Analysis of ChatGPT Performance in Computer Engineering Exams

Roberto Rodriguez-Echeverría, Juan D. Gutiérrez, José M. Conejero, and Álvaro E. Prieto

***Abstract*—The appearance of ChatGPT at the end of 2022 was a milestone in the field of Generative Artificial Intelligence. However, it also caused a shock in the academic world. For the first time, a simple interface allowed anyone to access a large language model and use it to generate text. These capabilities have a relevant impact on teaching-learning methodologies and assessment methods. This work aims to obtain an objective measure of ChatGPT's possible performance in solving exams related to computer engineering. For this purpose, it has been tested with actual exams of 15 subjects of the Software Engineering branch of a Spanish university. All the questions of these exams have been extracted and adapted to a text format to obtain an answer. Furthermore, the exams have been rewritten to be corrected by the teaching staff. In light of the results, ChatGPT can achieve relevant performance in these exams; it can pass many questions and problems of different natures in multiple subjects. A detailed study of the results by typology of questions and problems is provided as a fundamental contribution, allowing recommendations to be considered in the design of assessment methods. In addition, an analysis of the impact of the non-deterministic aspect of ChatGPT on the answers to test questions is presented, and the need to use a strategy to reduce this effect for performance analysis is concluded.**

***Index Terms*—Artificial Intelligence, ChatGPT, education, experiment.**

I. Introduction

THE influence of Artificial Intelligence (AI) on Computer Engineering education was already evident by the end of the 20th century. In the 1997 edition of the *Jornadas sobre la Enseñanza Universitaria de la Informática* (JENUI) [1], 25% of the papers included it directly in their title. These papers shared with the educational community the different ways this discipline entered educational programs. Just a quarter of a century later, the situation has changed so much that the issue at hand is what consequences the use of AI will have in all areas of higher education. For example, [2] proposes using AI models to assist in assessing complex programming assignments.

In 1943, McCulloch and Pitts introduced the perceptron [3], launching a field of knowledge with immense potential. In a narrative ellipsis worthy of Kubrick, this work allowed OpenAI to introduce ChatGPT [4] at the end of 2022, an interface for accessing its large-size language model (LLM), GPT 3.5. Such an achievement would not have been possible without the introduction of the transformer [5], a deep learning model presented by Google in 2017 based on the concept of attention, which has proven to be fundamental in the field of LLMs.

An LLM is an AI model trained using large text corpora. These models use deep learning techniques to generate text resembling human writing. Some examples of LLMs are OpenAI GPT-3 [6], Meta OPT [7], or BLOOM [8]. The latter is unique in being an open-access alternative in all its aspects, while the others are proprietary developments. Moreover, these models can satisfactorily perform tasks such as machine translation, text generation, text classification, and question answering.

The evaluation of ChatGPT capabilities has been in the spotlight from the start. Two months after its release, ChatGPT had already been tested in medical [9] and law admission exams [10]. In the first case, its performance was comparable to that of a third-year medical student. In the second case, it surpassed 50% of the questions.

In [11], a systematic review of the use of chatbots in education is carried out, analyzing the areas in which they have been used, their pedagogical role, their use in tutoring tasks, and their potential in personalized education. Combining a chatbot with a highly reliable LLM seems promising for use in education.

ChatGPT's ability as a writing assistant tool is tested in [12]. The author challenges himself to generate an academic paper with its help. The result suggests that it is a helpful tool for increasing user productivity. Moreover, it will be necessary to find new ways of student assessment focusing on aspects AI cannot replace, such as creativity and critical thinking. However, enhancing human capabilities using such technologies has the same implications as using drugs or stimulants to artificially enhance the physical performance of athletes, something that [13] focuses on.

Roberto Rodriguez-Echeverria is with Applied Information Technology Research Institute, Universidad de Extremadura, Av. Universidad s/n 10003, Cáceres, Spain (e-mail: rre@unex.es).

Juan D. Gutiérrez is with Department of Electronics and Computing, Universidad de Santiago de Compostela, Lugo, Spain (e-mail: juandiego.gutierrez@usc.es).

José M. Conejero is with Applied Information Technology Research Institute, Universidad de Extremadura, Av. Universidad s/n 10003, Cáceres, Spain (e-mail: chemacm@unex.es).

Álvaro E. Prieto is with Applied Information Technology Research Institute, Universidad de Extremadura, Av. Universidad s/n 10003, Cáceres, Spain (e-mail: aeprieto@unex.es).

DOI (Digital Object Identifier) Pending.

