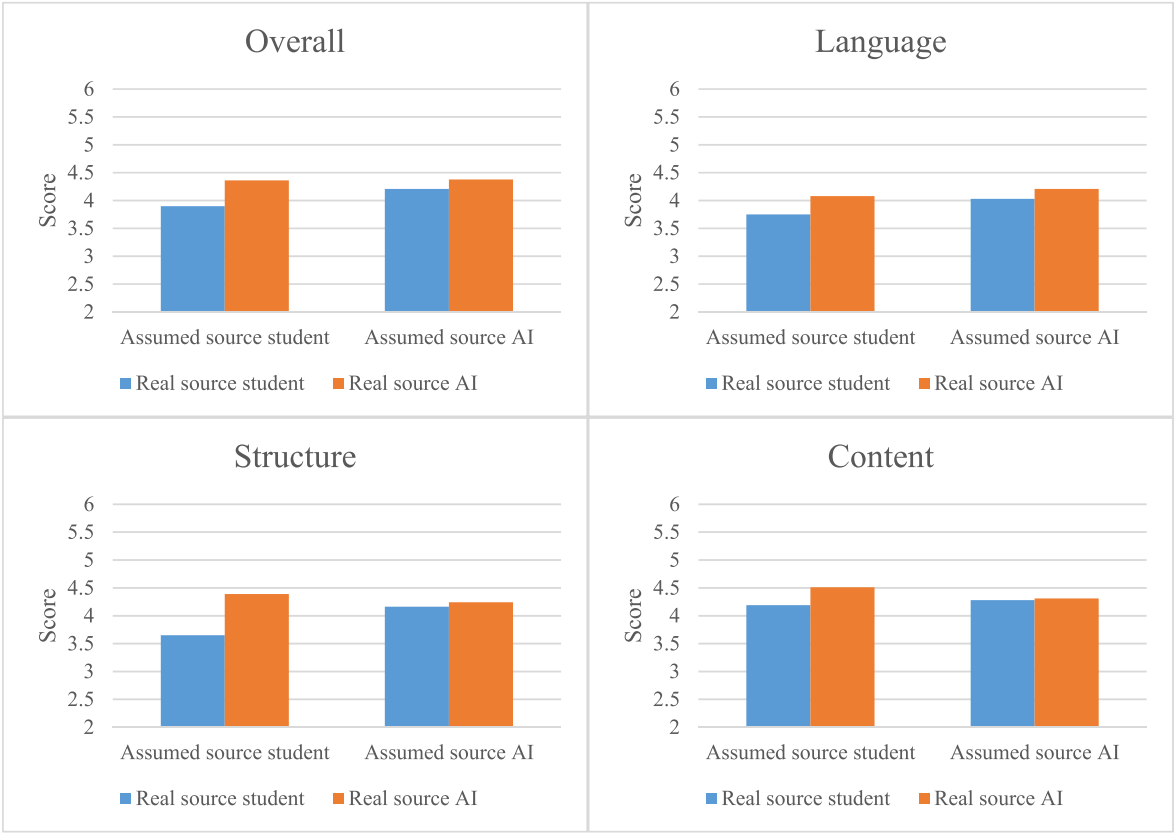


Fig. 1. Interaction effects of text quality * real source on text assessment.

Table 4
Percentage of correct and incorrect classifications.

Text quality			Source assumed by teachers	
			Student	AI
Low	Real Source	Student	83.0 %	17.0 %
		AI	89.0 %	11.0 %
	Total		86.0 %	14.0 %
High	Real Source	Student	63.0 %	37.0 %
		AI	35.5 %	64.5 %
	Total		49.3 %	50.7 %
Overall	Real Source	Student	73.0 %	27.0 %
		AI	62.3 %	37.8 %
	Total		67.6 %	32.4 %

(399) = -12.15, $p < .001$, $d = -0.61$). Secondly, a 2 (actual source: human vs. artificial) x 2 (assumed source: human vs. artificial) x 2 (quality: high vs. low) Chi-Square-test revealed differential results for low-versus high-quality texts: We found real source and assumed source to be independent for low-quality texts ($\chi^2(1) = 2.99$, $p = .084$, $\varphi = -0.09$), meaning that teachers were not able to correctly identify the source of the low-quality texts. For high-quality texts, the Chi-Square-test indicated the dependence of real and assumed source ($\chi^2(1) = 30.26$, $p < .001$, $\varphi = 0.275$). Thus, experienced teachers were better able to correctly identify the source of high-quality texts independent of whether they were AI-generated or student-written.

