

# Using ChatGPT for Generating and Evaluating Online Tests

Valeriu Manuel IONESCU

Department of Electronics, Communications and Computers  
University of Pitesti  
Pitesti, Romania  
valeriu.ionescu@upit.ro

Madalin Ciprian ENESCU

PhD Student - Political Sciences National School of Political  
and Administrative Studies  
Bucharest, Romania  
enescumadalin89@yahoo.com

**Abstract**— Creating multiple choice online tests is a time-consuming task involving manual creation of both questions and possible answers in order to simplify and automate the process of test evaluation. This system is prone to cheating if the person being evaluated obtains the questions and answers. In some domains essay type questionnaires are needed, but they need time to be evaluated as the process cannot be automated. This paper investigates the use of ChatGPT-3 natural language processing to generate the quiz questions based on human prompt and to automatically grade the answers of an essay type test. This paper presents a NodeJS test implementation of such a system, the results obtained and the challenges of this process.

**Keywords**— ChatGPT, automated grading, quiz generator, quiz grading, artificial intelligence, NodeJs

## I. INTRODUCTION

Artificial intelligence is today present in all aspects of IT, from hardware (algorithms included in smartphones for image processing [1] and in data centers for computation acceleration [2]) to software [3, 4]. One use is the need to understand human communication in order to implement better AI assistants. People tend to use phrases or a combination of phrases and keywords to search the internet [5]. An AI assistant will understand the purpose of the search and will perform the search optimization for the user to produce relevant results.

The introduction of ChatGPT Natural Language Processing (NLP) model [6] introduced by OpenAI in 2022 [7] had a profound impact in multiple domains but especially education and creative arts. It is a large language model using generative pre-trained transformer (GPT) architecture, that is pre-trained on a large dataset (with data up to 2021), and uses human feedback to improve itself. Observing this impact, multiple companies have started to develop their own natural language processing models [8]. Microsoft Bing Chat (codename Sydney) can already be accessed in searches and has an API available to the public. It is based on OpenAI GPT-4 [9] with better speed and reasoning compared to the implementation in ChatGPT. Chatsonic [10] is another conversational AI that uses GPT-4 but it has the capability to search via Google Search the internet for recent data, offering information more recent than ChatGPT limitation of 2021. Google Bard is another conversational AI introduced in 2023 [11] that takes its information from the web and uses Google's Language Model for Dialogue Applications. Due to the fact that it's use was on invitation only until May 2023 (opened to the public after Google I/O in May 10 2023), and the fact that is currently based on a light model that needs less

computational power makes it a less expired, but interesting, choice.

In this paper we designed an application that uses ChatGPT API to generate quiz questions based on a user prompt, presents the quiz to the user using a NodeJS server, and automatically collects the results in a database (DB) and evaluates the results using the ChatGPT LLM. The novelty of the paper consists in involving ChatGPT not only in quiz question generation but also in the automatic quiz answer grading, in this way improving the speed of quiz creation. The paper also presents a literature overview on uses of ChatGPT for speeding up quiz creation.

The reminder of the paper presents related works in Section II, the application design in Section III, Application implementation in Section IV, Results and problems in Section V and in the end the conclusions are presented.

## II. RELATED WORKS

The use of ChatGPT in evaluations and quiz creation was researched extensively. There are multiple areas where natural language processing can be used, for example: informatics, marketing, education, politics but also in inter-disciplinary domains.

In informatics, in the paper [12] an essay type questionnaire was given to students and ChatGPT for determining if there are differences between the answer styles of a human and a GPT code. Similarly, in the paper [13] a machine learning model is trained to differentiate between machine generated responses and human responses. The paper [14], in an extensive analysis shows the following are important in obtaining better responses from ChatGPT: the question has to be formulated clear and with specific prompts; the context information dramatically improves the quality of the answers; the output format should be specified; limitations should be placed on the output; break down larger questions in smaller questions. Many of these guidelines were followed in the development of the application. The paper [15] observes the problems of ChatGPT, the danger of automatically trusting the results that are given and focuses on Unit Test Generation noting an improvement over previous Unit Test Generation tools

