education sciences

MDPI

Article

From Human to Machine: Investigating the Effectiveness of the Conversational AI ChatGPT in Historical Thinking

Sergio Tirado-Olivares , Maria Navío-Inglés, Paula O'Connor-Jiménez and Ramón Cózar-Gutiérrez *

LabinTic, Lab of Technology Integration in Classroom, Faculty of Education of Albacete, University of Castilla-La Mancha (UCLM), 02071 Albacete, Spain; sergio.tirado@uclm.es (S.T.-O.); maria.navio@uclm.es (M.N.-I.)
* Correspondence: ramon.cozar@uclm.es

Abstract: In the digital age, the integration of technology in education is gaining attention. However, there is limited evidence of its use in promoting historical thinking. Students need to develop critical thinking skills to address post-truth and fake news, enabling them to question sources, evaluate biases, and consider credibility. With the advancement of artificial intelligence (AI), historical thinking becomes even more crucial, as chatbots appear capable of analysing, synthesizing, interpreting, and writing similarly to humans. This makes it more difficult to distinguish between human and AI-generated resources. This mixed study explores the potential of AI in developing an argumentative historical text compared to future teachers. After 103 preservice teachers were instructed in historical thinking, they assessed a text written by a human and an AI-written text without knowing their authors. The obtained results indicate that participants assessed the AI text better based on historical thinking skills. Conversely, when asked about the capability of AI to develop a similar text, they emphasized its impossibility due to the belief that AI is incapable of expressing personal opinions and reflecting. This highlights the importance of instructing them in the correct use and possibilities of AI for future historical teaching.

Keywords: historical thinking; artificial intelligence; preservice teachers; history teaching

check for updates

Citation: Tirado-Olivares, S.; Navío-Inglés, M.; O'Connor-Jiménez, P.; Cózar-Gutiérrez, R. From Human to Machine: Investigating the Effectiveness of the Conversational AI ChatGPT in Historical Thinking. *Educ. Sci.* **2023**, *13*, 803. https://doi.org/10.3390/educsci13080803

Academic Editors: Gary K. W. Wong and Ho-Yin Cheung

Received: 11 July 2023
Revised: 1 August 2023
Accepted: 3 August 2023
Published: 5 August 2023

1. Introduction

The integration of technology into the field of education has gained significant attention and its use has become regular in the classroom [1]. Today's students in primary and secondary education levels are considered digital natives, as technology plays a fundamental role in their daily lives. In this context, where information is instantly accessible and technology is part of our routines, it becomes fundamental for students to leverage their critical thinking skills and be cognizant of the era of post-truth and the prevalence of fake news [2,3]. The possibility of encountering fake news forces students to question the sources, analyse the information they come across, and consider its reliability.

In order to face this reality, students need to develop the necessary skills for it. This makes the implementation of methodologies that promote their development essential. Regarding this idea, in history teaching, there is currently a pedagogical trend that encourages the development of historical thinking. This promotes the improvement of critical thinking and reflective capacity [4,5], which should be applied when stumbling upon a resource of any kind. Nonetheless, the new advances carried out in the field of conversational artificial intelligence (hereinafter, AI) make it even more difficult. Chatbots are now more capable of writing texts almost as a human being would [6], which complements their ability to cope with huge amounts of information [7]. Consequently, now, both teachers and students must face the fact that there is a tool that can be used to generate new information that may not necessarily be accurate [8]. However, apart from the aspects mentioned, it should be noted that conversational AI can also help teachers in their tasks and can foster the student's learning process [9,10], so it can be a helpful tool when used properly.

Due to the dangers and the advantages raised, the present study focuses on the challenge in identifying the potential of AI to develop an argumentative commentary text in contrast to future teachers, considering the dimensions of historical thinking. To this end, future teachers assessed a text written by another student and a text written by a chatbot and reflected on the experience after finding one of them belonged to the conversational AI ChatGPT.

1.1. Technology and Conversational AI in Education

Technology is one of the main bases of the 21st century and plays a crucial role in the lives of individuals today [11]. It is present in various aspects of life, including education. Emerging technologies have brought forth new possibilities in the teaching and learning process, as well as tools to potentiate this process. Thus, the incorporation of active methodologies and the generation and diffusion of knowledge have found an ally in new technologies. The importance that technology has gained in the educational field can be seen in the inclusion of digital competencies in the curricula of many countries, as well as in how teachers are continuously encouraged to incorporate it in the classroom [12]. Therefore, transforming students into digitally competent individuals has become a priority in the education system. Such is its importance that international frameworks such as Digital Competence of Educators (DigCompEdu) address the need for future teachers to acquire this competence [13]. Moreover, the Educause Horizon Report [14], when talking about the implementation of technology, also mentions the possibility of incorporating AI in education now.

The realization of the importance and possibilities of technology in education has led to its introduction in the classroom. As has been the case with other types of emerging technologies, the introduction of AI in the field of education has generated debate since its beginnings and a diversity of opinions and perspectives. As early as the 1980s, there were authors who defended its usefulness in education [15,16] as well as its entry into this field as something inevitable [17]. The term *artificial intelligence* was coined by McCarthy in the 1950s when he conjectured that human intelligence could be simulated by a machine. Winston, who indicates that there are many definitions for this concept, narrows down the definition and states that AI is "the study of the computations that make it possible to perceive, reason and act" [18] (p. 5). In the same way, Rouhiainen [19] also mentions the difficulty on defining the term and explains that AI concerns the technological systems that learn from data to make decisions and perform tasks mimicking humanlike intelligence. Thus, the permanent evolution of these systems, which continue to progress to this day, can be perceived.

