

You've Got a Friend in ... a Language Model? A Comparison of Explanations of Multiple-Choice Items of Reading Comprehension between ChatGPT and Humans

George Dueñas¹, Sergio Jimenez², Geral Eduardo Mateus Ferro³

¹Doctorado Interinstitucional en Educación, Universidad Pedagógica Nacional, Colombia

²Instituto Caro y Cuervo, Colombia

³Departamento de Lenguas, Universidad Pedagógica Nacional, Colombia
geduenasl@upn.edu.co, sergio.jimenez@caroycuervo.gov.co, gmateus@pedagogica.edu.co

Abstract

Creating high-quality multiple-choice items requires careful attention to several factors, including ensuring that there is only one correct option, that options are independent of each other, that there is no overlap between options, and that each option is plausible. This attention is reflected in the explanations provided by human item-writers for each option. This study aimed to compare the creation of explanations of multiple-choice item options for reading comprehension by ChatGPT with those created by humans. We used two context-dependent multiple-choice item sets created based on Evidence-Centered Design. Results indicate that ChatGPT is capable of producing explanations with different type of information that are comparable to those created by humans. So that humans could benefit from additional information given to enhance their explanations. We conclude that ChatGPT ability to generate explanations for multiple-choice item options in reading comprehension tests is comparable to that of humans.

1 Introduction

Chatbots are used in education because they “promise to have a significant positive impact on learning success and student satisfaction” and “are promising tools to provide continuing feedback to lecturers and students” (Winkler and Söllner, 2018). According to Wollny et al. (2021), chatbots have been utilized in education to support learning and teaching, enhance services offered by educational institutions, promote well-being, and offer feedback and motivation. However, their use in assisting with the development of evaluation items, whether formative or summative, has not been widely explored.

The process of providing explanations for multiple-choice items can be more time-consuming and labor-intensive than constructing the item itself. This often results in the creation of numerous items

that lack explanations. Despite this, it is important to note that an item accompanied by explanations is significantly more versatile and useful than one without. Furthermore, the process of constructing explanations can reveal issues with the items that may not have been immediately apparent. As such, the implementation of a tool to assist item constructors in developing explanations could greatly enhance both the quantity and quality of items produced. Recent advancements in language models, such as ChatGPT, which have been trained on large amounts of text, show promise in their ability to assist with this task. This paper aims to investigate the efficacy of these models in comparison to explanations generated by humans.

2 Background

2.1 Explanations in multiple-choice items

Haladyna et al. (2002) proposed 31 multiple-choice item-writing guidelines focused on classroom assessment, but it can be applied to items used in other circumstances. They grouped these guidelines in five categories: Content concerns, Formatting concerns, Style concerns, Writing the stem, and Writing the choices. The last category is the most extensive one with 14 aspects, and in it, there are three aspects that are directly related to the explanation of the options: “Make sure that only one of these choices is the right answer”, “Keep choices independent”; choices should not be overlapping, and “Make all distractors plausible”. The provision of detailed explanations for each option is crucial in ensuring that only one option is unequivocally correct, while the remaining options contain inaccurate information that may appear plausible at first glance. On the other hand, they can be considered as a type of feedback (see Hattie, 2012, chap. 7) for students who require it, since the explanation for each option should include the reason why it is either correct or incorrect.

2.2 *Evaluar para Avanzar*

In Colombia, in 2020, the formative evaluation strategy called *Evaluar para Avanzar*¹ (EpA) was created by the Ministry of Education (MEN) and the ICFES (Colombian Institute for the Evaluation of Education). The aim of this strategy is to face the challenges of the COVID-19 pandemic by contributing to the classroom evaluations for students in grades 3 through 11 by means of complementary diagnostic instruments to the standardized tests. The assessment consists of two booklets per grade level for each academic year, each containing 20 items. For this study we selected the areas of Reading and Critical Reading for 5°, 9°, and 11° grades (see section 3.1). These selections correspond to the years 2021 and 2022 and were chosen because they contained only text-based items (MEN, 2006).

The framework used by ICFES to write these items is Evidence-Centered Design (ECD) (Mislevy et al., 2003, 2017). This means that the EpA items show information about the claims and evidences in the items. In this case, this type of information is the same as the *Saber*² standardized tests. The multiple-choice items in *Saber* 3°, 5°, 7°, and 9° have three *claims*³: (i) retrieve literal information expressed in fragments of the text, (ii) understand the local and global meaning of the text through inferences of implicit information, (iii) take a critical stance on the text by evaluating its form and content (Jurado and Rodríguez, 2020).