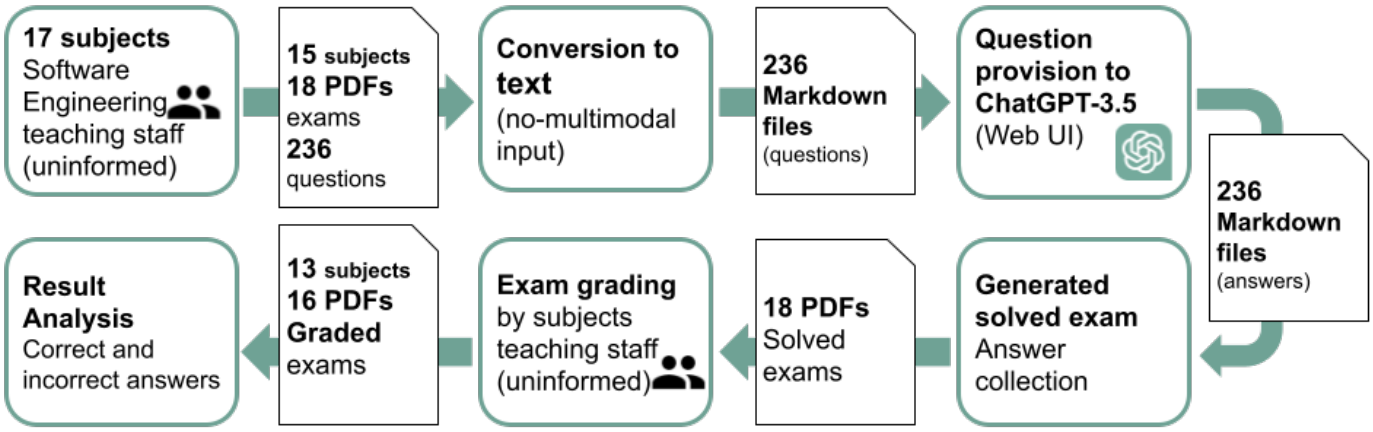

Fig. 1. Process overview.

After analyzing the advantages and implications of using chatbots in research and their limitations, the author shows the ethical considerations and possible biases that come with using technologies like ChatGPT in research, concluding that such technologies can revolutionize academic research.

Meanwhile, in [14], the author determines that ChatGPT can exhibit creative traits in its results, endangering the integrity of online exams and, therefore, their evaluation. The possibility of illicitly using ChatGPT in exams requires rethinking assessment methods to remain fair for all students.

The impact of large language models on the reality of education in general is a subject of study worldwide. The difficulties of detecting and preventing academic dishonesty are studied in [15]. This work suggests strategies universities can adopt to ensure an ethical and responsible use of these tools. Among them are developing policies and procedures, training and support, and various methods to detect and prevent cheating. With another focus, the enormous applicability of transformative AI tools like ChatGPT, emphasizing their possible positive and negative impact in various sectors, is studied in [16]. Despite recognizing its limitations and potential ethical issues, this work considers the productivity improvements obtained by using these technologies in different areas. Lastly, those interested in obtaining a pragmatic perspective, far from the biases inherent in extreme positions, on the challenge facing education will find in [17] a reflection that addresses the advantages, disadvantages, potentialities, limits, and challenges of generative artificial intelligence technologies in education.

This work is an extension of the one presented in [18], where the impact of ChatGPT on assessment methods of a Computer Engineering degree is analyzed. For this purpose, an experiment was designed to evaluate if this model could pass the exams of 15 subjects within Software Engineering. Although most academic subjects include assessment methods beyond exams, such as the development of projects, this initial work only analyzed its impact on exams. As an extension, a study of the frequency of responses provided by ChatGPT to multiple-choice questions is presented to assess the impact of the non-deterministic aspect of this tool in this type of study.

The rest of the work is organized as follows: Section II describes the process and guidelines followed in developing the experiment. Then, the results are analyzed in detail in Section III, and a series of recommendations are presented in Section IV. Finally, the main conclusions obtained and future lines of work are presented in Section V.

II. Methodology

In this section, the research questions, the main steps in the development of the experiment, and the guidelines followed to provide the exam questions to ChatGPT, given its characteristics, are presented. In addition, several dimensions for categorizing the exam questions are defined to conduct a more detailed analysis. Finally, the method followed to reduce the impact of the non-deterministic aspect of ChatGPT on the answers to multiple-choice questions is specified.

A. Research Questions

In this work, we aim to answer the following questions:

1) Can ChatGPT pass these exams?
2) What is the ratio of correct to incorrect answers?
3) Does the type of exam question matter?
4) Does the type of knowledge application matter?
5) How well has it performed in each subject?

B. Main Steps

The development process of the experiment presented in this work has followed the following six steps, as shown in Fig. 1.

1) Within our institution, we have contacted the teaching staff of all the subjects in Software Engineering (17) to collect actual exams and other assessment methods from the 2021-2022 academic year. They were just informed of the intention to conduct a joint assessment of the evaluation methods of the subjects in the field, but no mention of ChatGPT was made. 15 of the 17 considered subjects responded to this request, but only 13 provided corrections.
2) All the exams were organized by subject and briefly described to indicate the type of questions they contained (multiple-choice, short answer, problems), whether they contained figures, or whether they required some contex-

tual information.

3) A textual version of each exam was manually generated so ChatGPT could directly process them. All the modifications performed in each case were minimal so that the final result was as similar as possible to the original version of the exam. These modifications could be of different types. Sometimes, it was enough to divide a question into different parts so that ChatGPT could respond to all of them sequentially in a conversation. In other cases, where a figure accompanies the question, this has been replaced by a textual description.
4) The modified versions of the exams were provided to ChatGPT and formatted as conversations following the guidelines described later. For this work, we used the version of ChatGPT released on December 15, 2022[1].
5) Based on the answers obtained from ChatGPT for each exam, a completed exam (questions and answers) was generated and sent to the teaching staff of each subject for grading.
6) Finally, a detailed analysis of ChatGPT's results for each exam was performed. Considering the grade obtained, we have analyzed the performance question by question and the comments made by the teaching staff.

