



Ca' Foscari  
University  
of Venice

Corso di Laurea Magistrale  
in  
Scienze del Linguaggio

Tesi di Laurea

The influence of AI on EFL teaching materials  
development: evaluating reading comprehension  
exercises generated by ChatGPT

**Relatore**

Ch. Prof. Graziano Serragiotto

**Correlatore**

Ch. Prof. Monica Banzato

**Laureanda**

Linda Dal Negro

Matricola 879023

**Anno Accademico**

2023/2024

*Alle mie insicurezze che troppo spesso mi hanno portata a  
credere di non poter fare qualcosa,  
Agli ostacoli superati e a tutti quelli che mi attendono in futuro,  
E, soprattutto, a coloro che hanno creduto in me quando io ero  
la prima a non farlo:  
Questo traguardo è una piccola dimostrazione che, con  
impegno e costanza, anche ciò che sembra irraggiungibile può  
diventare realtà*

## **Abstract**

In recent years, the emergence of artificial intelligence (AI) is having a significant influence on a range of disciplines, including education. ChatGPT is an AI-powered chatbot whose impact on language instruction has been extensively researched by experts. Nonetheless, few studies have focused on the potential of this system to generate EFL teaching materials. As a result, the current research study seeks to assess the quality of a ChatGPT-generated reading comprehension exercise by comparing it to a human-written reading task on the same topic. Twenty English instructors are required to evaluate the reading texts prepared for Italian lower secondary school students by completing a questionnaire based on the criteria stated in Clark et al.'s study (2021). The research instrument was intended to gather quantitative data as well as qualitative remarks that justified participants' previous choices. The data is analysed with the ultimate goal of addressing the research question: are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content? On the one hand, the study's findings indicate that there is no substantial difference between AI-developed and human-written exercises because participants' evaluations show no important variances. On the other, teachers agree that the AI-generated reading task is slightly more difficult than the other assignment in terms of formality and text organisation. This study thus provides insight into the potential use of ChatGPT as a system for developing teaching materials, and its implications may be relevant not only to language teachers and instructors in general, but also to scholars who may conduct future research on the relationship between AI-technology and language teaching.

# Table of contents

<b>I. Introduction</b> .....	1
<b>II. Literature Review</b> .....	3
2.1 ChatGPT and foreign language teaching.....	3
2.2 Reading comprehension exercises development .....	7
2.3 Evaluation of AI-generated materials .....	9
<b>III. Presentation of Research</b> .....	11
3.1 Purpose of the study.....	11
3.2 Human-generated materials .....	12
3.3 AI-generated materials .....	14
<b>IV. Methodology</b> .....	18
4.1 Setting .....	18
4.2 Participants .....	19
4.3 Assessment criteria .....	23
4.4 Research Instrument.....	25
4.4.1 Consent acknowledgment .....	25
4.4.2 Demographics.....	26
4.4.3 Reading exercise .....	27
4.4.4 Form .....	27
4.4.5 Content.....	28
4.5 Administration.....	29
<b>V. Results</b> .....	30
5.1 Quantitative data.....	30
5.1.1 Form .....	31
5.1.2 Content.....	35
5.2 Qualitative data .....	40
5.2.1 Form .....	41
5.2.2 Content.....	45
<b>VI. Discussion</b> .....	50
6.1 Research findings .....	50
6.1.1 Grammar .....	52
6.1.2 Level of detail.....	54
6.1.3 Genre .....	55

6.1.4 Repetition .....	56
6.1.5 Factuality.....	58
6.1.6 Consistency .....	59
6.1.7 Common sense .....	61
6.1.8 Coherence.....	62
6.2 Previous research.....	63
6.3 Limitations .....	64
6.4 Future research .....	66
6.5 Implications .....	67
<b>VII. Conclusion .....</b>	<b>69</b>
<b>Bibliography .....</b>	<b>71</b>
<b>Appendix.....</b>	<b>77</b>

## **I. Introduction**

According to Adiguzel, Kaya & Cansu (2023, p. 2), the term “Artificial Intelligence” was first defined by John McCarthy in 1955 as “making a machine behave in ways that would be described as intelligent if a human was so behaving”. In spite of dating back to many years ago, this definition reflects some characteristics of modern AI-powered technologies that are currently revolutionising our daily lives. Indeed, artificial intelligence systems are not endowed with reason, nor are they inherently intelligent. These machines are instead designed to resemble human intelligence thanks to specific algorithms which allow the system to produce human-like output. Therefore, AI could be better defined as “any theory, methodology, or technique that facilitates the analysis, simulation, exploitation and exploration of human thinking processes and behaviours by machines, particularly computers” (Lu, 2019). In recent times the launch of ChatGPT 3.5 has brought about significant changes in several domains, particularly in the education field. According to Baskara e Mukarto (2023, p. 346): “The model’s capacity to generate text that resembles human language has significant implications for language learning, communication, and education.” The Generative Pre-Trained Transformer (GPT) is a deep learning model for Natural Language Process (NLP) which has been instructed on a large amount of data and is therefore capable of generating high quality answers based on the user’s prompts. As reported by Kim et Al. (2023, p. 80):

“ChatGPT, an AI-powered chatbot developed by OpenAI, was released in November 2022, and has sparked huge interest for its exceptional quality of response compared to other contemporary chatbots [...] ChatGPT can be used for various tasks including text generation, machine translation, content correction, problem solving, summarisation, grammar correction, and answering questions. The chatbot operates in a conversational manner, responding to each prompt entered by the user.”

The advent of AI has indeed modified traditional learning standards as it presents new advantages and challenges both for teachers and students. The current study aims to examine the influence of ChatGPT on language teaching and learning, with particular reference to English as a Foreign Language (EFL) teaching. Although recent research has focused on the utilisation of ChatGPT as a language learning tool, few studies have investigated its potential to develop EFL reading materials, hence, exploring the capabilities of AI-powered systems as teaching assistants emerges as a very pertinent area of research. The following research question

guides the research project: “Are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content?”. In order to assess the quality of the reading comprehension exercises produced through AI technology, human evaluation is required. English teachers in Italy will be asked to compare the reading passage and the corresponding activities contained in a middle school textbook with the materials developed by ChatGPT. As a starting point, the language model will be provided with specific prompts to generate a reading passage for middle school students on the same topic as the one chosen from the textbook. In addition, the system will be asked to produce some open-ended as well as some multiple-choice questions based on the previously generated passage. A questionnaire incorporating both human-written and ChatGPT-generated activities will be utilised to acquire quantitative data on the participants’ thoughts using linkert-scale items. Finally, teachers will be asked about the reasons for their evaluations, thus some qualitative data comprising participants’ comments will be collected as well. This method allows to gather information on the quality of the reading comprehension exercises in terms of form and content without specifying the source of the materials in order to avoid any bias. Finally, the comparison of the data based on the participants’ evaluations allows to investigate the potential of ChatGPT as a tool for developing suitable exercises to boost middle school students’ reading comprehension skills in English.

The present study therefore aims to offer insight into the integration of AI in the process of EFL teaching materials development. After having described the focus of this work, in the second chapter the most relevant studies concerning the implementation of ChatGPT in education are summarised. There are three main aspects that will be considered, namely the link between AI and foreign language teaching, the production of reading comprehension exercises and the assessment criteria of such materials. The third chapter illustrates the purpose of the study in greater detail by presenting the reading passages and the corresponding questions. Furthermore, chapter four includes information about the research project’s participants as well as a description of the questionnaire. The quantitative data is examined in the fifth chapter, along with comments expressing teachers’ perspectives. This section contains graphs reporting the results of the evaluation completed through the questionnaire, with the primary goal of comparing human-written materials to AI-generated tasks. Finally, in chapter six, the research findings are discussed taking into account previous research presented in the second chapter, before drawing some broad conclusions.

## II. Literature Review

### 2.1 ChatGPT and foreign language teaching

In the last few years ChatGPT has had a remarkable impact on education. Consequently, the Generative Pre-trained Transformer has received significant attention from experts in the education field, this leading to a vast amount of research related to the potential benefits and challenges of AI in teaching and learning. According to George & George (2023), ChatGPT reduces the amount of work for teachers, so that they can concentrate on what is more important for the students. Firstly, this AI-powered system allows teachers to develop lesson plans based on the students' interests and skills, giving them immediate feedback as they use the materials. Secondly, ChatGPT is able to grade students' work more quickly than humans do, allowing teachers to keep from providing handwritten feedback. Finally, in higher education contexts it may also serve as a virtual assistant that can answer questions about courses and services in the school. Atlas (2023) has also addressed the advantages and challenges of ChatGPT in education. In addition to the benefits described above, language models may be used to translate educational materials as well as to develop syllabus, quizzes and exams. What is more, they are able to generate summaries and reports, but also to identify relevant literature and writing papers, thus reducing researchers' work. As regards the limitations of using language models, ChatGPT is trained on data which may present some bias, as a result the AI generated output reflects such discriminating notions. In addition, ChatGPT could be used to achieve objectives that are not in line with our values, such as controlling students' behaviour and thoughts.

Finally, these models are not able to understand human emotions, intentions and moral reasoning which can be challenging "when using them for tasks that require empathy, such as providing counselling or tutoring" (Atlas, 2023, p. 10). To sum up, the development of language models based on artificial intelligence has influenced all education contexts, opening up several opportunities not only for students but also for teachers and scholars. However, it is fundamental to use ChatGPT in a responsible way, being aware of both its potential and limitations.

With respect to the field of language teaching, research so far has addressed the impact of ChatGPT on students' skills as well as on teachers' work. When teaching English as a Foreign Language (EFL), not only it is crucial to foster learners' receptive and productive skills but also to increase their engagement in order to achieve their full potential. The mixed-method study



carried out by Song & Song (2023) has focused on the role of AI in enhancing students' academic writing skills and motivation. EFL learners in higher education contexts have been instructed by ChatGPT to improve their writing:

“ChatGPT provided real-time feedback on grammar, vocabulary usage, sentence structure, coherence, and organisation. The AI model, trained on extensive language data, identified errors, offered suggestions for improvement, and provided contextualised recommendations to enhance their writing skills [...] ChatGPT also offered writing suggestions, alternative phrasing options, and vocabulary expansion ideas to enhance expressiveness and language fluency.” (Ibid., p. 6)

The outcomes of the study revealed that students' writing performance and motivation have improved thanks to AI-assisted language learning. The comparison between pre-test and post-test scores as well as the semi-structured interviews with the participants have demonstrated that AI tools have had a positive impact on learners' engagement and internalisation of writing skills, by providing personalised learning experiences. In addition, the integration of ChatGPT in the learning process allowed to develop a more relaxed classroom environment and to lower students' anxiety-level, as reported by the qualitative results of the study.

Kosta & Toncelli (2023) have explored opportunities and concerns related to the integration of ChatGPT in language education. Although language models' implementation in the classroom allows for innovative language teaching, there are also some challenging aspects of AI-powered tools to consider. For instance, academic integrity is threatened by the high risk for cheating through ChatGPT. As reported by Cotton et al. (2023, p. 6) in Kosta & Toncelli: “One major fear for educators, is that students will no longer write their own assignments”. As a consequence, universities and institutions are developing AI-generated texts detection systems to prevent students from cheating. However, such programs are not completely reliable, hence, teachers and professors may try alternative assessing methods. On the one hand, educators are encouraging students to use ChatGPT as a pedagogical tool to produce source-based materials instead of using it inappropriately. On the other hand, teachers are developing new assessing approaches based on problem solving activities and critical thinking in order to discourage the use of AI during examinations and tests. Moreover, Kosta e Toncelli (2023) have collected students' quotations indicating their attitudes towards language models after implementing ChatGPT in a listening and speaking course for international students. Even though the majority

of the learners was enthusiastic about using ChatGPT, some of the participants' voices were critical of AI-powered tools: "It's better to gather information myself instead of using tricky shortcuts. After all, ChatGPT involves that information from the internet and could not really give "real" feedback on how my learning has improved." (Ibid., p. 11). In addition, some quotes highlighted students' concerns about the replacement of human actions by language models: "Yes, it is a good tool on enhancing learning, but it can only provide some basic information on our learning. If we dig deeper, we still need to use our own ability on doing research." (Ibid., p. 11). As a result, the outcomes of the study emphasise that both teachers and students present various attitudes towards ChatGPT, whom implementation in the language education field displays benefits and challenges.

In order to investigate the influence of ChatGPT on language learners' motivation, Yildiz (2023) has also provided university students with AI-assisted instruction. Firstly, the language model was used during language classes to practice writing and listening in English as well as to provide feedback and suggestions for improvement. Secondly, it was trained to develop language exercises such as reading comprehension passages including specific vocabulary in English. The aim of the study was to compare students' answers to the "Motivational Strategies for Learning Questionnaire" developed by Pintrich & De Groot (1990). The control group was instructed through traditional language teaching while the experimental group was supported by ChatGPT. Findings suggested that ChatGPT assistance is beneficial for students: "...post-motivation test subcategories revealed that experimental group was more motivated after the ChatGPT integration to the post-lesson activities (especially from the perspective of self-regulation, intrinsic value, and test anxiety)." (Yildiz, 2023, p. 593).

The potential of languages models as helpful tools for teachers has also been investigated. Koraishi (2023) has provided ChatGPT with some instructions to generate a lesson plan tailored to specific objectives and instructional needs for EFL teaching. The final result was in line with the given prompt as well as with traditional lesson plan stages, even though some modifications may be necessary: "It is vital to emphasise that the purpose of employing ChatGPT in lesson planning is to facilitate the process in conjunction with the teacher's guidance, as the AI's capabilities may not yet suffice independently" (ibid., p. 62). However, the fact that ChatGPT can be asked to provide some adjustments to its output is noteworthy. Teachers have a chance to give additional prompts to ChatGPT either to ask for clarification or to add further details to

the lesson plan. The opportunity to cooperate with AI-powered systems is indeed turning out to be very useful for educators, as it is a time-saving process to develop personalised and well-structured teaching materials. According to Koraishi (2023), ChatGPT may also support EFL teachers in assessing students' performance in accordance with specific criteria. To provide evidence to this assumption, ChatGPT has been required to evaluate an independent written task according to the IELTS international proficiency exam standards. "The feedback generated aligns with the IELTS rubric in key aspects such as organisation and task achievement. It provided an approximate score band in addition to offering insights and a rationale behind the assigned score" (ibid., p. 67). AI-generated feedback is therefore in alignment with the given prompts, demonstrating that ChatGPT has been trained on hyper-specific assessment criteria. Nevertheless, teachers are supposed to collaborate with ChatGPT by enquiring about specific problems and corrections for further analysis, in order to provide students with exhausting information about their writing.

As reported in Atlas (2023, p. 52) AI-powered language models can be used to check grammar, punctuation and spelling in students' writing tasks: "These tools can assist students and instructors in improving the quality and clarity of written work, and can help to identify and correct grammatical errors, spelling mistakes, and other issues that can detract from the overall effectiveness of the writing". However, ChatGPT may also serve as a tool to promote personalised learning, adaptive testing and intelligent tutoring. Personalised learning refers to the adaptation of teaching materials to the unique needs of the learners. As a matter of fact, every student has different interests and learning styles, hence, ChatGPT may be used by the teacher or the students themselves to draft specific learning materials on the basis of their needs and characteristics. As regards adaptive testing, it refers to the development of testing items based on the student's level of knowledge. For instance, ChatGPT may be asked to create specific questions tailored to the students' performance on a pre-test, allowing teachers to generate different tasks according to the level of understanding of each student. This may be particularly useful to enhance the educational experience of those students who present special needs. Finally, intelligent tutoring can be defined as: "the use of technology to provide personalised, step-by-step guidance and real-time feedback to students" (ibid., p. 82). In this view, ChatGPT is considered as a tutor that can help students understand their improvements by providing personalised recommendations and instructions.

Finally, research in the education field has investigated the role of language models such as ChatGPT in promoting language teaching and learning. AI-powered technology is indeed modifying traditional teaching by providing innovative approaches to address students' needs. Nevertheless, at the present time teachers still have a fundamental role in the classroom as it is highly suggested to cooperate with ChatGPT in order to integrate the language model in the teaching process.

## 2.2 Reading comprehension exercises development

Pre-Trained Large Language Models are becoming increasingly useful to foster learners' language skills, as reported by the research projects mentioned above. However, the aim of the present study is to examine the influence of ChatGPT on the development of reading comprehension abilities by creating engaging reading passages and exercises. Although it is fundamental to provide students with appropriate reading materials according to their language level and interests, few studies have focused on the role of language models in generating personalised reading activities. According to Xiao et Al. (2023, p. 610): "English teachers believe that offering up to date, engaging reading exercises tailored to each student's capabilities and interests can spark their enthusiasm for learning and ultimately boost their English proficiency". Indeed, teachers are constantly struggling to find authentic sources which could serve as reading passages for EFL learners. In addition, they usually need to develop the corresponding reading activities from scratch, which is an extremely time-consuming task. Consequently, Xiao et Al. (2023) have attempted to develop some reading comprehension exercises for Chinese middle school students using ChatGPT: "Reading comprehension exercises typically consist of two components: a lengthy, coherent passage and several multiple-choice questions that align with its content" (ibid., p. 610). In order to assess the AI-generated output, human evaluators including English teachers and English native speakers have been required to compare such reading materials with human-written passages and exercises dealing with the same topic. ChatGPT was instructed to create highly specific reading comprehension exercises in English through the conversation prompts. The genre, length and difficulty level of the text along with the topic of the reading passage were included in the instructions. Moreover, the prompts contained both the question type and the number of options for every multiple-choice question. Firstly, evaluators were required to read the passage generated by ChatGPT as well as the human-written one in order to evaluate their quality on a

scale from 1 to 5 in terms of: readability, correctness, coherence, engagement and overall quality. Secondly, they had to consider different aspects to assess the reading exercises, namely usefulness and suitability of the questions for the training of middle school students as well as the extent to which the questions match the passage content. Participants were also asked to rate the extent to which questions appeared to be written by language models in order to understand whether they could identify any difference between AI-generated and human-generated exercises. The outcomes of the study suggest that the scores of the passages generated by ChatGPT are even higher than the scores of the texts devised by humans. Conversely, multiple-choice questions developed through AI were not considered on the same quality level as human-drafted exercises, as the latter outperformed the materials produced by ChatGPT on all four dimensions. Overall findings indicate that the integration of AI technology in the education field for the purpose of reading comprehension exercises development may be helpful for both teachers and students: “Teachers view this system as a valuable tool that can significantly reduce cost and time while providing students with more diverse and personalised learning materials” (ibid., p. 617).

Another experimental study conducted by Shin & Lee (2023) aims to compare the L2 assessment materials developed by ChatGPT with human-written reading exercises. The majority of the participants was selected among a group of undergraduate students who had been studying to qualify as teachers; furthermore, a smallest number of in-service English professors took part in the research project. Participants were asked to respond to a likert-scale questionnaire in order to assess the reading passages and the multiple-choice questions without being informed about the source of the materials. Four criteria were used for the purpose of evaluation, namely the naturalness of flow of the text, the naturalness of the English expressions, the attractiveness of multiple-choice options and the overall completion level of the testing item. Overall results suggest that there was no significant difference between AI-generated and human-developed materials in terms of naturalness of flow and English expressions. Instead, several participants reported the lack of distractors in the multiple-choice questions devised by the model, hence, the attractiveness of such items was rated significantly low, as can be inferred by the following comments: “There are no compelling option choices for this question”; “A significant modification is required for this question. For example, the phrase in the fifth option is not even mentioned in the passage” (ibid., p. 34). The same trend emerges in relation to the completion level of the testing item. In addition, data analysis

demonstrates that in-service professors gave lower ratings to the naturalness of flow and the completion level of the items than pre-service teachers, thus, participants' proficiency level emerges as a crucial factor to consider for the purpose of research.