An example for the first claim (i) is item 1 of the 2022 grade 5 booklet, which reads as follows: *Según el texto, ¿en qué momento ocurre la historia?* [According to the text, when does the story take place?]

- A. Ocurre en este instante. [It happens right now.]
- *B. Ocurrió en un tiempo lejano. [It happened in a distant time.]
- C. Ocurrió hace poco. [It happened recently.]
- D. Ocurrirá luego. [It will happen later.]

The test-taker must locate explicit information related to time that allows them to know when the

¹<https://www2.icfes.gov.co/en/caja-de-herramientas1>

²<https://www.icfes.gov.co/es/web/guest/evaluaciones>

³It is a statement we'd like to be able to make about what a student knows or can do on the basis of observations in an assessment setting (Mislevy et al., 2003).

events described in the story occurred. The context begins with the expression “A long time ago...”, indicating that the events being described occurred in the distant past. The example in Table 2 is related to claim (ii), since the test-taker must deduce the meaning of a certain expression according to the given context.

In the case of *Saber* 11°, it also has three *claims*: (i) identify and understand the local contents that make up a text, (ii) understand how the parts of a text are articulated to give it a global meaning, (iii) reflect from a text and evaluate its content (Donoso, 2021). Therefore, each item contains the following information: a claim, an evidence, correct and incorrect options, as well as explanations for both the correct and incorrect options.

Regarding the difficulty of the items, they can be ranked in the following way: the items in the first claim would be *easy*, as they require retrieving explicit information from the contexts; the items in the second claim would be *intermediate*, as they require making inferences based on the information from the contexts; and finally, the items in the third claim would be *difficult*, as they require evaluating the information from the contexts.

2.3 ChatGPT on items and related

In November 2022, OpenAI released ChatGPT, which is a general-purpose language model “trained to follow an instruction in a prompt and provide a detailed response”⁴. This tool can write texts that are human-like and, at the same time, can “understand” the instructions it receives in some natural languages such as Spanish. So anyone, classroom teachers, learners and/or test developers, can use this artificial intelligence by creating an account at <https://chat.openai.com/>. Then all we have to do is ask questions and wait for this tool to generate answers and explanations, as simple as having a conversation. This methodology has been employed by researchers who have tasked ChatGPT with answering test items in order to evaluate its performance. Some examples come from areas such as law (Choi et al., 2023), medicine (Saraju et al., 2023; Gilson et al., 2023; Kung et al., 2023; Fijačko et al., 2023), among others (Guo et al., 2023; OpenAI, 2023).

Choi et al. (2023) used ChatGPT to produce answers to multiple choice and essay questions of four separate final exams for law school courses:

⁴<https://openai.com/blog/chatgpt>

Constitutional Law; Federalism and Separation of Powers; Employee Benefits and Taxation; and Torts. The AI-generated answers were shuffled with student exams and graded blindly by three professors. They concluded that ChatGPT passed all exams and performed better on the essay components than on the multiple choice. [Sarraju et al. \(2023\)](#) created 25 open-ended questions addressing fundamental preventive concepts related to Cardiovascular Disease Prevention. Each question was given to ChatGPT 3 times and the responses were recorded. Then 3 reviewers graded each set of responses as “appropriate”, “inappropriate” or “unreliable”. ChatGPT gave 21 appropriate and 4 inappropriate answers. The authors observed great potential for interactive AI to assist clinical workflows (increased patient education and ease of patient-clinician communication), but these approaches must be further explored. [Fijačko et al. \(2023\)](#) given to ChatGPT 96 stand-alone and 30 scenario-based questions related to life support exams (BLS and ACLS). The authors concluded that although ChatGPT did not pass any of these exams, it has the potential to be a valuable resource for studying and preparing for life support exams.

In addition to evaluating the performance of ChatGPT, some studies have also assessed the explanations generated for its responses to test items. Our research aims to replicate this approach. [Gilson et al. \(2023\)](#) and [Kung et al. \(2023\)](#) analyzed the response explanations for user interpretability of the United States Medical Licensing Examination exams. The authors demonstrated that ChatGPT attained a score that is comparable to that of a third-year medical student and that its performance was either at or near the passing threshold for all the exams. Finally, [Guo et al. \(2023\)](#) evaluated, among other aspects, the answers created by ChatGPT with nearly 40K questions written in English and Chinese, and their corresponding answers created by human experts, creating the Human ChatGPT Comparison Corpus (HC3) dataset, coming from domains such as computer science, finance, medicine, law, psychology), and open-domain. The authors conclude that it is easier to distinguish the content generated by ChatGPT when an answer is provided for comparison than when the answer is provided alone and that those answers are considered more useful than those of humans.