Political science is another domain where essay type answers are often needed because of the complexity of the aspects investigated. In the research [16], the authors note that in China the officials are attentive to the requests of the citizens on the addressed questionnaire platforms. This collection process can be enhanced by the use of ChatGPT. Charles F. Manski states in [17] that both objective and subjective information about users' perceptions results from a questionnaire. The research seeks to obtain and measure

subjective expectations, since the social sciences do not have the same vision as the behavioral sciences. Verbal and probabilistic forms, however, have a series of relative merits. Finally, in [18] the Generative Pre-trained Transformer (GPT) model is used in fields such as: biology and environmental science, where ample benefits were obtained both for these but also in research and education. The risks that may appear are analyzed and despite them, the question arises whether the ChatGPT model can offer an innovative approach in the research of inter countries power dynamics. In this sense, the analysis of relevant texts and documents, as well as the identification and understanding of the power dynamics between countries, for example US and China.

In this paper, we investigate NLP use for both generating quiz content and analyzing essay type answers to the generated questions.

On the aspect of quiz creation, there are already commercial applications that use ChatGPT API to generate test questions:

- the ChatGPT creator's OpenAI playground [19] that can create quizzes on request. The output should be checked for correctness and in order to avoid question repetition;
- Quizalize, [20] that offers question creation based on user input and can export Google Forms for offering online tests
- AI Quiz Builder [21] that allows test creation and integration with ScoreApp to give the questions a good and interesting format.

The conclusion for the quiz generation by using NLP is that it is great for generating content but still needs human supervision.

Regarding the aspect of quiz answer summarization and evaluation, there are far fewer papers written. In the paper [22] the authors use ChatGPT to perform text classification on three problems: big-five personality prediction, sentiment analysis, and suicide tendency detection, and the results are good but not as good a specialized model. The communication with ChatGPT is split in stages and the information is formulated in smaller questions. Again, it is observed that the formulation of the questions is critical in obtaining relevant answers and that the answers can have different formats, making their parsing (using regular expressions) hard to complete.

In [23] the summarization capability of ChatGPT is used automatic code summarization for Python code in comparison with some dedicated code summarization models. The results showed that code summarization is worse than by using dedicated models with problems like: the randomness of the answers poses sometimes a threat to result validity therefore again the request to ChatGPT had to be tweaked to generate best results; too many details are usually given making code summarization hard to read and, finally, choosing a more popular programming language gives better results compared to a less used language due to the reduced training dataset.

A more hands-on approach is made in [24] where paper summarization is made using Python and OpenAI API in order to generate the summary of a text document.

In [25] the Large Language Models (LLMs) ability to comprehend user intent and provide reasonable responses to text analysis is investigated. The results show a high level of faithfulness to the text analyzed but it can become overconfident in both correct and incorrect predictions and possibly misguide the users.

Finally, in [26] the authors investigate the way ChatGPT handles programming challenges is investigated and the

observation is made that the API and web code have different versions and the results given by the web version are better than the ones generated by the API, while a problem was found in the balance of focus and comprehension: the more details are given, the more focus is lost.

In [27] ChatGPT is used for grading papers as an assistant offering feedback to the papers written by the students. The actual grade is still given by the teacher but the role of supervisor increases the speed of grading and reduced the time needed to compose an answer that helps the student understand the errors committed. False positives (red flags incorrectly indicating fraud) were one aspect that a teacher needs to pay attention. In [28] the attention is drawn to the bias in automatic evaluation generated by less representative training data or by the algorithm implementation that often lacks of transparency.

The conclusion for the papers investigating automatic quiz evaluation is that the summarization part of natural language processing can be used but there are issues with the verbosity of the result (in both validity and clarity) and the prompt needs to be tweaked to obtain the best results (providing more information than necessary, makes the results worse).