Finally, the automatic evaluation of reading comprehension exercises with ChatGPT has also been investigated. In addition to the generation of multiple-choice questions to foster learners' skills in the German language, Säuberli & Clematilde (2024) have developed a new evaluation metric which combines answerability and guessability in order to evaluate the quality of the exercises in terms of text informativity. Human assessment is yet a crucial component of the evaluation process, however, a protocol including the new testing metric has been used for automatic evaluation as well. By providing GPT-4 and Llama 2 Chat with specific prompts, the systems have been required to evaluate the previously generated questions. Results indicate that ChatGPT is able to generate reliable observations, outperforming the other language model: "High IAA between GPT-4 and human annotators also suggest that using GPT-4 as an evaluator is a viable option. In contrast, results from Llama 2 were less consistent with humans, both at the dataset level and the response level" (ibid., p. 8). As a matter of fact, recent research has focused on the potential of AI technology to radically change the role of the language teacher. Not only ChatGPT has been tested as a tool to develop reading comprehension exercises, but also it has been used to evaluate learning materials' quality. The present study therefore aims to provide further evidence of the model's applicability to the education field with particular reference to the development of reading materials for EFL middle school students.

### 2.3 Evaluation of AI-generated materials

With respect to the quality of the AI-produced output, there are some issues to consider in relation to the assessment criteria commonly used for evaluation. As reported in Belz et. al (2023, p. 3): "While some aspects of evaluations such as type and size of rating scale, evaluation mode etc., are relatively easy to determine from papers, the confusion over which evaluations assess which aspect of quality, and the paucity of detail about experimental design in the great majority of papers, at present mean we do not have a basis for establishing comparability". At the present time it is therefore quite difficult to find objective definitions of every quality criterion. Although terms like "readability", "language naturalness", "coherence" and "grammaticality" are frequently used to describe Natural Language Processing, different papers provide variable definitions of these words, as a result it is rather challenging to reproduce and

compare similar research projects. The issue of finding clear and well-defined criteria for evaluation has been addressed also by van der Lee et. al (2021, p. 6): “Many studies take some notion of ‘text quality’ as their primary evaluation measure, but this goal is not easy to assess, since text quality criteria differ across tasks. [...] In short, there is no standard evaluation model for NLG, nor agreement in terminology, and explanatory details for the criteria are often lacking”.

In order to classify the characteristics of evaluations, Belz et al. (2020) have proposed a set of values divided according to three main categories, namely quality criteria, evaluation mode and experimental design. As for the first section, it considers what type of quality is being assessed, what type of the system output is being assessed and whether the frame of reference for evaluation is internal or external to the system. Evaluation modes instead can be defined as objective or subjective, absolute or relative, extrinsic or intrinsic. Finally, experimental design properties include several elements such as scale or rating instrument, information about evaluators, form of response, etc. As regards the type of output to be assessed, it is worth considering the distinction between form and content of the materials generated through AI. Clark (2023) adopted the aforementioned distinction to compare human-written and machine-generated texts with respect to the production of stories, news articles and recipes. The study aimed to provide participants with a set of elements displaying both human-devised texts and GPT2- or GPT3-generated passages. Evaluators were required to comment on the source of every text in order to distinguish between human-authored and machine-developed pieces of writing. Final outcomes indicate that participants correctly identified the source of the text most of the times when choosing from GPT2- and human-written items, however, GPT-3 texts were less detectable. The study also aimed to label evaluators’ comments according to specific criteria. On the one hand, sentences concerning the form of the text fell into three different groups namely grammaticality, level of detail and genre. On the other, comments referring to the content of the passage were divided into five categories labeled as repetition, factuality, consistency, common sense and coherence. This categorisation is simple yet rather specific as a description of every criterion is provided along with some examples of the comments falling in each group. In conclusion, research so far has emphasised that much work remains to be done in the development of precise, well-defined evaluation criteria to be used for the purpose of assessing AI-generated materials. Thus, there are currently some issues to consider with respect to the replicability of research concerning objective evaluation of automatic-produced output.

### **III. Presentation of Research**

This chapter illustrates the main purpose of the study by providing a general overview of the research project. Furthermore, both human-written and AI-generated learning materials are presented along with the instructions given to ChatGPT in order to obtain specific output.

#### **3.1 Purpose of the study**

The role of AI-powered tools in fostering EFL students' reading comprehension skills appears as an increasingly relevant topic in the research field of language education, as shown in chapter 2 of the present paper. By developing suitable learning materials following precise prompts, indeed, ChatGPT may prevent language teachers from doing time-consuming tasks such as generating highly-specific reading exercises in English. Research involving ChatGPT so far has emphasised the potential of the language model to generate high-quality output, hence, this study endeavours to verify the hypothesis that AI-developed items can be rated the same as human-written materials. The current study therefore aims to investigate the quality of the reading comprehension exercises generated by ChatGPT 3.5 in order to answer the research question: "Are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content?". Thus, this work focuses on the comparison of human-written and AI-generated materials which may be used to boost middle school students' skills in English. Firstly, the human-written reading task was selected from the middle school English textbook "Step Up 3" which is commonly used in Italian schools. Secondly, the reading exercises produced by ChatGPT have been developed following the same prompts reported in a similar study conducted by Xiao et al. (2023), namely "Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications". English teachers in Italy are required to assess the quality of the teaching materials under investigation which consist of two reading passages and the corresponding exercises, namely three multiple-choice questions and eight open-ended questions for each reading passage. The evaluation process is conducted through a questionnaire which has been devised on the basis of the criteria indicated in "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text" by Clark (2021). The questionnaire aims to gather teachers' opinions on the form and content of the reading comprehension exercises without specifying which items have been developed by ChatGPT, so



as to avoid any bias due to preconceived ideas about AI-generated output. The analysis of both the quantitative data and the comments collected through the research instrument allows to compare human-generated and AI-developed materials on the basis of the English teachers' evaluations.

### 3.2 Human-generated materials

As the present study focuses on the evaluation of reading comprehension exercises developed for middle school students, the following human-written reading passage has been chosen to be assessed by language teachers. The source of the materials is the textbook "Step Up 3" by Oxford University Press, hence, EFL students at the third year of lower secondary school in Italy are supposed to understand the text and do the exercises. The title of the reading passage which can be found in Unit 6 is "Things you should do every day to be happy and healthy".

"Do you sometimes feel tired and stressed? In winter, do you often have colds, coughs and sore throats? If you've answered 'yes' to any of these questions, then we've got help for you! Take a look at our list of simple things to do every day to be happy and healthy.

#### 1. Go outside!

You shouldn't spend all your free time in your room sitting in front of a screen! Every day you should spend some time outside in nature. A short walk in a park or a forest or other green spaces reduces stress and helps your immune system.

#### 2. Get moving!

How do you get to school? Do you take a bus or do you go by car? If you walked part of the way or cycled to school every day, you would be healthier and happier. Simple activities like walking and cycling are great forms of exercise. (But remember when you cycle you should always pay attention to the traffic and wear a helmet! You don't want to have an accident and break your leg or bump your head!)

#### 3. Sleep

Teenagers need eight to ten hours' sleep every night. Studies show that teenagers who don't sleep enough are irritable, get headaches, and their grades at school suffer. It's important to have regular sleep patterns. You should always try to go to sleep and wake up at the same time, even at the weekend! You shouldn't have your mobile phone or

tablet in your bedroom when it's time to sleep. You should turn off all electronic devices at least one hour before bedtime.

#### 4. Laugh and hug

Laughter is important for your mental and physical health. Every day you should do things that make you laugh - watch your favourite comedy programme or funny videos and spend time with friends who make you laugh. Everybody always feels better after they've laughed! Hugging increases the levels of serotonin in your brain. Serotonin is a chemical that reduces stress and makes you feel happier. People who laugh get fewer colds in winter! Oh, and hugging dogs, cats and teddy bears is good for you, too!"

The following reading activities are reported in Unit 6 of the textbook "Step Up 3", they are presented at the bottom of the page, below the reading passage. Hence, students are supposed to read the text and then answer the questions.

*"Choose the correct options.*

1. Which piece of advice is NOT included in the article?

- A) Spend time in nature.
- B) Sleep eight to ten hours.
- C) Eat lots of fruit and vegetables.
- D) Watch funny videos.

2. The article says you shouldn't:

- A) go for a walk in the park.
- B) walk to school.
- C) hug animals.
- D) cycle without a helmet.

3. The article says you should:

- A) spend your free time at home.
- B) go to school by car.
- C) wake up late at the weekend.
- D) try to laugh every day.

*Answer the questions.*

1. How does spending time in nature help your health?
2. What should you remember to wear if you cycle to school?
3. How many hours should you sleep at night?
4. What happens if teenagers don't get enough sleep?
5. How long before bedtime should you turn off your electronic devices?
6. Why is it important to laugh?
7. What is serotonin?
8. What other things do you think you should do to stay happy and healthy?"

### 3.3 AI-generated materials

ChatGPT has been instructed to develop a reading comprehension exercise similar to the previously mentioned human-written reading task in English. To ensure reliability, the language model is addressed in a conversational manner by adapting the prompts used by Xiao et. al in the investigation conducted during the Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). The study aimed to compare the quality of the reading exercises developed for Chinese middle school students, as reported in section 2.2 of the present paper. According to Xiao et. al (2023, p. 621) it is fundamental to provide ChatGPT with certain information in order to avoid the generation of toxic learning materials:

“To ensure the safety of middle school students and avoid ethical issues, we have implemented measures to prevent the generation of toxic text. In our prompts, we explicitly specify that the generated content must not contain violence, racism, or other harmful elements for young language learners.”

Hence, the choice of addressing ChatGPT with prompts that follow the ones presented in this study not only comes from the belief that the present research addresses the same issue, but also it is due to the close attention given to the matter of receiving ethical output. In order to compare the two reading activities, the AI-generated passage deals with the same topic as the human-developed text and presents a similar length as well. Additionally, the text should correspond to A2 English level of the CEFR conceived by the Council of Europe (2001), as third-year

middle school students in Italy are supposed to master the skills associated to A2 level according to national directions.

*Request:*

“You are a helpful assistant to generate reading comprehension materials for Italian middle school English learners. Your responses should not include any toxic content. Please generate a passage that satisfies the following requirements. Length: 300 words; English level: A2; Topic: Things you should do every day to be happy and healthy”

*ChatGPT-generated answer:*

“Every day, there are simple yet essential things you can do to ensure both happiness and good health. These habits, when practiced regularly, contribute to a fulfilling and balanced life.

Firstly, starting your day with a nutritious breakfast is crucial. Eating a healthy meal in the morning fuels your body and provides you with the energy you need to tackle the day ahead. Make sure to include foods rich in vitamins, minerals, and fiber, such as fruits, whole grains, and dairy products.

Secondly, staying active is vital for maintaining both physical and mental well-being. Aim to engage in some form of exercise every day, whether it's going for a walk, riding a bike, or playing your favorite sport. Physical activity not only keeps your body strong and flexible but also releases endorphins, which are natural mood boosters. In addition to staying active, it's essential to take care of your mind. Dedicate some time each day to relaxation and stress management techniques, such as deep breathing, meditation, or simply spending time in nature. These practices help alleviate stress and promote a sense of calm and inner peace.

Furthermore, maintaining good hygiene habits is essential for overall health and well-being. Remember to brush your teeth twice a day, wash your hands regularly, and shower or bathe daily. Keeping yourself clean helps prevent the spread of germs and reduces the risk of illness.

Lastly, getting enough sleep is crucial for both your physical and mental health. Aim for 7-9 hours of quality sleep each night to allow your body and mind to rest and recharge. A good night's sleep improves mood, cognitive function, and overall

productivity.

By incorporating these simple yet effective habits into your daily routine, you can cultivate happiness and maintain your health and well-being for years to come. Remember, small actions taken every day can lead to significant positive changes in your life.”

The reading exercises developed by ChatGPT will be considered for the purpose of evaluation along with the AI-generated text. The model is capable of producing various activities referring to a given passage, however, the prompts given to the system present specific reference to question-type and number of exercises required. This reading task thus displays similar activities to its human-written counterpart in order to allow for comparison between the two items.

*Request:*

“You are a helpful assistant to generate reading comprehension exercise questions for Italian middle school English learners. Your responses should not include any toxic content. Please generate 3 multiple choice questions as well as 8 open-ended questions for the following reading comprehension exercise: ‘*passage*’ ”

*ChatGPT-generated answer:*

*“Choose the correct options.*

1. What is the primary purpose of starting your day with a nutritious breakfast?

- A) To lose weight
- B) To fuel your body and provide energy
- C) To skip meals throughout the day
- D) To satisfy cravings

2. Which of the following is NOT mentioned as a relaxation and stress management technique in the passage?

- A) Deep breathing
- B) Meditation
- C) Playing video games
- D) Spending time in nature

3. How many hours of sleep should you aim for each night according to the passage?

A) 5-7 hours

B) 7-9 hours

C) 9-11 hours

D) 12-14 hours

*Answer the questions.*

1. Why is it important to engage in physical activity every day according to the passage?

2. How can maintaining good hygiene habits contribute to your overall health and well-being?

3. What are some examples of relaxation and stress management techniques mentioned in the passage?

4. How does getting enough sleep contribute to both physical and mental health?

5. What are some example of foods rich in vitamins, minerals, and fiber mentioned in the passage?

6. Why is it important to take care of your mind in addition to staying active according to the passage?

7. How do small actions taken every day contribute to significant positive changes in one's life according to the passage?

8. Why is maintaining good hygiene habits essential for overall health and well-being according to the passage?"

AI-powered language models therefore allow to develop reading comprehension exercises which respect specific requirements according to the students' needs through a time-saving process of materials development. In this case ChatGPT has been instructed to generate a reading passage and the corresponding activities on the topic of happiness and health; the first answer provided by the model in a few seconds was considered as a satisfactory reply, thus, it was used for the purpose of research without asking ChatGPT for further details. Nevertheless, teachers who need to costumize AI-generated text have the chance to ask the system about more detailed output in order to redefine the learning materials until they obtain adequate results.

## IV. Methodology

This chapter aims to describe the study's methodology with particular reference to the setting, the participants in the research project and the instrument adopted for data collection. Information about the environment in which the study has been carried out is therefore provided along with some graphs reporting data on the participants' demographic characteristics. As regards the creation of the questionnaire, both the development process and the criteria considered for analysis are presented before describing the administration phase.

### 4.1 Setting

As the present study is conducted in Italy, it focuses on the evaluation of reading comprehension exercises developed for middle school students who are studying English in this country. The Italian education system includes Early childhood education and care, primary, secondary, post-secondary, tertiary and adult education. Lower secondary education lasts for 3 years and is attended by pupils aged 11 to 14. Lower secondary education and primary education make up the first cycle of education which is compulsory and lasts eight years altogether.<sup>1</sup> English is among the compulsory subjects taught in lower secondary school, along with another foreign language (usually Spanish, French or German). As a result, focusing on the production of English-language teaching resources seemed to be appropriate, given that this language is used in all Italian middle schools. Furthermore, lower secondary school students already have a basic knowledge of the English language, having studied it for five years in elementary school. This enables English teachers in Italy to take a step further and focus on improving pupils' reading comprehension skills through more challenging reading assignments, which may not be used in primary school because children are still in the early stages of language acquisition. Another factor to consider is that children in Italy must have a certain level of English by the third year of middle school in order to pass an exam before beginning the second cycle of education in high school. As described in the previous chapter, students are tested on their reading, writing, and speaking skills, which should equate to the CEFR's A2 level, via a written exam and an interview. In terms of the written component, students must apply their reading comprehension skills because they may be required to answer questions on a given test or compose a letter in

---

<sup>1</sup> Source: EU official website. Available at: <https://eurydice.eacea.ec.europa.eu/national-education-systems/italy/overview>. Last accessed: July 2024.

response to certain prompts. As a result, providing pupils with reading assignments during the third school year proves to be an extremely helpful activity in preparing them for the final exam. Furthermore, according to Italian education system regulations, the teaching curriculum is the same in all lower secondary schools in the country, implying that pupils across Italy are expected to have learned the same competences by the end of the education cycle. Instead, language instruction in upper secondary education settings differs by institution type. Thus, students aged 14 to 19 in Italy can attend a liceo, a technical institute, or a vocational institute based on their interests and skills. Given that these schools follow diverse language education learning programmes, high school pupils' English skills in Italy vary in terms of both level of knowledge and themes covered in class. The decision to focus on the development of English teaching materials for use in lower secondary schools rather than upper secondary education contexts arises from the belief that at this stage, the quality of the assignments can be evaluated in relation to the students' level of competence, which is less variable. For this reason, as well as the other motives listed above, the use of AI-powered tools to generate English reading comprehension exercises for middle school students has been considered a pertinent area of research.

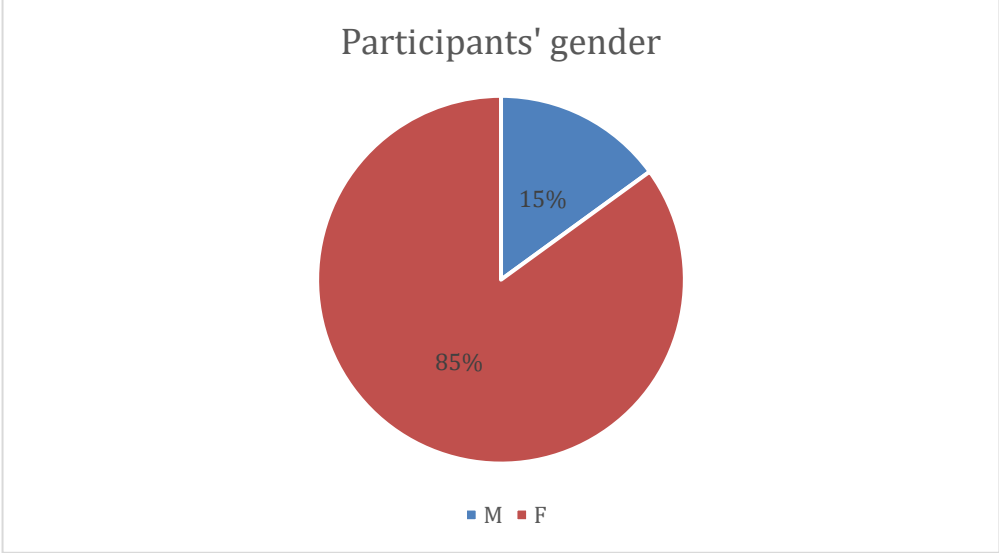
## 4.2 Participants

This section illustrates the demographics of the research project participants. The data given in the pie charts below was acquired using a questionnaire, which will be described in greater detail in the following sections. The graphs show data from 20 English instructors who are currently teaching in Italian schools and were asked to evaluate the quality of the reading comprehension tasks chosen for analysis and comparison. People were thus asked to provide information on their gender, age range, place of birth, languages spoken, number of years of teaching experience and teaching context before responding to questions evaluating the quality of the learning materials. Even though the participants in the study project may work in various educational contexts, they must be familiar with the Italian education system and the English teaching curriculum used in lower secondary schools. Furthermore, because they are needed to critically assess the reading comprehension exercises in English, participants must be able to identify good and negative features of each reading passage in order to provide an opinion on their quality. For all of these reasons, it was critical to select just English teachers working in Italian schools to participate in the study rather than involving other language teachers.