Our objective is to compare the explanations generated by ChatGPT in Spanish for multiple-

choice reading comprehension items with those created by human item-writers (hereafter referred to as ‘humans’). As [Guo et al. \(2023, p. 2\)](#), we also want to know if ChatGPT can be “honest (not fabricate information or mislead the user), harmless (shouldn’t generate harmful or offensive content), and helpful (provide concrete and correct [item explanations])” to the humans.

We also want to evaluate whether ChatGPT can classify these items into one of the three *claims* used to build EpA, which are based on ECD.

In a similar way to the works of [Gilson et al. \(2023\)](#) and [Kung et al. \(2023\)](#), our work aims to provide qualitative and quantitative feedback on the performance of ChatGPT and assess its potential to help classroom teachers, learners, and test developers. To the best of our knowledge, no existing research has compared the explanations generated by ChatGPT with those created by humans for multiple-choice reading comprehension items. Given that our items span multiple school grades and text types in Spanish, we believe that this presents a unique and challenging opportunity to evaluate the capabilities of ChatGPT.

The rest of the paper is organized as follows. In Section 3 we provide a detailed description of the method and data used. In Section 4 we present and discuss the main results. Finally, in Section 5 we provide some conclusions and perspectives.

3 Methods

3.1 Data

The data consists of a set of human-written textual explanations for each of the item options from grades 5°, 9°, and 11° of the years 2021-2 and 2022-1 of EpA strategy respectively. We supplied ChatGPT with the context associated with the items and prompted it using natural language to generate responses and explanations for the corresponding items and options. The motivation for choosing these school grades is twofold: same grade level and consecutive booklets, but from different years, and each grade is the completion of a cycle in the Colombian educational system. In 5°, basic primary education is completed; in 9°, secondary basic education is completed, and in 11°, secondary education is completed. The latter is the one that allows a student to enter higher education.

Each grade level is accompanied by a booklet containing 20 items, which are organized into context-dependent groups of either 3 or 5 items per

context. Items that depend on an image context have been discarded, as ChatGPT does not support this particular format. Table 1 shows the different school grades, the types of contexts, the length of each context, and the respective assigned items. Thus, there are 46 items in 2021, and 42 in 2022, making a total of 88 items. Items grouped by claim (i.e., difficulty), year, and grade, according to ECD proposed by ICFES, are below. The underlined items were subsequently discarded due to the fact that the responses provided by ChatGPT did not align with the established answers in the booklets.

- **2021, 5°:** *claim 1:* 1, 2, 4, 8, 12; *claim 2:* 6, 9, 13, 15; *claim 3:* 3, 5, 7, 10, 11, 14.
- **2021, 9°:** *claim 1:* 7, 10, 11, 12; *claim 2:* 8, 9, 13, 15, 18; *claim 3:* 6, 14, 16, 17, 19, 20.
- **2021, 11°:** *claim 1:* 8, 14, 16, 19; *claim 2:* 1, 2, 3, 6, 13, 18, 20; *claim 3:* 4, 5, 7, 15, 17.
- **2022, 5°:** *claim 1:* 1, 2, 5, 9, 17; *claim 2:* 3, 4, 6, 10, 11, 12, 18; *claim 3:* 7, 8, 19, 20.
- **2022, 9°:** *claim 1:* 1, 2, 4, 11; *claim 2:* 3, 5, 12, 13, 14; *claim 3:* 15.
- **2022, 11°:** *claim 1:* 2, 3, 13, 18; *claim 2:* 8, 9, 10, 15, 16, 17, 19; *claim 3:* 1, 11, 12, 14, 20.

