https://doi.org/10.36315/2023v2end127
Education and New Developments 2023

VALID BUT NOT (TOO) RELIABLE? DISCRIMINATING THE POTENTIAL OF CHATGPT WITHIN HIGHER EDUCATION

José L. González-Geraldo[1], & Leticia Ortega López[2]
[1]Department of Pedagogy, University of Castilla-La Mancha (Spain)
[2]University of Castilla-La Mancha (Spain)

Abstract

"Education is a complex concept that has been the subject of ongoing discussion and exploration by scholars and educators alike". Who could disagree with this sentence? As clear as empty but, nevertheless, plausible and convenient. Due to the title of this communication, none would be surprised if we, the authors, confess that these quoted first words were written by the AI of the moment, ChatGPT, when we kindly ask for a 400 word abstract regarding a congress about education. The rest, as Mario Bunge maybe would have said, is just -at the very best- mere literature.

With the increased use of the platform created by OpenAI, raising voices expressed their concerns about its lack of accuracy. Far from an appropriate way of imitating smooth human communication, the main problem relies in the inability to guarantee quality results. In this sense, a pilot study was designed within University of Castilla-La Mancha (UCLM, Spain) in which different definitions of education were presented to would-be social educators and would be teachers in two different faculties of education.

These definitions presented different levels of complexity and precision (from an eight-year-old child to a full professor). In addition, the researchers used artificial intelligence to mimic, through specific and direct commands, the same level of complexity and precision.

Subsequently, all the definitions were randomly arranged in a document in which the sample (n = 130) had to rate on a Likert-type scale their degree of agreement and the level of complexity they considered for each definition. After this first stage, they were also given a second record sheet in which they were informed that one or more -but not all- of the previous definitions had been elaborated by an artificial intelligence, requiring them to indicate, for each one, whether they considered that a person was behind it or not.

The results, currently being coded through the SPSS, are likely to contrast with what ChatGPT itself -surprisingly with no further evidence- predicts: "The results of the study indicated that the students had a moderate agreement with definitions made by people and a low agreement with definitions made by artificial intelligence. Additionally, the students perceived the definitions made by people to be more complex than the definitions made by artificial intelligence". We -the authors- are eager to corroborate whether the title of this communication should have been phrased as a solid statement or not.

Keywords: *Higher education, artificial intelligence, AI, ICT, theory of education.*

1. Introduction

In November 2022, OpenAI unveiled its latest and most revolutionary creation to the world: ChatGPT, a natural language model based on the GPT-3.5 architecture, which promised to be the most advanced and powerful in its class to date. Today, April 2023, the world is playing with the GPT-4 version (OpenAI, 2023) as the possible emergence of GPT-5 is announced in the not-too-distant future. Beyond the significant impact this entails, its speed is so concerning that there are many voices calling for, at the very least, a moratorium on its development (Future of life Institute, 2023).

In short, ChatGPT is a language model based on the Generative Pre-trained Transformer 3 architecture, which uses deep learning techniques to process text sequences and autonomously generate new ones based on user inputs (Gómez-Cano et al., 2023).

Alongside more visual tools such as Dall-E, it is part of what is known as Generative AI, as it uses complex natural language processing algorithms to interpret the meaning of words, grammar, and the context of a conversation, to produce a coherent and relevant response (Cortés-Osorio, 2023). This ability is possible thanks to the use of large training data sets and advanced algorithms that allow it to learn complex patterns and relationships in natural language. In addition to this unsupervised training, there is another stage of selective fine-tuning that makes ChatGPT a much more dialogical tool than its mere GPT engine.

This tool is ideal for applications such as virtual assistants, chatbots, and automatic text generation effectively, but it cannot be considered the definitive tool - for now - as it is in a continuous process of adjustment regarding its accuracy. However, the fact that we wonder if GPT-5 will achieve Artificial General Intelligence (AGI) is a sign that, at least, we are starting to get close (Tamim, 2023).