There are many technological tools whose functioning is based on artificial intelligence. Among them, we find chatbots, which are programs whose function is to simulate a conversation with the user [20], similarly to how an interaction between people would occur, and which recently have shown great advancement in generating humanlike speech. To achieve this, there are different models that regulate language processing and its generation. Currently, one of the most prominent models is the generative pretrained transformer (GPT). When talking about processing information, a generative model goes beyond working with the given information and generates new data [21]. Similarly, the GPT model also generates new data. To be able to do so, it undergoes a process based on an autonomous training phase with unlabelled information, and another supervised training phase that focuses on improving the performance of specific tasks [22,23]. Ultimately, it generates speech that is very close to that of a human being [6]. The most updated version of this model is employed by the ChatGPT chatbot developed by the company OpenAI [24]. The constant evolution of models such as GPT and conversational AI systems such as ChatGPT involves the development of tools capable of providing immediate responses to almost any question and completing complex tasks. This leads to the immediate generation of various types of text in which the chatbot's language usage closely resembles that of a person [7].

The introduction of AI in the educational field affects various aspects such as teaching methods, the role of the teacher, and the teaching–learning process [25]. Focusing on chatbots, the literature review conducted by Okonkwo and Ade-Ibijola [26] shows that, in most cases, they are used to improve the teaching–learning process, in which AI has already provided very positive effects [27]. In this regard, studies such as those by Lin and Chang [28], Murad et al. [29], and Troussas et al. [30] suggest that the introduction of conversational AI in the classroom contributes to improving interest in learning, academic outcome, and the development of certain cognitive skills. Thus, there are already research studies exploring the possibilities of this type of AI to generate learning guidance in which the chatbot acts as a support for knowledge improvement [31,32]. Some of them also focus on the use of chatbots as support for educational activities such as answering students' questions [10], helping to understand class content such as programming concepts [31], and assessment [33]. Additionally, they could also assist in administrative tasks, which would reduce the amount of time teachers need to spend on such tasks and allow them to invest more time in their students [9].

However, according to Okonkwo and Ade-Ibijola [26], further research is still needed to expand the framework for the proper introduction of chatbots in the educational field, as well as to continue studying the functionality of these type of tools. Such technological resources are constantly evolving, and therefore, as can be seen in what has been explained previously, their possibilities of usage and their effects may change. Moreover, in order to ensure a proper introduction of AI in education, it is important for teachers to first understand its potential and capabilities so they can get the most out of them. Therefore, it is necessary to assess their knowledge of chatbots, which inevitably involves understanding to what extent they are familiar with their main characteristic: generating humanlike discourse. In this regard, in recent years, some studies have focused on analysing the ability of conversational AI systems to generate textual productions as a human would, as well as the capability of humans to distinguish between AI-written texts and human-written texts (e.g., [34–36]). Some of them identify patterns that can help people differentiate between how AI and humans write. These patterns are usually related to aspects such as the use of personal pronouns, expressing feelings, and utilising rude expressions, which may identify a human's writing style [36]. However, according to the studies mentioned before, those patterns seem to not be enough, since some people show difficulties when attributing authorship of the texts to the chatbot or the human and do it wrong despite them.

Besides the necessity of further research [26] and the difficulties that people face now when differentiating texts written by humans and AI [37], it should be noted that the effects of conversational AI have not been studied in some educational areas. To the knowledge of the authors, there is no evidence of its use and the analysis of its effects in the teaching of history. Consequently, investigating the didactical possibilities of chatbots in the history classroom is required.

1.2. Technology, AI, and History Teaching: Promoting Higher-Order Skills and Historical Thinking

Even though the use of technology-mediated active methodologies has been widely studied in education, there is limited evidence in the field of history teaching [38]. There are some studies that analyse the consequences of using technological tools such as websites, applications, virtual reality, or video games in the teaching of history (e.g., [1]) and the perception of the teachers and preservice teachers about it (e.g., [39]). Despite this, as stated previously, research production in this field is less frequent.

Regarding the rise of those active methodologies, a pedagogical trend that establishes that studying history should go beyond the memorization of data and theoretical contents and dates and focus on historical thinking emerges [40]. This methodology follows a common practice of historians, in which the analysis, the interpretation, and the reconstruction of evidence from the past take place [41]. This methodology, when applied to history learning, makes students put themselves in the historian's shoes and replicate the process, which also gives importance to the procedural content that helps to understand the past, thereby gaining a better understanding of the present [42]. In other words, by applying historical thinking, students engage in analysis, contextualization, research, and critical reflection, enhancing important cognitive skills. To develop historical thinking, students are encouraged to work with sources that can be subject to analysis and enable this reconstruction of the past. In doing so, they work on second-order concepts such as causality, processes of change and continuity, and relevance [41]. At the same time, they develop skills such as critical thinking, creativity, and reflective capacity [4,5]. This not only optimizes the learning of some concepts and events but also helps students acquire the necessary skills for the 21st century [38]. According to the definition of *liquid modernity* by Bauman [43], we are in a time in which everything is constantly changing, characterized by instability and individualism. Post-truth has also become one of the threats we must face today [44], along with fake news, widely spread by new technologies. Therefore, knowing how to think critically and reflectively, analyse and contrast information, and distinguish reliable sources becomes necessary to understand what happens in the society we are part of and to navigate within it. However, at this point, AI's ability to write similarly to a person [6,7] can make this task more complex.

Thus, and in accordance with what has been stated before about the benefits that AI has demonstrated in the teaching and learning process, it may contribute to the implementation of active methodologies such as historical thinking. Through historical thinking, different skills such as critical thinking as well as a deeper understanding of historical contexts and concepts such as processes of continuity and change are promoted [5,41]. Regarding the qualities and abilities of chatbots recently developed, such as ChatGPT, it seems reasonable to think that AI could serve as a guide for this process. However, this needs to be studied, and for a proper implementation of chatbots, first, students should know both their benefits and their dangers, and teachers should become familiar with the capabilities and functionality of conversational AI systems. In order to reach this as well as to know the extent to which chatbots can be a tool for improving historical thinking and the writing of historical argumentative texts, one of the first steps should be to analyse AI's ability to produce historical argumentative texts. The writing skills they have developed, in addition to the access they have to a vast quantity of data nowadays, might help learners improve their analysis of historical sources. Moreover, by using chatbots, they could also develop skills that are fundamental in historical thinking, such as critical thinking, when evaluating the assistance that the chatbot offers. Again, this means that further investigation on chatbots' introduction in education is still needed [26].