3.2 Data Extraction

As the booklets with the contexts and items are public and available in PDF files, we manually copied each context and its respective items, and pasted them into a plain text file. Subsequently, we checked that the texts matched, since sometimes the texts copied from the PDF file pasted with errors in some characters. Afterwards, the explanation for each incorrect option was manually extracted, since these explanations were grouped together in a single paragraph. When the paragraph began with the following statement: “The options X, Y, and Z are not correct, because...” (where X, Y, and Z represent the incorrect options), that part was added to each of the explanations of the three options. The motivation behind this was to expand the explanation created by humans to compare each of these explanations with its corresponding one created by ChatGPT. Finally, this same information was pasted back into a spreadsheet⁵, where the

⁵<https://docs.google.com/spreadsheets/d/1CTXEJn0dT-4xzUYrwZZJDPMe-XnyvAHivYgPxG4ejCY/edit?usp=sharing>

information created by ChatGPT was also added.

As to ChatGPT, we collected explanations for each option between December 2022 and January 2023. In those months, ChatGPT was only available through its website, so we collected the information as follows:

1. The item context is copied and pasted into the input box, and the explanation given was omitted. We used a chat session for each context and its respective items so that the ChatGPT memory retention function only takes into account context-related and item-related information. This was done because, like humans, when constructing items, they take into account what has been said in the context and in the other items in order to fulfill what Haladyna et al. (2002) have outlined.
2. The first multiple-choice item is copied and pasted into the input box. The answer and the explanation are saved.
3. ChatGPT is asked the following: “In which of the following categories would the above question be classified?”, where the categories are the *claims* used to build EpA. The answer and the explanation are saved.
4. Step 2 is performed again⁶ to subsequently ask ChatGPT the following: “Why is option X incorrect?”, where X corresponds to each of the incorrect options of the respective item.
5. The next multiple-choice item is copied and pasted into the input box. The answer and the explanation are saved.

3.3 Comparing explanations

Given that the explanations authored by humans have been publicly available since 2021, no experiment was conducted to ascertain which explanation - human-authored or ChatGPT-generated - was deemed more suitable by teachers or other individuals. Instead, for each pair of explanations per option for the items, a manual review was carried out to identify the differences and similarities between them as long as ChatGPT selects the correct option (the key) according to the one established in

⁶This step must be done again because ChatGPT is susceptible to the immediately preceding text, so if asked why X option is incorrect, its response will be based on the question of the three categories. ChatGPT again selected the same option for each item.

2021-2					2022-1				
Gr	Text Type	Words	Items	Total	Gr	Text Type	Words	Items	Total
5	expository	274	1-5		5	narrative	296	1-4	
5	narrative	337	6-10	15	5	descriptive	311	5-8	
5	expository	344	11-15		5	expository	225	9-12	16
9	narrative	373	6-11		5	expository	203	17-20	
9	descriptive	108	12-15	15	9	narrative	208	1-5	
9	descriptive	291	16-20		9	narrative	269	11-15	10
11	expository	392	1-4		11	narrative	490	1-3	
11	argumentative	180	5-8		11	argumentative	264	8-11	
11	narrative	143	13-16	16	11	argumentative	302	12-16	16
11	narrative	420	17-20		11	argumentative	512	17-20	
				46					42

Table 1: Types of (con)texts by grade and year, length of each (con)text, and their respective assigned items. Gr stand for the school grade.

the booklet. The review was performed by the first author of this study, a linguist and native Spanish speaker with experience in writing reading comprehension items for national tests in Colombia.

This review involves a comparative analysis of each pair of explanations associated with each item option, irrespective of whether the option is correct or incorrect. To do that, six colored tags have been created to annotate the differences and similarities between the explanations. The explanations of the underlined items above were not compared because the answer given by ChatGPT did not match with that established in the booklets. The first tag, denoted as MIC and colored red, highlights text passages that explain the respective option and where there is agreement between humans and ChatGPT - that is, the passages convey the same meaning despite being phrased differently.

The second tag (AII) is colored green. This text has been created by humans and it is generally used to refer to part or all of the context expanding its explanation or providing additional information (such as the function of words, punctuation marks, or titles in a text). The meaning of this text does not have a match with any part of the explanation created by ChatGPT.

The third tag (AIC) is colored blue. This text has been created by ChatGPT and it is generally used to refer to part or all of the context expanding its explanation or providing additional information (such as the function of words, punctuation marks, or titles in a text). The meaning of this text does not have a match with any part of the explanation created by humans.

The fourth tag (EIC) is colored gray, where the texts highlighted intend to expand the explanations by making use of or referring to some part of the Context. In this texts, humans and ChatGPT have matched, that is, they have the same meaning although written differently.