Close, but not quite yet. At present, the reliability of the bibliographic sources and academic references that ChatGPT includes in the elaborations it carries out after each prompt is still weak and even astonishing. When it is said that the Chat "hallucinates" or "raves," it means that it is providing inaccurate or imprecise information, "pretending to know" due to its inability to understand certain situations, and its limited capacity for verifying information sources (Gravel et al., 2023). Although this data set is enormous, it is still not possible to guarantee that it contains all the relevant and up-to-date information on all existing topics in the world. Not only because of its training but also for other reasons: context window and sampling techniques. We must not forget that, despite its enormous potential for classifying and relating data, ChatGPT lacks the ability to understand the context and underlying intention behind each question. It does not understand what it reads, nor can it reason it out (Barros, 2023) as it only classifies and chooses information based on its storage and our requirements.

On the other hand, we find in this platform an amazing quality opportunity in terms of its ability to replicate human language, which is the reason and source of the revolution and controversy that is causing the use of this new tool in students -in principle- and in academics -in the future- due to the possible impossibility of differentiating between human and machine (Mitrović, Andreoletti, & Ayoub, 2023). The reason for its ability to mimic human language so accurately is because ChatGPT's architecture is based on a neural network of transformers (Srivastava, 2023; Vaswani et al., 2017), which processes text input and learns to predict the next word or phrase (token), depending on the context of the conversation. This prediction process is based on the relationship between the previous words and sentences (context window), which together with a controlled random creativity technique (sampling: temperature and *top-p*, mainly) allows it to generate coherent and relevant responses, thus producing more diverse and creative outputs, avoiding at the same time repetition.

All this rationale leads us to focus our project within two premises or objectives:

1) Check whether university students who are familiar with the concept of education -from a more formal conception towards a more holistic one- are able to differentiate definitions of education made by an Artificial Intelligence from those made by humans (questionnaire 3).
2) Corroborate or refute the reliability of the conclusions that ChatGPT predicted -hallucination- regarding the results of the analysis of the data collected in our research (questionnaires 1 and 2). When it was asked to give us an abstract for this research, and without any data, ChatGPT (v. 3.5) stated the following in past tense: a) Students had a moderate agreement with definitions made by people and a low agreement with definitions made by artificial intelligence, and b) Additionally, the students perceived the definitions made by people to be more complex than the definitions made by artificial intelligence.

2. Method

2.1. Sample

The sample (n = 130) consists of students from Primary Education Degree (n = 43) and Social Education Degree (n = 87) at the University of Castilla-La Mancha (UCLM, Cuenca, Spain). 77.7% (n = 101) were women and 18.5% (n = 24) were men. A small percentage preferred not to answer this item (3.8%, n = 5).

The sampling was intentional for convenience, as we needed students familiar with the educational concept in all its dimensions: from the more formal academic approach, provided by would-be teachers, towards a more holistic conception given from the perspective of would-be social educators.

2.2. Instrument

Data collection was carried out through three different questionnaires, which had to be answered by the students independently and in a predetermined order: questionnaire 1, questionnaire 2, and questionnaire 3.

These questionnaires consisted of 16 definitions of education. Half of these definitions (8) were created by humans, and the other half (8) were developed by ChatGPT-3.5. These definitions showed different profiles and varied between complex, simple, incomplete, poetic, childish, aseptic, professional, formal, and metaphorical elaborations. Having said this, the key point was that the AI was intended to mimic the same human profiles.

In questionnaire 1, students had to show their level of agreement with the 16 definitions through a Likert scale with scores from 1 to 5 (1 being completely disagree and 5 being completely agree).

Questionnaire 2 consisted of the same 16 education definitions in the same order, but this time students had to indicate the complexity rating they assigned to the definitions. In the same way, they did so through a Likert scale with scores from 1 to 5 (1 being very simple and 5 being very elaborate).

Questionnaire 3 also consisted of the same 16 education definitions and in the same order, but this time students only had to mark which definition or definitions they believed had been developed by Artificial Intelligence and not by a human. Students were warned that at least 1 definition had been created by ChatGPT, but not all of them, so they had a range in which they could mark a minimum of one definition and a maximum of 15 definitions.