### III. APPLICATION DESIGN

The application presented in this paper intends to automate the implementation the full process of quiz creation, presentation to the user, answer collection and evaluation of the answer. The user becomes an initiator and supervisor of the process instead of the creator and executor, in this way the quiz creation and evaluation speed are increased. In this case the ChatGPT-3 NLP model was used. The process has the following steps (that will have the implementation described in Section IV):

#### A. Ask the user for a subject that will be used to generate a prompt for the NLP

As presented in the section 2 of this paper, the questions should be chosen carefully, according to the desired test area of the questionnaire, placing restrictions on the output but not being too lengthy for the NLP to lose focus. ChatGPT usually gives conversational answers. In this example we desire less conversation and a more structured result. There are two main directions at this step:

- create true/false answers that can be placed in questionnaires along with the question and record the answers. In this case less accent is placed on the form of the response and more attention should be paid to the generated answers. Test types could be: matching, multiple choice, true-false.
- create questions that can be sent to the users in order to gather a free form answer and evaluate the answer. In this case more accent is placed on the form of the question; Test types could be completion, and essay. This is the direction chosen for this paper, because we want to let the NLP to interpret the essay format answer.

For example, if we want to create questions that target a specific content and we want answers to be given in a specific form, one sample question should be designed, for example, as: "What are the top 10 questions in the operating systems course, arranged in a list?", where the "arranged in a list?" is added automatically. ChatGPT replies with a list of possible questions that can be asked, for example: "What is an operating system and what is its role in a computer system? What are the main components of an operating system? How does process scheduling work in an operating system?" and

many more. As the questions are generated in various formats each time the question is asked, it makes the test less susceptible to being copied in future tests, while still targeting the same proposed area of interest. The questions generated by ChatGPT are stored in a database and constitute the bases of a questionnaire, with free form answers. They can be inspected by a human so that no nonsense questions are asked.

#### B. Presenting the generated questions to the user using a web interface

This step does not involve ChatGPT and takes the generated questions from the database and offers them to the user as a quiz with essay type answers. The user answers are collected in a different table in the database and will be later sent to ChatGPT for analysis.

#### C. The evaluation of the user answers by ChatGPT

Some of the possibilities at this step are:

- ask ChatGPT to summarize multiple answers given to a specific question in order to detect trends/patterns in the given results (such as answers placing accent on certain aspects or missing some important data that should be included in an answer). This type of data extraction is suitable for evaluating the results of a feedback quiz where essay form answers are expected. The automatic process can reduce the processing time for vast quantities of data.

- ask the NLP to evaluate the answer based on a specific scale. This is the direction chosen for the implementation described in this paper. For example, it can be as simple as: "Is the answer: {user essay type answer from the database} a correct answer for the question {question generated by ChatGPT}?" For example: <Is the following answer: "Virtual memory is a temporary storage area that is called upon whenever a program requires more RAM than is installed in a computer." a correct answer to the question "What is virtual memory?">. In this case ChatGPT gives, for example, as answer: "Yes, the given answer is correct. Virtual memory is a technology used by operating systems to expand the capacity of a computer's RAM memory when it is insufficient to run running programs. By using virtual memory, the operating system can allocate disk space to serve as an extension of RAM. Thus, when a program requires more memory than is available in RAM, the additional data is transferred between physical memory and virtual memory, allowing the program to run without encountering out-of-memory errors." The result can further be a simple binary estimation "correct/incorrect" or a graded estimation (in multiple formats such as x out of 10 or as a percentage 50%) and is accompanied by an explanation the grading or the percent of completeness of the user answer.

In this case the question can be rephrased as "In what percentage does the answer {user answer} responds to the question {question generated by ChatGPT}?" For example, for the previous question, the reply is: "The mentioned answer 100% answers the question "What is virtual memory?". The explanation provided defines virtual memory as a temporary storage area accessed when a program requires more RAM than is available on that computer. This definition outlines the primary role and function of virtual memory within an operating system.", which can be used to quantify the test results in a meaningful manner.

Finally, the process ends with the evaluation of the user answers stored in the database.

#### IV. APPLICATION IMPLEMENTATION

The application is a proof-of-concept application that was designed in NodeJS and is used to interact with ChatGPT API, generate questions, insert and extract data from a database and generate the web page for testing. It needs refinement to be implemented in a real usage system for automation of quiz creation and evaluation. The code and the database schema are available on GitHub [29]. The data flow for quiz creation and result analysis is presented in Fig. 1.