The fifth tag (CI) is colored brown. This text has been created by humans to close the option explanation in an individually or generally (when it is the last option) way. One way to identify the closure is by using connectors such as *therefore*, *thus*, or *so* (*por lo tanto* or *por tanto* in Spanish).

The sixth tag (CC) is colored fuchsia. This text has been created by ChatGPT to close the option explanation. In the case of ChatGPT, it adds or repeats the reason why the option is (in)correct, but with other words. One way to identify the closure is by using connectors such as *therefore*, *thus*, or *so* (*por lo tanto* or *por tanto* in Spanish).

Table 2 shows an example of the explanations labeled by option from their comparison for item 6 of grade 5 of 2022-1 including a Venn diagram for visualizing color tags.

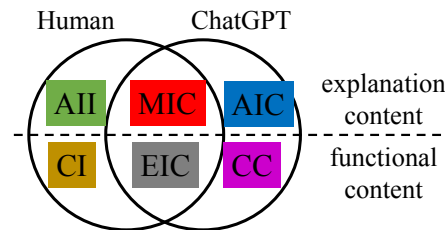
4 Results and Discussion

We utilized a traditional test approach of simply presenting the item to ChatGPT. Out of the total 88 items, ChatGPT answered 78 correctly, resulting in the exclusion of 10 items. Table 3 shows the average length of explanations in number of words per color tag for correct and incorrect answers, by school grade, and by humans and ChatGPT. It also includes effect sizes and p-values of the differences.

An analysis of the common information pro-

What does the author refer to when using the word “wonders” in the last paragraph of the text?

- A. That the tools were very old.
- *B. To the ability of hunters to build effective tools.
- C. To the ease with which hunters caught mastodons.
- D. That the tools were very beautiful.



Human explanation	ChatGPT explanation
<p>A. The attribution of wonders is not given to the tools for their antiquity. The value is given to the ability of the hunters and the effectiveness of the tools, for this reason, the options A, C and D are not valid.</p>	<p>The option A, “That the tools were ‘very old’”, is incorrect as an answer to the item ... In the last paragraph of the text, the author mentions that ... These tools were “wonders”, suggesting they were very good or effective ... Instead, the option A suggests that the tools were very old, which ... Therefore, the option A is incorrect as an answer to the item.</p>
<p>*B. In this case, the item requires students to deduce what the text attributes the character of “wonders” in the text, according to the statement ... that appears at the end of the news. The semantic feature of wonders does not refer only to the tools but also to the ability of hunters to build effective tools, therefore, the correct option is B.</p>	<p>The author refers to the ability of hunters to build effective tools when using the word “wonders” in the last paragraph of the text. In the last paragraph of the text, the author mentions that ... These tools were “wonders”, suggesting that they were very good or effective ... Therefore, it is possible to infer that the author uses the word “wonders” ...</p>
<p>C. The attribution of wonders is not given to the tools for the ease with which hunters caught mastodons. The value is given to the ability of the hunters and the effectiveness of the tools, for this reason, the options A, C y D are not valid.</p>	<p>The option C, “To the ease with which hunters caught mastodons”, is incorrect as an answer to the item ... In the last paragraph of the text, the author mentions that ... These tools were “wonders”, suggesting that they were very good or effective ... Instead, the option C suggests that hunters caught mastodons easily, which ... Therefore, the option C is incorrect as an answer to the item.</p>
<p>D. The attribution of wonders is not given to the tools because the tools were beautiful. The value is given to the ability of the hunters and the effectiveness of the tools, therefore, the options A, C y D are not valid.</p>	<p>The option D, “That the tools were very beautiful”, is incorrect as an answer ... In the last paragraph of the text, the author mentions that ... These tools were “wonders”, suggesting that they were very good or effective ... Instead, the option D suggests that the tools were very beautiful, which ... Therefore, the option D is incorrect as an answer to the item.</p>

Table 2: Tagging example of the explanations for item 6 of 5th grade in 2022-1 (texts translated from Spanish).

vided in explanations by humans and ChatGPT (MIC and EIC tags) reveals that ChatGPT’s writing tends to be more verbose than that of humans in all scenarios except for 9th grade in the MIC tag. The test results presented in the penultimate row of Table 3 indicate a significant difference in

the length of prose between humans and ChatGPT when expressing the same content. However, the effect size of this difference is small. With respect to these differences in additional explanations (last row), ChatGPT texts are considerably larger than those of humans, exhibiting a large effect size.

