Delft University of Technology

ChatGPT in Education: Empowering Educators through Methods for Recognition and Assessment

de Winter, J.C.F.; Dodou, D.; Stienen, A.H.A.

Article

ChatGPT in Education: Empowering Educators through Methods for Recognition and Assessment

Joost C. F. de Winter [1,*] [iD], **Dimitra Dodou** [2] [iD] and **Arno H. A. Stienen** [2]

1   Department of Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
2   Department of Biomechanical Engineering, Faculty of Mechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands; d.dodou@tudelft.nl (D.D.); a.h.a.stienen@tudelft.nl (A.H.A.S.)
*   Correspondence: j.c.f.dewinter@tudelft.nl

Abstract: ChatGPT is widely used among students, a situation that challenges educators. The current paper presents two strategies that do not push educators into a defensive role but can empower them. Firstly, we show, based on statistical analysis, that ChatGPT use can be recognized from certain keywords such as 'delves' and 'crucial'. This insight allows educators to detect ChatGPT-assisted work more effectively. Secondly, we illustrate that ChatGPT can be used to assess texts written by students. The latter topic was presented in two interactive workshops provided to educators and educational specialists. The results of the workshops, where prompts were tested live, indicated that ChatGPT, provided a targeted prompt is used, is good at recognizing errors in texts but not consistent in grading. Ethical and copyright concerns were raised as well in the workshops. In conclusion, the methods presented in this paper may help fortify the teaching methods of educators. The computer scripts that we used for live prompting are available and enable educators to give similar workshops.

Keywords: chatbot; student assessment; education

1. Introduction

Over the last year, considerable discussion has emerged surrounding ChatGPT, the chat interface that was launched in November 2022. The discussion within educational circles has been on the possibility for students to misuse ChatGPT in coursework [1,2]. Here, a concern arises from the fact that ChatGPT can complete writing assignments [3,4] or programming tasks [5,6] at a level that is sufficient or even above average. Therefore, the use of ChatGPT by students may jeopardize the attainment of learning goals [7,8].

Considering that ChatGPT is widely used among students and researchers (e.g., [9–12]), a question that arises is how educators might have to modify their course content, learning outcomes, and assessment methods to render them 'ChatGPT-proof' (for suggestions, see [13,14]). However, this prevailing attitude places the locus of control outside the educators.

An alternative perspective is one that assumes an empowering role for educators. Possible applications include the use of ChatGPT for generating lecture topics, demonstrations, exam questions, homework assignments, ideas, and product designs [15–23]. These approaches would encourage students to collaborate with AI, something that may equip them better for their future societal roles [24]. In this paper, two alternative ways of empowering educators are introduced, namely keyword detection and using ChatGPT for assessing student texts.

2. Empowering Educators: Recognizing Possible ChatGPT Use

It is challenging for educators to determine whether a student has used ChatGPT [25,26]. However, ChatGPT use can be detected, at least at a group level. More specifically, we discovered that ChatGPT tends to use certain words with striking frequency. This is illustrated with the following prompt:

Prompt:

Please rewrite the following concisely, for academic work; try this 3 times.

*The use of our method was very important. It enabled new applications and generated a better understanding.*

GPT-3.5 (API, temperature setting of 0):

Our method was <u>crucial</u> in <u>facilitating</u> new applications and <u>enhancing</u> comprehension.

The <u>utilization</u> of our method played a <u>pivotal</u> role in enabling novel applications and <u>fostering</u> a deeper understanding.

The <u>significance</u> of our method cannot be overstated, as it facilitated the emergence of new <u>applications</u> and contributed to a more <u>comprehensive</u> understanding.

Here, target words have been underlined. An examination of the Delft University of Technology (TU Delft, the Netherlands) repository indicated frequent use of these terms by students in their MSc thesis abstracts in 2023. For example, as of 23 September 2023, the term 'fostering' was used 65 times in 2023, whereas in 2021, it appeared only 10 times (see Figure 1, for an overview).
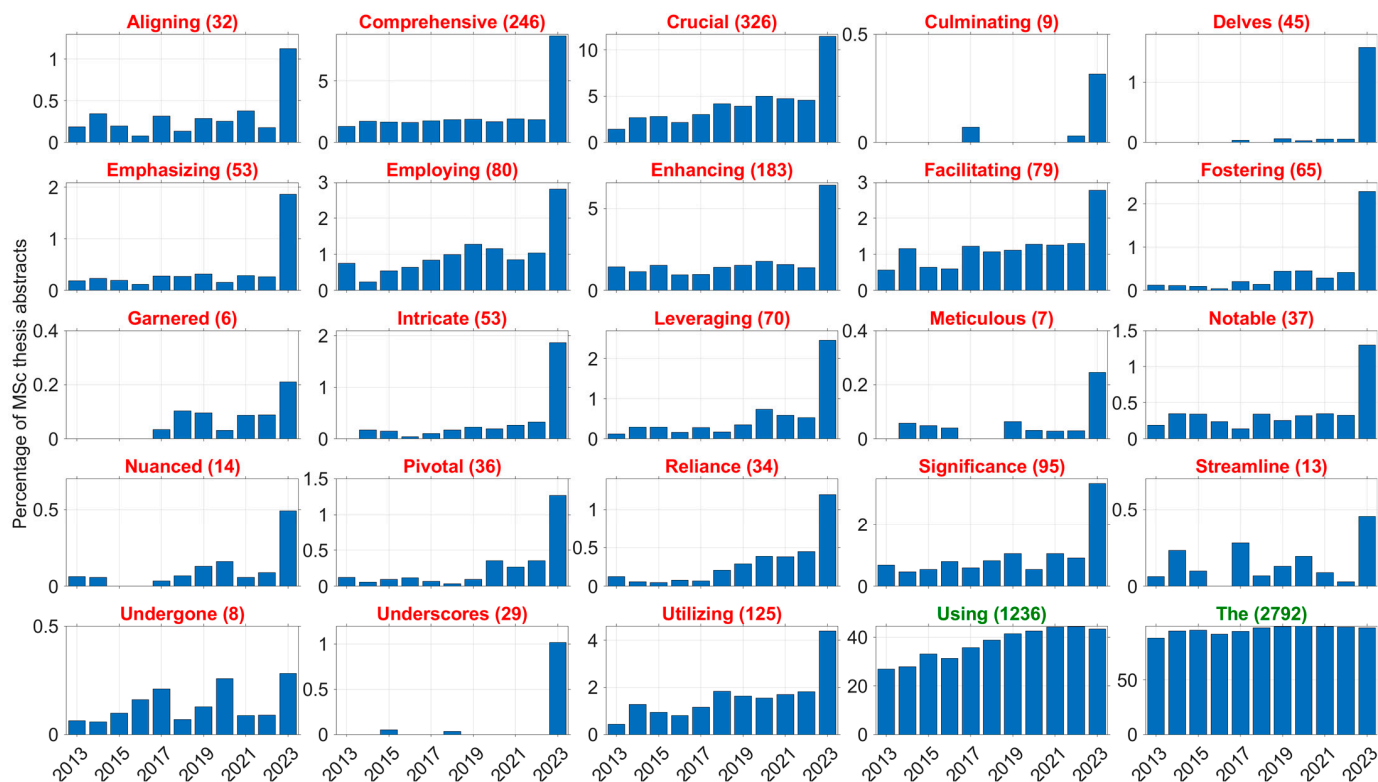
Figure 1. Prevalence of target words (red) and control words (green) in the metadata of MSc theses in the TU Delft education repository. The number in parentheses represents the total number of abstracts in 2023 (source: https://repository.tudelft.nl/islandora/search/?collection=education; accessed on 23 September 2023).