The free OpenAI key was used for testing (not paid) and is subject to the limitations established by OpenAI. For the database implementation, MySQL Community edition was used. For the command line query of ChatGPT functions the *node-chatgpt-api* package was used [30]. It allows rapid development of applications that use different models like *gpt-3.5-turbo* or *text-davinci-003*. The library also allows the use of Bing Chat AI using GPT-4 and can be implemented for this application with minimal changes (by changing *clientToUse*: 'bing' in *Settings.js* file). The AI freedom can be controlled with changes in the same file to the *toneStyle* option (creative, precise, fast, balanced) and will be tested in the future.

The browser client option of this package was not used as it exposes the user access token to a third-party server. The *mysql2* package will have to be installed in order to access the MySQL database. The tests were made in Windows 10 but with minimal modifications the code can be used in other operating systems too. The *prompt-sync* package was used for reading the user input.

As seen in Fig. 1, there are three modules designed in NodeJS and implemented:

- the QUERY module that takes the query from the user and inserts the ChatGPT collected data into a MySQL database;

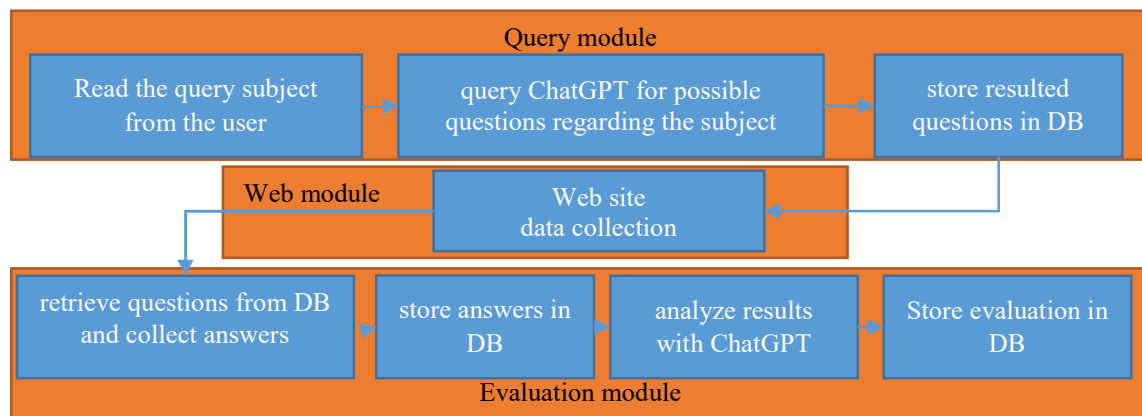


Fig. 1. The flow of question generation, answer collection and answer evaluation

-the WEB module that creates a web server to present to the user the questions from the list generated by ChatGPT, capture the answer and send it to the database in a different table;

-the EVALUATION module that makes the automatic evaluation of the result using again ChatGPT and places the evaluation into the database.

The MySQL database is composed of three tables *Questions (Intrebare)*, *Answers (Raspunsuri)* and *Questionnaire (Chestionar)*.

*Questions (Intrebare)* (Fig. 2) is the initial table that will be filled with the questions that were sent to ChatGPT and generated the quiz question list. The non nullable column is the primary key. The id is used in the web page module to select the quiz that will be asked.

Column	Type	Default Value	Nullable
id	int		NO
intrebare	varchar(200)		YES

Fig. 2. The *Questions (Intrebare)* table (MySQL Workbench 8.0)

*Answers (Raspunsuri)* (Fig. 3) table stores the questions generated by ChatGPT and approved by the user. The non nullable column *idraspunsuri* is the primary key. The *idintrebare* column from *Answers (Raspunsuri)* is a foreign key to *id* from the *Questions (Intrebare)* table. This column is the one that is used to randomly extract the questions belonging to the specified quiz.