5.2.2. *Confidence of source identification*

On average, teachers estimated the confidence in their source identification with 80.6 % for AI-generated texts and 79.6 % for student-written texts. We performed a multiple regression analysis to predict their confidence (see Table 5). The real source (student = 0; AI = 1), the assumed source (student = 0; AI = 1), and the text quality (low level = 0; high level = 1) were included as predictors as well as the first- and second-order interaction terms between the three variables. Results showed that the assumed source significantly predicted confidence: Teachers were less confident in their judgment when they assumed the text to be AI-generated. The interaction term of assumed source and text quality was also significant (see Fig. 2), indicating that for assumed AI texts the confidence was even lower when text quality was low. For assumed student-written texts, teachers were more confident when text quality was low than when it was high. The real source did not influence the confidence and neither did the interaction of real and assumed source. The second-order interaction of all three variables was also not significant.

Table 5
Regression of confidence estimation on real source, assumed source, and text quality.

Predictors	*B*	SE_B	*p*
(Intercept)	82.05	1.11	<.001
Real source	1.83	1.54	.235
Assumed source	−8.81	2.68	<.001
Real source * Assumed source	0.39	4.19	.926
Text quality	−3.00	1.68	.075
Real source * Text quality	1.52	2.61	.561
Assumed source * Text quality	7.47	3.40	<.05
Real source * Assumed source * Text quality	−5.63	5.13	.273

Real source: student = 0; AI = 1; assumed source: student = 0; AI = 1; text quality: low = 0; high = 1

5.2.3. *Text assessment*

In four additional multiple regression analyses, we investigated the effects of real source, assumed source and text quality as well as their interaction terms on the holistic (overall score) and analytic (language score, structure score, content score) assessments of the texts (see

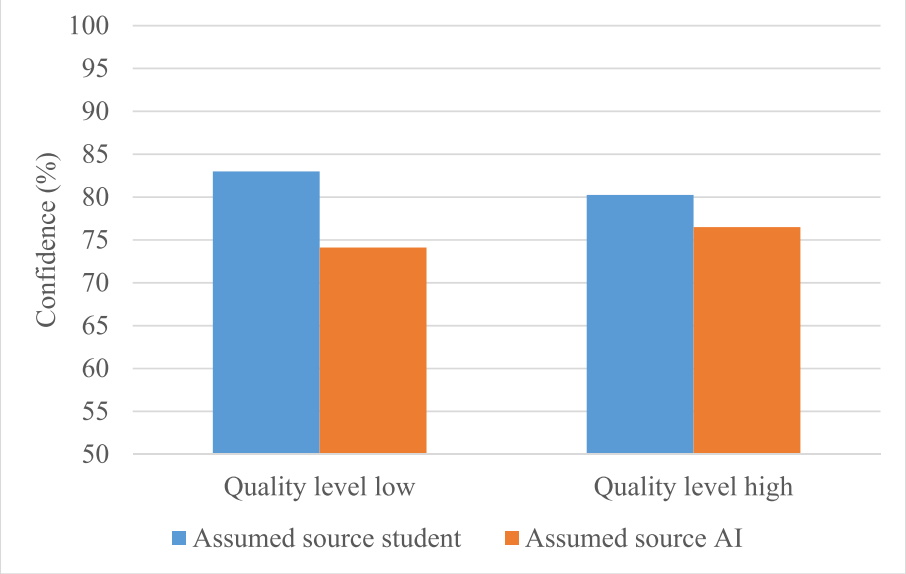


Fig. 2. Interaction effects of text quality * assumed source on estimated confidence.