Table 1 shows that the prevalence of target words is not just a phenomenon at TU Delft but also amongst academic publishers and other universities. For example, in IEEE Xplore, the use of the word 'delves' increased by 169% from 2021 to 2023, while 'enhancing' in combination with 'crucial' increased by 155% in the same period. In addition to the results shown in Table 1, we found similar trends for other publishers such as Taylor & Francis, PLOS, as well as other student thesis repositories (Croatian and Swedish universities).

Table 1. Number of records with target words ('delve', 'enhancing' in combination with 'crucial') as well as a control phrase ('room temperature') based on a full-text search in selected academic publishing repositories (ScienceDirect, SpringerLink, IEEE Xplore) and Dutch Master thesis repositories (TU Delft, Leiden University, University of Twente).

	Number of Records					Percentage of Total Records					2021 to 2023
	2019	**2020**	**2021**	**2022**	**2023**	**2019**	**2020**	**2021**	**2022**	**2023**	**Difference**
ScienceDirect (all records)											
delve	3009	3565	4408	5079	9290	0.38%	0.44%	0.51%	0.57%	1.15%	123%
enhancing AND crucial	73,181	86,296	105,811	122,667	151,861	9.3%	10.6%	12.4%	13.7%	18.7%	52%
"room temperature"	135,495	144,032	153,344	164,241	153,716	17.3%	17.8%	17.9%	18.4%	19.0%	6%
Total records	78,2741	811,097	855,974	893,214	810,323						
SpringerLink (articles and conference papers)											
delve	1922	2229	2757	3459	5624	0.46%	0.50%	0.57%	0.69%	1.22%	115%
enhancing AND crucial	30,925	36,054	43,938	51,000	66,857	7.4%	8.1%	9.0%	10.1%	14.5%	61%
"room temperature"	49,169	50,886	52,334	54,227	50,778	11.8%	11.4%	10.8%	10.8%	11.0%	2%
Total records	417,195	445,504	486,641	502,535	460,972						
IEEE Xplore (all records)											
delve	1178	1215	1726	2186	3125	0.40%	0.45%	0.56%	0.67%	1.51%	169%
enhancing AND crucial	12,688	13,892	17,587	24,367	30,155	4.3%	5.1%	5.7%	7.5%	14.6%	155%
"room temperature"	8861	7735	8489	8641	6082	3.0%	2.9%	2.8%	2.7%	2.9%	6%
Total records	296,317	271,174	307,734	324,713	207,026						
Delft University of Technology (Master theses)											
delve OR delves OR delved OR delving	157	220	222	299	691	5.0%	7.1%	6.5%	8.9%	23.6%	265%
enhancing AND crucial	455	527	556	507	898	14.6%	16.9%	16.2%	15.0%	30.7%	89%
"room temperature"	314	279	291	234	178	10.1%	9.0%	8.5%	6.9%	6.1%	−28%
Total records	3124	3115	3425	3374	2923						
Leiden University (Master theses)											
delve	243	241	227	225	236	10.4%	10.5%	10.6%	10.5%	14.3%	35%
enhancing AND crucial	580	605	515	483	475	24.9%	26.3%	24.1%	22.5%	28.8%	19%
"room temperature"	17	25	18	21	11	0.7%	1.1%	0.8%	1.0%	0.7%	−21%
Total records	2333	2304	2137	2150	1650						
University of Twente (Master theses)											
delve OR delves OR delved OR delving	65	75	83	80	205	5.4%	5.9%	6.0%	6.3%	18.7%	210%
enhancing AND crucial	151	180	192	172	267	12.4%	14.3%	13.9%	13.6%	24.3%	74%
"room temperature"	43	32	52	28	24	3.5%	2.5%	3.8%	2.2%	2.2%	−42%
Total records	1214	1262	1377	1263	1098						

Note. Search date: 1 October 2023. Search areas: Full text, including abstract and references. The search functions are "smart", meaning that they capture derivatives of the search term, except for the Delft University of Technology and the University of Twente, which do not capture derivatives. The extent to which derivatives are captured varies per database.

Note that these patterns occur at a group level, and the use of the keywords shown in Figure 1 does not prove use of ChatGPT at an individual student level. Still, the present results could empower educators at universities. By being vigilant at detecting AI use, educators could ensure that academic integrity is maintained. For example, if ChatGPT use is suspected (not necessarily proven), follow-up questions could be asked and the author could be encouraged to engage in more original thinking and writing.