For more detailed information on the technical and organizational aspects of the methodology followed the interested reader may review the materials available in our repository[2].

C. Adaptation Guidelines

Following, the guidelines applied to provide the exam questions to ChatGPT are presented:

1) The questions are provided in Spanish to maintain maximum consistency with the original exam and to obtain responses in the same language.
2) If the exam only contains unrelated multiple-choice questions, each question is asked in a separate conversation. This way, no artificial context related to the order of the questions is artificially created.
3) If the exam contains short answer questions with multiple sections, each section is provided separately while maintaining the same conversation. That way, we prevent failures in ChatGPT's response generation due to excessive length or any other limit.
4) In questions including code and referring to a specific line, the lines of code are numbered to make such reference to the line by its number.
5) For questions including a figure representing a data structure, a textual description of its main elements and relationships is provided. For example, in the case of a graph, its sets of nodes and edges can be provided. In some cases, it is also possible to use textual syntax to define diagrams, such as Mermaid[3] syntax.
6) Questions with a non-explicit context, for example, those referring to a problem or project carried out in the subject, are asked without providing additional information to ChatGPT.
7) Data tables are provided in CSV[4] format, although other formats like Markdown[5] would be possible.
8) In fill-in-the-blank questions, the tilde (~) indicates where the answer should go.

D. Category of Questions

In order to perform a detailed analysis of the results obtained, we have categorized the questions based on three dimensions: type of question, type of knowledge, and type of application.

Within the first dimension, the possible types of exam questions are multiple-choice questions, short-answer questions, or problems. The multiple-choice question has a single correct answer from four options. The short answer question consists of developing some theoretical content of the subject. Finally, problems propose exercises or the application of theoretical content to practical cases.

Regarding the type of knowledge, we have only considered two categories: literal definition and applied definition. These categories correspond to the first and last basic cognitive level of Bloom's taxonomy: knowledge and application. Some examples of the first type are defining a concept, indicating a specific term, or listing properties. The second type refers to the application (usage) of knowledge to a specific case or example specified in the exam.

It is important to note that the first and second dimensions are independent. In contrast, the third dimension obtains a finer grain categorization for questions classified as applied definition.

Finally, regarding the type of application, we have considered the following nine types:

1) *Analysis of Programming Languages* (APL). The question contains a code snippet that must be read and understood to answer the question.
2) *Generation of Programming Languages* (GPL). The question requests an answer in source code. For example, write a specific SQL query.
3) *Operational Semantics* (OS). The question contains a code snippet whose execution must be understood to answer the question.
4) *Mathematical Calculation* (MC). The question requires some explicit or implicit mathematical calculation.
5) *Algorithm or Method* (AM). The question requires following an algorithm or a method with multiple steps.
6) *Analysis of Algebraic Expressions* (AAE). The question contains algebraic expressions that must be read and understood to answer the question.
7) *Generation of Algebraic Expressions* (GAE). The question requests an answer in the form of an algebraic expression.
8) *Diagram Analysis* (DA). The question contains a diagram that represents, for example, a data structure or model that must be read and understood to answer the question.
9) *Diagram Generation* (DG). The question requests an an-

[1] https://help.openai.com/en/articles/6825453-chatgpt-release-notes
[2] https://github.com/i3uex/jenui23_chatgpt
[3] https://github.com/mermaid-js/mermaid
[4] https://datatracker.ietf.org/doc/html/rfc4180
[5] https://en.wikipedia.org/wiki/Markdown

swer in the form of a diagram.

The exam questions can belong to several types of applications simultaneously. For example, a question that presents an incomplete code snippet may request the generation of the missing code (analysis and generation of programming languages).

E. Non-determinism reduction in multiple-choice questions

ChatGPT is designed to introduce a degree of non-determinism when answering to obtain variability in its responses. Given the same question, ChatGPT can provide different responses each time it is asked. This feature, therefore, must be considered when analyzing the results obtained by ChatGPT from experiments like the one proposed in this work. For this reason, the effect of this indeterminism in the responses generated for the multiple-choice questions is analyzed.

In the case of multiple-choice questions, this non-deterministic effect can be easily analyzed since the space of possible answers is limited to the four options provided in the exam. Taking advantage of this characteristic of the multiple-choice questions, we have added a step in the methodology that consists of asking each question one hundred times. This way, we can obtain the distribution of the different responses provided by ChatGPT for the same question. The analysis of this distribution allows us, in many cases, to obtain a predominant response that we can consider as definitive for that question.

For the implementation of this process, we used the ChatGPT API[6]. Each multiple-choice question was asked one hundred times using a Python script, using a new conversation in each iteration. In addition, ChatGPT was instructed to respond only with the letter of the answer without including additional information. Finally, for each question, the number of times each of the four possible answers was provided has been counted for later analysis.

III. Results

A. Responses to Research Questions

The following is an analysis of the results based on the research questions.