The length of the explanations can be attributed to two factors: firstly, humans tend to be direct in their responses and may not offer additional information, as evident from the minimal or non-existent explanations provided by 9th and 11th grades for the AII tag. Furthermore, the results for the CI tag suggest that humans do not provide closure explanations. Secondly, ChatGPT explanations tend to be longer due to various factors, such as repetition the option, quoting fragments of the context, reiterating the status of the option (correct or incorrect) or all of the above. Thus, ChatGPT performed information expansion for almost half of the explanations (as seen in the AIC tag) and added a closure for one-fifth of the explanations.

When comparing the length of explanations for correct vs. incorrect options the only significant differences were observed in AII, AIC and CI tags. This let us conclude that humans prefer to provide additional explanations to correct options, while ChatGPT does it to incorrect options, both with a medium effect size in the difference of their respective preferences. Similarly, for the functional content, humans preserve such preference, while ChatGPT do the same but not significantly.

When comparing the length of explanations for correct versus incorrect options, significant differences were observed only in the AII, AIC, and CI tags. This allows us to conclude that humans prefer to provide additional explanations for correct options, while ChatGPT does so for incorrect options, both with a medium effect size in the difference of their respective preferences. Similarly, for functional content, humans maintain this preference, while ChatGPT does the same but not significantly. More importantly, both humans and ChatGPT agree on not having differences in the length of the main common explanations for correct or incorrect options.

Another important factor is the ability of ChatGPT of identifying the correct option. Among the 31 items to grade 5, ChatGPT fails to provide a correct answer for item 20 in the 2022 dataset. In its explanation for the four options, ChatGPT notes that “There has been no mention of option X being wrong at any point”, and gives an explanation for each option. Although ChatGPT did not choose the correct option for this item, its explanations were compared to determine their accuracy. ChatGPT considered that each incorrect option (distractor) could be correct. However, upon comparing these

explanations to those created by humans, it was determined that ChatGPT’s explanations were incorrect. A more extensive discussion of these results exceeds the scope of this paper and is left for future research as the information is available in the spreadsheet. As a result, this item was deemed invalid and excluded from the analysis.

For the remaining 30 items, it can be seen that the explanation given by humans and ChatGPT have matched in the meaning, but with different words (MIC). Regarding the tags AII and AIC, it is evident that AIC is more commonly used. This suggests that ChatGPT often includes additional information to provide a more comprehensive explanation, based on either a specific portion or the entire context. Something similar occurs with the CI and CC tags, where the latter occurs slightly more frequently than the former. It is evident that approximately one-third of the explanations for the correct options have a concluding statement provided by both humans and ChatGPT. Finally, regarding the EIC label, humans and ChatGPT explanations rarely coincide and expand by referencing parts of the context. Similar to the previous findings, the comparison of explanations by grade, tag, and correctness reveals that ChatGPT tends to provide more information than humans, albeit not uniformly for all options. Another factor in which ChatGPT may fail is its inability to affirm that other options are incorrect, even when it has correctly chosen the correct option. In item 14 of grade 5 from 2021, ChatGPT indicates that “Option A is not incorrect” and provides an explanation to support this assertion.

Among the 25 items to grade 9, ChatGPT fails to provide a correct answer for item 15 in the 2021 and item 1 in the 2022. Regarding item 15, ChatGPT selected option B, which is incorrect. This may be because ChatGPT omitted or confused some words. The text states that “Simone reflects on her own life as a woman and after this reflection, publishes the book *The Second Sex*”, but as the human who built this item says in his/her explanation: “[said quote] is not a reflection on sex”. Regarding item 1, ChatGPT did not select any of the four options as correct and in its explanations it stated that the word whose meaning is contrary to the word “illegal” is “legal”, thus omitting the correct option: “allowed”. This may be because ChatGPT discarded the presented context in which the word “allowed” (permitted) fits as the semantic opposite