Table 6
Regression of assessment scores on real source, assumed source, and text quality.

Predictors	*B*	SE_B	*p*
Overall score			
(Intercept)	2.93	0.09	<.001
Real source	−0.37	0.12	<.01
Assumed source	0.33	0.21	.118
Real source * Assumed source	−0.08	0.33	.821
Text quality	1.58	0.13	<.001
Real source * Text quality	1.76	0.21	<.001
Assumed source * Text quality	0.29	0.27	.282
Real source * Assumed source * Text quality	−0.49	0.41	.231
Language score			
(Intercept)	2.76	0.09	<.001
Real source	−0.25	0.13	<.05
Assumed source	0.33	0.22	.137
Real source * Assumed source	−0.24	0.35	.481
Text quality	1.52	0.14	<.001
Real source * Text quality	2.05	0.22	<.001
Assumed source * Text quality	0.31	0.28	.266
Real source * Assumed source * Text quality	−0.46	0.42	.278
Structure score			
(Intercept)	3.04	0.10	<.001
Real source	−0.72	0.14	<.001
Assumed source	0.26	0.25	.298
Real source * Assumed source	−0.16	0.39	.673
Text quality	1.38	0.16	<.001
Real source * Text quality	2.16	0.24	<.001
Assumed source * Text quality	0.40	0.31	.200
Real source * Assumed source * Text quality	−0.56	0.47	.240
Content score			
(Intercept)	3.27	0.10	<.001
Real source	−0.18	0.14	.180
Assumed source	0.24	0.24	.318
Real source * Assumed source	−0.00	0.37	.998
Text quality	1.60	0.15	<.001
Real source * Text quality	1.25	0.23	<.001
Assumed source * Text quality	0.25	0.30	.399
Real source * Assumed source * Text quality	−0.45	0.45	.314

Real source: student = 0; AI = 1; assumed source: student = 0; AI = 1; text quality: low = 0; high = 1

Table 6). As expected, text quality significantly predicted all assessment scores, indicating the successful manipulation of the text quality. The real source predicted all scores, with the exception of content significantly favoring AI-generated compared to student-written texts. However, the significant interaction effects of real source and text quality showed that high-quality texts received even higher scores when they were AI-generated, whereas low-quality texts tended to receive higher scores when they were written by students (see Fig. 3). The assumed source did not predict any of the assessment scores, neither did the interaction of real and assumed source or the interaction of all three variables.

5.3. Discussion

In Study 2 we aimed at investigating the ability of more experienced and proficient teachers to correctly identify AI-generated texts among student-written texts. Furthermore, we looked into their assessment of the texts and the factors that may influence it. Overall, we found that text quality played an important role for the teachers' judgments. Hence, the teachers were unable to identify the source of the texts correctly, especially with respect to low-quality texts. This finding indicates that teachers had a general idea of what a typical AI-generated text might look like compared to a student text. However, they were unaware that generative AI could also produce deficient texts and imitate student writing at lower proficiency levels.

They overestimated their own ability to identify the correct sources, particularly when they thought a text was written by a student and even more so when the presumed student text was of low quality.

With respect to their assessment of the texts, the results showed that high-quality texts received even higher scores when they were AI-generated, whereas low-quality texts tended to receive higher scores when they were written by students. This could partly be a reflection of the actual text quality range being more widespread for AI-generated texts than for the randomly chosen student-written texts in the corpus. However, it still shows that generative AI is rewarded when assessing different aspects of text quality.

In Study 2, we were able to replicate some of the central findings from Study 1. More experienced and proficient teachers also have a hard time identifying AI-generated texts, even though the results were somewhat promising, suggesting that expertise may be a factor in distinguishing human writing from generative AI. Compared to Study 1, the results were more differentiated, especially with respect to text quality. The findings indicate that it is difficult for teachers to detect a text that was purposely prompted to be of low quality.

In Study 2, we changed not only the population (experienced vs.