Can ChatGPT pass these exams? Without specific training in the assessed subjects' content, ChatGPT has passed 8 out of 15 exams. The last column of Table I shows the grades obtained in each exam. The minimum grade to pass each exam is 5, while the maximum is 10. Exams from those subjects with a dash (-) as the grade have not been corrected by their corresponding instructors. A single grade is shown if we have analyzed a final exam or two grades if we have analyzed two partial exams. This result alone suggests that current LLMs already have a significant and tangible impact on the evaluation methods of many subjects evaluated.

What is the ratio of correct to incorrect answers? To obtain finer-grained data on ChatGPT's exam performance, we can analyze how many correct answers it got. As shown in Fig. 2, ChatGPT could correctly answer 56% of the questions. Specifically, out of 230 questions, it answered 129 correctly. If the assessment were a single exam and all questions had the same value, ChatGPT would have passed the assessment of all the knowledge considered in the experiment. Let this hypothesis reflect ChatGPT's impact on exams as an evaluation method.

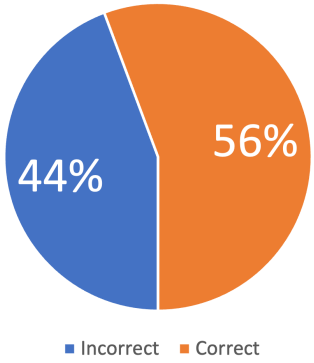

Fig. 2. Incorrect and correct answers.

Does the type of exam question matter? Fig. 3 shows the quantities of each question type and the number of correct and incorrect answers.

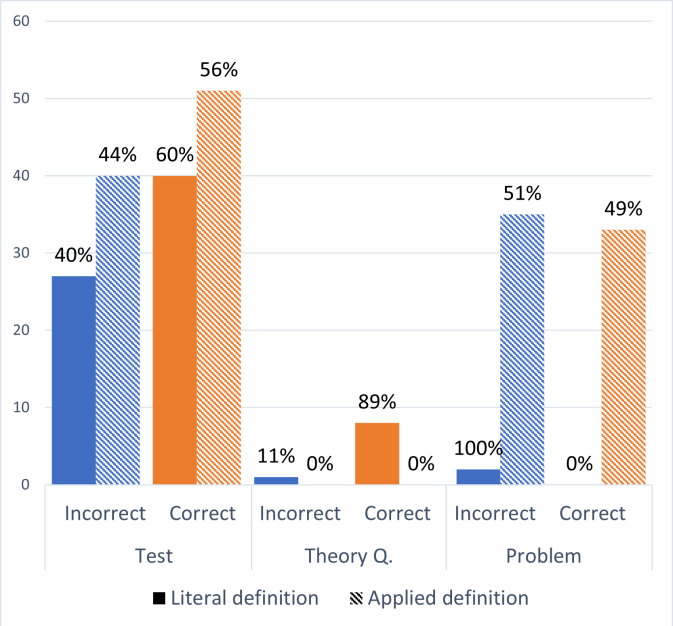

Fig. 3. Incorrect and correct answers by question type.

As can be observed, the vast majority of questions are multiple-choice (155), followed by problems (70), with a much smaller number of short-answer questions (9). However, looking at the percentage of correct answers for each type of question, we can see that short-answer questions get the highest value, 89%, followed by multiple-choice questions, 58%, and finally, problems, which present a success rate below 50%. Finally, let us consider the dimension of the type of knowledge. The first thing we can see is that all short answer questions have been classified as literal definitions. At the same time, the problems mainly belong to the category of applied definition. In the case of multiple-choice questions, there

[6] https://platform.openai.com/docs/api-reference

are questions in both categories of this dimension. ChatGPT has a close success rate, 60% versus 56% in both categories.

TABLE I
SUBJECTS

Year	Subject	Acronym	Question type	Grade
1	Estructuras de datos y de la información	EDI	Multiple choice (MC)	4.58
1	Introducción a la programación	IP	MC & Problems	3.90 + 7.88
2	Análisis y diseño de algoritmos	ADA	MC & Problems	5.88 + 5
2	Bases de datos	BD	MC & Problems	2.32
2	Desarrollo de programas	DP	Short answer	7
2	Inteligencia artificial y sistemas inteligentes	IASI	Problems	1.25
2	Programación concurrente y distribuida	PCD	MC & Problems	5.05
3	Diseño y administración de bases de datos	DADB	MC & Problems	3.39 + 2.19
3	Diseño e interacción en sistemas de información	DISI	MC	2.88
3	Diseño y modelado de sistemas software	DMSS	Short answer & Problems	6.45
3	Ingeniería de requisitos	IR	Short answer	-
3	Programación en Internet	PI	Short answer	7.5
3	Teoría de lenguajes	TL	MC & Problems	2
4	Arquitecturas software en entornos empresariales	ASEE	Short answer	5.25
4	Gestión de proyectos software	GPS	Short answer	-

As illustrated in Fig. 4, considering only the dimension of type of knowledge, literal definition questions present a higher percentage of correct answers than applied definition questions, 62% versus 53%.