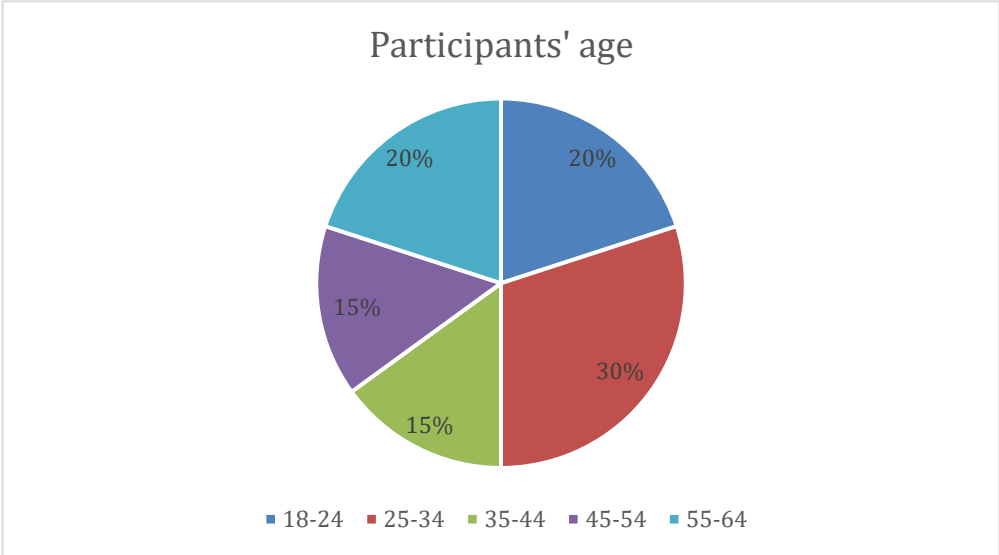
According to the first graph, just 15% of the participants are men, while women make up the majority of responders.

Figure 1 - data about participants' gender



When it comes to participant age, those between the ages of 25 and 34 make up the bulk of responders, accounting for 6 out of 20 people, or 30%. Younger teachers (18-24) account for 20% of the participants, as do those aged 55 to 64. Finally, 15% of people selected the age range 35-44, while the same percentage chose the 45-54 age group.

Figure 2 – data about participants' age



As seen in the graph in figure 3, practically all of the participants were born in Italy, with the exception of 2 people out of 20 (10% of responders).

Figure 3 – data about participants' place of birth

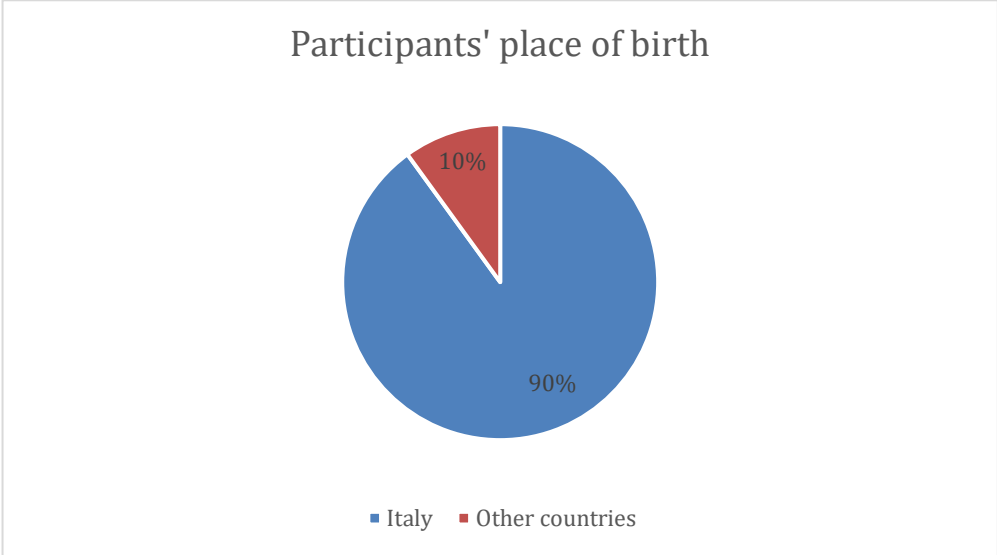
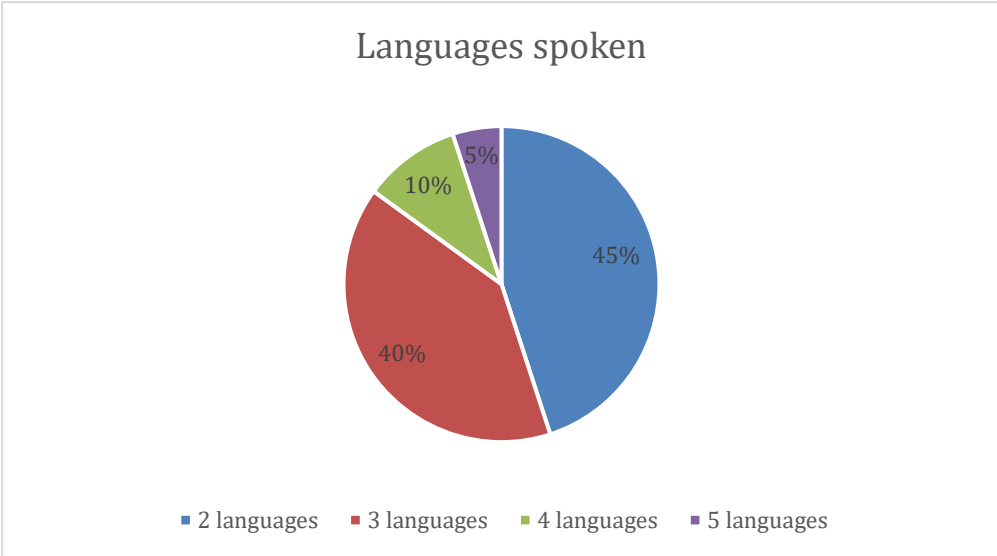


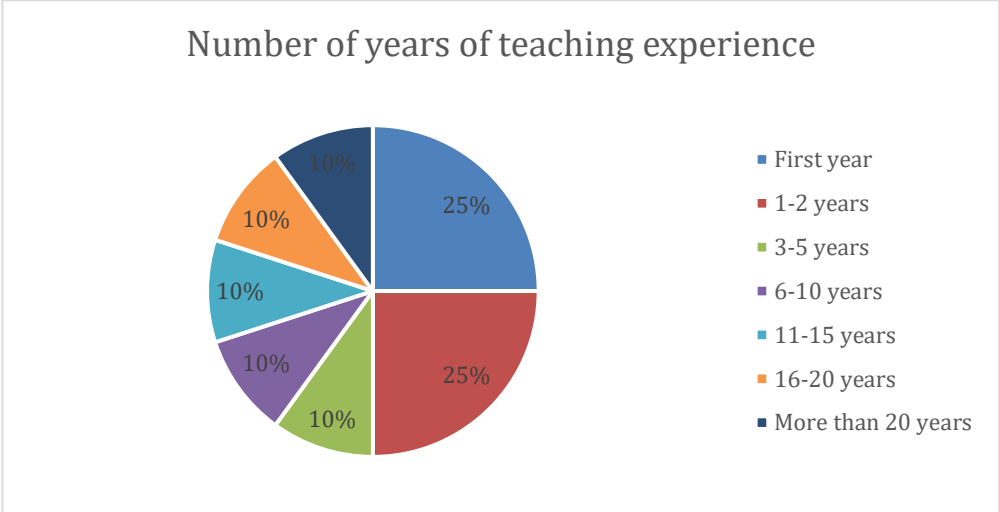
Figure 4 demonstrates that nearly half of the participants (45%) speak just two languages, whereas 40% know three languages: English, Italian, and another foreign language from French, Spanish, or German. Furthermore, just 15% of respondents master four or five languages.

Figure 4 – data about the languages spoken by the participants



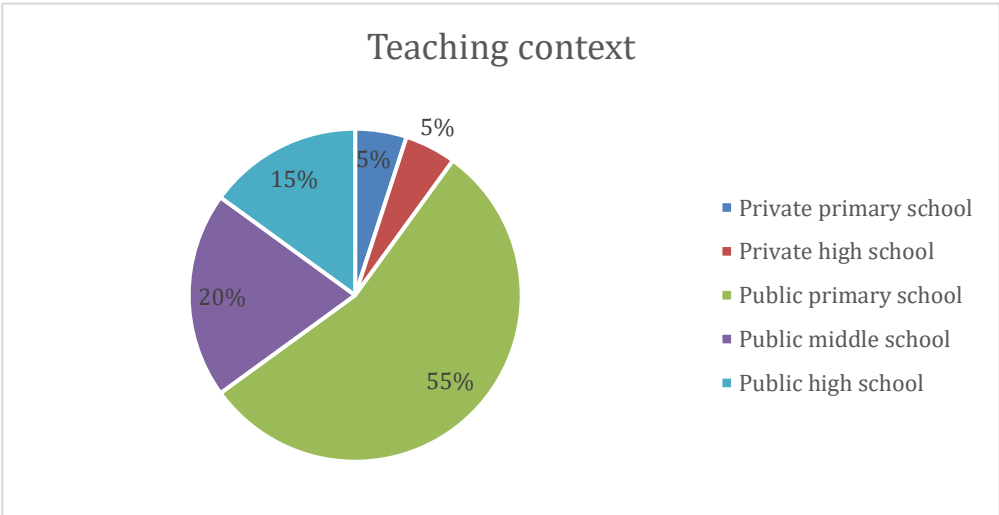
The pie chart in figure 5 clearly illustrates that participants have varying levels of teaching expertise, even if the bulk of respondents (50%) have only been teaching English for 1-2 years or are in their first year of teaching.

Figure 5 – data about the number of years of teaching experience



According to the data in figure 6, the majority of respondents work in public schools in Italy. As a result, only a small minority (10%) work at private institutions. Furthermore, 55% of the participants work in primary schools, while 20% are middle school teachers. Just 15% of teachers work in public upper secondary school settings.

Figure 6 – data about the teaching context



### 4.3 Assessment criteria

The present study aims to determine whether exist any difference between ChatGPT-generated and human-written reading comprehension exercises, as a result the participants in the research project must evaluate the quality of the reading assignments according to specific parameters which are presented in this section. To collect respondents' opinions on the form and content of the reading exercises, the criteria specified in the study by Clark et al. (2021, p. 7294) detailed in chapter two of the present thesis were chosen for the purpose of analysis. The sections “category”, “label”, and “description” in figure 7 are derived from the table established in the aforementioned study project, whereas the last column classified as “questions” contains phrases created from scratch for this research.

Figure 7 – table taken from “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text” (Clark et al. 2021) and additional questions

Category	Label	Description	Questions
Form	Grammar	The spelling and grammar of the text, punctuation and formatting issues	<ol style="list-style-type: none"> <li>1. The text presents no mistakes in terms of grammar.</li> <li>2. The text presents no mistakes in terms of punctuation.</li> <li>3. The text presents no mistakes in terms of spelling.</li> </ol>
Form	Level of detail	Is the text simple or does it go more in-depth?	<ol style="list-style-type: none"> <li>4. The text goes into detail about the main topic.</li> </ol>
Form	Genre	If the text is the genre/domain/style/formality that the reader expects, adheres to style norms.	<ol style="list-style-type: none"> <li>5. The text genre is the one expected by the reader.</li> <li>6. The text adheres to style norms.</li> <li>7. The level of formality is the one expected by the reader.</li> </ol>
Content	Repetition	Word/phrases/content repeated itself.	<ol style="list-style-type: none"> <li>8. There are no repetitions of words.</li> <li>9. There are no repetitions of phrases.</li> </ol>

			10. There are no repetitions of content.
Content	Factuality	The accuracy of the text, whether it describes things that are “true”.	11. The text presents facts that are true.
Content	Consistency	How the text relates to the context and other pieces of the text.	12. The text is consistent with the context and other pieces of the text.
Content	Common sense	Whether the text “makes sense” within the world that it is written.	13. The text makes sense within the world that it is written.
Content	Coherence	The structure and coherence of the text. Order issues go there.	14. The text is clearly structured and organised. 15. The text is logically ordered.

As shown in the chart, the categories in the first column were described not only by identifying distinct labels that correspond to certain features of form or content, but also by providing a more complete explanation of each piece in order to further define the evaluation criteria. For instance, in order to assess the form of the reading passage it is worth analysing the text genre which considers: “if the text is the genre/domain/style/formality that the reader expects, adheres to style norms” (Clark et. al, 2021). As a result, the decision to use the criteria offered in the table above as a starting point for developing the questions to be included in the questionnaire stems from the belief that the description of each label might serve as guidance for developing specific evaluation parameters. With respect to the text genre, three distinct items (number 5, 6 and 7 in the table) were developed on the basis of the previously mentioned definition:

- 5. The text genre is the one expected by the reader.
- 6. The text adheres to style norms.
- 7. The level of formality is the one expected by the reader.

The explanations in the third column therefore facilitated the development of the fifteen elements listed in the fourth section, which were generated by the researcher in order to obtain a set of phrases to be used for evaluating both human-generated and ChatGPT-developed reading comprehension exercises.

#### 4.4 Research Instrument

The research instrument built specifically for the current study will now be presented. For a variety of reasons, the questionnaire was selected as the best research instrument for gathering information on participants' perceptions of reading materials. To begin with, the low cost of the data collection and analysis process along with the minimal training required by the person administering the questionnaire emerge as two major advantages attributed to this research tool. Additionally, questionnaires can be delivered in various ways, and they can reach larger numbers of respondents than would be possible using other instruments, such as interviews (Jones et. al, 2008). The questionnaire created for this research endeavour has four main sections: consent acknowledgement, demographics, evaluation of the human-written reading passage, and evaluation of the AI-generated materials. Furthermore, each of the final two parts was divided into three components. First, the reader is presented with the reading text and the exercises that accompany it. Second, the quality of the text form is assessed using the aforementioned criteria, and finally, the content of the reading assignment is evaluated as well. The various sections are discussed further in the following paragraphs. It is worth noting that the sections "reading passage", "form", and "content" will be presented only once, despite the fact that they were repeated twice in the original questionnaire since the questions to evaluate the two reading passages are identical. The questionnaire was administered online, thus it was generated as a word document and then replicated as a Google Module that English teachers could simply access via a link.

##### 4.4.1 Consent acknowledgment

The first section is written in Italian and contains the title of the questionnaire, *Valutazione di materiali per lo sviluppo della capacità di comprensione scritta in inglese*, as well as some general information regarding the study project. Its purpose is to tell participants that the researcher is conducting the study for her master thesis, and they are encouraged to contact her via email if they have any concerns or issues. The goal of the questionnaire is also introduced in this part, and respondents are informed that the data acquired as a result of their responses will be used anonymously solely for research purposes, in accordance with *D.lgs 163/2017, Ex art. 13 D.L. 196/2003, and Ex art. 13 Regolamento Europeo 2016/679*. Finally, participants are required to insert their email to access the next section.

#### 4.4.2 Demographics

Given that the second section of the questionnaire attempts to collect demographic data from participants, teachers must supply some personal information by answering six questions. Fernandez et. al (2016) emphasise the need of minimising sensitive or re-identifiable information while collecting demographic data for research purposes. While the chance to collect all possibly important data in one occasion is appealing given the realities of research, collecting demographic questions if one does not intend to use them for future analysis is improper and potentially problematic. As a result, the information requested in this study is regarded as essential for defining the characteristics of research participants as well as discussing the study's findings in order to answer the research question. To begin, respondents are questioned about their gender and given three options: "male", "female", and "prefer not to say". To respect all gender minorities, the last option was added to the list of possible answers. As reported in Hughes et al. (2016, p. 140), according to APA (2015, p. 20) gender identity is defined "as a person's deeply-felt, inherent sense of being a boy, a man, or male; a girl, a woman, or female; or an alternative gender [...]. These gender identities may or may not correspond to a person's sex assigned at birth or to a person's primary or secondary sex characteristics". Thus, the decision to enquire about participants' gender rather than their sex given at birth stems from a desire to avoid being offensive. The second question covers participants' ages, and they are invited to select an age range in order to avoid having to reveal their exact age. The following question focuses on participants' places of birth, as respondents must indicate whether they were born in Italy or another nation, and in this case, they must identify the country they were born in. In the fourth question, participants are asked which languages they speak, and they are given six options, namely Italian, English, Spanish, German, French, and other languages. Although in previous questions respondents could only select one choice, in this case they could select more than one, given that they are expected to know at least Italian and English. The final two questions are designed to collect information regarding the research participants' teaching experiences. First, they are asked to specify how long they have been teaching English by selecting one of seven age ranges. Second, they must select their work context by indicating whether they work in a private or public school. In addition, in this question, they must identify whether they work in primary, middle, or high school. All these questions are mandatory to access the next section.

#### 4.4.3 Reading exercise

After submitting personal demographic information, participants are required to read a human-written reading comprehension passage, which is followed by some multiple-choice items and open-ended questions (the materials are fully described in Chapter 3 of this thesis). As a matter of fact, respondents were told the goal of the questionnaire in the first section, so that they understood they had to evaluate two reading passages and the related exercises. As a result, the content was preceded by a single sentence encouraging respondents to read it: *“Please, read the following reading passage and the corresponding reading comprehension activities”*. Furthermore, no reference is made to the source of the reading materials to be examined, which also applies to the AI-generated reading passage that will be presented to the participants later. This is critical to avoiding prejudice due to preexisting notions about ChatGPT and the use of AI in the development of educational materials. Thus, teachers evaluated each text without realising that one of the two reading passages was generated by AI to ensure reliability.

#### 4.4.4 Form

In this section of the questionnaire, participants are invited to evaluate the reading comprehension exercises according to the specific criteria illustrated in figure 7 (section 4.3 of the current thesis). The instructions provided are the following: *“Now you will be asked to agree or disagree with the following statements in order to evaluate the form of the text in terms of grammar, level of detail and genre. Remember that the reading exercise has been developed for Italian middle school English learners”*. At this point, respondents are presented with the items from 1 to 7 in the fourth column of table 7. Given that participants must declare their agreement or disagreement with each of the statements listed above, the likert scale was selected as the best alternative for collecting data on their thoughts. According to Taherdoost (2020, p. 4): *“Likert scale is simple to construct and likely to produce a highly reliable scale. Besides, from the perspective of participants, it is easy to read and complete”*. In this part of the questionnaire a 5-point likert scale is used to investigate participants’ beliefs on the quality of the reading passage form. Number 1 is connected with *“strongly disagree”* and number 5 with *“strongly agree”*, therefore participants who select numbers 4 or 5 want to indicate their agreement with the statement, instead numbers 1 and 2 are chosen to convey dissent. Consequently, given that the propositions were developed to express a positive attitude towards



the aspect to be analysed (i.e. “the text presents no mistake in terms of grammar” and “the text adheres to style norms”, etc.), those who agree with the statement are underlining the good quality of the reading exercise. Indeed, disagreeing with the statements is related with poor quality. Questions 1–3 examine the reading exercise in terms of grammar; question 4 focuses on the degree of detail; and questions 5–7 concern the text genre. Furthermore, participants are encouraged to justify their decisions in order to clarify the reasoning for their evaluation: *“Please, leave a comment to explain your choices. If you want, you can use some examples taken from the text”*. This question aims to collect qualitative data about teachers’ perspectives on the form of the reading comprehension assignment. This information is critical for understanding the motivations behind the participants’ evaluations and comparing the positive and negative features of the human-written exercise to the AI-generated reading materials.