Column	Type	Default Value	Nullable
idraspunsuri	int		NO
idintrebare	int		YES
raspunsurigt	varchar(200)		YES

Fig. 3. The *Answers (Raspunsuri)* table (MySQL Workbench 8.0)

*Questionnaire (Chestionar)* (Fig. 4) table stores the user answers in the web page and, in the evaluation module, it is used to store the evaluation given by ChatGPT in the grade column. The non nullable *idchestionar* column is the primary key. The grade column was chosen to be a *varchar(45)* in order to allow in the future to store literal evaluations like “correct” or “incorrect”, not only numbers in the 0-10 range or percentages. The *idraspunsuriq* column from *Questionnaire (Chestionar)* is a foreign key to *idraspunsuri* from the *Answers (Raspunsuri)* table.

Column	Type	Default Value	Nullable
grade	varchar(45)		YES
idchestionar	int		NO
idintrebareq	int		YES
idraspunsuriq	int		YES
raspunschestionar	varchar(2000)		YES
useremail	varchar(45)		YES

Fig. 4. The *Questionnaire (Chestionar)* table (MySQL Workbench 8.0)

### A. The QUERY module

This module starts by launching a command line interface to run *node-chatgpt-api*. A certain amount of time is allowed to pass (in this case 3 seconds) before the shell is initialized and the connection with ChatGPT is made. The interaction with ChatGPT is handled invisibly to the user.

The user is then asked to insert the question that will be sent to ChatGPT to generate the quiz questions. When addressed to ChatGPT, an “as a list” is added in the end of

the message in order to format the output as we need to process it. From *node-chatgpt-api* all the text formatting was removed (*tryBoxen* function) to simplify the parsing process. The output should be a returned as numbered list and the code identifies the lines that start with a number. The question list is presented to the user and it should be confirmed before being inserted into the database along with the question that generated the list. The database inserts are made using SQL statements. The confirmation before data insert is necessary because of the unpredictability of the answer format. Sometimes the ChatGPT generates data in the requested format, in other situations it places them as an enumeration or does not respond and asks more data. For example, we asked for cat, dog, boy and girl names. While the first 3 categories did not generate any problems regardless of the number of times asked, when asking to generate girl names at the third attempt it asked the purpose of the question and what is our intention to do with the data. If the answers are not ok or not present, the user will not confirm the list and the process is resumed from the beginning.

### B. The WEB module

This module creates a web server by using the *createServer* method and is able to handle the POST with the user data. On web server start the administrator is asked to specify the question id of the quiz that will be used in the quiz. The HTML page is generated when the user accesses it and filled with a random question. When the user fills the form with the essay answer, on POST clicking, the body is parsed and inserted into the database. A new page will be generated on page refresh. Preventing the POST of empty data is made using the “required” HTML5 attribute and the email field uses the email type (Fig. 5).

What is virtual memory and how does it work in an operating system?

answer:

Please fill out this field.

Fig. 5. The web interface with the essay field to be completed and the protection for submitting a blank answer or a bad email

Currently the user is identified by the email used and only a single question is pulled from the database. In the future a better presentation of this page should be created and a table of users should be made in order to differentiate them. However, the purpose of this paper is to query, collect the answers and evaluate them by using a minimalist implementation.

### C. The EVALUATION module

The module starts by creating a command line interface to call the ChatGPT API and will read one by one all the user answers that do not have values in the Grade field from the *Questionnaire (Chestionar)* table. The ChatGPT result reading was implemented to handle multiple problems:

-the length of the answer should be longer than the question length. From testing, short answers we given when ChatGPT does not understand the question so these are to be avoided;

-ensure that we are receiving the evaluation as a numeric value in either the form of “5 out of 10” there we used a regex to match the numbers in the string output by ChatGPT. The

regular expression pattern “`^d+/g`” matches one or more digits (`^d+`) globally (`g`) in the given text. If there are at least two numbers found, the first will be the required evaluation grade.

- ensure we can handle an evaluation in percentage format. Similarly a regex with “`^d+(?:\\.\\d+)?%`” was used to find the value. The percentage is priority to the numbered format.

- eliminate all other evaluation types (link “mostly/partially correct”), when the regex does not find numbers.