It is very important to emphasize and respect the order in which the questionnaires were administered to the students. Until the last third part, no mention about Artificial Intelligence was given. Therefore, we ensured that both perceptions -the level of agreement (questionnaire 1) and the level of complexity (questionnaire 2)- were implemented without possible bias towards the source of these definitions.

After the administration of the questionnaires and the collection of the information, we proceeded to input the data into the statistical software (*SPSS* v.28).

3. Results and conclusions

The results of the present study show that among the eight possible definitions of AI present, students identified on average four definitions (x̄ = 4.34), and that the level of agreement within these four possible definitions averaged two real hits (x̄ = 2.19). In addition, subjects never identified any of the AI definitions above 50% agreement (questionnaire 3).

Regarding ChatGPT's predictions, the real data show that students gave higher scores to the agreement rating of AI (questionnaire 1, x̄ (AI) = 3.57; x̄ (HUM) = 3.13) as well as to the degree of complexity of AI definitions (questionnaire 2, x̄ (AI) = 3.17; x̄ (HUM) = 2.74).

With these data, we can affirm that ChatGPT was completely wrong in its prediction of the outcome of this study, thus refuting predictions a) and b) discussed in the first section.

In addition, apart from ChatGPT's predictions, other analysis was carried out to contrast means between groups, taking the mean degrees of agreement and complexity (questionnaires 1 and 2) as the dependent variable and sex and degree of the sample as independent variables. Since the principle of sample normality were not met (Komogorov-Smirnov, p < 0.001), non-parametric tests were used, with the U Mann-Whitney test being the most appropriate due to the two levels of the independent variables.

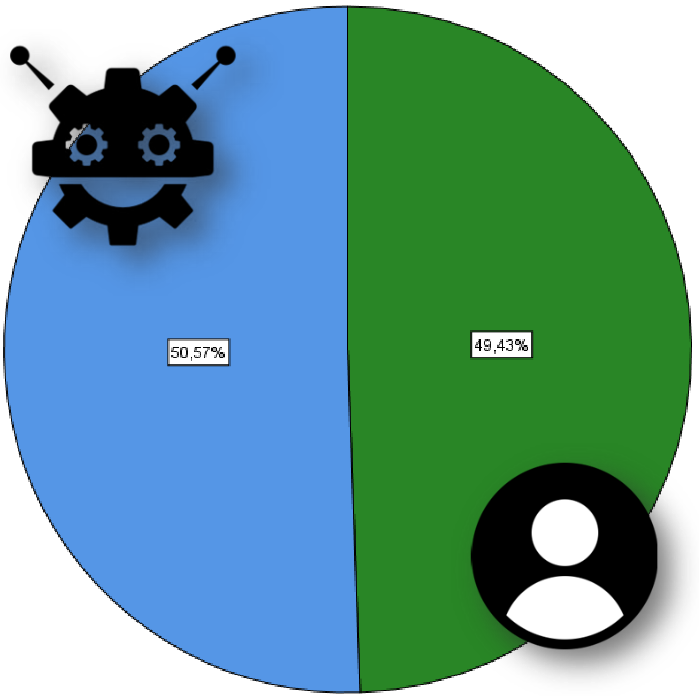
The results of the real data indicate that the degree of agreement in women (*Mdn*(IA) = 3.63/*Mdn*(HUM) = 3.13) does not differ significantly from that of men (*Mdn*(IA)= 3.75/*Mdn* (HUM)= 3.19), in relation to the definitions elaborated by the IA (*U* = 1133.00, *z* = -0.497, *p* > 0.05) nor in relation to those elaborated by humans (*U* = 1204.50, *z* = -0.05, *p* > 0.05).

In addition, the results indicate that the degree of complexity in women (*Mdn*(IA) = 3.13/*Mdn*(HUM) = 2.63) does not differ significantly from that of males (*Mdn*(IA)= 3.25/*Mdn*(HUM) = 2.63), neither in relation to the definitions elaborated by the IA (*U* = 1114.50, *z* = -0.613, *p* > 0.05) nor in relation to those elaborated by humans (*U* = 1132.50, *z* = -0.50, *p* > 0.05).