#### 4.4.5 Content

The linkert scale has also been used to assess the content of the reading exercise. At the beginning of this section, participants are instructed as follows: *“Now you will be asked to agree or disagree with the following statements in order to evaluate the content of the text in terms of repetition, factuality, consistency, common sense and coherence. Remember that the reading exercise has been developed for Italian middle school English learners”*. Statements ranging from 8 to 15 (see figure 7) are presented to respondents, who rate their level of agreement with each statement on a scale of 1 to 5. Each question is designed to examine a specific aspect of the reading passage content. Items 8-10 considers the repetition of words, phrases, and content; question 11 involves factuality; question 12 analyses consistency; and question 13 speaks to common sense. Finally, questions 14 and 15 deal with coherence. Additionally, like in the previous section, participants must explain their selections by providing a comment, which is essential to go into detail regarding the differences between the two reading assignments. The instructions provided to the respondents are the same as those mentioned above: *“Please, leave a comment to explain your choices. If you want, you can use some examples taken from the text”*. There is no word limit for the final question, so participants can provide a quick explanation of their responses in a few sentences or write a more extensive analysis of the material. It is also worth mentioning that all items, included those needed to assess the reading exercise form, are mandatory to complete the questionnaire. After participants have evaluated the quality of the first reading activity, the AI-generated text is provided, and readers are kindly

asked to read it and analyse it. All questions presented in sections 4.4.4 and 4.4.5 are thus repeated twice, because ChatGPT-generated reading comprehension exercise must be evaluated using the same criteria as the human-written assignment.

#### 4.5 Administration

The research participants are English teachers currently working in Italian schools who took part in the research project on a voluntary basis. As a matter of fact, the researcher contacted twenty teachers of her acquaintance and they agreed to answer the questionnaire. Therefore, the sampling method used for this research is convenience sampling. This approach is described by Fraenkel et al. (2012, p. 99): “Many times it is extremely difficult (sometimes even impossible) to select either a random or a systematic nonrandom sample. At such times, a researcher may use convenience sampling. A convenience sample is a group of individuals who (conveniently) are available for study.” The benefits associated to convenience sampling are listed in (Golzar et al., 2022):

“First, the researchers can consume less effort to select the participants compared to other non-random sampling techniques. Second, convenience sampling requires the researcher to select participants at a very low cost. Third, the researchers invest less time since the sample taken from the target population is readily accessible.”

However, there are significant drawbacks to consider, especially the difficulties of controlling bias and of generalising conclusions beyond the sample (Acharya et al., 2013). As a result, the current study does not intend to extend the findings to the entire population of English teachers in Italy; rather, the study aims to report the opinions of the teachers who volunteered to participate in the research and use them to account for the differences between human-written and AI-generated reading materials. Participants can easily complete the questionnaire online via a link because the questions were reported on a Google Module. The questionnaire was delivered to the teachers at the beginning of May 2024, and the researcher collected all 20 responses by the end of the month. Some instructors waited longer to respond to the questionnaire because they are highly busy towards the end of the school year. Furthermore, they noted that it takes between 20 and 30 minutes to read the materials and answer the questions, so they wanted to take their time and deliberate on the answers they would provide.

## V. Results

The current chapter illustrates the data acquired via the questionnaire after all 20 participants had submitted their responses. The data was reported on Microsoft Excel to easily calculate indices of central tendency and dispersion (mean and standard deviation), with the primary purpose of determining participants' level of agreement with the statements reported in the questionnaire. The tables created in Excel to report all of the participants' responses are shown in the appendix (figure 26, 27 and 28). Every column corresponds to one of the questions, and each row contains a response from one of the participants. Figure 26 summarises the participants' demographic information, whereas figures 27 and 28 illustrate the English teachers' replies to evaluate the first and second reading comprehension passages (sections *testo 1* and *testo 2*). As a result, the purpose of this chapter is to first show the questionnaire's findings by examining the quantitative data using descriptive statistics. This section addresses participants' evaluations of the reading activities' form before focussing on the materials' content. The following section presents and analyses qualitative data in the form of comments written by teachers to further explain their selections. The analysis is conducted with the primary goal of addressing the research question: "Are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content?". The investigation indeed focuses on comparing the two reading passages and their related exercises in order to identify variations and explain the reasons for any potential inequities.

### 5.1 Quantitative data

The questionnaire described in Chapter four was primarily used to obtain quantitative data, therefore linkert-scale items were employed to assess participants' level of agreement with the 15 assertions reported in Figure 7 (fourth column). A five-point linkert scale was used to indicate agreement or disagreement with the information regarding the form and content of the reading passages. To evaluate the data, mean and standard deviation (SD) were determined in Excel. To better grasp the results, it is important to realise that number 5 denotes the maximum level of agreement, whilst number 1 corresponds to the lowest level, indicating disagreement with the statement. Andreda (2020, p.49) defines the mean and standard deviation as follows:

“If we regard distance from the mean as a positive number, the SD conceptually tells us how far from the mean the average person is. This indicates that if the SD is large, the values are widely scattered around the mean. In contrast, if the SD is small, the scatter is also small. Thus, the mean tells us what the average value is and the SD tells us what the average scatter of values is, around the mean.”

Tables and charts containing the mean and standard deviation of each item are provided below. Indeed, greater mean values correlate with higher levels of agreement, indicating that the reading passage is of high quality. Instead, lower mean values indicate disagreement and poor reading exercise quality. Small standard deviation, on the other hand, indicates a smaller spread of values around the mean, implying that values are not as distant from the mean. Higher standard deviation numbers represent a broader scatter around the mean, resulting in values that are less close to the mean.

#### 5.1.1 Form

This section presents descriptive statistics about the form of the reading comprehension passages. Regarding the human-written reading exercise (figure 8), item #3 has the highest mean and lowest standard deviation, hence teachers agree that the text has no spelling errors. This is also true for the passage created by ChatGPT (figure 9), as the identical item has a high mean value (4,35) despite a significant standard deviation (1,18). However, the two texts have some variances. Item #2 in the human-generated reading activity has the lowest mean, whereas in the AI-generated exercise, it has the highest value. As a result, teachers discovered that the first text contains several punctuation errors that do not appear in the second text. Furthermore, while item #7 in Figure 9 has the biggest standard deviation, it also has the lowest mean, which is not the case for the other reading passage. Consequently, many participants remarked that the level of formality provided by ChatGPT is not the one expected by middle school students. Another difference concerns item #5, in which teachers believe that the text genre of the AI-generated reading passage fails to meet the expectations of the intended readers, i.e. lower secondary school students. In fact, the mean score is 3,75, which is lower than the equivalent item in the human-written paragraph (4,36). Finally, the standard deviation value of all items is often larger in the AI-generated passage than in the other one, indicating that these values are widely distributed around the mean.

Figure 8 – *descriptive statistics of the human-written exercise, form (all items)*

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
<b>Human-written exercise</b>	The text presents no mistakes in terms of grammar	The text presents no mistakes in terms of punctuation	The text presents no mistakes in terms of spelling	The text goes into detail about the main topic	The text genre is the one expected by the reader	The text adheres to style norms	The level of formality is the one expected by the reader
Mean	4,20	3,80	4,50	4,05	4,36	4,00	4,10
Standard Deviation	1,01	1,28	0,69	0,83	0,88	0,92	0,79

Figure 9 – *descriptive statistics of the AI-generated reading exercise, form (all items)*

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
<b>ChatGPT-created exercise</b>	The text presents no mistakes in terms of grammar	The text presents no mistakes in terms of punctuation	The text presents no mistakes in terms of spelling	The text goes into detail about the main topic	The text genre is the one expected by the reader	The text adheres to style norms	The level of formality is the one expected by the reader
Mean	4,25	4,35	4,30	3,95	3,75	4,20	3,45
Standard Deviation	1,16	1,18	1,17	1,22	1,16	0,95	1,54

The overall results concerning the form of the reading exercise are provided in the table below. Although the human-written exercise has a higher mean than the AI-generated text and a lower standard deviation, the difference in mean values is minimal, and the degree of agreement exceeds 4 in both passages. As a result, teachers agree that both reading comprehension tasks are of good quality in terms of written text form, although the first passage received a slightly higher grade.

Figure 10 – *descriptive statistics of the form (total)*

	<b>Human-written exercise</b>	<b>AI-generated exercise</b>
Mean	4,14	4,04
Standard deviation	0,93	1,18

The following charts are created to compare the human-written and the AI-generated reading assignments using the criteria outlined in Chapter 4. Thus, the mean and standard deviation of the items related to each criterion, namely grammar, level of detail, and genre, are pooled together to understand the influence of each component on the overall findings. As a starting point, figure 11 illustrates the opposite trend from the general results examined above, i.e., the mean value of the AI-generated text (4,30) is greater than that of the human-written passage (4,17) in terms of grammar. Nonetheless, the standard deviation is larger too, therefore the responses are not close to the mean value.

Figure 11 – *descriptive statistics, grammar*

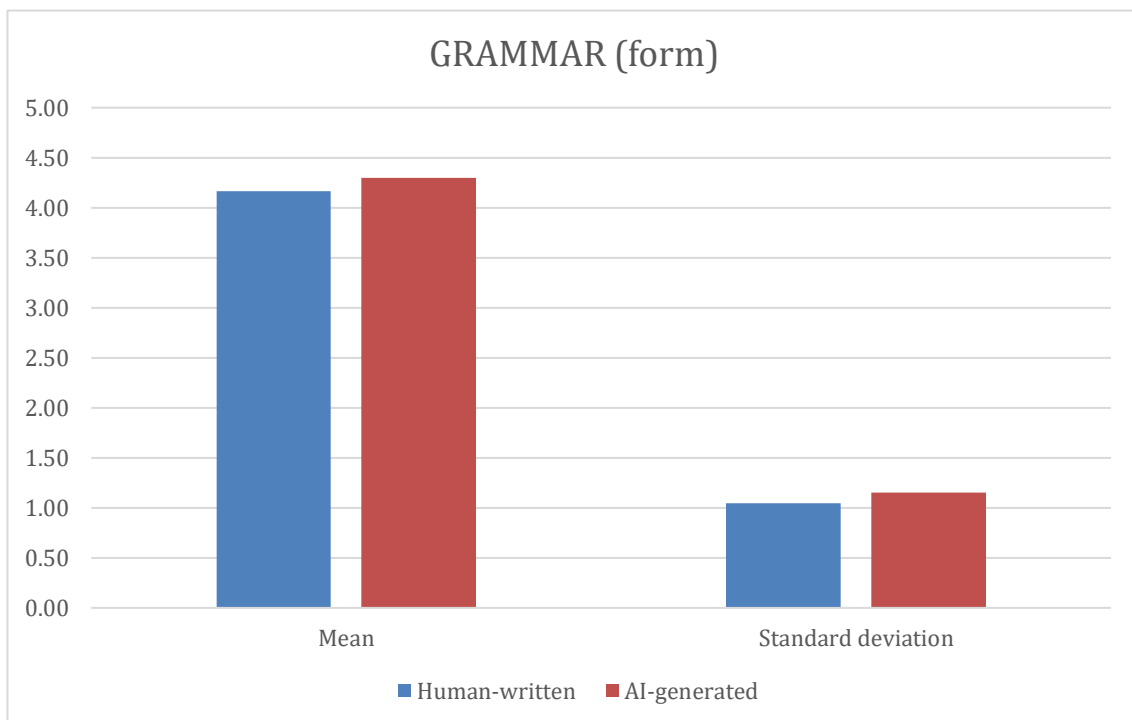
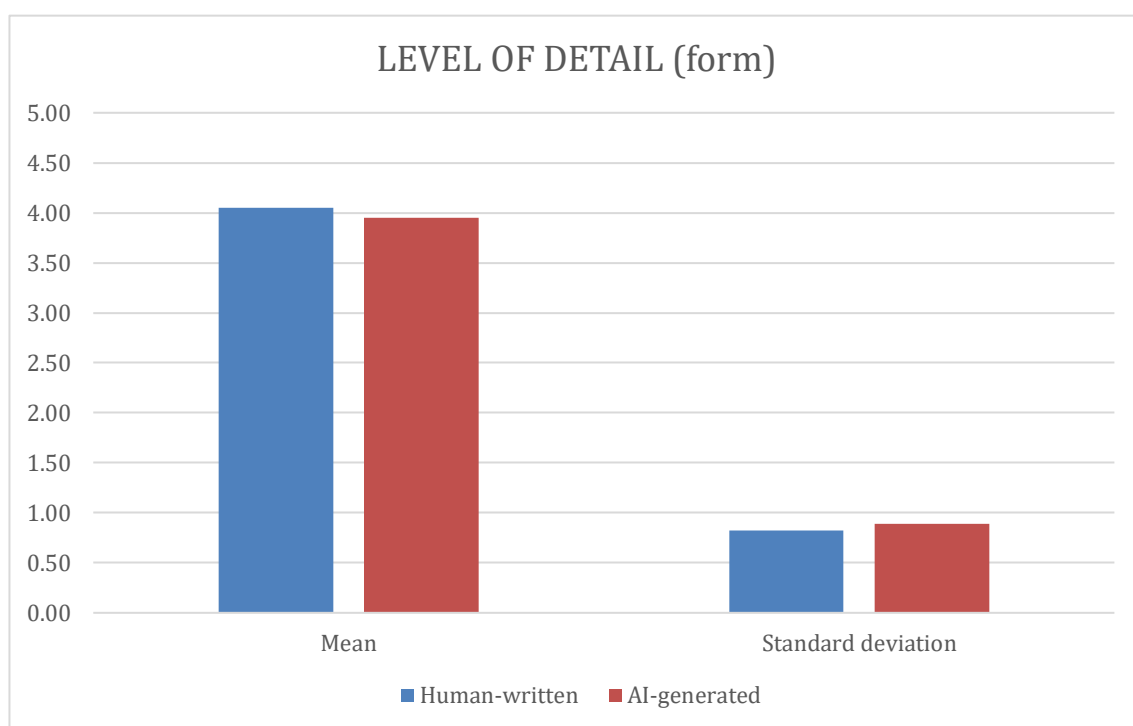


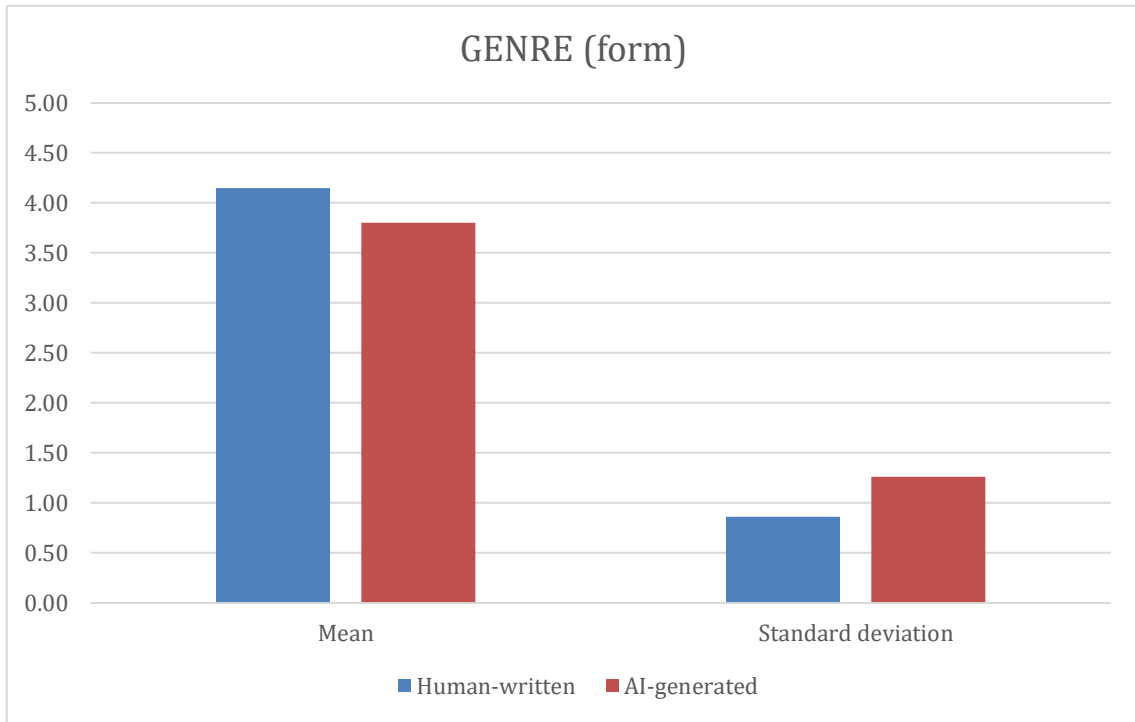
Figure 12 depicts descriptive statistics for the level of detail. Participants felt that both reading activities go into detail about the main topic, even though the human-written exercise presents a slightly higher mean value (4,05) and a lower standard deviation (0,83) than the text created by ChatGPT.

Figure 12 – *descriptive statistics, level of detail*



According to the statistics presented in Figure 13 about the genre of the reading activities, the human-written activity has a mean of 4,15, which clearly outperforms the mean of the AI-generated text (3,80). Teachers' responses demonstrate that their degree of agreement with the statements is higher for the first text than for the second. As a result, the genre of the human-generated text is more suited to middle school students than that of the AI-developed passage. The smaller standard deviation for the first activity compared to the second one verifies this trend.

Figure 13 – *descriptive statistics, genre*



### 5.1.2 Content

In terms of the content of the reading comprehension passages, the questionnaire findings are presented in the charts below (figures 14 and 15). To begin, the mean values for items #8 and #9 in the human-written reading comprehension task are lower (3,35 and 2,95, respectively) than those in the AI-generated assignment. This is also true for item #10, which has the highest mean in the ChatGPT-generated exercise (4,60). For these items, the standard deviation is lower in the AI-generated text, indicating that participants' responses are close to the mean values. As a result, teachers found that the first text presents more repeated elements than the second one. With respect to the human-written text, item #11 achieves the highest level of agreement while also having a low standard deviation. Nonetheless, the passage produced using ChatGPT also has a high score (4,40) with the same standard deviation. Teachers therefore agree that both passages depict true facts. Furthermore, both reading comprehension exercises are judged as logically organised, since the mean of item #15 is the same in both texts (4,35) and has a similar standard deviation.