3. Empowering Educators: Using ChatGPT for Assessing Student Works

A second approach that empowers educators is to use ChatGPT to assess student work. Similar ideas have been proposed before. Examples include the use of ChatGPT for grading essays [27,28] and programming assignments [29,30]. Nguyen et al. [31] evaluated ChatGPT's ability to assess answers provided in a learning game called Decimal Point. They found that GPT-4 accurately assessed 75% of the students' answers. Nysom [32] introduced a pipeline where ChatGPT generates feedback that is approved by the lecturer before being released to the student. De Winter [33] demonstrated that ChatGPT is able to evaluate scientific abstracts. The scores produced by ChatGPT were found to be moderately correlated with citation counts and altmetrics. However, the scores were also found to be sensitive to the order in which the evaluation items were entered into the prompt. To resolve this, De Winter [33] used a self-consistency prompting method, where the items were repeatedly presented in a different order and the scores were averaged.

The concept of ChatGPT-generated assessment raises questions about fairness, accountability, and transparency [34]. Concerns may arise when students do not know how feedback has been generated, or in scenarios where the role of educators becomes marginal-

ized [35]. To address these questions, it is important to first explore the practicality of using ChatGPT for student assessment and have educators experiment with this idea.

We conducted two workshops at TU Delft educational events, organized at the end of August 2023 (right before the start of the academic year) and in early November 2023 (right before the second quarter). The two workshops were attended by educational experts and educators, including lecturers, program coordinators, directors, educational and assessment advisors, and educationists, in groups of 60 and 36, for Workshop 1 and 2, respectively. Workshop 1 was organized only for the staff from the Faculty of Mechanical Engineering, whereas Workshop 2 was meant for staff from all faculties of TU Delft. At the start of each workshop, we asked the audience to raise their hand if they had used ChatGPT before. About half of the participants and nearly all participants raised their hand in Workshops 1 and 2, respectively. In Workshop 2, there was a stronger representation of educationists as compared to Workshop 1, which could explain the familiarity of the audience with ChatGPT.

3.1. Workshop Introduction

The workshops commenced with a 30 min presentation. In our presentation, we reviewed the international adoption of ChatGPT. Our message was that ChatGPT has rapidly gained in popularity, owing to its user-friendly web interface.

We explained that the GPT-4 update, launched in March 2023, scores highly on benchmarks, often outperforming the average student [36–39]. This has led to high expectations and concerns about ChatGPT and AI in general. The first half of 2023 was marked by various events, including interviews with OpenAI's research team with a futuristic tone [40], an appearance by OpenAI's head in the U.S. Senate [41], a ban on ChatGPT in Italy [42], and an open letter from scientists and entrepreneurs calling for a six-month halt of the development of models more powerful than GPT-4 [43].

Our presentation also outlined the general working mechanisms of ChatGPT. Using two slides from Karpathy [44], we explained that ChatGPT is a model trained to predict tokens on a token-by-token basis. In doing so, the model demonstrates an understanding of linguistic structures, but lacks the ability to perform computations. For example, while ChatGPT can describe how to solve a Tower of Hanoi puzzle, or generate computer code to do so, it cannot execute code and solve this puzzle itself [45]. Furthermore, ChatGPT is not designed to engage in forward planning. It is possible, for example, that ChatGPT generates an answer, and upon a subsequent prompt, "Is this answer correct?", replies, "No, this answer is not correct". During token generation, the statistically most probable token is selected, which may result in a diverging or hallucinatory output [45,46].

We elucidated techniques to circumvent ChatGPT's limitations. One technique involves offering examples and encouraging sequential reasoning (multiple-shot prompting, chain-of-thought prompting; [47,48]). Another technique is by connecting ChatGPT with calculators. We provided examples where GPT-3.5 and GPT-4 were able to explain engineering problems in rigid-body dynamics, although they did make errors. We also demonstrated that in algorithmic computations, such as basic matrix multiplication, ChatGPT-4 makes various mistakes. Yet, when connected to the Wolfram plugin [49], matrix multiplication was executed flawlessly.

We explained that ChatGPT can be seen as a linguistic interface capable of translating prompts into computational code and subsequently presenting the code's output in accessible textual format to the user [50]. Another exemplification of interfacing with computers is a talking chess robot at TU Delft. While ChatGPT performs poorly at chess, it can translate evaluations of chess positions provided by a chess engine into a coherent narrative.

3.2. Workshop 1—Findings

Participants were given a handout on A4 paper containing a 460-word summary of a randomly selected MSc thesis from the TU Delft repository. The student's text is kept confidential in this paper. Participants were asked to skim through this summary. Additionally,

they were given an overview of the workshop objectives: to assess the feasibility of using ChatGPT in evaluating student work and to gain insight into its limitations. We mentioned that evaluating student work using ChatGPT involves ethical considerations, which could be touched upon in the upcoming workshop.

Next, participants were asked to create a prompt for evaluating the student's summary, under a hypothetical scenario wherein the student would submit this work to a supervisor. Participants were asked to email their prompts to the email address of the host. Incoming emails in Microsoft Outlook were automatically read, and the prompts were displayed on the screen. Participants had 7 min to send in a prompt. The received prompts are shown in Box 1.

Box 1. Prompts received from the participants in Workshop 1.

> 1. *What is the core message of this summary?*
> 2. *Is this summary clear and brief? Does it capture the research questions and performed research?*
> 3. *What is the contribution of the thesis?*
> 4. *How would you grade this thesis on a scale from 1 to 10?*
> 5. *Please evaluate whether this thesis meets basic research standards: introduction of problem, relevant literature, methodology, results explained and conclusions*
> 6. *Is this an original research?*
> 7. *Does this essay meet the learning objectives of course*
> 8. *Did chatgtp write this?*
> 9. *Has this summary been generated by chatgpt?*
> 10. *What are the outcomes of all project phases*
> 11. *Give the definition of a good summary and look how many requirements the student checks, based on that percentage give a grade and feedback.*
> 12. *Is this innovative work?*
> 13. *How well does this summary reflect the content of the thesis?...*
> 14. *is this abstract relevant for the firld.*
> 15. *Does the flow of the thesis make sense?*
> 16. *Can you write the summary in less words?*
> 17. *Is the goal and conclusion clear and do they align? Please also add a grade from 1 to 10, based on the Dutch grading scheme.*
> 18. *Is the structure of the summary suitable for an MSc thesis?*
> 19. *Grade the following abstract of a Mechanical Engineering Master thesis, using a percentage grade. For each criterion, list the grade and one sentence of explanation.*
>
>   a. *Theoretical knowledge*
>   b. *New design*
>   c. *Creativity and skills*
>   d. *Research significance*
>   e. *Structure*
>   f. *Level of English*

About a third of the attendees sent in a prompt. Possible explanations could be ambiguity surrounding the use of their prompts, limited time available to generate a prompt, and working in groups in order to send a joint prompt.