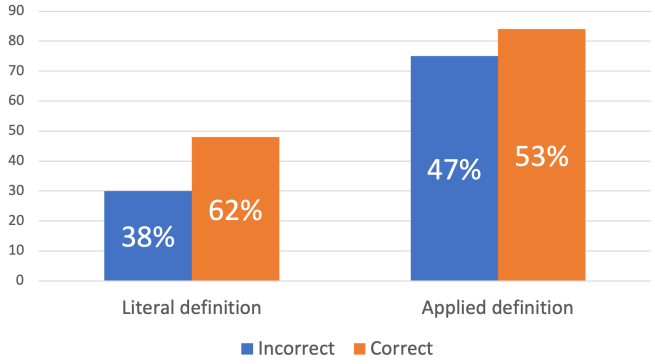

Fig. 4. Incorrect and correct answers by type of knowledge.

As preliminary conclusions, these results seem consistent with the capabilities of LLMs at the time of this study, which have a much more excellent formal knowledge of language than functional knowledge [19]. Therefore, it is normal for them to be able to repeat a literal definition but have more problems when applying the definition of a concept to a specific example. However, we were surprised that the success rate was lower in multiple-choice questions within the literal definition category. This result can have multiple explanations, such as the statement being slightly ambiguous or overlapping answers. However, we have yet to detect many questions about this problem. After a deeper analysis, we believe the errors in the multiple-choice questions may be associated with the *mispriming* effect [20], which consists of using some distractor type in a question. In this case, the possible answers may be distracting and lead ChatGPT to return a wrong answer.

Although we do not have undeniable evidence of this phenomenon, we have tried several multiple-choice questions to provide only the statement without including the answers, and, in that case, ChatGPT has responded appropriately.

Does the type of application of knowledge influence the outcome? As shown in Fig. 5, the most common types of application are code analysis, the application of an algorithm or method, and code generation, in that order, while the types of operational semantics, mathematical calculation, and diagram analysis appear in less than 20 questions.

In terms of the success rate (correct to incorrect rate), relatively high success values are obtained in the analysis (62%) and generation (70%) of programming languages, while the application of an algorithm or method results in more incorrect than correct answers (46%). Notably, there is a high rate of incorrect answers appearing in the categories of mathematical calculation and diagram analysis.

As expected, ChatGPT performs well on application questions that require formal linguistic competencies of a language, as is the case with most questions of analysis and generation of programming languages. On the other hand, it shows much poorer performance on questions requiring functional language competencies, such as mathematical calculation or applying a multi-step algorithm or method. Finally, regarding diagram analysis, no valid conclusion can be drawn, as we cannot ensure that the transcription made of these diagrams was the most appropriate.

How well has ChatGPT performed by subject? Apart from the final grade obtained, it is worth conducting a more detailed analysis of its performance by subject from the viewpoint of the type of exam questions used, according to the proposed classification. For reasons of brevity, this paper only details the results of one subject. Specifically, the subject IP has been chosen, and two partial exams with disparate results have been analyzed. It is the only subject where ChatGPT did not show uniform performance. It does not pass the first partial, while it obtains a high grade in the second. Each partial comprises a test of 8 multiple-choice questions and a problem.

Table II shows the results in more detail for the multiple-choice questions. As can be seen, ChatGPT answers all ques-

tions correctly in the second partial. At the same time, there is an equal number of correct and incorrect answers in the first.

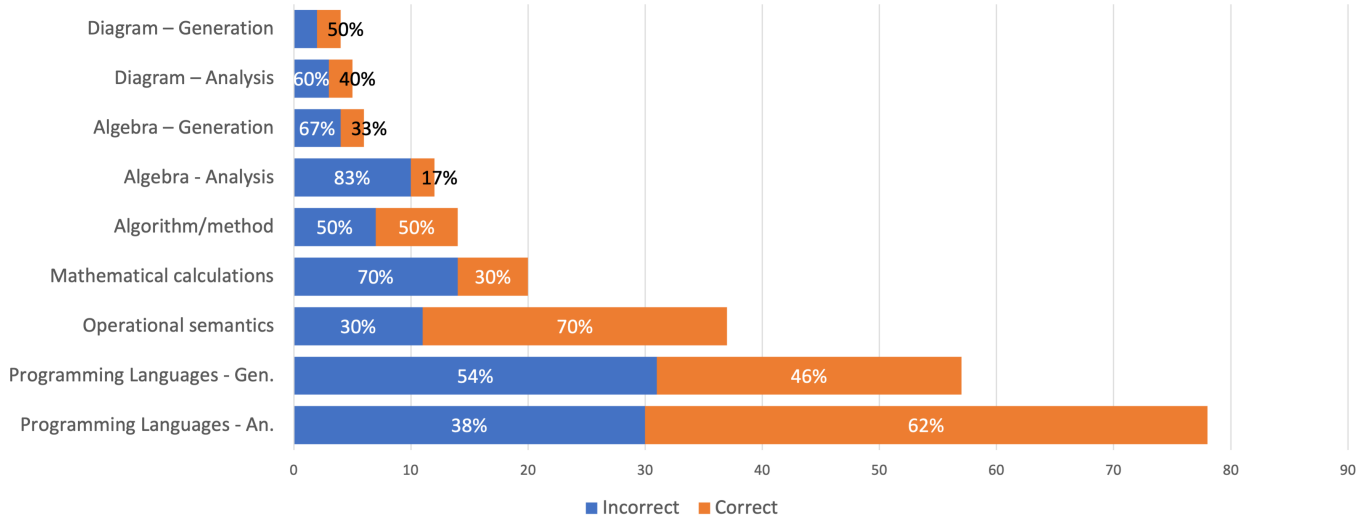

Fig. 5. Incorrect and correct answers by type of application of knowledge.