Therefore, considering everything noted above, there is undeniable evidence, on the one hand, of the benefits of chatbots in the teaching–learning process and, on the other hand, of the possibilities that emerging technologies offer to engage with historical thinking. However, there is a lack of research on the utilisation of chatbots for promoting historical thinking. To harness the advantages of conversational AI for the development of historical thinking in students, it is first necessary to train future teachers, since they will be responsible for their education. Given these circumstances, this research aims to examine the potential of chatbots to write an argumentative historical text based on historical thinking in contrast to future teachers. This is followed by the analysis of both texts by the preservice teachers without knowing their authors and, later, the critical reflection of the analysis carried out after discovering that one of the texts was written by ChatGPT. Thus, the research objectives that conduct the present study are:

- To evaluate the performance of AI in developing an argumentative historical text based on the dimensions of historical thinking in contrast to preservice teachers.
- To analyse preservice teachers' perceptions and beliefs about the capability of AI to develop an argumentative historical text similarly to a human.

2. Materials and Methods

2.1. Design

To analyse the pre-established research objectives, a mixed study based on both quantitative and qualitative data was carried out. This study design aimed to compare the historical thinking ability shown by preservice teachers in contrast to the conversational AI ChatGPT after students were instructed in this historical competence.

2.2. Participants

This study was conducted with university students. In particular, a total of 103 preservice primary education teachers from the Faculty of Education of Albacete of the University of Castilla-La Mancha (Spain) participated in the study. All participants were enrolled in the subject entitled "Social Science II. History and its didactics", and they were studying in the same academic year. The whole study was conducted inside the classroom during the compulsory practical lessons in which participants learn about the current methodological approaches to be implemented in history teaching.

Before the intervention, all students were informed that they were going to participate in an investigation and were required to read and sign an informed consent form after they had the opportunity to ask any relevant questions they had. Data from those students who did not sign the informed consent were not used in the study. Then, the ethical standards established by the APA were always complied with.

2.3. Procedure

The experimental phase of the study consisted of six sixty-minute sessions (see Figure 1). The first four sessions comprised an introductory theoretical lesson followed by three practical sessions. During these sessions, students received instruction on what historical thinking is and its application in history teaching through the analysis of historical resources. The objective was to teach students the skills to write argumentative texts while considering the various dimensions of historical thinking based on Seixas and Morton [41]. Following these sessions, the fifth session of the research involved a task where students were required to compose an argumentative text. The specific objective of this session was to assess the students' ability to construct a good historical narrative based on the translated abstract of the article authored by Kesternich et al. [45], which explores the consequences of World War II. In this session, exam conditions were simulated. Students were strictly prohibited from using any materials other than a pen or pencil and a rubber. The use of any additional materials would lead to the exclusion from the data collection for this study.

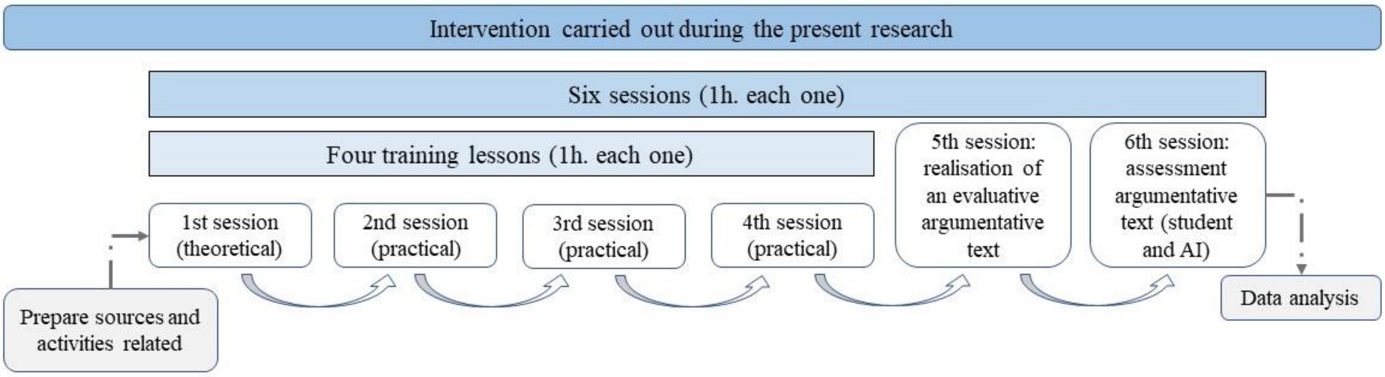
Intervention carried out during the present research
Six sessions (1h. each one)
Four training lessons (1h. each one)
1st session (theoretical)
2nd session (practical)
3rd session (practical)
4th session (practical)
5th session: realisation of an evaluative argumentative text
6th session: assessment argumentative text (student and AI)
Prepare sources and activities related
Data analysis

Figure 1. Summary of the procedure followed in the present research.

Prior to the final session, the researchers in this study evaluated the argumentative texts produced by the participants and selected one that demonstrated above-average development. In addition, the researchers requested the conversational AI ChatGPT (version 3.5) to generate an argumentative text using the same historical source and the same guidelines provided to students: "From this abstract, write a text as if you were an undergraduate student for primary education in which you use both the dimensions of historical thinking and you: (a) argue and reflect on the relevance and consequences of the Second World War based on this text; (b) indicate what ethical critique you would make of this event and whether you think its impact continues to the present day". Thus, the instructions provided to both preservice teachers and ChatGPT were identical.