Figure 14 – *descriptive statistics of the human-written exercise, content (all items)*

	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15
<b>Human-written exercise</b>	There are no repetitions of words.	There are no repetitions of phrases.	There are no repetitions of content.	The text presents facts that are true.	The text is consistent with the context and other pieces of the text.	The text makes sense within the world that it is written.	The text is clearly structured and organised.	The text is logically ordered.
Mean	3,35	2,95	3,80	4,60	4,15	4,10	4,40	4,35
Standard Deviation	1,39	1,47	0,95	0,60	0,67	0,72	0,99	0,81

Figure 15 – *descriptive statistics of the AI-generated exercise, content (all items)*

	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15
<b>AI-created exercise</b>	There are no repetitions of words.	There are no repetitions of phrases.	There are no repetitions of content.	The text presents facts that are true.	The text is consistent with the context and other pieces of the text.	The text makes sense within the world that it is written.	The text is clearly structured and organised.	The text is logically ordered.
Mean	4,10	4,30	4,60	4,40	4,55	4,20	4,30	4,35
Standard Deviation	1,34	1,12	0,66	0,60	0,82	0,60	1,06	0,93

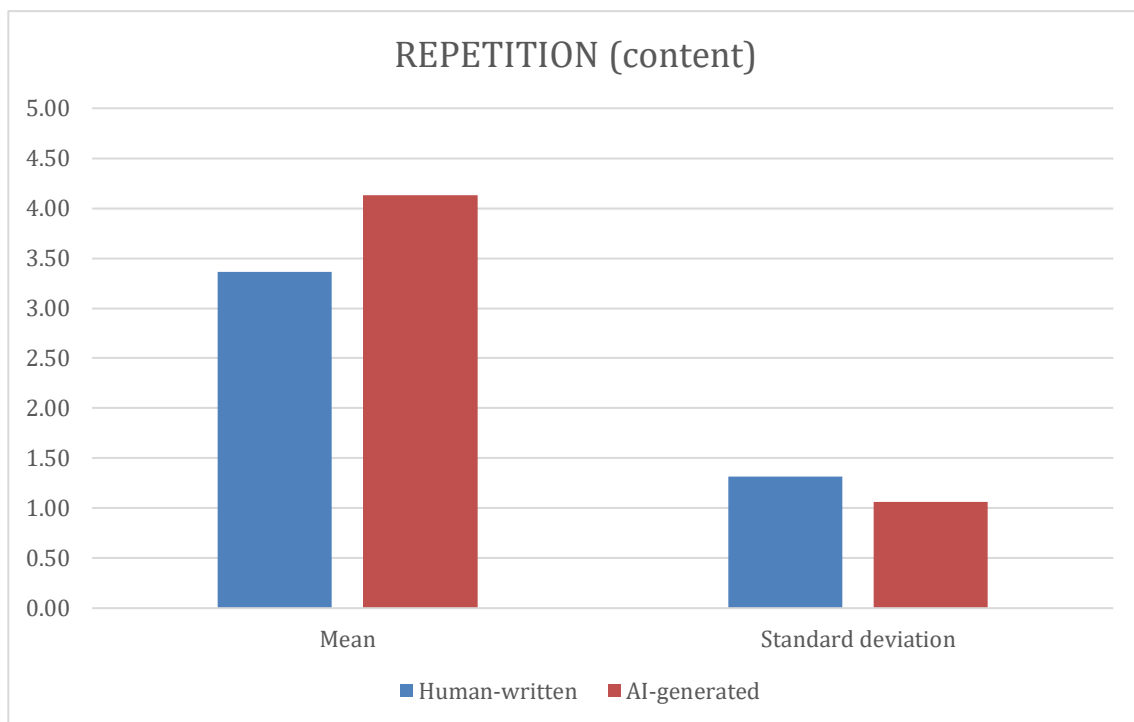
Figure 16 illustrates the overall results concerning the content of the reading comprehension passages. The human-written exercise has a lower mean and higher standard deviation than the AI-generated activity, as seen in the table. As a result, participants agreed more with the statements about the content of the reading exercise created by ChatGPT than with the human-generated text.

Figure 16 - *descriptive statistics of the content (total)*

	<b>Human-written exercise</b>	<b>AI-generated exercise</b>
Mean	3,96	4,31
Standard deviation	1,11	0,93

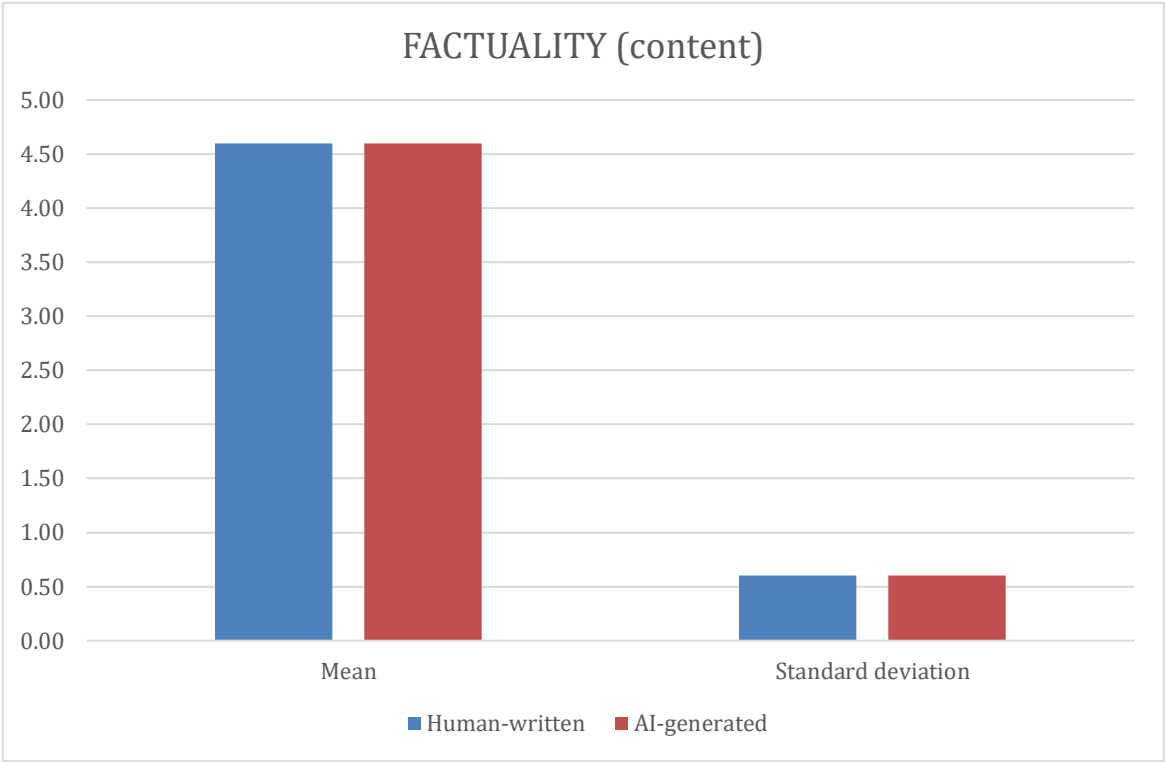
The following charts show descriptive statistics for items ranging from 8 to 15, classified according to five particular criteria listed in Chapter four. Indeed, the reading passages' content was evaluated in terms of repetition, factuality, consistency, common sense, and coherence. Considering repetition of words, phrases and content all together, the chart in figure 17 clearly illustrates that the mean value of the human-generated text is lower than that of the AI-generated text. Furthermore, the two results are significantly different (3,37 and 4,41 respectively). Finally, the standard deviation for the AI-generated assignment is lower than that of the other exercise. As a result, teachers' responses about this activity are closer to the mean, demonstrating the data's reliability.

Figure 17 – *descriptive statistics, repetition*



The chart below shows participants' levels of agreement with the factuality items. The mean and standard deviation for both reading passages are exactly the same, at 4,60 and 0,60. As a result, both texts are regarded as very high quality reading comprehension activities in terms of the truthfulness of the themes covered.

Figure 18 – *descriptive statistics, factuality*



The content of the reading passages was also evaluated for consistency and common sense. According to figure 19, descriptive statistics demonstrate that the AI-generated reading assignment has a slightly higher mean value than the human-generated text. As a result, participants think that the ChatGPT-generated exercise is the most consistent with the context and other parts of the text. Nonetheless, the reduced standard deviation in the human-written text indicates that the data for this exercise is not widely scattered around the mean; at the same time, the AI-generated data shows the opposite pattern. This is true also for figure 20, which depicts how teachers evaluate common sense. According to the chart, participants are more likely to agree with the common sense item in the AI-generated section than with the human-written assignment. Consequently, the ChatGPT-developed exercise has a mean of 4,20, but the

mean of the other exercise is 4,10. As a result, while both reading activities adhere to common sense standards, the AI-generated text had a higher level of agreement.

Figure 19 – *descriptive statistics, consistency*

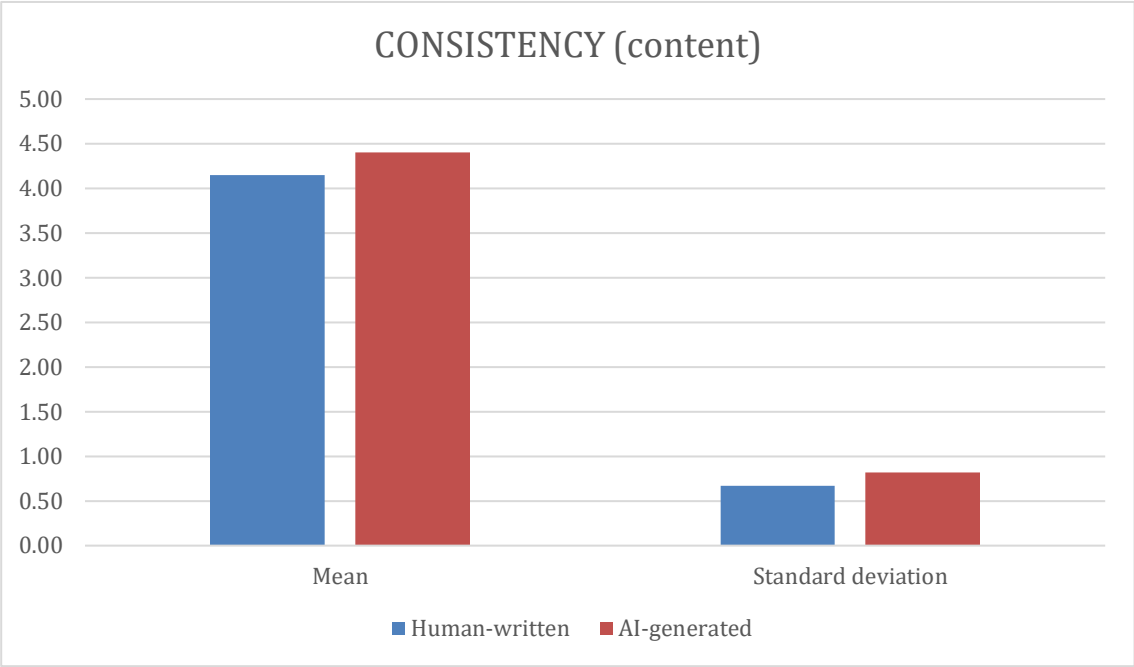
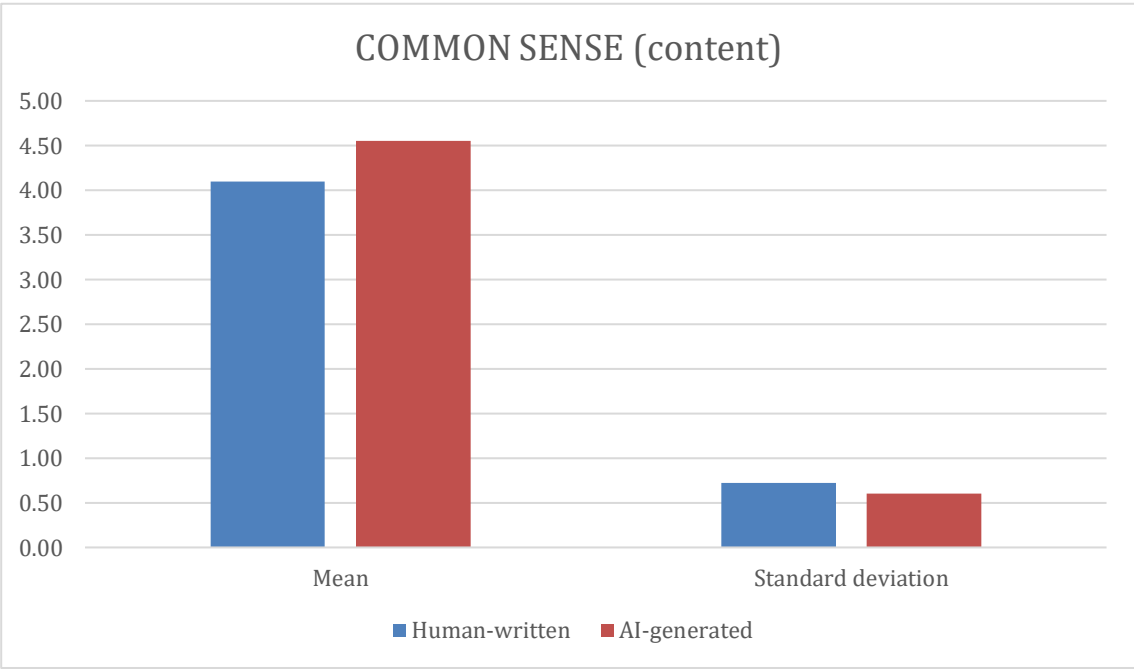
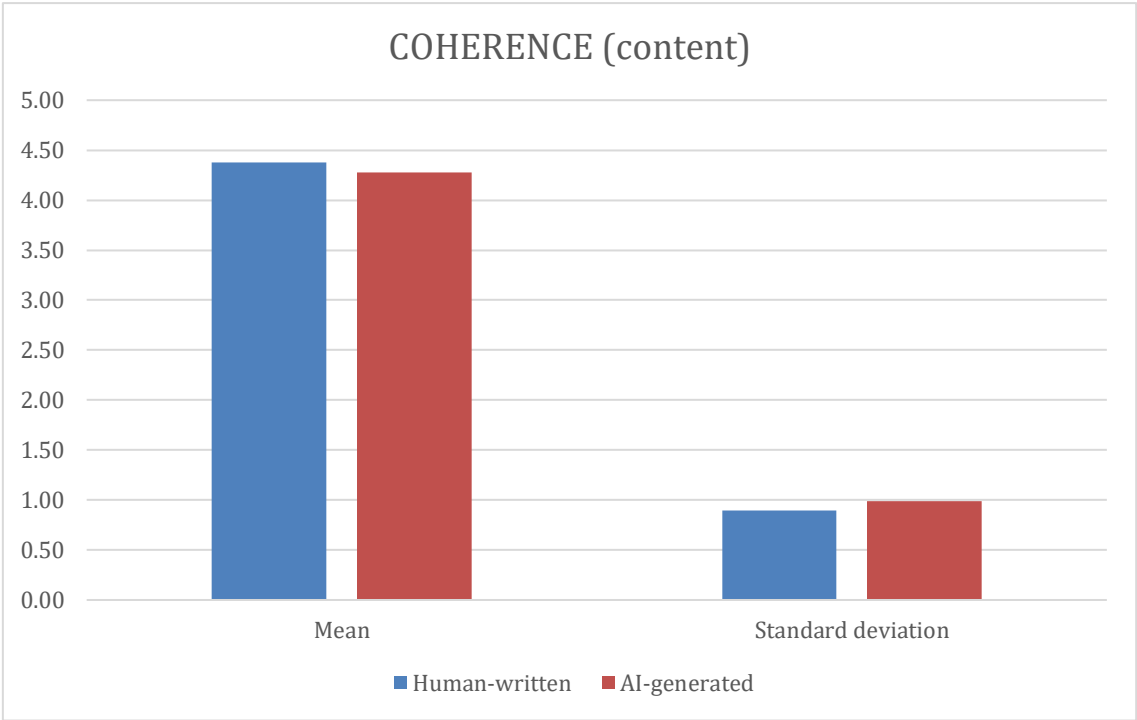


Figure 20 – *descriptive statistics, common sense*



Finally, the reading comprehension exercises were evaluated in terms of coherence. The chart in Figure 21 shows that the mean value of the two passages is similar, despite the fact that the level of agreement is slightly higher in the human-written text (4,38). Furthermore, this passage has a lower standard deviation (4,90), indicating that participants’ responses are less dispersed around the mean.

Figure 21 – descriptive statistics, coherence



5.2 Qualitative data

Given that the current study’s goal is to determine how human-written and AI-generated reading activities differ from one another, the research instrument was designed to collect qualitative data to better understand teachers’ perspectives on potential inequalities between the two reading comprehension exercises. The most relevant comments are included in the tables presented in the sections below. Some of the comments were excluded from the tables since they were deemed irrelevant for research reasons. In reality, some of the participants did not provide any analytical explanation for their replies, instead they wrote extremely simple remarks such as “the text is correct” or “the text is ok for students”. Teachers’ statements are

reported in the tables below exactly as they were written in the questionnaire; as a result, some responses are in Italian and others in English, as participants were given the option of using either of the two languages. All the replies were classed employing the same criteria used for quantitative data, which means that comments on the form and content of the reading passages are divided into distinct categories to ease the analytical process.

### 5.2.1 Form

As shown in Figure 22, participants' comments on the form of the human-written exercise are mostly concerned with the grammatical aspects of the reading passage. In fact, grammar is addressed in six of the fourteen comments selected for examination. More exactly, the bulk of the responses emphasised that the text presents some grammatical errors, while only one comment (item #4 in the table below) highlighted the passage's favorable characteristics. More specifically, teachers focused on the punctuation element, stating that the text should be enhanced by adding commas to make it clearer. Furthermore, one of the participants suggested adjusting the usage of pronouns when addressing the reader, while another teacher stated that some expressions in the text may be regarded as too difficult for the intended readers. Nevertheless, aside from these minor details, the text does not present significant issues. Only three of the participants made comments on the level of detail in the reading text. Although one of the responses claims that some parts of the passage are slightly redundant, the text does not contain any errors in terms of level of detail, since the other comments emphasise that it is clear and accessible to students. This is also true for the reading exercise genre, as the five comments on this feature indicate that the text is similar to a normal reading passage seen in English textbooks for middle school pupils. Furthermore, the participants found the language to be fairly informal, which was understandable considering that the target readers were middle school students.

Figure 22 – *participants' comments on the human-written exercise*

Grammar	1. Mi è sembrato un testo informale, non sempre corretto nella forma e nella punteggiatura (“Simple activities like walking and cycling are great forms of exercise” mancherebbero delle
---------	--

	<p>virgole “Simple activities, like walking and cycling, are great forms of exercise.”).</p>
	<p>2. Il testo presenta alcuni errori grammaticali e di punteggiatura.</p>
	<p>3. There are a few grammatical mistakes. In the second paragraph, the phrase “Every day you should spend some time outside in nature” would be clearer if it were written as “Every day, you should spend some time outside in nature.” In the third paragraph, “Teenagers need eight to ten hours’ sleep every night” could be improved by adding a comma after “night” for better clarity: “Teenagers need eight to ten hours’ sleep every night.” In the same paragraph, “Studies show that teenagers who don’t sleep enough are irritable, get headaches, and their grades at school suffer” could be revised for clarity: “Studies show that teenagers who don’t get enough sleep are irritable, experience headaches, and suffer in their school grades.” Punctuation: Commas could be added for clarity and flow in certain places, such as after introductory phrases or to separate items in a list. There is a slight inconsistency in using “you” and “one” to address the reader. It’s better to choose one and stick with it throughout the text. the reading exercise has been developed for Italian middle school English learners so the informality of the texts is predictable.</p>
	<p>4. The structure of the text is simple in grammar and vocabulary. Accessible even for those less skilled in the language.</p>
	<p>5. Non avrei usato la forma contratta, ma dipende tutto dal contesto.</p>
	<p>6. There are some expressions that middle school students may not understand, for example “bump your head”.</p>

Level of detail	7. The text is clear and could be perfectly accessible to middle school students.
	8. The article was very interesting and well written with concepts explained clearly.
	9. Some parts of the text are slightly redundant and repetitive.
Genre	10. It is an informal piece of writing. So, abbreviations, words contractions, or other features are expected.
	11. The text is well-written, there are no mistakes and the style is not too formal as the text is written for middle school students.
	12. Il testo presenta tutte le caratteristiche di una lettura indicata per ragazzi della scuola media.
	13. Il testo, dopo una veloce lettura mi appare coerente e senza errori.
	14. Il testo rappresenta in tutto e per tutto un tipico testo scolastico che uno studente è abituato a vedere nei libri di testo. Non sono presenti errori di vario genere.