Grade	Option	Explanation content				Functional content			
		Human		ChatGPT		Human		ChatGPT	
		AII	MIC	MIC	AIC	CI	EIC	EIC	CC
5°	correct	10(17)	38(22)	46(22)	39(28)	3(5)	1(3)	2(6)	7(10)
5°	incorrect	4(11)	37(14)	46(21)	43(34)	1(3)	1(3)	6(3)	9(11)
9°	correct	2(7)	59(22)	56(25)	20(24)	0(0)	1(7)	3(14)	17(23)
9°	incorrect	0(0)	62(16)	54(18)	23(28)	0(8)	0(0)	0(0)	10(9)
11°	correct	1(3)	48(23)	62(28)	15(21)	0(0)	3(10)	3(12)	18(13)
11°	incorrect	3(9)	51(15)	53(24)	34(25)	0(3)	1(3)	1(5)	9(10)
average	correct	5(12)	48(24)	54(25)	25(27)	1(3)	2(7)	3(11)	13(16)
average	incorrect	3(9)	49(18)	51(21)	34(30)	0(2)	0(2)	0(3)	9(10)
	Effect size correct vs. incorrect†	0.456	0.547	0.460	0.577	0.463	0.481	0.481	0.483
	p-value	0.034	0.222	0.291	0.039	0.017	0.126	0.120	0.089
average	both	3(10)	49(20)	52(22)	32(30)	1(3)	1(4)	1(6)	10(12)
	Effect size of diff. [p-value]‡‡		0.111	[0.047]			0.170	[0.004]	
	Effect size of diff. [p-value]‡‡		-0.717	[<.001]			-0.642	[<.001]	

†Effect size for Wilcoxon test calculated as $\frac{z}{\sqrt{N}}$

‡‡Effect size for Mann-Whitney test calculated as $\frac{U}{n_1 \times n_2}$

Table 3: Average (STD) number of words per color tag (significant differences having $p < 0.05$ showed in boldface).

of the word “illegal”. In item 20 of grade 9 of 2021, ChatGPT states for options A and C that “There is not enough information in the text to determine if the option . . . is incorrect or not. The text does not explicitly mention who it is intended for”.

Among the 36 items to grade 11, ChatGPT fails to provide a correct answer for 5 items of 2021 (1, 14, 16, 17, 20) and 2 items of 2022 (12, 20). In item 1, ChatGPT selected option A, which is incorrect. This option asserts that the argumentative relationship between the two presented statements is one of premise and evidence, while the correct option provided by the human is D, which asserts that the relationship is conjecture and counterevidence. Regarding the remaining items, items 14 and 16 belong to the same context, and their responses are derived from the same fragment. Furthermore, these items ask for information that is explicit in the text (claim 1), making it uncommon for ChatGPT to provide incorrect answers. Similarly, items 17 and 20 are associated with the same context (but different from before), although their answers are derived from different fragments, and they pertain to different claims (2 and 3, respectively). In general terms, ChatGPT could have provided incorrect responses, but the comparison of these types of explanations precisely calls for a more in-depth analysis, which falls outside the scope of this work.

In three items of the grade 11, ChatGPT correctly

identified the correct option but did not provide a conclusive explanation for why the other options were incorrect. In item 13 of 2021, ChatGPT indicates that “Option X is not incorrect” and adds that “this option does seem to be a reason for the character’s feeling of unease”. This type of explanation may be due to the format of the question, which uses a negation structure: “Which NOT”. For items 10 and 11 of 2022, which are based on a fragment of HAMLET’S MONOLOGUE, ChatGPT provided explanations that diverged from the expected pattern. For item 10, ChatGPT stated that both options A and C “is not incorrect, but rather one of the options that can adequately describe the above text”. The explanation then goes on to clarify why the adjective “philosophical” (option A) or “poetic” (option C) would be a better fit for the fragment. In item 11, something similar to item 13 of 2021 occurs, where ChatGPT selected the correct option, but regarding the incorrect options, it indicated that “it is not incorrect, but rather it is a statement that Hamlet mentions and reflects on in his monologue”. The question for this item also had the “Which NOT” structure: “Based on the above text, which of the following statements would Hamlet NOT agree with?”. It is worth noting that in ChatGPT’s explanation for the correct option, it also states that options A, B, and D are incorrect, since it states: “The other statements

(A, B and D) do seem to agree with the content of Hamlet’s monologue”. Given the above, only the explanation for the correct option was compared, while the explanations for the remaining incorrect options were not evaluated. Due to space constraints, we cannot fully explore the analysis of these differences in this study.

Regarding the classification of the items in the three claims, ChatGPT correctly classified 49 (55.68 %) items into the three claims. Below are the items that were classified in a wrong claim.