Looking at the type of knowledge, both partials contain only applied definition questions. As for the different application types, all except one require analyzing a code fragment. Therefore, the main differences between both partials are in the application types: operational semantics and algorithm/method. The second partial contains three algorithm application questions, all answered correctly; therefore, given the results and the small number of questions of this type, it also does not explain the difference in performance. However, in the case of operational semantics, we have a more significant number of questions with distinct results. Specifically, the first partial contains six questions of this type, and ChatGPT answers two correctly and four incorrectly. Although the number of questions to analyze is small, ChatGPT presents more difficulties when asked about the result of executing a code snippet.

TABLE II
IP RESULTS

P.	Type	1	2	3	4	5	6	7	8
1	APL	✓	✓	✓	✓	✓	✓	✓	✓
1	OS	✓	✓	✓		✓		✓	✓
1	C/I	I	I	C	C	C	C	I	I
2	APL	✓	✓	✓		✓	✓	✓	✓
2	OS						✓		
2	AM		✓		✓			✓	
2	C/I	C	C	C	C	C	C	C	C

B. Analysis of non-determinism in multiple-choice questions

In this analysis, 147 multiple-choice questions were considered out of the 155 available because all those that were interdependent and entailed a long conversation with ChatGPT were discarded. Fig. 6 shows the variation in correct and incorrect answers between the previous and new studies, taking the predominant answer as a reference.

As can be seen, in total terms, the reduction of ChatGPT's non-deterministic effect has led to an increase in the number of incorrect answers. Specifically, it has gone from 61 to 71 incorrect answers, from 41.5% to 48.3%, an increase of almost 7%. The last row of that figure shows the number of questions that have changed from correct to incorrect or vice versa and those that have changed from one incorrect answer to another (from incorrect to incorrect). As can be observed, 89 responses out of 147 have not changed, which means that 60.5% of the original answers coincide with the predominant one obtained. However, the remaining 39.5% (58 responses) have changed: 17 from incorrect to correct, 30 from correct to incorrect, and 11 from incorrect to incorrect.

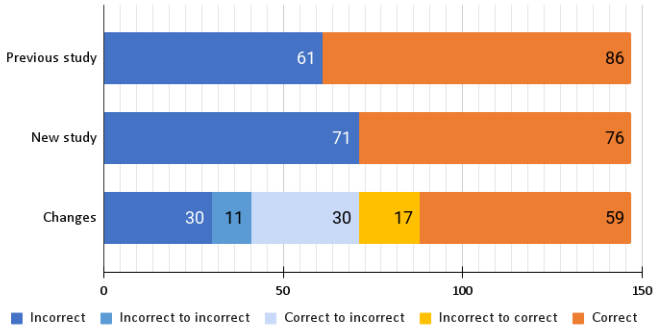

Fig. 6. Variation in correct and incorrect answers.

Fig. 7 shows a histogram with the distribution of all correct and incorrect answers according to the percentage value of the predominant answer, grouped into bins of 10%. As can be seen, correct answers occur more frequently in the higher partitions (80-100% value of the predominant answer). In the case of incorrect answers, they are quite uniformly distributed across all partitions, with a higher number in the intermediate partitions (50-80%). However, as can be seen, there are also extreme cases where the answer has 100% predominance but is incorrect. On the other hand, answers with low predominance, below 50%, are correct. On the other hand, Fig. 8 shows the frequency of the predominant response categorized by the type of change for the 58 questions that present a new answer.

This article has been accepted for publication in IEEE Revista Iberoamericana de Technologias del Aprendizaje. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/RITA.2024.3381842

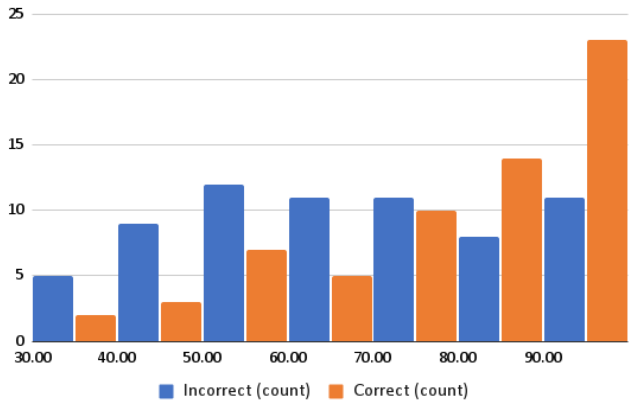

Fig. 7. Histogram of predominant response divided into correct and incorrect.

As can be appreciated, the majority of the changes from incorrect to correct have a higher frequency, above 55%, compared to the changes from incorrect to another incorrect (below 50%). This situation may indicate, in the former case (incorrect to correct), that the answer from the original study (incorrect) does not match the most frequent answer of the second study. Meanwhile, in the latter (incorrect to another incorrect), it may indicate several answers with a similar frequency, which results in a higher probability of error. Finally, the changes from correct to incorrect are the most numerous and appear much more distributed in the figure. Therefore, they contain both cases where the correct answer has a frequency close to the predominant one and cases where the correct answer has a very low or even nil frequency.