One of the workshop organizers read prompts aloud. From the list of incoming prompts and the audience's feedback, several prompts were selected and submitted live to the API (model gpt-4-0613) [51].

The audience was curious about ChatGPT's ability to provide a numerical evaluation. Thus, Prompt 4 was chosen. This prompt sparked intrigue regarding whether ChatGPT would assign a grade to the student's summary, especially considering that ChatGPT was provided with no information about the learning objectives:

> ***Prompt:***
>
> *This is a summary text written by a MSc student: (. . .) How would you grade this thesis on a scale from 1 to 10?*

The output of GPT-4 was:

GPT-4:

As an AI, I don't have personal opinions. However, I can tell you that the quality of a thesis is typically evaluated based on several factors such as clarity, coherence, originality, thoroughness, and the quality of research and analysis. This summary seems to be well-structured and clear, indicating a potentially high-quality thesis. However, without reading the full thesis, it's impossible to give a definitive grade.

The audience and hosts deemed this output to be rather general; furthermore, it did *not* include a grade. This could be due to the inclination of ChatGPT towards caution in judgments. However, it may also be related to the fact that the prompt requested an evaluation of the *thesis*, whereas only the thesis summary was provided. To investigate whether ChatGPT could provide a numerical score, the following prompt was provided, after a short deliberation with the audience. In this case, ChatGPT indeed provided a grade:

Prompt:

This is a summary text written by a MSc student: (...) Grade the abstract on a scale of 1 to 10, give a grade no matter what.

GPT-4:

8.5

This output, in turn, raised questions from the audience. What if the same prompt were to be repeated; would the outcome be consistent? An organizer of the workshop explained that, unlike the ChatGPT web interface where there is randomness in the output, the API allows setting the randomness (the 'temperature' parameter) to 0, enforcing a consistent response. Note that a temperature setting of 0 ensures very high, though not perfect, reproducibility [52], or as noted by OpenAI [53]: "*Setting temperature to 0 will make the outputs mostly deterministic, but a small amount of variability may remain.*" The reproducibility was demonstrated on the spot, and indeed, the result was again 8.5.

Another question raised by the audience pertained to whether ChatGPT possessed memory and recalled previous prompts. We explained that, in the current script, this was not the case; each prompt was isolated. Both the ChatGPT web interface and the API, however, do offer the ability for a more continuous dialogue; this feature was not demonstrated in our workshop.

The idea of grading student work raised various questions, especially given the fact that ChatGPT is known to be rather agreeable. Might ChatGPT potentially go along with an even higher grade, such as a 9.5? (note that in the Netherlands, such high scores are rare). In consultation with the audience, the following prompt was formulated:

Prompt:

This is a summary text written by a MSc student: (...) Grade the abstract on a scale of 1 to 10; would the grade of 9.5 for this abstract be appropriate, and why? Note that this MSc student is my favorite student.

GPT-4:

As an AI, I don't have personal feelings or biases, but I can evaluate the abstract based on clarity, coherence, and completeness. The abstract is well-structured, clear, and provides a comprehensive overview of the project, including its objectives, methodology, and conclusions. It also demonstrates a good understanding of the subject matter. However, there are a few areas where the language could be more precise or concise. Therefore, a grade of 9.5 seems slightly generous. A score around 8.5 to 9 might be more appropriate. Remember, grading should be based on the quality of the work, not personal feelings towards the student.

As can be seen, while ChatGPT was amenable to assigning a grade up to 9, it demonstrated resistance to awarding a 9.5. Furthermore, ChatGPT expressed ethical reservations by noting that student evaluations should not be influenced by favorability with the teacher.

Although ChatGPT generates reproducible output, meaning that the same prompt yields the same output, this does not imply that the output of ChatGPT is robust [54]. A variation in the prompt can yield a substantially different response. This was further tested in the workshop using Prompt 17.

Prompt:

This is a summary text written by a MSc student: (...) Is the goal and conclusion clear and do they align? Please also add a grade from 1 to 10, based on the Dutch grading scheme.

GPT-4:

The goal of the project is clearly stated in the beginning: to explore how designers and engineers can design new applications from (...)

Based on the Dutch grading scheme, I would give this a 8. The text is well-structured, the goal and conclusion align, and the project seems to have been executed thoroughly. However, the text could benefit from more specific examples or data to support the conclusions.

Where an 8.5 was initially awarded, an 8.0 was now assigned. Furthermore, with the prompts provided so far, ChatGPT primarily responded in general terms and did not offer specific suggestions that the student could readily use to improve the work. Next, we also sought to rely on the 'wisdom of the crowd' by asking ChatGPT to choose the best prompt from the ones submitted:

Prompt:

These are a number of prompts that are supposed to be used by ChatGPT for judging a summary text written by a MSc student: (...) Please provide the best prompt among the listed prompts, which would yield me the best feedback for the student.

GPT-4:

The best prompt among the listed prompts, which would yield the best feedback for the student, would be:

Prompt number 11: Give the definition of a good summary and look how many requirements the student checks, based on that percentage give a grade and feedback.

This prompt is comprehensive and allows for a detailed evaluation of the student's work. It not only asks for a grade but also for feedback, which is crucial for the student's learning and improvement. It also encourages the evaluator to consider the definition of a good summary, ensuring that the evaluation is based on clear criteria.

ChatGPT selected this prompt as the most suitable, and the workshop organizers concurred since it adhered to the principle of 'chain-of-thought prompting' [48]: the prompt asked ChatGPT to first generate criteria before providing an evaluation based on these criteria.

Prompt:

This is a summary text written by a MSc student: (...) Give the definition of a good summary and look how many requirements the student checks, based on that percentage give a grade and feedback.

GPT-4:

A good summary should be concise, clear, and accurately represent the main points and conclusions of the original text. It should be written in a way that someone who has not read the original text can understand the main ideas and findings. It should also be objective and free from personal interpretation or opinion.

Based on these criteria, the student's summary checks most of the requirements. It is clear, concise, and accurately represents the main points and conclusions of the project. It provides a good overview of the project's aims, methods, results, and conclusions. However, it could be improved by providing a bit more context about why the project was undertaken and why it is important.