- **2021, 5^o:** *claim 1:* 6, 11, 15; *claim 2:* 3, 5, 7, 10.
- **2021, 9^o:** *claim 1:* 6, 8, 9, 14, 15, 16; *claim 2:* 19, 20.
- **2021, 11^o:** *claim 1:* 4, 6, 13; *claim 3:* 2, 3, 18.
- **2022, 5^o:** *claim 1:* 3, 4, 7, 11, 12; *claim 2:* 8, 19.
- **2022, 9^o:** *claim 1:* 12, 15; *claim 3:* 13.
- **2022, 11^o:** *claim 1:* 8, 10, 14, 17; *claim 2:* 1, 11, 12, 20.

Among the items of the grade 5, 14 were classified incorrectly. Items 3, 5, 7, 10, 18, and 19 were reclassified from claim 3 to claim 2, while items 3, 4, 6, 11, 12, and 15 were reclassified from claim 2 to claim 1. Additionally, items 7 and 11 were reclassified from claim 3 to claim 1. Something similar occurs with grades 9 and 11, where in the former 11 items were misclassified, while in the latter 14 items were misclassified. We hypothesize that ChatGPT misclassifies certain items from claim 3 due to the presence of quoted expressions from the context in the question. It appears that ChatGPT interprets these quotes as literal text and consequently categorizes them under claim 1, but we are unable to delve further into the analysis of this classification by ChatGPT in this study and further investigation is needed.

5 Conclusion and Future Work

This study provides insights into the creation of explanations for multiple-choice item options in reading comprehension tests with the assistance of AI. By comparing explanations generated by ChatGPT with those created by humans, our analysis indicates that ChatGPT can produce explanations that could be considered equivalent and possibly

better than those created by humans, with potential benefits for both humans and language models. ChatGPT can offer more detailed and specific insights into the text, which can enhance the quality of explanations provided by humans. However, our findings also suggest that there is still room for improvement for both humans and language models. To address these limitations, future research could explore ways to combine the strengths of humans and language models to produce even more accurate and informative explanations. Therefore, ChatGPT has the potential to assist teachers and other professionals in the creation of high-quality assessment items through a well-designed prompt, which can help ensure that items have a single correct answer, independent options, non-overlapping options, and plausible options. Furthermore, ChatGPT ability to classify items based on ECD principles is promising, but further research is needed. For example, the evidences could be provided to language models and ask them to classify each item in one of them. Also, they could be asked to create the options and the respective explanations based on some kind of guidelines such as the one cited.

References

- Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. [Chatgpt goes to law school](#). *SSRN. Minnesota Legal Studies Research Paper*, (23-03).
- Lizeth Donoso. 2021. *Prueba de Lectura Crítica Saber 11.º, Saber TyT y Saber Pro. Marco de referencia para la evaluación*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- Nino Fijačko, Lucija Gosak, Gregor Štiglic, Christopher T Picard, and Matthew John Douma. 2023. [Can chatgpt pass the life support exams without entering the american heart association course?](#) *Resuscitation*, 185.
- Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2023. [How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment](#). *JMIR Med Educ*, 9:e45312.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. [How close is chatgpt to human experts? comparison corpus, evaluation, and detection](#).
- Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. [A review of multiple-](#)

- choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.
- John Hattie. 2012. *Visible learning for teachers: Maximizing impact on learning*, chapter The flow of the lesson: the place of feedback. Routledge.
- Fabio de Jesús Jurado and María Elvira Rodríguez. 2020. *Competencias Comunicativas en Lenguaje: Lectura y Escritura. Marco de referencia para la evaluación*. Instituto Colombiano para la Evaluación de la Educación (ICFES).
- Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.
- MEN. 2006. *Estándares Básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas. Guía sobre lo que los estudiantes deben saber y saber hacer con lo que aprenden*. Ministerio de Educación Nacional, Colombia.
- Robert J Mislevy, Geneva Haertel, Michelle Riconscente, Daisy Wise Rutstein, and Cindy Ziker. 2017. *Assessing Model-Based Reasoning using Evidence-Centered Design: A Suite of Research-Based Design Patterns*, chapter Evidence-centered assessment design. Springer.
- Robert J Mislevy, Linda S Steinberg, and Russell G Almond. 2003. On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1):3–62.
- OpenAI. 2023. Gpt-4 technical report.
- Ashish Sarraju, Dennis Bruemmer, Erik Van Iterson, Leslie Cho, Fatima Rodriguez, and Luke Laffin. 2023. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*, 329(10):842–844.
- Rainer Winkler and Matthias Söllner. 2018. Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting*.
- Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachler. 2021. Are we there yet? - a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.