Although it would be interesting to perform a detailed analysis of the questions whose answers have changed, this part of the study is beyond the scope of this work due to the need for more space for its development. Nevertheless, one of the most exciting results is identifying seemingly simple questions that ChatGPT frequently needs an apparent or easy-to-identify reason. A multiple-choice question from the subject Introduction to Programming (IP) has been selected to exemplify this case. Specifically, the first question of the first partial of IP, presented in Fig. 9, is a very illustrative example of how the use of distractors can affect ChatGPT.

As any reader familiar with programming can deduce, the correct answer is c, according to the expected output of the program's execution. While the first parameter (`a`) is passed by value (i.e., a copy of the actual parameter) to the `intercambiar` (swap) module, the second one (`b`) is passed by reference. Therefore, after the invocation of this module, only the value of the parameter `b` is modified, obtaining the desired output.

In the tests carried out for the original study of this work, ChatGPT gave an incorrect answer to this question: option d. At first, this result was attributed to the nature of the question (applied definition of operational semantics type), which requires the tool to be able to simulate the program's execution. However, after repeated attempts, ChatGPT provided the correct answer on one occasion: option c. Subsequent tests, however, again kept answering d as correct, showing ChatGPT's non-deterministic behavior.

The left column of Fig. 10 shows the distribution of answers provided by ChatGPT to that question. The most frequent answer is d (incorrect), with 63% appearances, followed by answer c (correct), with 24%. Therefore, ChatGPT also selects the correct option a significant number of times. The question arises as to why the incorrect option is selected more than the correct one. Could it be due to the way the question is formulated? After all, LLMs choose the next word in their response probabilistically.

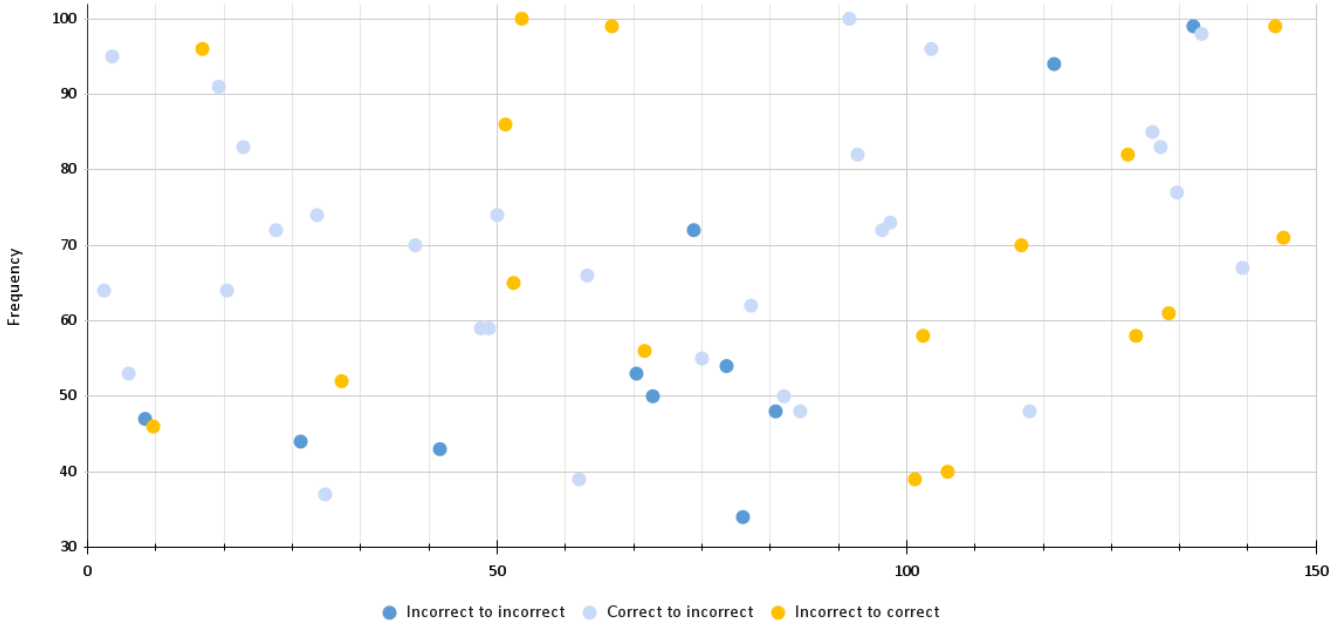

Fig. 8. Predominant answer frequency grouped by type of change.

This article has been accepted for publication in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/RITA.2024.3381842

Pregunta 1

Tenemos el siguiente módulo para intercambiar los valores de dos variables, pero no sabemos cómo son los parámetros:

```
void intercambiar ???????????{
        int aux;
        aux = x;
        x   = y;
        y   = aux;
}
```

Se ejecuta el siguiente algoritmo:

```
int main(){
        int a, b;
        a = 1;
        b = 2;
        intercambiar (a,b);
        cout << a << " " << b;
}
```

En la pantalla se escribe: **1 1**

¿Cómo era la cabecera de la función *intercambiar*?

a) `void intercambiar (int x, int y)`
b) `void intercambiar (int &x, int y)`
c) `void intercambiar (int x, int &y)`
d) `void intercambiar (int &x, int &y)`

Fig. 9. Question from the first partial exam of IP.