Grade: 85%

Feedback: This is a well-written summary that clearly and concisely presents the main points and conclusions of your project. However, it would be helpful to provide a bit more context about why the project was undertaken and why it is important. This would help readers who are not familiar with the topic to better understand the significance of your work.

This output was again highly general: while the feedback was credible, it was judged to be too broad for a student to act upon. Thus, the feedback could be a manifestation of the Forer effect [55] in the sense that similar feedback might equally apply to other abstracts. Additionally, the criterion "*It should also be objective and free from personal interpretation or opinion*" was not explicitly addressed in the feedback. We also attempted several other prompts in consultation with the audience, such as: "*Is this an original research?*". This again yielded fairly general and agreeable, but not necessarily incorrect, output.

Based on the feedback from the audience, we demonstrated that it is possible to guide ChatGPT or force a particular type of answer. Through the following prompt, we aimed to ask ChatGPT to explain what was not 'objective and free from personal interpretation':

Prompt:

At which points in this summary is the Msc student overly subjective or gives too many of his/her personal interpretations?

This was followed by a prompt prepared by a workshop organizer prior to the workshop. It was grounded in previous research wherein we converged on the use of highlights to compel ChatGPT to make specific points [56].

Prompt:

List 5 grammar mistakes and 5 content mistakes, in the form of bulleted highlights, max 85 characters each.

The outputs for the above two prompts provided more specific insights insofar as they highlighted sentences in the student's summary that could benefit from revision. The student's text did not contain any spelling errors; hence, ChatGPT's suggestions were of a stylistic nature, though still deemed valuable. Furthermore, ChatGPT offered points in the student's abstract that could be improved upon in terms of objective reporting and argumentation.

In conclusion, with targeted prompting, ChatGPT was found to be capable of generating useful suggestions. The impression from the workshop was that these suggestions would still require human verification before being suitable for grading or providing feedback.

We ended the first workshop with a different consideration: Could ChatGPT be suitable for formulating questions, for example, questions that could be posed in an MSc thesis examination? The output generated by ChatGPT for this purpose was deemed meaningful, albeit general.

Prompt:

Could you generate 10 questions which I could ask this MSc student on his Msc defence?; 5 questions about the summary, and 5 that are outside of the immediate scope but related.

GPT-4:

(...) 2. How did you determine the design criteria for the functional design? (...) 5. How could the design processes you used be improved or optimized for future projects?

3.3. Workshop 2—Findings

Instead of using a summary of a MSc thesis as in Workshop 1, we used a 590-word summary of a recently submitted paper by two of the current authors. For the purposes of the workshop, we introduced nine errors in the summary:

(1)   Three grammatical errors: '*weather*' instead of '*whether*'; '*than*' instead of '*then*'; and '*results suggests*' instead of '*results suggest*',

(2)   Three errors of internal consistency: '*200 images*' vs. '*210 images*'; '*−0.86*' in the text vs. '*−0.53*' in the table; and a correlation of 0.69 being characterized as '*strong*' vs. '*weak-to-moderate*',

(3)   Three content-related errors: '*YOLO (You Only Live Once)*', while in the context of the abstract, '*YOLO*' referred to the object detection algorithm '*You Only Look Once*' [57]; '*Spearman product-moment correlation*' instead of '*Pearson product-moment correlation*'; and '*GPT-4V output adds predictive value by incorporating context, something traditional computer vision methods do incorporate*', which needs the addition of '*not*' before '*incorporate*'.

The participants received an A4 printout of the summary with the three types of errors color-coded. The summary contained a figure and a table. However, since the abstract was submitted as a text-only file to the GPT API (gpt-4-1106-preview), only the figure caption but not the figure itself was submitted; moreover, the table was submitted unformatted. Participants had about 7 min to send in a prompt. The 12 prompts received via email are shown in Box 2.

Box 2. Prompts received from the participants in Workshop 2. Note. *Five prompts were* sent by the same participant, who consulted in a group; these prompts are annotated with (*S*).

---

1.   *What would be a good argument to disagree with the conclusion of this paper*
2.   *Please give this student feedback on this document, find grammatical mistakes and inconsistencies.*
3.   *Is this made by chatgpt?*
4.   *Assess societal relevance of the research (S)*
5.   *Assess academic quality (S)*
6.   *Fix spelling errors (S)*
7.   *Provide 4 criteria for evaluating papers. Evaluate and provide feedback to improve on each criteria. The feedback should be actionable but not prescribe exactly what to do*
8.   *Are the information in the table valid and correct?*
9.   *Find internal inconsistencies in te text (S)*
10.   *Check the document for inconsistencies and spelling errors*
11.   *Point out inconsistencies in the numerical data (S)*
12.   *Please point out the inconsistencies, what are the spelling mistakes, and what are the potential factual mistakes in the text?*

From the 12 prompts, it can be seen that 7 asked for identifying errors. After consultation with the audience, Prompts 2 and 12 were submitted to the API. Note that Prompt 2 applied the aforementioned chain-of-thought prompting method.

Prompt 2 yielded an output of 475 words, while Prompt 11 yielded a 387-word output. Although the outputs were rather verbose, a characteristic that seems typical for the gpt-4-1106-preview model [58], they were interpreted as useful. Specifically, a total of six and five out of nine errors were correctly identified by these two prompts, alongside a number of stylistic suggestions.

Next, the following prompt was proposed for receiving more succinct feedback:

Prompt:

Check the summary for any types of errors: regarding factual information, consistency-related, grammar-related, etc. List in bullet form, max 100 characters each.

GPT-4:

- "weather" should be "whether" (grammar error).

- "mean ranks for the 200 images" should be "210 images" (factual inconsistency).

- "Spearman product-moment correlation" is typically "Spearman's rank correlation" (factual error).

- "YOLOv4" mentioned but earlier "pretrained YOLO" without version (consistency error).

- "how humans rate risk from images" should be "how humans rated risk from images" (grammar consistency).

- "methods do incorporate" should be "methods don't incorporate" (factual error).

- "real-time feedback systems, will become" should not have a comma (grammar error).

- "large-language models" should be "large language models" (consistency with earlier usage).

- "the prediction strength of human risk" is unclear, might mean "the prediction of human risk" (clarity issue).