If a correct result is obtained, the column in the *Questionnaire (Chestionar)* table will be updated, if not, a new attempt will be made. Usually if ChatGPT receives the same question multiple times it will vary the answer and at one point will give in and give an evaluation.

## V. RESULTS AND DISCUSSION

The application modules were tested successfully for Romanian and English language quizzes, as seen in Fig. 6. Questions were asked from multiple domains such as computer science, politics or geography. Question generation takes seconds and essay answer evaluation is similarly fast. The person managing the quiz becomes a supervisor of the process in validating questions and supervising answer grading. The number of retries for a correct test generation is low (it usually generates correct quiz questions in the format requested or after minimal prompt changes), but the number of retries for the automatic evaluation is high, especially if similar answers need to be evaluated.

We observed that when asked multiple times to re-evaluate the same answer, after a few evaluations in the same range (for example it graded an answer as 5/10 and another time 6/10), the algorithm can sway its assessment format (no longer gives grades as requested but evaluates the results as correct or partially correct) or excuses itself for not understanding the question (or even the language!).

This behavior is helpful if a human detects an evaluation error from ChatGPT and confronts the algorithm to give an improved/different answer but less so if the question was meant as a simple verification of the exposed reasons.

Even if generative pre-trained transformers, like ChatGPT, do not explicitly address the problem of local minima, it looks like this is the modality used by ChatGPT to increase answer diversity and prevent itself from continuing to give similar answers (repetitive or low-quality responses).

The tests showed that the results should be supervised by humans as there are many errors that can be given. For example, for questions from Computer Science: “Give 5 examples of incorrect IPv4 addresses”, the addresses 0.0.0.0 and 255.0.0.1 were included, which are in fact valid IPv4 addresses, (the incorrect answer was given because the question was too vague, as those are reserved addresses) and from Geography: Bulgaria and Albania were identified as having a common border – in this case it is a clear error. In

some cases, the model was detected to lie, for example presenting two identical answers as different.

In conclusion the results we obtained are in concordance with other authors observations (as presented in the related works section), where they noted that: erroneous evaluations are offered and are upheld by the algorithm with justifications; regex parsing of the results is hard to make as the answer format can vary wildly (even if asked to a specific type of answer). It was observed that, while ChatGPT tries to imitate human chat format, in this case – given the specific task of grading answers from 1 (low) to 10 (high) – a human would follow the task to the end and not try to vary the evaluation format with every answer. For example, a teacher would not change its assessment style from numeric (for example in the 0-10 range) to literal: Correct / Incorrect / Mostly correct, after a few papers just for the sake of introducing diversity.

## VI. CONCLUSIONS

This paper analyzed the current research in the domain of quiz generation and evaluation using the ChatGPT-3 NLM model. The literature research showed that the Quiz generation aspect is well covered in the literature with many papers coming since ChatGPT was launched. The main problems identified were the sensitivity of the model to the prompt format as minute changes made the model to become verbose, lose focus and avoid completing the requested task.

The quiz evaluation was covered by far fewer papers with the main accent placed on the summarization capabilities. The papers that actually tested the use of NLPs for evaluation were few and discussed the unpredictability of the model due to its habit of varying the answer format even for similar questions and the bias in the answers due to the black box nature of the algorithm. In both aspects, quiz generation and especially quiz evaluation, human supervision is advised. The implemented application and our testing confirmed this behavior.

ChatGPT was found to have many limits: the number of tokens for the GPT-3 and 3.5 models present in the free versions can limit the size of the text summarized; the answers can be inaccurate but still justified with confidence by the model, the year 2021 limitation of the model information and, if not checked, the fact that the answers are usually verbose. There are alternatives present for ChatGPT and others are in the works.