In the last session, students were asked to assess both argumentative texts without informing them that one of them was generated by a chatbot. The first argumentative text given to the participants was written by a trainee teacher from another group outside the intervention. Therefore, the second of these texts was the one written by ChatGPT. After that, a discussion space was created based on four open-ended questions. During the discussion space, students became aware that one of the texts had been generated by a conversational AI as these questions were only revealed to the students after they had completed the evaluation of the texts. The following section provides detailed information on the instrument used to assess the texts, as well as on the open-ended questions. A summary of the procedure followed in the present research can be seen in Figure 1.

2.4. Instruments and Measurements

The evaluation instrument utilized in this study was developed by Sáiz and Gómez [46] and has undergone validation and implementation by the authors in analysing historical narratives of preservice teachers. It consists of seven dimensions derived from the historical thinking dimensions described by Seixas and Morton [41] as outlined in Table 1. These dimensions were assessed using a four-choice Likert scale, ranging from 0 (indicating no level of historical thinking reflected in the dimension) to 3 (indicating an optimal demonstration of historical literacy and discursive ability in the dimension). Students employed this instrument to evaluate both texts. After that, the participants were asked the following four open-ended questions:

1. Which text do you believe deserves a higher score in an exam? Why?
2. Which one is able to convey more emotions and feelings? Why?
3. Do you think a machine or AI can make a comment such as these ones?
4. If I told you that one of these two texts was written by an AI, would you believe it? Which one do you think it would be?

Table 1. Scores given by preservice teachers to both AI text and student text in each historical dimension based on Sáiz and Gómez [46] and in total.

	Argumentative Text					
	Student Text		**AI Text**			
Dimensions	**Mean**	**SD**	**Mean**	**SD**	**Z**	***p***
Cause and consequence	1.98	0.816	2.29	0.736	−3.31	0.001
Continuity and change	1.86	0.525	2.30	0.669	−4.72	<0.001
Substantive historical content	1.72	0.733	2.28	0.617	−5.36	<0.001
Historical significance	2.17	0.706	2.34	0.635	−1.92	0.055
Historical awareness	2.05	0.746	2.39	0.744	−3.48	<0.001
Presence of substantive metaconcepts	1.39	0.645	1.89	0.685	−5.16	<0.001
Complexity of narratives	1.92	0.860	2.22	0.79	−3.01	<0.001
Total score	1.87	0.39	2.25	0.43	−6.22	<0.001

Subsequently, data gathered were coded and analysed using two different software tools. For the quantitative data, statistical software SPSS vs. 24 was used. A descriptive data analysis and an inferential analysis were carried out with it. A Wilcoxon signed-rank test was carried out to analyse the first research objective as we were working with ordinal data. On the other hand, for the qualitative data—the content of the answers to the open-ended questions—specific software for qualitative research Atlas.ti (vs.9) was used. An inductive analysis was carried out as there was no prior categorisation of the codes used [47]. Thus, a situational design was used with the aim of analysing the second research objective from the perspective of those preservice teachers who participated in the study. Two of the study researchers carried out the content analysis of the data, and any doubts or disagreements were resolved by discussion. In the results section, the codes have been underlined for better identification. Finally, since it was possible to compare the responses to the open-ended questions, a statistical inferential analysis was carried out through the chi-squared test. Answers to each question were coded to this aim. Firstly, if students thought human-written text deserves a higher score, it was coded as (1), if AI text deserved a higher score as (2), and when each text deserved the same score (0). The same coding was applied to which text was able to convey more emotions and feelings. For the third question, the coding was (1) if they considered AI able to write a similar text or (2) if not. For the last question, the coding was (1) if participants thought that the first text was written by an AI, (2) if they considered the second one was, or (0) if they still thought that neither was.

3. Results

The results are organised into sections corresponding to each research objective. The first analysis was done to compare the score given to each text by students. The second analysis aimed to study the students' perception about why one of the texts deserved a higher score if any, which one conveyed more emotions, and the participants' perception about the capability of a conversational AI to write a similar argumentative text.

3.1. Comparison between an Argumentative Text Developed by a Human and the Conversational AI ChatGPT

In relation to the first objective, scores given by preservice teachers for each historical thinking dimension and the total score given to the texts are presented in Table 1. In addition, to check if the differences obtained were significant, an inferential analysis using the Wilcoxon signed-rank test was carried out (Table 1).

As can be seen, students scored the ChatGPT text higher in almost all dimensions. The results obtained evidenced the capability of the conversational AI text to write not only a correct text but also one that encompasses all dimensions of historical thinking. In fact, with the exception of the historical significance dimension, which was also close to significance ($Z = -1.92$, $p = 0.055$), the Wilcoxon test indicated that scores given to the AI text by the participants in the rest of the historical dimensions were statistically significantly higher than scores given to the student text (cause and consequence: $Z = -3.31$, $p = 0.001$; continuity and change: $Z = -4.72$, $p < 0.001$; substantive historical content: $Z = -5.36$, $p < 0.001$; historical awareness: $Z = -3.48$, $p < 0.001$; presence of substantive metaconcepts: $Z = -5.16$, $p < 0.001$; complexity of narratives: $Z = -3.01$, $p < 0.001$). These differences also led to a statistically significant difference in the total score ($Z = -6.22$, $p < 0.001$).

3.2. Is Conversational AI Able to Write Argumentative Historical Text Based on Historical Thinking? Preservice Teachers' Perceptions

After preservice teachers indicated the ability of ChatGPT to produce texts encompassing all dimensions of historical thinking unknowingly, the open-ended questions provided an opportunity to evaluate the preservice teachers' own view of this matter. Firstly, to answer the question "Which text do you think deserves a higher mark in the exam? Why?" participants used the scores given in the Likert scale to answer that question. Thus, while only 4.9% of the participants scored the two texts equally or the first text better (18.4%), 76.7% of the participants scored the second text higher.