Participants were also asked to evaluate the form of the AI-generated passage. According to the data in figure 23, the majority of the responses (nine) focus on the text's genre, while only six are concerned with the passage's grammatical elements and level of detail. Teachers did not identify any grammar errors, yet this reading passage was considered substantially more challenging than the previous text. In fact, comments #5 and #6 on the level of detail suggest that this exercise is more sophisticated than the others in terms of structure and language. One of the participants also stated clearly that this exercise level was significantly higher than that of the other. Nonetheless, the majority of comments on the text genre focused on the reading passage's formality, emphasising that it is a formal text with highly specific terms. Even if not all the participants claim that the AI-generated passage may result in being too challenging for middle school students, two of the teachers openly expressed their concern that the exercise may be too advanced for lower secondary school pupils (see comments #8 and #15). Furthermore, one of the teachers observed that the text contains some elements that are not typical of formal language (such as shortened verbs). These features are intended to appear in



a more informal reading passage, such as the preceding piece of writing, rather than a formal one. Finally, certain references to the text’s structure can be found in the responses, even if more accurate explanations are provided in the comments on the content of text (see figure 25). Participants generally agree that the section is well-structured and ordered, which is significant given that it is an informational text (the genre was properly identified in item #10).

Figure 23 – *participants comments on the AI-generated exercise*

Grammar	1. The text is well-written and does not contain any obvious grammatical mistakes, punctuation errors, or spelling errors. The text adheres to style norms commonly found in informative or self-help content. It uses clear and concise language, presents information in a structured manner, and offers actionable advice.
	2. This part is correct in terms of punctuation.
	3. Some terms might be a little difficult for Italian middle school English learners.
Level of detail	4. The text is clear, much more detailed and well-explained than the first one.
	5. This text is more difficult to read than the other and it contains specific terms.
	6. Il testo corrisponde a un livello più avanzato, leggermente più complesso in termini di struttura, scelta del vocabolario e divisione testuale.
Genre	7. It is a formal text, yet sometimes there are parts where other styles are encountered (such as shorten versions of verbs).
	8. Il secondo testo, seppure uguale al primo in termini di contenuto, è presentato in maniera più formale, meno alla portata di ragazzi della scuola secondaria di primo grado, a mio parere.
	9. The text is complicated on a formal and lexical level.
	10. The text adheres to style norms commonly found in informative or self-help content. It uses clear and concise language, presents information in a structured manner, and offers actionable advice.

	11. Ho incontrato un testo più formale, con un linguaggio più elevato e termini più specifici.
	12. The second text is more difficult than the first.
	13. The text is clear and well organised. Vocabulary is used accurately.
	14. Livello alto, testo ben strutturato, lessico impegnativo, dipende sempre a chi è rivolto il testo, l'età degli studenti
	15. I think the text has many words that middle school students don't know, like ensure, tackle, furthermore... and because of that the formality level is too high.

### 5.2.2 Content

Figure 24 depicts a table with participants' comments on the content of the human-written reading comprehension exercise. The bulk of responses (six out of fifteen) address the issue of coherence. As a result, items from #10 to #15 stress that the material is well-organised and has a clear structure; one of the participants also noted how the numbered list makes the content appropriate for middle school students. According to the teachers' responses, the titles of the various paragraphs also help to produce a logical piece of writing. Four comments address phrase and content repetition, with each asserting that the text contains repeated elements, particularly in terms of verb forms. However, the participants believe that such repeated expressions help to convey the correct meaning of the text, improving its clarity and emphasis (see items #1 and #3). The remaining five remarks are concerned with factuality, consistency, and common sense. These comments often emphasise the positive aspects of the text, providing that the topic addressed in the reading exercise is described as engaging and relevant to the target readers. Furthermore, participants wrote that the passage is consistent with the content it intends to express and adheres to common sense standards, resulting in a rationally structured piece of writing (see item #8). As a result, all the remarks on the content of the human-written text have an overall positive connotation, given that also the repetition of content is considered predictable in light of the text's main objective.

Figure 24 – *participants' comments on the human-generated exercise*

<p>Repetition</p>	<p>1. Il testo tende a ripetere più volte i concetti, soprattutto tramite l'utilizzo delle stesse forme verbali. Segue un filo logico ed è ben organizzato. Anche grazie alle ripetizioni dei verbi risulta molto chiaro.</p> <p>2. In some parts of the text there are repetitions.</p> <p>3. There are some repetitions of words and phrases in the text, but they are used for emphasis and clarity. For example, "every day" is repeated in several sentences to reinforce the idea of daily habits for happiness and health. This repetition helps drive home the importance of these activities. In terms of content, the text provides common sense norms for maintaining physical and mental well-being. It covers various aspects that are all widely recognized as beneficial practices for overall health. The text is clearly structured. It presents four simple habits to promote happiness and health, each with its own section and explanation. The order of the habits also makes logical sense, starting with going outside, followed by physical activity, sleep, and finally, laughter and hugging.</p> <p>4. Il testo mi sembra semplice, chiaro e ben strutturato con qualche ripetizione.</p>
<p>Factuality</p>	<p>5. The topic is current and interesting for children.</p>
<p>Consistency</p>	<p>6. I think the text is consistent with the contents it wants to explain.</p>
	<p>7. L'elenco numerato e aperto da un titolo che sintetizza il contenuto del paragrafo, fa sì che il testo risulti ben organizzato e alla portata di studenti della scuola</p>

	secondaria di primo grado. Rende il contenuto più immediato per i ragazzi.
Common sense	8. Properly developed for Italian middle school English learners.
	9. Mi è sembrato un testo coerente con il senso comune, con una struttura logica e argomentativa adeguata alla formalità.
Coherence	10. The text is written clearly and is easy to read for the learners.
	11. The text's structure is organised in a very mindful way because it's clearly drafted. It is divided into different sections which made the text easier to read.
	12. The structure of the text is clear as there are different titles for each paragraph.
	13. The text is ordered from the most to the least important things that you could do to be happy and healthy.
	14. The text is well structured.
	15. The text is coherent and has no repetitions.

Figure 25 summarises participants' observations on the content of the AI-generated exercise. The majority of the replies in this table, as in the one depicted in figure 24, relate with the coherence issue. However, the content of the six remarks on this feature differs slightly from the ones discussed above. Although some participants indicated that this is a well-structured piece of writing with linkers to reinforce the structure, five of them stated that the material is too difficult for the target audience (see comments #13, #14 and #15). According to the teachers, the text lacks subtitles, which would make reading easier for middle school students. Furthermore, several sentences are excessively long, and the text includes specialised terminology that could hinder students' comprehension of the reading passage. Overall, participants' comments appear to indicate that the argumentative framework of this reading exercise is less evident than that of the previous activity. There are three remarks on the subject of repetition. However, just one person noticed that some repeated expressions appear in the text (item #1). According to item #3, the text does not contain repetitive content, and its

organisation is clear, but it is perhaps too difficult for readers, considering that it is likened to an academic essay. Furthermore, the text not only contains generally true facts and is consistent with the topic that it seeks to cover, but it also follows common sense norms, as seen by the remarks on factuality, consistency, and common sense.

Figure 25 – *participants' comments on the AI-generated exercise*

Repetition	1. There is some repetition of phrases, such as “every day”, which is typical in informative writing to reinforce key points as before.
	2. The text is coherent and has no repetitions.
	3. Non sono presenti ripetizioni, né di contenuto né di vocabolario scelto. L’organizzazione segue quella di un essay accademico, chiaro ma probabilmente un po’ più complicato per studenti di livelli bassi.
Factuality	4. The text presents facts that are generally true, such as the importance of nutrition, physical activity, stress management, hygiene, and sleep for maintaining good health and happiness as before.
Consistency	5. I think the text is consistent with the contents it wants to explain.
	6. The text is organized into five paragraphs, each discussing a specific daily habit for happiness and health. This structure flows well and allows the reader to easily grasp the main points and builds upon each previous point.
	7. The text is more complex and delves deeper into the topic, but it's difficult to understand if you don't know English very well.
Common sense	8. The habits are introduced in a logical order, starting with nutrition, followed by physical activity, stress management, hygiene, and sleep.

	9. Le idee proposte rispettano il senso comune, hanno una struttura argomentativa meno chiara del testo precedente.
Coherence	10. The text is boring and uninteresting despite the topic. The concepts are expressed in sentences that are too long.
	11. The text is better organised, also thanks to the use of some linkers.
	12. Il testo è presentato in maniera ordinata e ben scandito dai connettivi logici.
	13. The structure of the text is clear but it is a bit too difficult for middle school learners as there are no subtitles. Some words are not easy to understand for middle school students (furthermore, to tackle, fulfill).
	14. The text hasn't a clear structure, there are many paragraphs that don't have a logical order. There isn't any title that would make the reading easier.
	15. Alcune parti del testo risultano troppe complesse per dei ragazzi che frequentano la scuola media

## VI. Discussion

This chapter aims to present the research findings by critically analysing the data illustrated in the previous section to answer the research question: “Are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content?”. Furthermore, the hypothesis presented in chapter three that reading comprehension exercises generated by ChatGPT can be evaluated similarly to human-written activities will be investigated. Before delving into the differences between the form and content of the reading assignments, an overall comparison is made between the human-written reading comprehension exercise and the AI-generated activity. Furthermore, the results of the study project are compared to past research reported in the second chapter of the current thesis. Additionally, the study’s limitations are presented along with some input to conduct future research concerning the impact of AI in the field of language education. Some reference to the educational implications of the present research project is finally made.

### 6.1 Research findings

The current study compares a human-written reading comprehension assignment to an AI-developed exercise using evaluations from qualified Italian English teachers. As a result, the primary purpose of the questionnaire designed to assess the reading tasks was to discover any potential discrepancies between them and to explain the reasons for such differences. In terms of quantitative data, the results shown in Chapter 5 report the mean score and standard deviation of the questionnaire items, which summarise participants’ responses indicating their level of agreement with the statements concerning both the form and content of the reading exercises. The overall scores in figures 10 and 16 show no significant difference between the two reading assignments, whose mean values are both high (i.e., they are all near to 4, indicating a fairly high level of agreement with the claims). Given that teachers evaluated the reading exercises in terms of form and content without knowing that one of the texts was generated by ChatGPT, the outcomes indicate that they considered the AI-developed assignment on par with the human-written activity. As a result, research findings highlight the enormous potential of the AI-powered chatbot ChatGPT 3.5 to generate high-quality reading comprehension materials for EFL middle school students; thus, the AI-technology that powers ChatGPT may be regarded as a valuable source for developing personalised teaching materials to boost pupils’ reading

comprehension skills. Standard deviation scores were also analysed to better understand how values vary around the mean. However, the form and content of the reading assignments correlate with varied SD values (see figures 10 and 16). In the human-written activity, participants' responses assessing the form of the text were closer to the mean than those concerning the content of the passage. On the contrary, the AI-developed exercise had lower standard deviation values with respect to its content rather than its form.

Apart from a minor variation in the scatter of participants' responses around mean values, results show that teachers found no significant divergence between the two reading assignments, implying that ChatGPT can be effectively used by language instructors to generate reading comprehension activities using AI. This is fundamental for teachers to save time while acquiring customised resources that meet their needs and specifications, as reported in Gupta & Bhaskar (2020, p. 700) and Ahmad et al. (2017): "Through the use of AI, teachers can have customised content that is aligned with their students' needs. [...] Moreover, the use of AI can assist them in content development, assignment designing, and assessments, thereby providing them sufficient time for quality teaching". Furthermore, Jusak (2024) stated that, given the growing relevance of AI-generated content, it is vital to understand how to best utilise it. In fact, AI can produce language-based products like exercises, quizzes, and short stories that are appropriate for different levels of proficiency in language learning. The current study's findings, hence, are consistent with existing literature on the potential of ChatGPT to develop teaching resources. For example, Karaman and Göksu (2024, p.120) examined the quality of AI-generated lesson plans: "The means of achievement test applied at the end of the process revealed that although there was no significant difference in academic achievement between the groups, the lesson plans suggested by ChatGPT were as effective as the existing lesson plans applied in schools to increase academic achievement". In addition, Mikeladze (2023) claimed that designing products with ChatGPT also allows teachers a great deal of flexibility. They can modify their approach to suit different levels, techniques, logistics, technological integration, instructional styles, evaluation procedures, and intended outcomes. This flexibility allows teachers to meet the needs of their students, resulting in an inclusive and dynamic learning environment. However, a more in-depth analysis based on the evaluation criteria given in Clark et al. (2021) is undertaken in the following sections, beginning with the results relating to the form of the reading texts and then focussing on their content. In this way, teachers' feedback is used to provide a more complete description of the AI-generated text and to highlight any minor



variations between reading activities. To delve deeper into teachers' observations and explain the reasons for their responses, qualitative remarks are analysed and interpreted alongside quantitative data.

### 6.1.1 Grammar

As a starting point, the form of the reading comprehension exercises is analysed in terms of grammar, level of detail and genre. Clark et al. (2021, p. 7285) draw a contrast between a text's form and content: "Form qualities focus on the format, style, and tone of the text, while content focuses on the text's meaning." Grammar is the first criterion to be examined, which also includes punctuation and spelling issues (items #1, #2 and #3 in the questionnaire). Ur (2012, p. 75) defines grammar as: "the way words are put together to make correct sentences" whereas Harmer (2015) says that grammar is concerned with syntax, which is a set of rules governing the order in which the various elements of a phrase can appear. However, the author claims that grammar is equally related to the formation of words and how they may change form to communicate meaning. Khanh (2020, p. 187) provides a more thorough definition: "Based on different perspectives of many language scholars, grammar can be generally considered as the knowledge concerned with types of words, phrases, clauses, sentences, and the internal rules of how sentences in a language is constructed." As a matter of fact, the results in figure 11 reveal that the AI-generated assignment surpassed the human-written exercise in terms of grammatical features evaluation. This is primarily due to punctuation issues, as participants discovered some problems with this component in the first text (also see figures 9 and 10). Given that the human-written assignment was found in a school textbook, teachers' disagreement with the statements concerning the first text is unexpected. Nonetheless, it is worth noting that the variation in mean value between the two exercises is minimal, and it could be attributed to teachers' personal preferences in terms of writing style. To further understand participants' beliefs, the qualitative statements in figure 22 are analysed. According to their comments, the bulk of punctuation issues are related to commas, which might be placed after introductory sentences and used to separate parts in a list to better explain specific concepts and increase clarity and flow. Furthermore, although some comments mention the incorrect use of pronouns and contractions, these small features have little bearing on the text's intelligibility. This is inferred by the fact that incorrect pronoun usage is merely defined as "a slight inconsistency" while contractions are defined as being appropriate for some settings (e.g., middle school textbooks).

Thus, participants' observations suggest that tiny changes can be made to provide a better written reading exercise, even if the text is still entirely readable as supplied in the questionnaire and there are no significant mistakes that necessitate drastic grammar changes. On the one hand, these slight criticisms of the human-generated text demonstrate that participants were totally uninformed of the two reading assignment sources; on the other hand, they suggest some adjustments that some teachers may consider important to improve text clarity.

Regarding the AI-generated assignment, participants' comments show that the material contains no grammar problems. This data is consistent with the information stated above, since the mean value of the human-written text is lower than that of the passage created by ChatGPT. These findings highlight the grammatical correctness of ChatGPT-generated output, implying that the language model can generate clear and well written reading comprehension activities with no spelling or punctuation errors. Nonetheless, it is worth noting that AI-generated output typically lack punctuation elements to express emotions and concepts in a personal manner; in contrast, human-written materials do contain such features that allow them to communicate their emotions and engage the reader. For instance, De Cesare (2023) compared human-written and AI-generated biographies, concluding that ChatGPT's output uses a lower set of punctuation marks than the biographies accessible on Wikipedia. Additionally, according to Guo et al. (2023, p. 6):

“ChatGPT expresses less emotion in its responses, while human chooses many punctuation and grammar feature in context to convey their feelings. Human uses multiple exclamation marks, question marks, ellipsis to express their strong emotion, and use various brackets to explain things. By contrast, ChatGPT likes to use conjunctions and adverbs to convey a logical flow of thought.”

In fact, punctuation marks are rarely employed in the ChatGPT-generated reading exercise under investigation in the present study, although this is not the case in the human-written text. As a result, while teachers' evaluations of the grammar features emphasise that there are no errors in the AI-generated text and that clarity is ensured, it is worth noting that the reading assignments differ significantly in terms of punctuation, resulting in the text generated by ChatGPT being more objective and detached. Therefore, participants' evaluations of the grammar component may be influenced by the varied style associated with the specific use of punctuation and grammar features.

### 6.1.2 Level of detail

The second criterion to be examined with the primary goal of evaluating the form of the reading comprehension assignments, is level of detail (item #4). Descriptive statistics indicate that both the human-written and AI-generated texts go into detail about the main topic; nevertheless, qualitative comments highlight that the text made by ChatGPT is considered slightly more sophisticated and detailed than the other activity. This information contradicts the data provided by the calculations of indices of central tendency, as the mean value of the ChatGPT-generated text is the lowest. This is most likely due to participants' considerations regarding the target readers' language ability. Although ChatGPT can provide very detailed output, this may result in a text that is too complex for middle school students, as evidenced by comments 5 and 6 in figure 23, which highlight that the second text is more difficult to read and corresponds to a higher level than the first. While this may be viewed as a positive aspect of the reading assignment in higher education contexts, it may not be suitable to offer such a complex reading passage to middle school pupils. This could be the reason behind the somewhat lower mean value associated to the AI-generated text. The issue of supplying pupils with the proper input in order for language acquisition to occur was addressed by several experts in the research field of education. Krashen is the pioneer of the input hypothesis, which was presented in his theory of the Natural Approach (1982):

“In order for acquirers to progress to the next stages in the acquisition of the target language, they need to understand input language that includes a structure that is part of the next stage. [...] An acquirer can “move” from a stage  $i$  (where  $i$  is the acquirer's level of competence) to a stage  $i + 1$  (where  $i + 1$  is the stage immediately following  $i$  along some natural order) by understanding language containing  $i + 1$ .”

According to this hypothesis, it is critical to supply students with linguistic input that contains structures they are unfamiliar with in order for them to acquire the language. Nonetheless, the oral or written input they receive must not be too complicated; otherwise, the language acquisition process may become too difficult. In recent years, additional research has been undertaken on this topic. For example, Namaziandost et al. (2019, p. 7) tested the input hypothesis by giving students slightly more difficult reading materials than their level of language knowledge, concluding that: “materials of EFL English textbooks could be one level

higher than the current level of the students to help to motivate and challenge them”. As a result, while it is critical for language teachers to provide students with reading comprehension tasks that allow them to learn new language structures, they must also be mindful of the risk of producing materials that are too challenging. According to participants’ observations, the reading exercise under investigation in the current study is regarded as a little too detailed and complex for the target readers. Therefore, it is worth noting that one of the shortcomings of AI-powered language models is that they may provide teaching materials that are not exactly at the next stage in terms of the acquirer’s level of competence.