The application described in this paper was designed as proof-of-concept in order to automate the question generation, result collection and evaluation. The modules were successfully tested but it was found that the process sometimes needed multiple retries in order to complete its task because sometimes ChatGPT avoids giving the answers at all or using a requested format. The modules implemented in the application will need be improved in the future for example for ease of use (integrated in a single application and constructing a graphical user interface), or security (to prevent SQL injection from the web interface).

```
Type your question:top 10 questions in political science regarding Europe
current:
Sure, here's a list of top 10 questions in political science regarding Europe:

1. What factors contributed to the rise of populism in European politics?
10. How has transatlantic relations and the changing global order influenced Europe's foreign policy priorities and strategic outlook?

current->1. What factors contributed to the rise of populism in European politics?
->10. How has transatlantic relations and the changing global order influenced Europe's foreign policy priorities and strategic outlook?

Are results ok?(Y/N):Y
? message:
ChatGPT Response inserted:1. What factors contributed to the rise of populism in European politics?
```

Fig. 6. The application’s text interface has the following areas 1: answer generated by ChatGPT, 2: answer components identified by the application (marked by “->”) and 3: user is asked to confirm the list that will be inserted into the database, 4: answers inserted into the database

As a future work other chat bots will be tested (Bing is already implemented in *node-chatgpt-api* package [30]) and other chat models. For the aspect of evaluation correctness when faced with essay type answers, the problem observed was related to incorrect knowledge concerning objective information (such as IP addresses, height of mountains, country locations, etc.) in all tested areas (operating systems, computer networks, political science, history and geography). As the model trains with more data these errors should, in time, improve the correctness of the answers, but, in this moment, they cannot be trusted automatically. When asked to evaluate short answers and to understand them ChatGPT showed a good result and made verbose suggestions along with the assessment.

As a future development, ChatGPT result parsing will have to be improved for the implemented application, as the AI returns various numeric estimation formats when requested to evaluate an answer to a question, such as: 9/10, 9 out of 10, 90%, or even text evaluations (such as "in great part", "does not respond" or even refuses to evaluate in cases where the answer is too distant from true). Better result handling will result in fewer retries. Finally, a graphical interface will be developed for the application in the future (to increase the ease of use) and the quiz web interface will be improved.