When asked to justify their answers, they highlighted that they had focused on different aspects such as whether the text had a good structure, which involved cohesion, coherence, and being well written and organized: "The second one has a better structure, the ideas are more cohesive and the structure makes it easier to read and the ideas are detected in the first reading" (S.13). Moreover, they gave a higher score when they considered that the text provided relevant information; this means that the text talked about at least all the items that were asked and exposed important information: "The text has all the information that is asked and does not talk about other things that are not relevant (. . .)" (S.9). Another aspect considered by the preservice teachers was whether the text was more complete. This characteristic entails giving the required and additional information, more data, and showing a greater knowledge of history: "I think that the second text merits a higher score as it specifies in a detailed manner the historical events that occurred in that time, giving more information" (S.40). Likewise, they gave a higher score to the text that showed critical thinking, so it had a critique, personal opinion, or different approaches to the same topic "The second text because it includes value judgements by means of ethical criticism (. . .)." (S.65). Finally, the preservice teachers rated more positively the text in which there was a greater inter-relationship between the historical facts: "I consider that he or she has better selected the information and related it to other aspects that do not appear in the text" (S.28), as well as whether the text showed a relation with current affairs: "(. . .) It mentions aspects that have had repercussions up to the present day, that is, the impacts that the situation has had in the current days" (S.25).

In addition, regarding the second question "Which of the two is able to convey more emotions and feelings? Why?", the responses revealed that 70.9% of participants believed that the human-generated text had a greater capacity to convey emotions, while 29.1% chose the AI-generated text. These data suggest that the majority of participants, albeit unknowingly, perceived human-generated text as more effective in eliciting emotional responses compared to AI-generated text. Moreover, the chi-squared test of independence was performed to examine the relation between the text rated higher and the capacity to convey emotions. This test revealed no significant relationship between the two variables (X^2 (2) = 0.335, p = 0.846). Therefore, the higher score given to the AI text was not related to a higher capacity to convey emotions from the participants' point of view. However, it is noteworthy that a significant percentage of the participants already acknowledged AI's ability to effectively convey emotions, indicating a growing recognition of AI's proficiency in this domain.

To choose the best text in relation to this point, the preservice teachers considered different aspects such as whether the text talked about the consequences of the War, and the social impact that it had: "I think the text that can transmit more feelings is the first one, as it talks more about the peoples' consequences" (S.68). Another aspect used to choose the best text was whether the text mentioned tragic situations, which involve talking about the deaths, painful events, or crimes, among others: "The first text: talks about numerical datum of the deceased people, refers to the hard situation of the families, names physical and mental diseases, and so on" (S.65). Moreover, the preservice teachers said that one of the texts evoked more emotions whenever they identified an allusion to empathy in it: "The first one because I think that it is based more on sentimental aspects, putting ourselves in other people's shoes and realising about what happened" (S.29). Finally, the preservice teachers took into account whether the text talked about society, which implies the families, people, or how society lived those hard moments: "I think that it is the first one because it talks from a perspective centred on the society (. . .)" (S.9).

In contrast, question three revealed notable inconsistencies in the ratings allocated within the rubric. Specifically, the question inquired about the capacity of AI to generate a text of comparable quality. The responses from students demonstrated a division of opinions. A majority of 53.4% expressed scepticism, while 45.6% held the belief that AI could indeed produce a text of similar nature. A chi-squared test of independence was calculated to determine whether there was a relationship between the higher text rating and the students' perception of the capability of a conversational AI to write a similar text. Again, no relationship was found (X^2 (2) = 0.945, p = 0.623). This suggests that what students think about the capability of an AI to write a similar text is not related to scoring the AI text better.

In this sense, participants, in line with the answers to the previous question, commented that AI could not make a text such as these two because an AI text would lack emotional character. The preservice teachers considered AI incapable because it could not transmit feelings or, at least, not as much as a person: "I do not think so, because a machine can give a lot of datum, but it does not think by itself, and it has not got a historical conscience and cannot transmit feelings neither awaken emotions" (S.23). Moreover, the preservice teachers believed that a text developed by these technologies would lack ethical and critical perspective, which means that they would not give different points of view or personal opinion, when being asked to reflect on moral aspects and criticism about the events that had taken place: "It could argue the causes and consequences, but I do not think that it was able to do an ethical criticism, giving its personal opinion about the consequences of the War" (S.11).

However, the preservice teachers considered that a text composed by AI would present more historical content, for example, because the machine would be able to take information from different sources and relate and contrast it: "Yes. Referring to datum, a machine could conduct it and even do it more complete than ourselves. (. . .)" (S.85). Lastly, many preservice teachers believed that AI is advanced, which alludes to the modernity and the advances in this technology. They consider AI has evolved to such an extent that they see it as plausible for AI to write texts such as the ones assessed: "In my opinion, it would be able to do a commentary text because technology advances really fast and, after all, machines are programmed by humans to be able to do whatever we want" (S.55). This shows that we are at a time of great and profound change, which means there is no single, clear vision among the future teachers participating in this study.

Finally, when participants were tasked with determining which of the two texts they believed to be generated by a conversational AI (question 4), a significant majority of students correctly identified text 2 as having been created by an AI (85.3%). In contrast, only 6.8% attributed text 1 to an AI, while the remaining 3.9% still maintained the belief that neither text could have been autonomously produced by an AI. Deciding which one had been developed by an AI was not related to scoring it higher or not (X^2 (4) = 4.25, p = 0.374) nor by the capability to convey emotions (X^2 (2) = 4.71, p = 0.095). This evidenced that the students' decision of which one was written by an AI was not related to having scored it higher or the emotions the text was able to transmit.