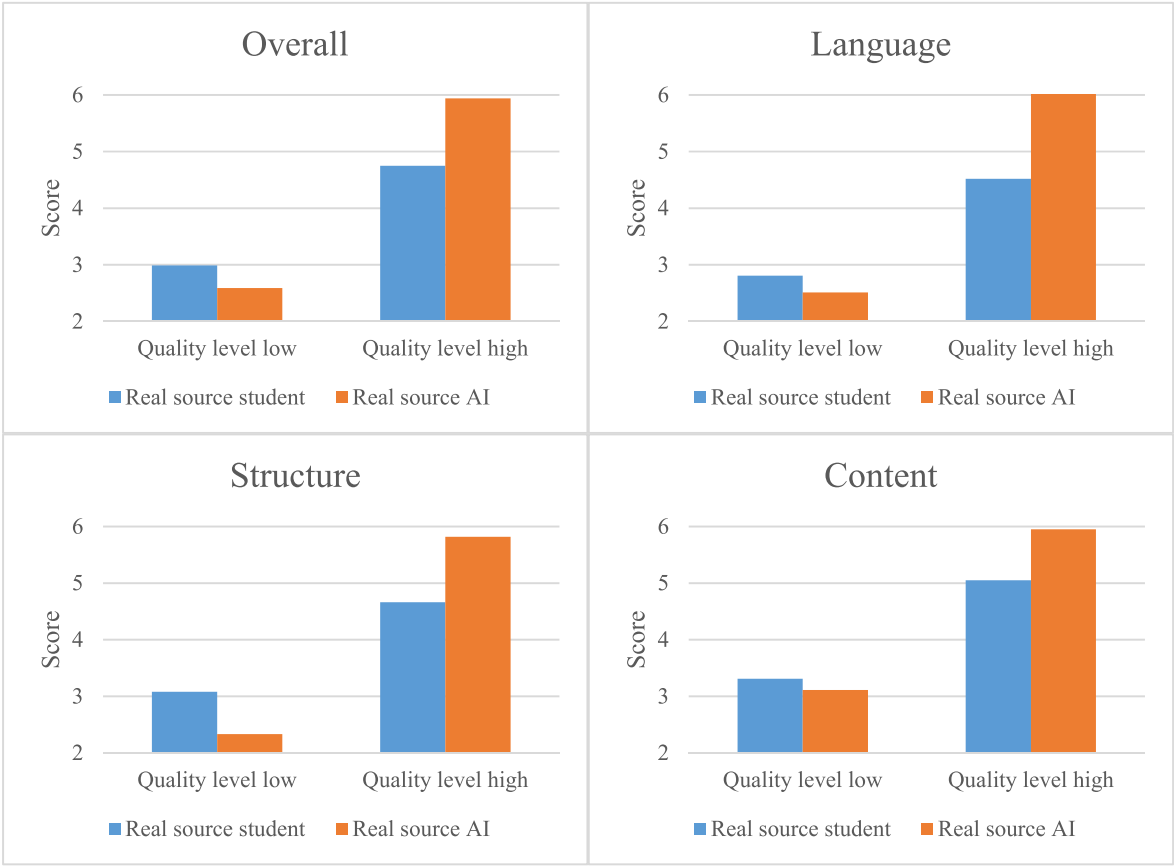


Fig. 3. Interaction effects of text quality * real source on text assessment.

preservice teachers) but also certain parts of the procedure like informing the participants of the percentage of AI-generated texts (open vs. 50:50) and the selection of student texts (random vs. preselected). This is a limitation of our study, as we do not know for sure what change has caused divergent findings. As for the disclosure of the distribution of the sources, we recommend that future research refrain from providing such information. First, as we saw in Study 1, participants do not always adhere to this prompt. Second, as we saw in Study 2, leaving out this information lead to a strong overestimation of the number of student-written texts. Thus, leaving out the information would provide a more realistic estimation of the actual ability to identify AI-generated texts.