### 6.1.3 Genre

Participants in the research study evaluated the reading comprehension exercises in light of the texts’ genre, which may be defined as: “a label which denotes a set of conventions in the way in which information is presented.” These conventions encompass both formatting and language style. (Dewdney et al., 2001, p. 1). Moreover, Swales (2011, p. 46) emphasise that: “the principal criterial feature that turns a collection of communicative events into a genre is some shared set of communicative purposes.” Although there are various definitions of genre, scholars appear to agree that this term is linked to a set of shared elements found in texts belonging to the same genre. The current study examines two reading activities from the same genre that deal with the same topic (i.e. things you should do every day to be happy and healthy). One of the research participants accurately identified the text type and stated it directly: “The text adheres to style norms commonly found in informative or self-help content” (see figure 23, comment 10). However, to facilitate analysis, the criterion designated “text genre” in this study also included teachers’ responses concerning style and formality. Considering teachers’ evaluations with respect to items #5, #6 and #7 of the questionnaire, the AI-generated text was not deemed on par with the human-written exercise. As a result, participants appeared to find the human-written reading activity more acceptable than the ChatGPT-developed exercise in terms of the stylistic features associated to the text genre. To further understand the causes for the lower mean value of the AI-generated text, qualitative comments in figures 22 and 23 are reviewed. Participants’ remarks emphasise that the second exercise evaluated in the questionnaire is more formal than the first, noting that higher levels of formality are connected with a more specific lexicon and more advanced linguistic expressions (see comments 8, 9, 11, and 15 in figure 23). As a result, the AI-generated reading activity may not meet the target

readers' needs. The quantitative statistics given in figures 8 and 9 support this information, since participants' replies indicate that the formality level of the ChatGPT-developed text is less compatible with middle school students' expectations. In contrast, the human-written text has a higher mean value, and again, qualitative comments reflect this trend, given that the text's informal register is judged suitable for lower secondary school students (see comments 10, 11, and 14 in figure 22). Informal writing is distinguished by the use of the active voice as well as the use of contractions, abbreviations, and phrasal verbs, whereas formal texts lack these features and are associated with the use of the passive voice and more specific vocabulary (Sheikha and Inkpen, 2012). These characteristics of informal reading tasks help to create interesting materials that engage students directly, making them want to read and make an effort to grasp the content of the text. In contrast, more formal and objective pieces of writing are typically related with students' boredom and demotivation. Consequently, teachers' evaluations of the AI-generated exercise most likely highlighted their concern about providing lower secondary school learners with a formal piece of writing, pointing out that pupils may struggle to understand the materials.

Nonetheless, teachers agree that style features are consistent with the text genre, and this also applies to the human-written assignment. In fact, both texts showed a high mean value for this feature (see item #6 in figures 8 and 9). This information is considered rather surprising, considering that the second reading comprehension activity is related to an impersonal and objective style, whereas the human-written work contains more personal and colloquial components. However, participants may have found both styles acceptable since both texts successfully transmit the information supplied in the text in a clear and accurate manner. Indeed, while a more informal piece of writing may be more appropriate for the intended audience, both text styles are adequate. Findings thus demonstrate that ChatGPT-generated reading assignments contain stylistic aspects that are appropriate for the text genre, despite the fact that the exercise's formality level may have a negative impact on the target readers' motivation and overall comprehension of the text.

#### 6.1.4 Repetition

Although repetition of vocabulary, phrases and content generally correlates with poor writing quality, some scholars argue that it can be employed for specific communicative purposes. For instance, Sigar & Saeed (2022, p. 3446) provide a definition of repetition emphasising its

fundamental role in the text: “Repetition is the process of repeating a word stated earlier in the text either in the same form or with some modifications. It is one of the key signaling devices available for setting up relations and, thereby, for the patterning of discourse or text.” Additionally, they claim that repetition is multifunctional, given that: “It can be used for clarification, emphasis, continuity, syntactic, and stylistic roles.” (ibid. p. 3453). Abdullayev et al. (2024) confirm that phonetic and lexical repetitions in literature promote cognitive meaning and establish a connection between poetic picture and meaning. By doing so, they create coherence based on the harmony of picture, content, and word, with the ultimate goal of perception. Finally, Halliday and Hasan (1976) present a taxonomy of cohesive devices, which includes the repetition of words or phrases as a rhetorical strategy for lexical coherence. Depending on the context, repetition can thus be employed for a variety of objectives and ought not to be associated just with poor writing quality.

Participants in the research study were asked to evaluate the content of the reading assignments using five criteria, the first of which was repetition of content, phrases, and words. Teachers’ levels of agreement with items #8, #9, and #10 of the questionnaire indicate whether the reading passages contain repetitive parts (low mean value) or do not offer them (high mean value). The results in figure 17 reveal that the human-written text has a substantially lower mean than the AI-generated reading passage; thus, participants agree that repetition is commonly used in this text. Teachers’ qualitative remarks help to better grasp their perspectives on the function of repetition. As reported in figure 24 (comments from 1 to 4), the human-written reading comprehension task presents repeated words and phrases to add emphasis and clarity to the text (e.g. “every day” is repeated to reinforce the idea of daily habits). Furthermore, participants agree that repetition helps to create a well-structured text, and that using the same verb forms is essential for conveying the text’s meaning clearly. Tannen (1973) highlights that the automatic nature of repetition makes comprehension easier by offering semantically less complex language. If some of the words are repeated, less new information is delivered than if all words included new data. This redundancy allows the reader to better comprehend the meaning of the text; hence, giving middle school students reading assignments with repeated concepts might help them understand the message and remember the various elements presented in the passage. Instead, participants’ remarks about the AI-generated text reveal that the reading exercise lack repetition of words and content. Only one of the respondents recognised one repeated phrase in the passage, pointing out that repetition is deemed

appropriate in informative writing because it allows for the reinforcement of previously presented topics (see comment 1 in Figure 25). Quantitative statistics also demonstrate that the ChatGPT-developed reading activity does not include a large number of repeated items. On the one hand, this aspect of the written passage reveals that the language model is capable of producing high-quality texts employing a diversity of words and expressions that allow to avoid repetition and present thoughts in a unique way; on the other hand, it shows that this assignment corresponds to a slightly higher level than the previous one, as it lacks repeated elements that might enhance students' comprehension of the text. Findings therefore suggest that lower secondary school pupils may find it easier to understand and complete the human-written reading exercise rather than the AI-generated activity.

#### 6.1.5 Factuality

According to Saurí & Pustejovsky (2012, p. 263) event factuality (or factivity) is defined as “the level of information expressing the factual nature of eventualities mentioned in text. That is, expressing whether they correspond to a fact in the world, a possibility, or a situation that does not hold”. Moreover, Thomson et al. (2023) state that factual accuracy in writings created using Natural Language Generation techniques can be assessed in terms of either real-world accuracy or data fidelity. The former pertains to determining whether the information in the text is true in the real world, whereas the latter considers whether the information in the text may be deduced from the system's input data or not. Nevertheless, from a pragmatic perspective, human annotators find it easier to identify real-world errors, as they don't need to check if the input data contains anything they already know to be true. As a result, because the current study focuses on human evaluation to determine the level of factuality in the reading activities, real-world accuracy is taken into account for analysis.

As seen in Figure 18, both human-written and AI-generated assignments had high mean values for factuality, with exactly the same low SD, indicating that participants' replies are all close to the mean. This trend is supported by the qualitative comments in figures 24 and 25, in which teachers report that the texts are current, interesting and conform to general truths. As a result, participants agree that both reading comprehension activities being investigated deal with information that is true in the real world. This aspect of the text created by ChatGPT helps to demonstrate that the language model creates reading assignments whose content is congruent with real-life circumstances that students may encounter in their daily lives. Actually,

informational writing is the most suitable text genre for introducing the topic suggested in the ChatGPT prompts (i.e., things you should do in everyday life to be happy and healthy). Şaşmaz & Çifci, (2023, p. 5) state that: “in informative text descriptions, the descriptions and information presented by the author are realistic” while Duke (2000, p. 205) explains that informational texts have: “a function to communicate information about the natural or social world”. As a result, in order to provide reliable information to readers, informative writing must include true facts. According to the current study’s findings, ChatGPT can provide such reliability. Not only did the model offer a reading comprehension exercise that was relevant to the topic stated in the prompts, but it also provided information that is true in the real world, thereby meeting the requirement of informative writing. Furthermore, young pupils may benefit from reading realistic and accurate texts that convey information relevant to their own personal experiences. According to Caswell and Duke (1998), as quoted in Duke (2000, p. 202): “Scholars have pointed out that informational texts can play an important role in motivating children to read in the first place. Some young children find a way into literacy through informational texts that they do not find through narrative and other forms of text”. For this reason, English teachers must have up-to-date and stimulating learning resources that engage pupils and pique their interest in studying English. ChatGPT may therefore assist them in developing reading comprehension tasks that contain factual information about real-life scenarios.

#### 6.1.6 Consistency

In order to compare the reading comprehension tasks, participants were requested to evaluate the learning materials’ consistency. Farkas (1985, p. 353) defines consistency as “the orderly treatment of a set of linked elements” and emphasises that “it is either ‘uniform’ or ‘harmonious’, depending on whether a set of linked elements is indivisible or divisible into subsets.” Furthermore, Kılıymaz et al. (2020, p. 209) explain that consistency is the analysis of sentences and paragraphs in the text and it is also associated with the mental competence of the reader. Consistency is thus concerned with the links between the different parts of the written passage (sentences, paragraphs, etc.), which must be ordered and linear in order for the reader to understand the text. As a result, providing students with well-structured reading assignments that include semantic connections to build consistency among the various elements of the text



is vital, because it allows to delve into the deep textual structure. In other words, consistency is a rhetorical device to develop semantic uniformity.

According to the current study's findings, participants noticed that the AI-generated text has a high level of consistency, which is equally true for the human-written exercise. In fact, both assignments have a high mean value, with the ChatGPT-generated exercise obtaining the highest level of agreement (see figure 19). This data shows that AI-powered language models generate output with semantic consistency, which means that the many components of the text work together to create a uniform piece of writing that conveys the passage's overall meaning. Despite the fact that both reading assignments are in line with the context and other elements of the text, teachers' qualitative comments reveal some variations between the two reading passages. In fact, the human-written text presents various daily habits in the form of a numbered list, with each point introduced by an appropriate title. Instead, the AI-generated section is broken into five paragraphs that are logically structured using some linkers, thus allowing to "easily grasp the main points and build upon each previous point" as indicated by remark 7 in figure 25. Nevertheless, participants reported that using numbered lists makes the text more accessible to middle school students, whereas the usage of linkers and sequential words embedded in the paragraph may not be appropriate for their needs. This distinction is presented also in Seki (2000):

"When readers see a bulleted or a numbered list on a page, they immediately know that a series of points is presented at that point before they begin reading the words. That is, such a list provides readers with a visible structure of the text and this allows them to assimilate the information smoothly. In contrast, when a list is embedded within a paragraph, readers cannot see its underlying structure until they read the words."

As a result, the structure of the AI-generated text could turn out to be too complex for the target readers, who may be unable to distinguish between the various sections of the text. Instead, the numbered list does not impede students' understanding of the text's meaning; rather, it allows them to divide the reading passage into parts and focus on one issue at a time. Finally, while the ChatGPT-developed reading assignment has a high level of consistency, the way the various textual components are linked together may hinder pupils' comprehension of the text.

### 6.1.7 Common sense

Common sense is another criterion evaluated in the questionnaire. Zang et al. (2013) define common sense knowledge as: “a tremendous amount and variety of knowledge of default assumptions about the world, which is shared by (possibly a group of) people and seems so fundamental and obvious that it usually does not explicitly appear in people’s communications.” Additionally, Zhang et al. (2017) describe common sense knowledge as possibilistic since it refers to things that everyone would expect to hold in a given context but without the necessary strength of logical entailment. Common sense thus relates to beliefs and expectations about real-world situations, and both reading comprehension tasks under consideration appear to adhere to common sense norms. The graphs in figure 20 show that participants have a high degree of agreement with item #7 of the questionnaire for both reading passages. Teachers’ qualitative observations regarding common sense reveal that the human-written text meets the target readers’ expectations, as does the AI-generated task. Indeed, both assignments provide some practical tips for living a healthy lifestyle. Students who read the title and attempt to guess the content of the reading passages will most likely consider some of the advice found in the actual texts, as the exercises’ content is consistent with common sense knowledge. This is a crucial feature of the reading task because it allows students to predict the information described in the text before they begin reading. According to Nation. (1993, p. 37):

“Prediction is usually drawn from the topic or heading of the text or from the first sentence of a paragraph. The justification for such an activity is that prediction stimulates and helps develop relevant previous knowledge which will make comprehension of the spoken or written text easier and more meaningful.”

Providing students with texts that adhere to the standards of common sense knowledge, allowing them to predict the content of the text, is a potential way to improve their comprehension skills. According to the study’s findings, language teachers can use ChatGPT to create reading exercises that engage students and motivate them to learn English by using prediction activities before reading the text. The fact that the AI-generated reading activity under investigation was evaluated in the same way as the human-written task in terms of common sense demonstrates ChatGPT’s potential.

### 6.1.8 Coherence

Finally, the content of the reading comprehension exercises was assessed for coherence. Participants' level of agreement with items #8 and #9 of the questionnaire about the structure of the text and the order in which information is presented are combined to aid analysis because these aspects of the reading passage are all related to coherence. According to Graesser et al. (2003), a text is perceived to be coherent by the reader when the ideas are connected in a meaningful and organised way. Coherence relations are constructed in the reader's mind using linguistic and discourse markers. Furthermore, Reinhart (1980) explains that a coherent text must meet three conditions, namely connectedness, consistency, and relevance. The first is a condition governing the linear concatenation of sentences in a text whereby sentences must be formally connected. The second is concerned with the connections between the text's various elements in order to achieve semantic uniformity, while the third refers to both semantic and pragmatic conditions. Consequently, coherence regards both the formal structure of the text and its content.

As shown in Figure 21, both reading assignments are perceived as coherent by teachers, with the human-written text having a slightly higher mean value. Nonetheless, the qualitative remarks show that the reading assignments under investigation also present some differences. The comments from #10 to #15 in Figure 24 indicate that the human-written exercise is well-organised and does not present significant issues. Even if information is presented by breaking the text down into sections, the passage is perceived as a cohesive piece of writing because the various daily habits appear in order of importance. Moreover, the titles associated with the different sections help to provide students with a text that expresses information in an understandable and clear manner. However, responses evaluating the AI-generated reading assignment show the opposite trend. On the one hand, participants agree that the text is clear due to the use of linkers and discourse markers, which improve cohesion. On the other hand, unlike the human-written task, this activity does not include any titles that could help students understand the text. Furthermore, participants noted that not only are the sentences too long and complicated, but also some of the vocabulary in the text may be inaccessible to middle school students due to their low language knowledge. As a result, the analysis of the qualitative comments provided by the teachers suggests that ChatGPT is capable of producing well-structured and cohesive texts, though the overall level of the AI-generated reading assignment may not be exactly in line with the target readers' abilities.

## 6.2 Previous research

The overall findings of this research project appear to be consistent with existing literature on the potential of AI-powered language models to generate insightful learning materials for teachers and students, as shown in sections 6.1 of the current thesis. However, given that the primary goal of this study is to create reading comprehension exercises to improve Italian middle school students' English skills, a comparison of the current study's findings to previous studies on human evaluations of ChatGPT-generated reading comprehension assignments is provided.

As reported in section 2.1 of this thesis, Xiao et al. (2023), Shin e Lee (2023), and Säuberli & Clematilde (2024) compared human-written and AI-generated exercises with the primary goal of evaluating the quality of the reading tasks and the possible application of language models to develop teaching materials that can be implemented in the classroom. Although these studies focus on different foreign languages and are conducted using various evaluation criteria, the final results generally highlight the positive features of ChatGPT-generated output, emphasising the high potential of language models to develop customised teaching materials to help students master the language and improve their reading comprehension skills. Because the variables in the research projects summarised in section 2.1 differ from those in the current study, only a broad comparison of their results is possible. In fact, one of the studies used the advanced ChatGPT-4 rather than ChatGPT-3.5, and some of the research focused on evaluating multiple-choice questions, whereas the current study's primary focus is on the quality of the reading passage. As a starting point, it is worth noting that ChatGPT-developed materials sometimes outperformed human-written exercises, indicating that evaluators were unable to distinguish between AI-generated and human-written assignments. For instance, Xiao et al. (2023, p. 617-617) indicate that: "the quality scores of ChatGPT passages are higher than the scores of human-written passages across all selected dimensions". Additionally, they explain that: "extensive evaluations were conducted among various groups of representative human evaluators, and the high quality of the generated reading exercises was widely acknowledged". These findings are consistent with the outcomes of the current study, confirming the idea that language models can serve as teaching assistants and valuable teaching resources for language instructors. Furthermore, the questionnaire used to collect teachers' opinions on the quality of reading assignments was developed using the criteria outlined in Clark et al.'s (2021) study, which sought to determine whether human evaluators could distinguish AI-generated output

from human-written materials. Although the study does not specifically focus on the evaluation of reading exercises, but rather on other types of content such as recipes and newspaper articles, the findings show that human evaluators were unable to correctly identify the text source, even after being trained with examples. These results demonstrate the high quality of the output generated by language models, and they are somewhat similar to the findings of the current investigation. Clark et al (2021, p. 7285) reported that: “formality of the text, spelling and grammar errors, and clarity were all cited to justify both human and machine judgments”, which is also true for the this research, as participants indicated that both the human-written assignment and the AI-generated reading exercise could be improved.

The present investigation confirms the findings of previous studies on the characteristics of AI-generated teaching materials. On the one hand, English teachers and instructors may benefit from the high quality of output generated by AI-powered language models such as ChatGPT, which allows them to easily obtain personalised reading comprehension exercises by prompting the model with very simple questions. On the other hand, due to the widely acknowledged quality of ChatGPT output, detecting AI-generated texts is becoming increasingly difficult, which may impede teachers’ assessment of students’ progress, given that pupils are more likely to cheat and use ChatGPT if they know the model produces human-like results. It is undeniable that AI-powered tools will have a significant impact on the education field, and teachers will need to adapt their teaching styles to accommodate the introduction of AI technology in the classroom.

### 6.3 Limitations

The current investigation has some limitations that are related to both the researcher’s choices when designing the study and the characteristics of the participants. To begin, the study was conducted using ChatGPT-3.5, an older version of OpenAI’s AI-powered Chatbot. ChatGPT-4 is the most recent version of the model; it is more creative and collaborative than ever before. It can generate, edit, and iterate with users on creative and technical writing tasks, like composing songs, writing screenplays, or learning a user’s writing style.<sup>2</sup> It was released on March 2023 but was made freely available to the users on July 2024. As a result, when the AI-generated reading comprehension exercise was developed, the model’s most recent free version

---

<sup>2</sup> Source: OpenAI official website. Available at: <https://openai.com/index/gpt-4/>. Last accessed: September 2024.

was ChatGPT-3.5. Using GPT-4 technology could have resulted in a more accurate version of the exercise. Furthermore, the current study focused on evaluating ChatGPT's reading comprehension exercise by comparing it to a reading task from an English middle school textbook commonly used in Italian schools. The AI-generated reading assignment's topic and text genre therefore matched those in the textbook and research focuses on analysing a single reading passage without taking into account other topics or text genres. Additionally, this study only looked into the quality of EFL teaching materials and did not include other languages commonly taught in Italy. As a result, the ability of the AI-powered language model to generate reading comprehension exercises in languages other than English was not investigated. Furthermore, the primary goal of this study was to compare reading comprehension assignments designed for middle school students. As a result, the research findings cannot be applied to other educational settings because ChatGPT's ability to develop teaching materials for primary or secondary school students was not tested. Finally, the prompts used to interact with ChatGPT were the same as those reported in the study by Xiao et al. (2023), and the criteria used for analysis were derived from Clark et al.'s investigation. As a result, it is worth noting that addressing the model with different prompts and analysing the AI-generated output based on other criteria may produce slightly different outcomes.