Among the statements found among the students, the following stands out: the text written by an AI was that one with a better writing, considering different aspects such as the text structure, clarity, and organisation of the arguments provided: "In my opinion, the second text is the one that has been written by a machine because it has a very well elaborated redaction. (. . .)" (S.78). In addition, students considered AI would use a more accurate language when writing a text, which involves using more formal, precise, correct, or technical vocabulary: "The second text because it is elaborated with more formal and specific words. (. . .)" (S.96). Similarly, the preservice teachers chose the text created by an AI based on which one included more content (referring to more data, details, or being more complete): "The second text as it is written specifying more datum and contextualising in a more detailed manner" (S.40). Finally, students again pointed out that AI could evoke no emotion, or at least not as much as a person: "it is the second one, because it transmits less emotions than the first text" (S.47). Thus, it is evident the future teachers believe in the ability of conversational AI to generate texts based on a greater amount of information, but without the embodiment of higher-order competences such as critical thinking or reflection.

4. Discussion and Conclusions

Chatbots have been advancing and improving over time. Thus, current conversational AI systems such as ChatGPT can generate texts that are very similar to those written by humans [6,7], using language just as a human would. This makes it even more important to be able to critically analyse the information and resources we encounter in the post-truth and fake news era, and promoting historical thinking in history class helps students develop that ability. Such are the advances in conversational AI that, together with the ability to process vast amounts of data [7], according to the results obtained in this study, the chatbot created by OpenAI could carry out an argumentative text based on the analysis of a historical source with better results than a university student. Studies such as those by Mitrović et al. [36] and Guo et al. [34] establish differences often related to the chatbot's incapacity to imitate some practices and abilities considered typical of a human. However, according to the analysis of the participants in this study, ChatGPT is able to replicate the process of critical reflection and analysis that students should engage in when they are in front of a source.

The preservice teachers evaluated the text regarding the instrument developed by Sáiz and Gómez [46] to analyse the level of historical competence. When assessing the texts, the chatbot managed to delve into second-order concepts and the foundations of historical thinking better than the preservice teachers, according to the results. This can be seen in the significant differences in almost every historical thinking dimension, in which ChatGPT scored better. It should be noted that the only item with no significant differences was the one referred to historical relevance. This dimension alludes to historical events and, consequently, to facts and data. Hence, these differences between dimensions are surprising since one of the main characteristics of ChatGPT is that it has access to large quantities of data to carry out its task [48], so the expected outcome would be for the AI to perform better than the student in this item as well. In relation to this, students also believed the chatbot would be able to focus on facts and data such as causes and consequences but not to reflect on it. Taking these results into account, as well as the benefits that the literature highlights about the use of AI and chatbots in the teaching–learning process [27], it seems reasonable to consider investigating the possibility of seeking methods for the proper incorporation of this type of AI as support for the development of students' historical thinking. As the results demonstrate, students believed the chatbot's level was higher than theirs when carrying out that activity and therefore, the learner's development of historical thinking is still improvable. Now, chatbots such as ChatGPT seem to be able to simulate a critical reflection process upon a text and go beyond facts and data. Given this and the higher score the preservice teachers gave to the AI in comparison to their colleague's text, knowing the extent to which this tool can produce an historical argumentative text could lead to its use in improving both their writing when they are asked to produce an historical argumentative text and their historical thinking skills.

Regarding the second objective of this investigation, when discussing the dissimilarities between conversational AI and humans, the main difference, according to some comparative studies, is the expression and transmission of emotions [34,36]. This aspect is usually directly related to human expression and is often a differentiating criterion. In this regard, over 70% of the participants identified the text commentary made by a student as the one that conveyed more emotions. However, it should not be ignored that nearly 30% attributed this capability to the activity done by AI. This, again, could support the idea that the use of language by chatbots is becoming increasingly similar to that of humans [6,7].

In relation to this last idea, the results show that, before knowing that one of the commentaries belonged to ChatGPT, there was a notable scepticism regarding the idea that AI could write such a text. Overall, 53% of the participants did not conceive the possibility of a chatbot being able to compose a text commentary based on the analysis of a source and the reflection upon it while considering historical thinking. Therefore, over half of the pre-service teachers were unaware of the extent of these tools' writing capabilities. This reinforces the idea that more studies on the introduction of chatbots in education are needed [26] and highlights the lack of training regarding a tool that has already been introduced in education and has shown benefits in students' instruction [27–30].

However, most of the future teachers were able to distinguish the text commentary written by the AI from the one written by another preservice teacher. It could be due to the existence of writing patterns that characterize ChatGPT compared to humans [36], as stated before. Nevertheless, it is worth noting that, as shown in previous studies, the distinction is not unanimous [37], and approximately 15% failed to attribute the texts to their author because of the similarities in the writing style.

Thus, according to the obtained results and the evaluations carried out by the future teachers, ChatGPT appears to be capable of writing an argumentative text based on a historical source and basing its writing on the dimensions of historical thinking. Without knowing the authors of the text, they rated the one elaborated by the chatbot better. Nonetheless, they are not aware of the capacity and possibilities of this tool, which has already been introduced in other areas of education, despite being able to detect some features and patterns typically associated with AI and human writing. Thus, the conducted investigation could be a starting point to delve further into the knowledge and training of future teachers regarding this tool, especially regarding the abilities of ChatGPT observed in this investigation, which could be useful for the improvement of historical thinking in history teaching.