6. General discussion

6.1. Summary

6.1.1. Source identification and confidence

The central goal of the study was to investigate to what extent pre-service and experienced teachers are capable to distinguish between AI-generated texts and student-written texts and how their assessment of the source relates to their text assessment. In both samples, the source identification showed to be far from perfect. Preservice teachers were unable to identify the source of the texts correctly, independent of whether they were student-written or AI-generated and independent of the text quality level. Experienced teachers were unable to correctly identify low-quality texts but more successful when it came to high-quality texts. This is at least partly due to the fact that they assigned most of the low-quality texts to being student-written. Thus, whereas novices have trouble identifying AI among student texts in general, more experienced teachers may simply not be aware of the possibility to generate texts of lower quality with AI. In conclusion, experienced teachers are more likely than pre-service teachers to recognize when good texts are produced by an AI. They are probably more capable of assessing students' writing skills and know the students' limitations when it comes to writing argumentative essays. They may also be aware of the typical texts produced by LLMs. However, experienced teachers are not superior when it comes to identifying the sources of weak texts. It is possible that they cannot imagine that computers can also be used to write weak texts including, for example, simple grammar or spelling mistakes. This result also underlines the need to familiarize teachers with the capabilities of AI.

This result is even more noteworthy as both samples are over-confident in their source identification, however, experienced teachers showed a rather appropriate confidence in their identification of student-written texts. Both groups were more confident when they assume texts to be written by students. This finding indicates a certain level of uncertainty when it comes to identifying AI-generated texts that holds true for both novice and experienced teachers.

6.1.2. Text quality assessment

Another goal of our study was to analyze the teachers' assessment of human-written and AI-generated texts. Whereas previous research showed that AI-generated texts were perceived as less well-written and less interesting than human-written texts (Gao et al., 2023; Graefe et al., 2018; Gunser et al., 2022), in both our studies, teachers did not assess AI-generated texts more positively than student-written texts. Experienced teachers even rewarded AI when text quality was high. The reason that our results diverge from prior research may be due to the deficiencies of learner texts, especially in a foreign language, compared to texts written by professional human writers (literary texts, scientific abstracts). Similar to the study of Gunser et al. (2022), the differential assessment of the two types of texts did not lead to better identifiability of the source. So whatever features of the texts may have caused the teachers to assess AI-generated texts more positively, this did not aid

their source identification.

6.2. Implications for educational research and practice

From an educational perspective, the finding that teachers cannot differentiate between student-written texts and texts generated by ChatGPT has several important implications for education. If teachers cannot reliably distinguish between human-generated and AI-generated content, it may pose challenges for assessing student work. So far, the AI detection software available does not seem to solve this issue either. For example, the AI classifier by OpenAI falsely categorizes many student texts as potentially AI-generated (Junge et al., 2023). Such findings support the claim that the new developments in the area of text-generative AI should and could not be met with more rigorous control mechanisms by teachers – technology-enhanced or not. Instead, we strongly recommend a meaningful exploration of strategies with and without the use of generative AI for assessment and learning in school.

Teachers need to carefully design their evaluation criteria to account for the presence of AI-generated materials. If AI-generated content is difficult to distinguish from student-generated content, it becomes easier for students to use AI to complete assignments or assessments without proper attribution. Educators may need to rethink their teaching and assessment strategies in light of the availability of AI-based tools. Whenever possible, instead of focusing on reproduction, educators might emphasize skills that AI cannot easily replicate (e.g., critical thinking, literature review). If the objective of an assessment demands it, teachers need to make sure that students do not have access to generative AI such as ChatGPT. Written assignments can be complemented by oral examinations in order to assess students' actual understanding of their own written work. Furthermore, there is a need for comprehensive education on academic integrity and responsible AI use to ensure that students are aware of the ethical implications of using AI tools in their coursework. Concerning technology integration in education, educators may need to incorporate AI literacy and critical thinking skills into the curriculum.

There are certain indications in our studies that high expertise facilitates the detection of AI-generated texts among student-written essays. More experienced and proficient teachers seem to be better able to identify typical AI-generated texts of high-quality. This suggests that further training could be a viable option to improve teachers' awareness of the use of LLMs in student writing. Continued education on the topic of generative AI in school should inform teachers of the possibilities that LLMs offer for cheating and deception (e. g., generating texts of lower quality and incorporating typical student errors). Teachers should be trained to recognize AI-generated content, understand its capabilities, and develop strategies for leveraging AI as a teaching tool. Teachers should also be educated on the ethical use of AI in the classroom, including the potential consequences of AI-generated content on student learning and development.