The above output captured five of our nine deliberately introduced errors. One participant remarked that the feedback was overly specific, focusing on textual errors. It was argued that good students would have already corrected such errors themselves before submitting a text to a supervisor. Therefore, the following prompt was proposed to gain insight into the higher-level evaluation abilities of ChatGPT:

Prompt:

Give your overall opinion on the relevance and originality of this extended abstract? Is it good or not; why, or why not? Max 150 words.

GPT-4:

The extended abstract presents a novel and relevant application of GPT-4V's image-to-text capabilities to assess risk in traffic images, an area with significant implications for autonomous driving and road safety. The comparison of GPT-4V's performance with human risk assessment and traditional computer vision approaches is original and provides insight into the potential of language models in understanding complex visual contexts. The strong correlation found between GPT-4V's assessments and human judgments suggests that GPT-4V can capture nuanced risk factors. The study's acknowledgment of the limitations, such as the sensitivity to prompt variations and the need for faster processing times, indicates a well-considered approach. Overall, the abstract outlines a promising direction for future research and the development of real-time feedback systems in vehicular contexts, making it a valuable contribution to the field.

One of the three workshop hosts was able to directly assess the quality of the above output, because he was the first author of the summary being evaluated. He expressed that he found ChatGPT's assessment unhelpful because it only generally described what the work was about and that it was innovative. This feedback strongly resonated with the superficiality we identified in Workshop 1. However, this does not mean that ChatGPT responded to the prompt incorrectly, but rather that the prompt did not guide ChatGPT in a direction that was considered useful.

As the plenary session drew to a close, our discussion shifted to focus on the ethical aspects of using ChatGPT. We discussed whether it is fair for educators to use ChatGPT if students are prohibited from doing so. Furthermore, the future role of educators in a scenario where students might rely on ChatGPT for feedback was explored. Concerns about data storage, confidentiality, and copyright, particularly in situations where student work is being performed at a company, were also addressed.

4. Discussion

ChatGPT is widely being used within the student population. This paper provides insights into how ChatGPT need not be perceived defensively, but rather in a manner that empowers educators.

Our analysis has identified specific keywords that may indicate that a student has used ChatGPT for text generation. Educators may be able to use these signals to ask follow-up questions in order to discover whether the student has relied on ChatGPT for content creation or merely for editorial assistance. Such questions could be asked during the oral defense of a MSc thesis, where it is important to evaluate if the student has a deep understanding of the content. Note that these keywords appeared at a group level, and their occurrence in a single text does not prove ChatGPT use by an individual. For example, a term like 'significance' is naturally common in research that involves statistical tests for significance. Moreover, an author could always argue that their writing skills or vocabulary have improved as a result of learning from ChatGPT, rather than through direct copying of ChatGPT output.

Secondly, by means of two workshops, we obtained ideas and discussion of how ChatGPT can be used for evaluating a student's or academic's work. We showed that ChatGPT is not reliable for student grading, as different prompts could result in the grade changing. Earlier work in which ChatGPT was used to assess scientific abstracts showed that only through repeated prompting, and subsequently taking an average, could a statistically reliable estimate of, for example, writing quality be obtained [33]. At the same time, ChatGPT was found to be useful in detecting weaknesses, awkward writing, as well as content mistakes. The workshop illustrated the critical role of the prompt in determining the outcome produced by ChatGPT. ChatGPT appears to provide feedback in generic language. However, with tailored prompts, specific weaknesses and errors in the text could be pinpointed.

The workshop served an additional objective: to familiarize educators with the capabilities of ChatGPT, including its accompanying API. The workshop demonstrated how different teachers provide different prompts, and that these differences are consequential for the output generated by ChatGPT. Together with the reviewed literature and the results of the workshop, a consensus emerged that ChatGPT can provide valuable feedback and error correction of scientific work, but requires human oversight or intervention.

Caution might still be needed when a student's work is completed in cooperation with third parties. According to OpenAI, data submitted by users through the API are not used for training the models; for the web interface, submitted data may be used for this purpose, though the user has the option to disable or opt out of this [59]. Policies with respect to data confidentiality and permitted uses of ChatGPT may differ between organizations, with some companies having banned ChatGPT [60]. Also, as indicated in the workshops and from general impressions gathered in discussions before and after the workshops, educators see potential in using ChatGPT, but some remain skeptical. Concerns about the influence of big tech, the impact on CO_2 emissions, and the loss of traditional values associated with the academic system were frequently mentioned.