Proposals for Future Research and Improvement

Throughout the research, several new questions and hypotheses emerged, with the potential to yield interesting results in the field of AI-mediated history teaching. Firstly, given the AI capacity for written argumentative text based on historical sources, it would be interesting to go deeper into this matter by analysing different AI texts. In addition, expanding the information provided to the AI during the development of argumentative texts in accordance with historical thinking could lead us to learn more about the potentialities and shortcomings of AI productions in terms of this historical competency. In line with this, it would also be interesting to analyse the texts created by chatbots other than ChatGPT and its underlying language model GPT.

On the other hand, expanding the scope of the study by including a larger number of participants from different educational levels could allow for focused investigations into variables such as gender and comparisons between participants from various educational contexts. These could lead to relevant results and conclusions for the educational community. Finally, it would be beneficial to conduct studies where participants directly interact with the AI and its generated texts. This hands-on engagement would provide students with a deeper understanding of the possibilities of AI in relation to historical thinking and its dimensions.

Author Contributions: Conceptualization, S.T.-O., M.N.-I. and R.C.-G.; methodology, S.T.-O. and R.C.-G.; software, S.T.-O., M.N.-I. and P.O.-J.; formal analysis, S.T.-O. and P.O.-J.; investigation, S.T.-O. and R.C.-G.; writing—original draft preparation, S.T.-O. and M.N.-I.; writing—review and editing, S.T.-O., M.N.-I., P.O.-J. and R.C.-G.; visualization, S.T.-O., M.N.-I., P.O.-J. and R.C.-G.; supervision, R.C.-G.; project administration, R.C.-G.; funding acquisition, R.C.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR under grant TED2021-131557B-I00, by the Ministry of Education, Culture and Sports of Spain under grant FPU20/02375, and by the University of Castilla-La Mancha under grant 2022-GRIN-34039.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Social Research Ethics Committee of the University of Castilla-La Mancha (CEIS-632710-Z1N4 / Date 7 April 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank all students who agreed to participate in the present study.

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

References

1. Yildirim, G.; Elban, M.; Yildirim, S. Analysis of Use of Virtual Reality Technologies in History Education: A Case Study. *Asian J. Educ. Train.* **2018**, *4*, 62–69. [CrossRef]
2. Chinn, C.A.; Barzilai, S.; Duncan, R.G. Education for a "post-truth" world: New directions for research and practice. *Educ. Res.* **2021**, *50*, 51–60. [CrossRef]
3. Kozyreva, A.; Wineburg, S.; Lewandowsky, S.; Hertwig, R. Critical ignoring as a core competence for digital citizens. *Curr. Dir. Psychol. Sci.* **2023**, *32*, 81–88. [CrossRef]
4. Cooper, H. Why are there no history text books in english primary schools? *Ensayos. Rev. Fac. Educ. Albacete* **2014**, *29*, 27–42.
5. Gómez-Carrasco, C.J.; Miralles-Martínez, P. ¿Pensar históricamente o memorizar el pasado? La evaluación de los contenidos históricos en la educación obligatoria en España. [Thinking Historically or Memorizing the Past? Assessing Historical Content in Compulsory Education in Spain]. *Rev. Estud. Soc.* **2014**, *52*, 52–68. [CrossRef]
6. Dale, R. GPT-3. What's it good for? *Nat. Lang. Eng.* **2021**, *27*, 113–118. [CrossRef]
7. Lund, B.D.; Wang, T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Libr. Hi Tech News* **2023**, *40*, 26–29. [CrossRef]
8. Kidd, C.; Birhane, A. How AI can distort human beliefs. *Science* **2023**, *380*, 1222–1223. [CrossRef]
9. Cunningham-Nelson, S.; Boles, W.; Trouton, L.; Margerison, E. A review of chatbots in education: Practical steps forward. In Proceedings of the 30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate. Motivate, Brisbane, Australia, 8–11 December 2019; Engineers Australia: Brisbane, Australia, 2019; pp. 299–306.
10. Sinha, S.; Basak, S.; Dey, Y.; Mondal, A. An educational chatbot for answering queries. In *Emerging Technology in Modelling and Graphics*; Mandal, J.K., Bhattacharya, D., Eds.; Springer: Singapore, 2020; pp. 55–60. [CrossRef]
11. Raja, R.; Nagasubramani, P.C. Impact of modern technology in education. *J. Appl. Adv. Res.* **2018**, *3*, 33–35. [CrossRef]
12. Scherer, R.; Siddiq, F.; Tondeur, J. The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Comput. Educ.* **2019**, *128*, 13–35. [CrossRef]
13. Redecker, C. *European Framework for the Digital Competence of Educators: DigCompEdu*; Publications Office of the European Union: Luxemburg, 2017. [CrossRef]
14. Pelletier, K.; McCormack, M.; Reeves, J.; Robert, J.; Arbino, N.; Al-Freih, M.; Dickson-Deane, C.; Guevara, C.; Koster, L.; Sánchez-Mendiola, M.; et al. *2022 EDUCAUSE Horizont Report. Teaching and Learning Edition*; EDUCAUSE Publications: Boulder, CO, USA, 2022.
15. Ennals, R. History Teaching and Artificial Intelligence. *Teach. Hist.* **1982**, *33*, 3–5.
16. Yazdani, M.; Lawler, R.W. Artificial intelligence and education: An overview. *Instr. Sci.* **1986**, *14*, 197–206. [CrossRef]
17. Roth, G.; McEwing, R. Artificial Intelligence and Vocational Education: An Impending Confluence. *Educ. Horiz.* **1986**, *65*, 45–47.
18. Winston, P.H. *Artificial Intelligence*, 3rd ed.; Addison-Wesley: Boston, MA, USA, 1984.
19. Rouhiainen, L. *Inteligencia Artificial. 101 Cosas Que Debes Saber Hoy Sobre Nuestro Futuro*, 1st ed.; Alienta Editorial: Barcelona, Spain, 2018.
20. King, M.R. The future of AI in medicine: A perspective from a chatbot. *Ann. Biomed. Eng.* **2022**, *51*, 291–295. [CrossRef]
21. Pavlik, J.V. Collaborating with ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *J. Mass Commun. Educ.* **2023**, *78*, 84–93. [CrossRef]
22. Budzianowski, P.; Vulić, I. Hello, It's GPT-2-How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, China, 4 November 2019; Birch, A., Finch, A., Hayashi, H., Konstas, I., Luong, T., Neubig, G., Oda, Y., Sudoh, K., Eds.; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 15–22.
23. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. *preprint*. Available online: https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=0CAIQw7AJahcKEwjwzb2hicCAAxUAAAAAHQAAAAAQAg&url=https%3A%2F%2Fs3-us-west-2.amazonaws.com%2Fopenai-assets%2Fresearch-covers%2Flanguage-unsupervised%2Flanguage_understanding_paper.pdf&psig=AOvVaw3b9TGDzA8vqLvSx7Uj3L3N&ust=1691137695453108&opi=89978449 (accessed on 16 April 2023).
24. OpenAI. Introducing ChatGPT 2022. Available online: https://openai.com/blog/chatgpt (accessed on 3 June 2023).
25. Murphy, R.F. *Artificial Intelligence Applications to Support K–12 Teachers and Teaching: A Review of Promising Applications, Opportunities, and Challenges*; RAND Corporation: Santa Monica, CA, USA, 2019. [CrossRef]
26. Okonkwo, C.W.; Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100033. [CrossRef]
27. Xie, H.H.; Hwang, G.; Wong, T. Editorial Note: From Conventional AI to Modern AI in Education: Re-examining AI and Analytic Techniques for Teaching and Learning. *Educ. Technol. Soc* **2021**, *24*, 85–88.
28. Lin, M.P.C.; Chang, D. Enhancing post-secondary writers' writing skills with a chatbot. *Educ. Technol. Soc.* **2020**, *23*, 78–92.
29. Murad, D.F.; Irsan, M.; Akhirianto, P.M.; Fernando, E.; Murad, S.A.; Wijaya, M.H. Learning support system using chatbot in "kejar c package" homeschooling program. In Proceedings of the 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 24–25 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 32–37. [CrossRef]