However, if we distinguish by grade, we observe how the degree of agreement of future teachers (*Mdn*(IA) = 3.88/*Mdn*(HUM) = 3.38) differs significantly from the degree of agreement of future social educators (*Mdn*(IA) = 3,50/*Mdn*(HUM) = 3,13) with respect to the definitions elaborated by AI (*U* = 1194.00, *z* = -3.358, *p* < 0.001, *r* = -0.29) and with respect to the definitions elaborated by humans (*U* = 1386.00, *z* = -2.403, *p* < 0.05, *r* = -0.21). In both cases, as we can see, the effect size (*r*) is small. In parallel, there are no significant differences in relation to the perception of complexity and grade.

Finally, also dividing the sample by groups, it is necessary to observe how there is only one case in which one of the groups, in this case would-be social educators (n = 87), believe that they detect a definition of AI (50.57% agreement, see Graphic 1). The definition is: "Continuous and voluntary intellectual process by which a person progresses thanks to the knowledge previously attained by others and thanks to which he/she can create new knowledge". Interestingly, and with this thought in mind we close our proposal, this was the definition elaborated by the highest-ranking professor in our pedagogy department.

Graphic 1. Percentages of the only definition that more than half of the sample (n = 87) -wrongly- detected as being made by an Artificial Intelligence. The definition was made by a high-rank full professor.



4. Conclusions

The results show that the university students of both degrees were not able to identify none of the definitions of education made by Artificial Intelligence. This leads us to the conclusion that, to date, we are not prepared -nor even alert- to distinguish creations made by AI from human elaborations. At least, regarding to college students. Future lines of research could be focused on other groups. What about teachers? Would they fail too?

In addition, seeing how ChatGPT was wrong in its predictions about our greater agreement with human elaborations, force us to think that we appreciate what we like to call “the algorithm's melody”, an -aseptic and formal tune- more than the possibly less perfect features that emerge from the human hand.

Accordingly, do we value mechanical correctness over those more spontaneous shades that characterize human creativity. In fact, the students considered the definition elaborated by the senior academic –full professor- as an elaboration made by an Artificial Intelligence, perhaps because of its formalism and correctness. Does this mean that we identify error with humanity and perfection with AI?

Another noteworthy fact about the results observed is the difference in the degree of agreement and the assessment of the complexity of ChatGPT's creations. This difference may stem from the conceptualization of education in each professional field: would-be teachers and would-be social educators.

Following this rationale, primary school degree may be focused on the competences on teaching and formal academic education, meanwhile those developed by social educators could look at education from a more holistic perspective, pursuing relational autonomy and the full development of all areas of life, a more complex view to mimic and a more difficult place to reach for an AI. A single word: Education. An infinite universe of definitions and conceptions for both: humans and machines.

References

Barros, J. (2023). Chatgpt: Más preguntas que certezas. *Tecnología, 15*, 50-54.

Cortés-Osorio, J. A. (2023). Explorando el potencial de ChatGPT en la escritura científica: ventajas, desafíos y precauciones. *Scientia et Technica, 28*(1), 3-5.

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter. Future of Life Institute.

Gómez-Cano, C., Sánchez-Castillo, V., & Clavijo, T. A. (2023). Unveiling the Thematic Landscape of Generative Pre-trained Transformer (GPT) Through Bibliometric Analysis. *Metaverse Basic and Applied Research, 2*, 1-8.

Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *MedRxiv*. https://doi.org/10.1101/2023.03.16.23286914

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*. https://arxiv.org/abs/2301.13852

Srivastava, M. (2023). A Day in the Life of ChatGPT as a researcher: Sustainable and Efficient Machine Learning-A Review of Sparsity Techniques and Future Research Directions. *OSFPreprints.* https://doi.org/10.31219/osf.io/e9p3g

Tamim, B. (2023). GPT-5 expected this year, could make ChatGPT indistinguishable from a human. *Interesting Engineering.* https://interestingengineering.com/innovation/gpt-5-chatgpt-indistinguishable-human

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Conference on Neural Information Processing Systems (NIPS).* USA.