6.3. Conclusion

In summary, the finding that teachers cannot differentiate between student-written texts and AI-generated texts underscores the need for a thoughtful and ethical integration of AI in education. It calls for a reevaluation of assessment practices, increased awareness of AI's capabilities and limitations, and a focus on student skills that AI cannot easily replace. Additionally, it highlights the importance of ongoing research and development in AI and education to ensure that these technologies benefit students and educators while upholding educational standards and integrity.

Statements on open data and transparency

This study was not preregistered. All data and syntax are available from the first author upon request. The data collection and processing adhered to the EU General Data Protection Regulation (GDPR). Participants gave their consent before participating in the study. They were informed of their right to withdraw their data at any time.

CRediT authorship contribution statement

Johanna Fleckenstein: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Jennifer Meyer:** Conceptualization, Writing – review & editing. **Thorben Jansen:** Conceptualization, Writing – review & editing. **Stefan D. Keller:** Resources, Writing – review & editing. **Olaf Köller:** Conceptualization, Resources, Writing – review & editing. **Jens Möller:** Conceptualization, Investigation, Resources, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-4 from OpenAI in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of the publication.

Declaration of competing interest

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

References

Cotton, D., Cotton, P., & Shipway, J. R. (2023). Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT. *EdArXiv*. https://doi.org/10.1080/14703297.2023.2190148

Educational Testing Service. (2022). *TOEFL iBT® independent writing rubric*. https://www.ets.org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*. https://doi.org/10.1101/2022.12.23.521610

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism, 19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D. C., & Gerjets, P. (2022). The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors?. In *Proceedings of the first workshop on intelligent and interactive writing Assistants* (pp. 60–61). Dublin, Ireland: In2Writing 2022.

Jansen, T., Meyer, J., Fleckenstein, J., Horbach, A., Keller, S., & Möller, J. (2024). Individualizing goal-setting interventions using automated writing evaluation to support secondary school students' text revisions. *Learning and Instruction, 89*, Article 101847. https://doi.org/10.1016/j.learninstruc.2023.101847

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*, 1–11. https://doi.org/10.1016/j.tate.2020.103216, 103216.

Junge, F., Lohmann, L., Fleckenstein, J., & Möller, J. (2023). Can AI classifiers identify texts as AI-generated or student texts?. In *Presentation at the biannual meeting of the Fachgruppe educational Psychology, Kiel, Germany*.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., & Seidel, T. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *EdArXiv*. https://doi.org/10.35542/osf.io/5er8f

Keller, S., Fleckenstein, J., Krüger, M., Köller, O., & Rupp, A. A. (2020). English writing skills of students in upper secondary education. Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing, 48*, Article 100700. https://doi.org/10.1016/j.jslw.2019.100700

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, Article 106553.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199.

Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., & Reble, R. (2022). Judgment accuracy of German student texts: Do teacher experience and content

knowledge matter? *Teaching and Teacher Education, 119*, Article 103879. https://doi.org/10.1016/j.tate.2022.103879

Rupp, A. A., Casabianca, J., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. In *TOEFL research Report TOEFL-RR-86 and ETS research Report; No. RR-19-12*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12249.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction.* , Article 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Yalçın, Ö. N., Abukhodair, N., & DiPaola, S. (2020). Empathic AI painter: A computational creativity system with embodied conversational interaction. In *Proceedings of the NeurIPS 2019 Competition and demonstration track, Vancouver: CA* (Vol. 123, pp. 131–141).

Zacharakis, A., Kaliakatsos-Papakostas, M., Kalaitzidou, S., & Cambouropoulos, E. (2021). *Evaluating human-computer Co-creative processes in music: A case study on the*.

Zesch, T., & Horbach, A. (2018). ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Japan* (pp. 2310–2316).