30. Troussas, C.; Krouska, A.; Virvou, M. Integrating an adjusted conversational agent into a mobile-assisted language learning application. In Proceedings of the 2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI), Boston, MA, USA, 6–9 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1153–1157. [CrossRef]
31. Okonkwo, C.W.; Ade-Ibijola, A. Python-bot: A chatbot for teaching python programming. *Eng. Lett.* **2020**, *29*, 25–34.
32. Thomas, H. Critical literature review on chatbots in education. *Int. J. Trend Sci. Res. Dev.* **2020**, *4*, 786–788.
33. Durall, E.; Kapros, E. Co-design for a competency self-assessment chatbot and survey in science education. In Proceedings of the International Conference on Human-Computer Interaction, Learning and Collaboration Technologies. Human and Technology Ecosystems, Copenhagen, Denmark, 19–24 July 2020; Zaphiris, P., Ioannou, A., Eds.; Springer: Cham, Switzerland, 2020. [CrossRef]
34. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; Wu, Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597.
35. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; de Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef]
36. Mitrović, S.; Andreoletti, D.; y Ayoub, O. ChatGPT or Human? Detect and explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. *arXiv* **2023**, arXiv:2301.13852.
37. Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing scientific abstracts generated by ChatGPT to original abstracts with detectors and blinded human reviewers. *NPJ Digit. Med.* **2023**, *6*, 75. [CrossRef] [PubMed]
38. Tirado-Olivares, S.; Cózar-Gutiérrez, R.; López-Fernández, C.; González-Calero, J.A. Training future primary teachers in historical thinking through error-based learning and learning analytics. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 44. [CrossRef]
39. Gómez Carrasco, C.J.; Rodríguez-Medina, J.; Chaparro Sainz, Á.; Alonso García, S. Digital resources and teaching approaches in preservice training of History teachers. *Educación XX1* **2022**, *25*, 143–170. [CrossRef]
40. Gómez-Carrasco, C.J.; Miralles-Martínez, P. *Los Espejos de Clío. Usos y Abusos de la Historia en el Ámbito Escolar*; Sílex: Madrid, Spain, 2017.
41. Seixas, P.; Morton, T. *The Big Six Historical Thinking Concepts*; Nelson: Toronto, ON, Canada, 2013.
42. Levstik, L.; Barton, K. *Doing History. Investigating with Children in Elementary and Middle Schools*; Routledge: New York, NY, USA, 2015.
43. Bauman, Z. *Modernidad Líquida*, 1st ed.; Fondo de Cultura Económica: Buenos Aires, Argentina, 2003.
44. Moreno-Vera, J.R. Epilogue: "We wanna learn like common people whatever common people did". In *Re-Imagining the Teaching of European History. Promoting Civic Education and Historical Consciousness*; Gómez-Carrasco, C.J., Ed.; Routledge: New York, NY, USA, 2023; pp. 219–223.
45. Kesternich, I.; Siflinger, B.; Smith, J.P.; Winter, J.K. The effects of World War II on economic and health outcomes across Europe. *Rev. Econ. Stat.* **2014**, *96*, 103–118. [CrossRef]
46. Sáiz, J.; Gómez, C.J. Investigar el pensamiento histórico y narrativo en la formación del profesorado: Fundamentos teóricos y metodológicos. [Researching historical and narrative thinking in teacher education: Theoretical and methodological basis]. *Revocation Elect. Int. For. Prof.* **2016**, *19*, 175–190. [CrossRef]
47. Bisquerra, R. *Metodología de la Investigación Educativa*; La Muralla: Madrid, Spain, 2004.
48. Cotton, D.R.E.; Cotton, P.A.; Shipway, R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**, 1–12. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.