5. Conclusions

The current paper offers various ideas and perspectives that may give educators more control over their teaching, especially in terms of recognizing ChatGPT use in texts, and the deployment of ChatGPT in assessing student work. The associated dilemmas, both in terms of the quality of the ChatGPT output and the ethics involved, have been addressed in this paper. We hope this paper may enlighten educators who contemplate using ChatGPT in education.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MATLAB script used during the workshops, the data underlying Figure 1 and Table 1, an update of the same data based on a new search conducted on 25 November 2023, and data for publishers and repositories other than those reported in Figure 1 and Table 1 are available at https://doi.org/10.4121/a41d11f9-83a3-4f12-b59f-540a1e18dc1d.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**. [CrossRef]
2. Sullivan, M.; Kelly, A.; McLaughlan, P. ChatGPT in Higher Education: Considerations for Academic Integrity and Student Learning. *J. Appl. Learn. Teach.* **2023**, *6*, 1–10. [CrossRef]
3. Herbold, S.; Hautli-Janisz, A.; Heuer, U.; Kikteva, Z.; Trautsch, A. A Large-Scale Comparison of Human-Written Versus ChatGPT-Generated Essays. *Sci. Rep.* **2023**, *13*, 18617. [CrossRef] [PubMed]
4. Yeadon, W.; Inyang, O.-O.; Mizouri, A.; Peach, A.; Testrow, C.P. The Death of the Short-Form Physics Essay in the Coming AI Revolution. *Phys. Educ.* **2023**, *58*, 035027. [CrossRef]
5. Kiesler, N.; Schiffner, D. Large Language Models in Introductory Programming Education: ChatGPT's Performance and Implications for Assessments. *arXiv* **2023**, arXiv:2308.08572. [CrossRef]
6. Savelka, J.; Agarwal, A.; An, M.; Bogart, C.; Sakr, M. Thrilled by Your Progress! Large Language Models (GPT-4) No Longer Struggle to Pass Assessments in Higher Education Programming Courses. In Proceedings of the 19th ACM Conference on International Computing Education Research (ICER '23 V1), Chicago, IL, USA, 7–11 August 2023; pp. 78–92. [CrossRef]
7. Malik, A.; Khan, M.L.; Hussain, K. How is ChatGPT Transforming Academia? Examining Its Impact on Teaching, Research, Assessment, and Learning. *SSRN* **2023**. [CrossRef]
8. Schreiner, M. OpenAI Calls GPT-4 Turbo Its "Smartest" Model, but What Does That Even Mean. Available online: https://the-decoder.com/openai-calls-gpt-4-turbo-its-smartest-model-but-what-does-that-even-mean (accessed on 18 November 2023).
9. Cardon, P.; Fleischmann, C.; Aritz, J.; Logemann, M.; Heidewald, J. The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age. *Bus. Prof. Commun. Q.* **2023**, *86*, 257–295. [CrossRef]
10. Forman, N.; Udvaros, J.; Avornicului, M.S. ChatGPT: A New Study Tool Shaping the Future for High School Students. *Int. J. Adv. Nat. Sci. Eng. Res.* **2023**, *7*, 95–102. [CrossRef]
11. Ibrahim, H.; Liu, F.; Asim, R.; Battu, B.; Benabderrahmane, S.; Alhafni, B.; Adnan, W.; Alhanai, T.; AlShebli, B.; Baghdadi, R.; et al. Perception, Performance, and Detectability of Conversational Artificial Intelligence Across 32 University Courses. *Sci. Rep.* **2023**, *13*, 12187. [CrossRef]
12. Jishnu, D.; Srinivasan, M.; Dhanunjay, G.S.; Shamala, R. Unveiling Student Motivations: A Study of ChatGPT Usage in Education. *ShodhKosh J. Vis. Perform. Arts* **2023**, *4*, 65–73. [CrossRef]
13. Lau, S.; Guo, P.J. From "Ban It Till We Understand It" To "Resistance Is Futile": How University Programming Instructors Plan to Adapt As More Students Use AI Code Generation and Explanation Tools Such As ChatGPT and GitHub Copilot. In Proceedings of the 19th ACM Conference on International Computing Education Research (ICER '23 V1), Chicago, IL, USA, 7–11 August 2023; pp. 106–121. [CrossRef]
14. Rose, R. ChatGPT-Proof Your Course. In *ChatGPT in Higher Education. Artificial Intelligence and Its Pedagogical Value*; University of North Florida Digital Pressbooks: Jacksonville, FL, USA, 2023. Available online: https://unf.pressbooks.pub/chatgptinhighereducation/chapter/chatgpt-proof-your-course (accessed on 18 November 2023).
15. Atlas, S. *ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI*; College of Business Faculty Publications: Kingston, RI, USA, 2023; Available online: https://digitalcommons.uri.edu/cba_facpubs/548 (accessed on 18 November 2023).
16. Dos Santos, R.P. Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agents-To-Think-With: A Comparative Case Study. *arXiv* **2023**, arXiv:2306.00724. [CrossRef]
17. Filippi, S. Measuring the Impact of ChatGPT on Fostering Concept Generation in Innovative Product Design. *Electronics* **2023**, *12*, 3535. [CrossRef]
18. Girotra, K.; Meincke, L.; Terwiesch, C.; Ulrich, K.T. Ideas Are Dimes a Dozen: Large Language Models for Idea Generation in Innovation. *SSRN* **2023**, 4526071. [CrossRef]
19. Liu, J.; Liu, S. The Application of ChatGPT in Medical Education. *EdArXiv* **2023**. [CrossRef]
20. McNichols, H.; Feng, W.; Lee, J.; Scarlatos, A.; Smith, D.; Woodhead, S.; Lan, A. Exploring Automated Distractor and Feedback Generation for Math Multiple-Choice Questions Via In-Context Learning. *arXiv* **2023**, arXiv:2308.03234. [CrossRef]
21. Mollick, E.R.; Mollick, L. Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts. *SSRN* **2023**. [CrossRef]
22. Mondal, H.; Marndi, G.; Behera, J.K.; Mondal, S. ChatGPT for Teachers: Practical Examples for Utilizing Artificial Intelligence for Educational Purposes. *Indian J. Vasc. Endovasc. Surg.* **2023**, *10*, 200–205. [CrossRef]
23. Yang, Z.; Wang, Y.; Zhang, L. AI Becomes a Masterbrain Scientist. *bioRxiv* **2023**. [CrossRef]

24. Yu, H. Reflection on Whether Chat GPT Should Be Banned by Academia from the Perspective of Education and Teaching. *Front. Psychol.* **2023**, *14*, 1181712. [CrossRef] [PubMed]

25. Pegoraro, A.; Kumari, K.; Fereidooni, H.; Sadeghi, A.R. To ChatGPT, or Not to ChatGPT: That Is the Question! *arXiv* **2023**, arXiv:2304.01487. [CrossRef]

26. Waltzer, T.; Cox, R.L.; Heyman, G.D. Testing the Ability of Teachers and Students to Differentiate Between Essays Generated by ChatGPT and High School Students. *Hum. Behav. Emerg. Technol.* **2023**, *2023*, 1923981. [CrossRef]

27. Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; Chen, G. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, 10–13 July 2023; pp. 323–325. [CrossRef]

28. Mizumoto, A.; Eguchi, M. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [CrossRef]

29. Gao, R.; Merzdorf, H.E.; Anwar, S.; Hipwell, M.C.; Srinivasa, A. Automatic Assessment of Text-Based Responses in Post-Secondary Education: A Systematic Review. *arXiv* **2023**, arXiv:2308.16151. [CrossRef]

30. Nilsson, F.; Tuvstedt, J. GPT-4 as an Automatic Grader: The Accuracy of Grades Set by GPT-4 on Introductory Programming Assignments. Bachelor's Thesis, KTH, Stockholm, Sweden, 2023. Available online: https://www.diva-portal.org/smash/record.jsf?pid=diva2%253A1779778&dswid=-1020 (accessed on 18 November 2023).