Although some attempts were made to prevent participants from being influenced by any bias while answering the questionnaire, their responses may still be biased. To prevent participants' preconceptions about AI-generated output from influencing their responses, they were not informed of the source of the reading materials they were analysing. Nonetheless, their evaluations of the two reading exercises may be influenced by their personal writing style preferences. Furthermore, because participants are currently employed in a variety of instructional settings (the study included not only middle school teachers but also primary and high school teachers), their work experience may influence their responses. In fact, high school teachers may find it slightly difficult to evaluate materials designed for middle school students because they are accustomed to working with more complex reading assignments appropriate for high school students. Finally, the data collection instrument (a questionnaire with both likert-scale items and open-ended questions) had a total of 34 questions, and participants were required to go through the reading exercises before responding. As a result, the questionnaire could take at least 20 minutes to complete, and some participants may have answered questions carelessly in order to avoid wasting too much time on it. It is also worth noting that only 20

teachers participated in the research project, and a larger number of respondents could have yielded more precise results.

#### 6.4 Future research

As described in Section 6.3, an updated version of ChatGPT is now freely available to users, and future research may consider evaluating AI-generated reading comprehension exercises developed by ChatGPT-4. For example, future studies on the quality of automatically generated reading exercises could focus on a comparison of teaching materials developed by ChatGPT-3.5 and those generated by the most recent version of the model. Furthermore, ChatGPT could be addressed with more specific prompts than those provided in this study, or such instructions could be implemented and improved by including additional details in the request. Furthermore, similar studies with a larger number of participants chosen using probability sampling would allow for a more thorough examination of the quality of AI-generated output. To learn more about teachers' view of the reading comprehension assignments, it would be useful to conduct interviews with participants and collect qualitative data, which allows for more specific information about their opinions. Another option for assessing AI-generated activities is to involve students in the evaluation process. Researchers could gather useful information about the quality of learning materials by assigning a reading exercise to students and then asking them questions about it. Students' responses to the reading exercise questions may also help determine whether the materials can be effectively used to improve learners' reading comprehension skills. Furthermore, the current research study does not take into account the development of teaching materials for use in settings other than Italian lower secondary schools. Thus, future research could focus on developing and analysing reading exercises for younger as well as older learners. Further research into this field could include using AI to generate reading tasks in other languages commonly taught in Italian schools (e.g., French, German, Spanish, Russian, and so on). Given that the current study focuses on the evaluation of a single reading task consisting of a text of a specific genre, it would be worthwhile to compare multiple AI-generated activities that include different text genres and address a variety of topics. Finally, the current study aimed to compare human-written and AI-generated materials, as well as provide insight into the quality of the reading comprehension exercise developed by ChatGPT in terms of both form and content of the written passage. Future research may focus on a similar

analysis that takes into account other aspects of the AI-generated task, such as the activities to complete after reading the text and examining their type and structure.

## 6.5 Implications

The current research project may have several implications not only for teachers, but also for academics who want to conduct additional research in the field of AI technology to investigate ChatGPT's growing influence on education. Language teachers may be interested in the current study's findings because they may benefit from incorporating AI-generated reading comprehension exercises into the classroom. On the one hand, this study supports the notion that NLP devices such as ChatGPT can produce well-written and fluent output, allowing teachers to quickly obtain personalised reading comprehension exercises. On the other hand, the current study shows that when asked to create teaching materials for specific target readers, ChatGPT may fail to generate language activities that exactly meet the students' needs and are appropriate for their language level. Teachers may therefore need to modify AI-generated reading passages, and the current study's findings suggest how to do so. Indeed, it identifies which aspects of the AI-generated passage teachers should pay close attention to in order to produce an exercise that is consistent with the students' prior knowledge and language levels. Given that it can be difficult for language educators to provide students with engaging reading activities to foster their language skills, ChatGPT can be a useful tool for obtaining personalised tasks that meet their educational needs. The current study goes into detail about the quality of the reading assignment generated by ChatGPT and provides an evaluation of its form and content. It sheds light on the issues raised by the comparison of human-written and AI-generated exercises, demonstrating that ChatGPT output can be evaluated on par with the other assignment, even if it may be slightly difficult for the target readers to fully understand it. Indeed, the results show that the final AI-generated exercise is well-developed, although some aspects could be improved. Teachers should therefore consider how to properly address ChatGPT in order to obtain reading exercises that meet certain requirements; for example, if they want to use the AI-powered model to generate reading comprehension assignments, the prompts could include more specific instructions that address the issues raised in this study. In this way, they avoid having to modify the final text. Finally, as stated in the previous section, scholars who wish to conduct research on the impact of ChatGPT on the education field may be interested in the results of this investigation because they could conduct similar studies to



confirm or refute the findings of the current study. This study may also be useful to academics interested in analysing the characteristics of AI-generated output, as it allows for a comparison of AI-generated and human-written materials. Exploring the features of AI-generated texts is critical for determining how closely the AI-generated output resembles human language and how it can be improved to produce better results.

## VII. Conclusion

Given that the rise of Artificial Intelligence is having a significant impact on a variety of fields and is revolutionising our daily lives, several research studies have recently looked into the benefits and drawbacks of using AI-powered devices. ChatGPT is a large language model for natural language processing whose impact on education has been extensively researched, with the primary goal of analysing and determining how to detect AI-generated output. The release of ChatGPT-3.5 and then ChatGPT-4 expanded the potential of AI tools by allowing the model to produce human-like output when given extremely simple prompts. The current study thus investigated the potential of ChatGPT-3.5 to generate reading comprehension exercises for Italian middle school students. The aim of this research was to answer the research question: are there any differences between AI-generated and human-generated reading activities for middle school EFL learners in terms of form and content? As a result, to assess the quality of the AI-generated activity, English teachers working in Italian schools were required to analyse both the AI-generated exercise and a human-written assignment on the same topic. The former was created by instructing ChatGPT with the same prompts as in a similar study by Xiao et al. (2023), whereas the latter was taken from a middle school textbook widely used in Italian schools. Both reading assignments include a written passage with tips for living a healthy lifestyle, as well as exercises with multiple-choice and open-ended questions. To compare the teaching materials, a questionnaire with various statements evaluated through likert-scale items was designed to collect participants' feedback on the quality of the reading tasks. The researcher used the same criteria for analysis as Clark et al. (2021) in which teachers' remarks were classified into two categories; observations on the form of the text were examined in terms of grammar, level of detail, and genre, whereas content observations were analysed in terms of repetition, factuality, consistency, common sense, and coherence. To better understand participants' perspectives on the AI-generated text, some qualitative data were also collected, including comments motivating their previous choices. After calculating indices of central tendency and dispersion of the quantitative data expressing teachers' level of agreement with the statements, the mean and standard deviation values of every item were reported on Microsoft Excel to better understand the trend generated by the replies. Results highlighted the high quality of the AI-generated exercise, which is associated with high mean values demonstrating teachers' positive attitudes towards the reading exercise. This is true also for the human-written text, hence, it is possible to claim that there are no significant differences

between the two reading comprehension activities. Nonetheless, an in-depth examination of each criterion provided insight into the characteristics of the reading assignment generated by ChatGPT. According to the final results, not only is the AI-generated text grammatically correct, but it was also found to be more accurate than the human-written text, despite the fact that the two passages have very different writing styles. The analysis of the level of detail and genre of the tasks confirms that the two exercises are exhaustive and well-written, though the text generated by ChatGPT is described as a little too formal for the intended readers. The same issue arose regarding the text's content. Thus, even though it received high marks, participants expressed concern about giving young students a reading comprehension exercise with specific and challenging vocabulary as well as a more complex text structure. In conclusion, the findings of the current study are consistent with previous research, demonstrating ChatGPT's ability to produce human-like output that is difficult to distinguish from human-written texts. Nonetheless, this investigation reveals one of the AI-powered language model's weaknesses, since its capacity to generate language teaching materials may not precisely meet the needs of the target students. In fact, it is critical to remember that each learner is unique and possesses distinct characteristics, as their language level and prior knowledge differ depending on their life experiences and personal background. Although ChatGPT may cause several changes in the education field, such as teachers being required to integrate AI technology in the classroom, this study also highlights the importance of teachers' roles in the development of teaching materials. Educators in the future may thus use ChatGPT to generate reading comprehension assignments; however, the language models at the moment cannot completely replace instructors in the task of creating exercises in English for the students, because teachers will always have to either instruct the model with very specific prompts or modify the final output in order to produce reading exercises that match the readers' expectations.

## Bibliography

Abdullayev, K., Mammadov, A., Mammadov, M., & Huseynov, S. (2024). Repetition in discourses across languages and genres. *Open Linguistics*, 10(1).

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: why and how of it? *Indian Journal of Medical Specialities*, 4(2).

Andrade, C. (2020). Understanding the Difference Between Standard Deviation and Standard Error of the Mean, and Knowing When to Use Which. *Indian Journal of Psychological Medicine*, 42(4), 409–410.

A. Shaji George, A.S. Hovan George, & A.S. Gabrio Martin. (2023). *A Review of ChatGPT AI's Impact on Several Business Sectors*.

Atlas, S. (2023). *ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI*. College of Business Faculty Publications.

Aydin Yildiz, T. (2023). The Impact of ChatGPT on Language Learners' Motivation. *Journal of Teacher Education and Lifelong Learning*, 5(2), 582–597.

Baskara, R., Mukarto, M. (2023). Exploring the Implications of ChatGPT for Language Learning in Higher Education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7(2), 343-358.

Belz, A., Mille, S., & Howcroft, D. M. (2020). *Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing*, 183–194.

Clark et al. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 1, 7282–7296.

Cotton, D., Cotton, P., & Shipway, J. R. (2023). *Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT.*

Council of Europe (CEFR) (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* New York: Cambridge University Press.

Dewdney, N., VanEss-Dykema, C., & MacMillan, R. (2001). The Form is the Substance: Classification of Genres in Text. *In Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management.*

Du, Y., & Gao, H. (2022). Determinants affecting teachers' adoption of AI-based applications in EFL context: An analysis of analytic hierarchy process. *Education and Information Technologies, 27(7), 9357–9384.*

Farkas, D. K. (1985). The Concept of Consistency in Writing and Editing. *Journal of Technical Writing and Communication, 15(4), 353–364.*

Fernandez et al. (2016). More Comprehensive and Inclusive Approaches to Demographic Data Collection. *School of Engineering Education Graduate Student Series, 60.*

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed).

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension.* Guilford Press. 82-98.

Golzar, J., Noor, S., & Tajik, O. (2022). Convenience Sampling. *International Journal of Education Language Studies, 1(2).*

Harmer, J. (2015). *The Practice of English Language Teaching eBook* (5th ed). P. Ed Australia.

Hughes, J. L., Camden, A. A., & Yangchen, T. (2016). Rethinking and Updating Demographic Questions: Guidance to Improve Descriptions of Research Samples. *Psi Chi Journal of Psychological Research*, 21(3), 138–151.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Jones, S., Murphy, F., Edwards, M., & James, J. (2008). Doing things differently: advantages and disadvantages of web questionnaires: *Nurse Researcher*, 15(4), 15–26.

Karaman, M. R., & Göksu, İ. (2024). Are Lesson Plans Created by ChatGPT More Effective? An Experimental Study. *International Journal of Technology in Education*, 7(1), 107–127.

Kim, S., Shim, J., & Shim, J. (2023). A Study on the Utilization of OpenAI ChatGPT as a Second Language Learning Tool. *Journal of Multimedia Information System*, 10(1), 79–88.

Khanh, L. C. (2020). *The Importance of Grammar in Language Teaching and Learning*.

Kıymaz, M. S. & Doyumğaç, İ. (2020). Consistency of the Narrative Texts Written by Middle School Students and Their Textual Creation Levels. *Educational Policy Analysis and Strategic Research*, 15(3), 207–232.

Koraishi, O. (2023). Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment. *Language Education & Technology (LET Journal)*, 3(1), 55-72.

Kostka, I., Toncelli, R. (2023). Exploring Applications of ChatGPT to English Language Teaching: Opportunities, Challenges, and Recommendations. *Teaching English as a Second or Foreign Language, TESL-EJ*, 27(3).

Krashen, S. D., Terrell, T. D. (1995). *The natural approach: language acquisition in the classroom*. Phoenix.

Lu, Y. (2019). Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*, 6(1), 1–29.

Mikeladze, T. (2023). Creating teaching materials with ChatGPT. *13th International Research Conference on Education, Language and Literature*. Proceedings books, 29-36.

Namaziandost, E., Esfahani, F. R., Ahmadi, S., & Yates, G. (2019). Varying levels of difficulty in L2 reading materials in the EFL classroom: Impact on comprehension and motivation. *Cogent Education*, 6(1).

Nation, P. (1993). Predicting the content of texts. *The TESOLANZ Journal*, 1, 37-46.

Patty, J. . (2024). THE USE OF AI IN LANGUAGE LEARNING: WHAT YOU NEED TO KNOW. *Pendidikan Dan Pengajaran (JRPP)*, 7(1), 642–654.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.

Priya Gupta, K., & Bhaskar, P. (2020). Inhibiting and Motivating Factors Influencing Teachers' Adoption of AI-Based Teaching and Learning Solutions: Prioritization Using Analytic Hierarchy Process. *Journal of Information Technology Education: Research*, 19, 693–723.

Reinhart, T. (1980). Conditions for Text Coherence. *Poetics Today*, 1(4).

Säuberli, A., & Clematide, S. (2024). Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models.

Saurí, R., & Pustejovsky, J. (2012). Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, 38(2), 261–299.

Şaşmaz, E., & Çifci, S. (2023). Expert Opinions on Improving Informative Text Writing Skills Through Descriptive Writing Practices. *International Journal of Education and Literacy Studies*, 11(1), 50–58.

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, 27(3), 27–40.

Sigar, A. H., & Saeed, B. M. (2022). role of lexical repetition in English written texts. *International Journal of Health Sciences*, 3446–3457.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14.

Swales, J. M. (2011). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Sheikha, F. A., & Inkpen, D. (2012). Learning to Classify Documents According to Formal and Informal Style. *Linguistic Issues in Language Technology*, 8.

Taherdoost, H. (2019). What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale/Attitude Scale/Likert Scale, *International Journal of Academic Research in Management*, 8(1), 1-10.

Tannen, D. (1987). Repetition in Conversation: Toward a Poetics of Talk. *Language*, 63(3), 574-605.

Thomson, C., Reiter, E., & Sundararajan, B. (2023). Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80.



Van Der Lee, C., Gatt, A., Van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67, 101-151.

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). *Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications*. 610–625.

Zang, L.-J., Cao, C., Cao, Y.-N., Wu, Y.-M., & Cao, C.-G. (2013). A Survey of Commonsense Knowledge Acquisition. *Journal of Computer Science and Technology*, 28(4), 689–719.

Zhang, S., Rudinger, R., Duh, K., & Van Durme, B. (2017). Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5, 379–395.

## Appendix

Figure 26 – demographic information about the participants

Partecipanti						
M	25-34	Italy	Italian English Spanish	This is my first year	Private primary school	
M	25-34	Italy	Italian English Spanish	1-2 years	Private high school	
F	25-34	Italy	Italian English Spanish French	This is my first year	Public primary school	
F	55-64	Italy	Italian English French	More than 20 years	Public middle school	
F	55-64	Italy	Italian English Spanish	6-10 years	Public primary school	
F	18-24	Italy	Italian English Spanish	1-2 years	Public middle school	
F	45-54	Italy	Italian English	16-20 years	Public primary school	
F	55-64	Italy	Italian English	16-20 years	Public primary school	
F	35-44	Italy	Italian English	6-10 years	Public primary school	
F	18-24	Italy	Italian English Spanish	1-2 years	Public middle school	
F	18-24	Italy	Italian English	This is my first year	Public primary school	
F	55-64	Italy	Italian English	More than 20 years	Public primary school	
F	35-44	Italy	Italian English Spanish German French	1-2 years	Public primary school	
F	25-34	Italy	Italian English	1-2 years	Public primary school	
F	25-34	Italy	Italian English French	This is my first year	Public primary school	
F	45-54	Germany	Italian	3-5 years	Public primary school	
F	35-44	Bermuda	Italian English	11-15 years	Primary school, middle school, high school	
F	18-24	Italy	Italian English	This is my first year	Public middle school	
M	25-34	Italy	Italian English Spanish	3-5 years	Public high school	
F	45-54	Italy	Italian, English, Spanish, French	11-15 years	Public high school	

Figure 27 – quantitative data about the human-written text

Testo 1														
Forma							Contenuto							
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
4	4	4	2	5	5	4	2	2	3	3	4	3	2	4
5	4	5	4	5	5	4	4	4	3	5	4	4	5	4
3	3	4	5	4	4	3	3	2	3	5	4	5	5	5
4	5	5	4	5	4	5	1	1	5	5	5	5	5	5
5	5	5	5	4	4	4	4	4	5	5	4	4	5	5
2	1	4	4	3	2	4	4	3	3	5	3	4	4	3
5	5	5	4	5	5	4	5	1	3	5	4	4	5	5
5	3	5	5	5	5	5	3	3	3	5	5	5	5	5
3	3	3	4	4	4	4	3	4	4	4	4	4	4	4
5	5	5	3	5	5	5	3	4	3	5	5	5	5	5
4	3	5	4	5	3	4	3	2	4	4	5	4	4	5
4	5	5	5	5	4	4	2	1	3	4	4	5	5	4
5	5	5	3	5	5	5	1	1	5	5	5	5	5	5
4	1	4	4	4	4	5	4	4	4	5	4	4	3	2
4	4	4	5	3	3	3	4	3	4	4	3	4	4	4
5	4	4	4	4	4	5	1	1	4	4	3	3	5	4
2	3	3	4	2	2	2	5	4	2	4	4	3	2	4
5	5	5	5	5	4	4	5	5	5	5	4	3	5	4
5	5	5	3	5	4	4	5	5	5	5	5	4	5	5
5	3	5	4	4	4	4	5	5	5	5	4	4	5	5

Figure 28 – quantitative data about the AI-generated text

Testo 2														
Forma							Contenuto							
Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
1	1	3	2	3	2	2	1	1	4	3	3	4	5	5
5	5	5	4	5	5	5	4	4	4	5	5	5	5	5
5	5	5	5	3	5	5	5	5	4	5	5	5	5	5
4	5	1	5	5	5	5	5	5	5	5	5	5	5	5
4	5	5	3	3	4	5	5	5	5	5	5	4	4	4
5	5	5	5	5	5	4	3	3	3	5	5	4	5	4
5	5	5	4	3	4	3	1	5	4	5	5	5	5	5
5	5	5	4	5	5	5	3	3	3	5	5	5	5	5
3	3	3	5	3	2	2	4	4	4	4	4	4	2	4
5	5	5	4	5	5	5	2	2	4	5	5	5	4	4
3	3	4	3	2	3	3	5	5	5	4	5	4	3	5
4	4	2	3	4	4	4	4	4	4	4	4	4	4	4
5	5	5	4	5	5	5	5	5	5	5	5	5	5	5
2	2	4	5	1	4	1	5	5	5	5	4	5	4	4
5	5	5	4	3	4	3	4	4	4	5	3	5	5	5
4	4	4	3	4	4	3	5	4	5	4	5	4	5	4
5	5	5	4	5	5	5	4	4	4	4	5	5	5	5
5	5	5	5	3	5	1	5	4	4	4	4	5	3	5
5	5	5	3	4	4	2	5	5	5	5	3	5	3	2
5	5	5	4	4	4	1	5	5	5	5	3	3	2	2