## REFERENCES

- [1] H. Li, X. Li, Y. Huang, Y. Wang and G. Guo, "Special Session - Test for AI Chips: from DFT to On-line Testing," 2021 IEEE 39th VLSI Test Symposium (VTS), San Diego, CA, USA, 2021, pp. 1-1, doi: 10.1109/VTS50974.2021.9441057.
- [2] N. P. Jouppi, C. Young, N. Patil, D. Patterson, et al., "In-Datcenter Performance Analysis of a Tensor Processing Unit," in Proceedings of ISCA '17, Toronto, ON, Canada, June 24-28, 2017, pp. 1-12. doi: 10.1145/3079856.3080246.
- [3] M. Harman, "The role of Artificial Intelligence in Software Engineering," 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE), Zurich, Switzerland, 2012, pp. 1-6, doi: 10.1109/RAISE.2012.6227961.
- [4] Z. Nan, "The application of artificial intelligence technology in the communication engineering industry," 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 252-254, doi: 10.1109/NetCIT54147.2021.00057.
- [5] F. Radlinski and T. Joachims, "Query Chains: Learning to Rank from Implicit Feedback," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 239-248. doi: 10.1145/1081870.1081899.
- [6] M. Abdullah, A. Madain and Y. Jararweh, "ChatGPT: Fundamentals, Applications and Social Impacts," 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), Milan, Italy, 2022, pp. 1-8, doi: 10.1109/SNAMS58071.2022.10062688.
- [7] T. Wu et al., "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development", IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122-1136, May 2023, doi: 10.1109/JAS.2023.123618.
- [8] E. Gent, "GPT-4 Ups the Ante in the AI Arms Race", IEEE Spectrum, <https://spectrum.ieee.org/gpt-4>, Accessed March 18, 2023.
- [9] D. Kovačević, "Use of ChatGPT in ESP Teaching Process," 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2023, pp. 1-5, doi: 10.1109/INFOTEH57020.2023.10094133.
- [10] N. Sitapure and J. S.-I. Kwon, "Exploring the potential of time-series transformers for process modeling and control in chemical systems: An inevitable paradigm shift?", Chemical Engineering Research and Design, vol. 194, pp. 461-477, 2023. doi: 10.1016/j.cherd.2023.04.028.
- [11] S. Pichai, An important next step on our AI journey, <https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/>, Accessed Feb 06, 2023.
- [12] H. H. Alamleh, A. A. S. AlQahtani, and A. Elsaid, "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning" in 2023 Systems and Information Engineering Design Symposium (SIEDS) (IEEE SIEDS'23), Charlottesville, USA, Apr. 2023. doi: 10.1109/SIEDS58326.2023.10137767.
- [13] S. Mitrovic, D. Andreoletti, and O. Ayoub, "ChatGPT or human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short Chatgpt-Generated Text," arXiv preprint arXiv:2301.13852v1 [cs.CL], Jan. 2023. [Online]. Available: <https://arxiv.org/pdf/2301.13852.pdf>. Accessed on: Jan. 30, 2023.
- [14] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope", Journal of Internet of Things and Cyber-Physical Systems, volume 3, pages 121-154, 2023, ISSN 2667-3452, doi: 10.1016/j.iotcps.2023.04.003
- [15] M.M. Rahman, Y.Watanobe, "ChatGPT for Education and Research: Opportunities, Threats, and Strategies". Appl. Sci. 2023, 13, 5783, doi: 10.3390/app13095783,
- [16] T. Chen, Z. Liang, H. Yi, S. Chen, "Responsive E-government in China: A way of gaining public support", Government Information Quarterly, 2023,doi: 10.1016/j.giq.2023.101809.
- [17] C. F. Manski, "Questionnaires: Collection of Subjective Expectations Data," in International Encyclopedia of the Social & Behavioral Sciences (Second Edition), 2015, pp. 776-780.
- [18] E. Agathokleous, C.J. Saitanis, C.Fang, Z. Yu, „Use of ChatGPT: What does it mean for biology and environmental science?", Science of The Total Environment, Vol. 888, doi: 10.1016/j.scitotenv.2023.164154.
- [19] OpenAI, "OpenAI playground", OpenAI Platform, <https://platform.openai.com/playground>, Accessed May 23, 2023.
- [20] Zzish LTD, "Quizalize", <https://www.quizalize.com/>, Accessed May 23, 2023.
- [21] S. Oddy, "AI Quiz Builder that makes a quiz in less than 3 minutes", <https://www.scoreapp.com/ai-quiz-builder-waitlist/>, Accessed May 23, 2023.
- [22] M. M. Amin, E. Cambria and B. W. Schuller, "Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT," in IEEE Intelligent Systems, vol. 38, no. 2, pp. 15-23, March-April 2023, doi: 10.1109/MIS.2023.3254179
- [23] W. Sun et al., "Automatic Code Summarization via ChatGPT: How Far Are We?", <https://arxiv.org/pdf/2305.12865.pdf>, Accessed May 23, 2023
- [24] L. Soares, "Summarizing Papers With Python and GPT-3", <https://medium.com/geekculture/a-paper-summarizer-with-python-and-gpt-3-2c718bc3bc88>, Accessed May 23, 2023.
- [25] Li, Bo, et al., "Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness", <https://arxiv.org/abs/2304.11633>, Accessed May 23, 2023.
- [26] H. Tian et al., "Is ChatGPT the Ultimate Programming Assistant -- How far is it?", <https://arxiv.org/pdf/2304.11938.pdf>, Accessed May 23, 2023.
- [27] B. Ellis, "How ChatGPT can help with grading", <https://blog.tcea.org/chatgpt-grading/>, Accessed May 23, 2023.
- [28] S. Coghlan, T. Miller and J. Paterson, "Good Proctor or Big Brother? Ethics of Online Exam Supervision Technologies". Philos. Technol. 34, 1581–1606 (2021), doi: 10.1007/s13347-021-00476-1
- [29] V.M. Ionescu, "chatgptautomation" GitHub, 2023. <https://github.com/manuelcore/chatgptautomation>, Accessed May 23, 2023.
- [30] W. Wanderer, "Node-ChatGPT-API," GitHub, 2023. Available: <https://github.com/waylaidwanderer/node-chatgpt-api>, Accessed May 23, 2023.