31. Nguyen, H.A.; Stec, H.; Hou, X.; Di, S.; McLaren, B.M. Evaluating ChatGPT's Decimal Skills and Feedback Generation in a Digital Learning Game. In *Responsive and Sustainable Educational Futures. EC-TEL 2023*; Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 278–293. [CrossRef]

32. Nysom, L. AI Generated Feedback for Students' Assignment Submissions. A Case Study in Generating Feedback for Students' Submissions Using ChatGPT. Master's Thesis, University College of Northern Denmark, Aalborg, Denmark, 2023. Available online: https://projekter.aau.dk/projekter/files/547261577/Lars_Nysom_Master_Project.pdf (accessed on 18 November 2023).

33. De Winter, J.C.F. Can ChatGPT Be Used to Predict Citation Counts, Readership, and Social Media Interaction? An Exploration Among 2222 Scientific Abstracts. ResearchGate 2023. Available online: https://www.researchgate.net/publication/370132320_Can_ChatGPT_be_used_to_predict_citation_counts_readership_and_social_media_interaction_An_exploration_among_2222_scientific_abstracts (accessed on 18 November 2023).

34. European Commission. Ethical Guidelines on the Use of Artificial Intelligence (AI) and Data in Teaching and Learning for Educators. Available online: https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1/language-en (accessed on 18 November 2023).

35. Aithal, P.S.; Aithal, S. The Changing Role of Higher Education in the Era of AI-Based GPTs. *Int. J. Case Stud. Bus. IT Educ.* **2023**, *7*, 183–197. [CrossRef]

36. De Winter, J.C.F. Can ChatGPT Pass High School Exams on English Language Comprehension? *Int. J. Artif. Intell. Educ.* **2023**. [CrossRef]

37. Guerra, G.A.; Hofmann, H.; Sobhani, S.; Hofmann, G.; Gomez, D.; Soroudi, D.; Hopkins, B.S.; Dallas, J.; Pangal, D.; Cheok, S.; et al. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. *World Neurosurg.* 2023; *Online ahead of print*. [CrossRef]

38. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]

39. Zhai, X.; Nyaaba, M.; Ma, W. Can AI Outperform Humans on Cognitive-Demanding Tasks in Science? *SSRN* **2023**. [CrossRef]

40. Sutskever, I. Ilya Sutskever (OpenAI Chief Scientist)—Building AGI, Alignment, Spies, Microsoft, & Enlightenment. Available online: https://www.youtube.com/watch?v=Yf1o0TQzry8 (accessed on 18 November 2023).

41. U.S. Senate Committee on the Judiciary. Oversight of A.I.: Rules for Artificial Intelligence. Available online: https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence (accessed on 18 November 2023).

42. Kreitmeir, D.H.; Raschky, P.A. The Unintended Consequences of Censoring Digital Technology—Evidence from Italy's ChatGPT Ban. *arXiv* **2023**, arXiv:2304.09339. [CrossRef]

43. Future of Life. Pause Giant AI Experiments: An Open Letter. Available online: https://futureoflife.org/open-letter/pause-giant-ai-experiments (accessed on 18 November 2023).

44. Karpathy, A. State-of-GPT-2023. Available online: https://github.com/giachat/State-of-GPT-2023/tree/main (accessed on 18 November 2023).

45. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of Artificial General Intelligence: Early Experiments With GPT-4. *arXiv* **2023**, arXiv:2303.12712. [CrossRef]

46. Chuang, Y.S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; He, P. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *arXiv* **2023**, arXiv:2309.03883. [CrossRef]

47. Chen, J.; Chen, L.; Huang, H.; Zhou, T. When Do You Need Chain-of-Thought Prompting for ChatGPT? *arXiv* **2023**, arXiv:2304.03262. [CrossRef]

48. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Nice, France, 2022; pp. 24824–24837. [CrossRef]

49. Davis, E.; Aaronson, S. Testing GPT-4 with Wolfram Alpha and Code Interpreter Plug-ins on Math and Science Problems. *arXiv* **2023**, arXiv:2308.05713. [CrossRef]

50. Lubiana, T.; Lopes, R.; Medeiros, P.; Silva, J.C.; Goncalves, A.N.A.; Maracaja-Coutinho, V.; Nakaya, H.I. Ten Quick Tips for Harnessing the Power of ChatGPT in Computational Biology. *PLoS Comput. Biol.* **2023**, *19*, e1011319. [CrossRef] [PubMed]
51. OpenAI. API Reference. Available online: https://platform.openai.com/docs/api-reference (accessed on 18 November 2023).
52. Ouyang, S.; Zhang, J.M.; Harman, M.; Wang, M. LLM Is Like a Box of Chocolates: The Non-Determinism of ChatGPT in Code Generation. *arXiv* **2023**, arXiv:2308.02828. [CrossRef]
53. OpenAI. Models. Available online: https://platform.openai.com/docs/models (accessed on 18 November 2023).
54. Tabone, W.; De Winter, J. Using ChatGPT for Human–Computer Interaction Research: A Primer. *R. Soc. Open Sci.* **2023**, *10*, 231053. [CrossRef]
55. Forer, B.R. The Fallacy of Personal Validation: A Classroom Demonstration of Gullibility. *J. Abnorm. Soc. Psychol.* **1949**, *44*, 118–123. [CrossRef]
56. De Winter, J.C.F.; Driessen, T.; Dodou, D.; Cannoo, A. Exploring the Challenges Faced by Dutch Truck Drivers in the Era of Technological Advancement. ResearchGate 2023. Available online: https://www.researchgate.net/publication/370940249_Exploring_the_Challenges_Faced_by_Dutch_Truck_Drivers_in_the_Era_of_Technological_Advancement (accessed on 18 November 2023).
57. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
58. Scheider, S.; Bartholomeus, H.; Verstegen, J. ChatGPT Is Not a Pocket Calculator—Problems of AI-Chatbots for Teaching Geography. *arXiv* **2023**, arXiv:2307.03196. [CrossRef]
59. OpenAI. How Your Data Is Used to Improve Model Performance. Available online: https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance (accessed on 18 November 2023).
60. Security Magazine. 32% of Organizations Have Banned the Use of Generative AI Tools. Available online: https://www.securitymagazine.com/articles/100030-32-of-organizations-have-banned-the-use-of-generative-ai-tools (accessed on 18 November 2023).