After thoroughly analyzing it, we realized that, although the module was called `intercambiar` (swap), the program did not perform an exchange as such of the value of both parameters. Answer d would be the one to select if looking for a header that exchanged the values passed as parameters. What would happen if the question was reformulated, replacing `intercambiar` with a meaningless name like `foo`? Fig. 10 shows the distribution of responses provided by ChatGPT to the same question by making only that substitution. As can be seen, answer c becomes the most frequent with this change, exceeding the percentage originally obtained by d, with 69%. None of the other two answers reach 20%.

As demonstrated in [20], it is possible to easily confuse a pre-trained language model (PLM) like ChatGPT through mispriming, which consists of adding terms to the statement of a question to steer the model towards a wrong answer. In conclusion, although ChatGPT does not respond correctly every time, the result obtained in this example clearly shows the role of the name of the module name as a distractor in this case.

IV. Recommendations

The variety in the type of questions on an exam allows for the evaluation of different levels of student knowledge in a particular subject. In this work, a series of recommendations are made about the design of questions so that their inclusion or not in a given exam can be evaluated. Of course, instructors must consider these recommendations in each context and assessment moment. For example, the risk of students' illicit use of ChatGPT in an exam may differ in face-to-face or virtual teaching modalities or in self-assessment tests compared to official exam calls.

As a first recommendation, it is proposed to avoid literal definition questions, mainly short answer questions, as much as possible. ChatGPT can answer complex questions of this type, even reasoning the answer and giving application examples. Our experiment had more problems with multiple-choice questions, perhaps because the answers themselves could confuse it, possibly due to the mispriming effect.

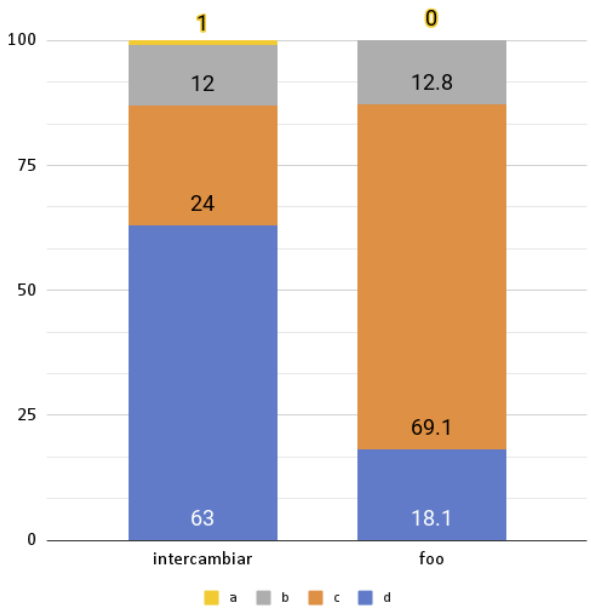

Fig. 5. Comparison of answer distribution with and without distractor.

As a second recommendation, it is also suggested that questions that only involve code analysis or code generation be reduced or avoided within application-of-knowledge questions. ChatGPT can perform quite exhaustive static code analysis and can generate code snippets to solve well-defined problems, such as those typically used in multiple-choice questions in an exam. After all, this type of competence for an LLM seems similar to what it needs to answer literal definition short answer questions. Therefore, it is advisable to mix this type of application with others, such as the necessity to apply a multi-step algorithm or method, understand the code's operational semantics, or even perform complex mathematical calculations. For example, if the question requests to generate the Alpha-Beta pruning code of a game tree generated with the minimax algorithm, it should also be requested to apply it to a specific game tree.

As a third recommendation, problems that involve the step-by-step application of an algorithm or a multi-step method are proposed. ChatGPT has performed relatively poorly in our study in this type of knowledge application, as can be seen by the results obtained in IASI, whose problems are practically all of this type.

The fourth recommendation is to incorporate questions with complex mathematical calculations where, for example, it is necessary to follow a specific method correctly. For instance, in the subject DABD, one of its answers contains the following error: "The FLIGHT table will have a record size of 34 bytes (4 + 3 + 3 + 4 + 8 + 70)."

As a fifth recommendation, we would like to note that the use of diagrams that represent specific instances of data struc-

tures or models on which the concepts learned in the subject have to be applied always poses an additional inconvenience for the use of ChatGPT, given that a prior translation into a textual format is necessary to include them in the prompt.

As a final recommendation, we have detected that it requires an additional effort to provide ChatGPT with exams that pose problems based on sufficiently elaborate and complex practical scenarios, e.g., the problems of BD or DABD. Apparently, those kinds of problems seem to affect ChatGPT's performance more clearly.

Finally, it is essential to remember that, given its constant evolution, the validity of these recommendations must always be checked against the latest version of ChatGPT available. For example, improvements in mathematical calculation are included in the January 30, 2023 version.

V. Conclusion

In this work, we are interested in evaluating the capabilities of ChatGPT for passing exams for a Computer Engineering degree. With this aim, we have designed and developed an experiment to provide ChatGPT with exams from 15 subjects. The results allow us to conclude that this type of technology already has a noticeable impact on the evaluation methods used in higher education. Therefore, it is necessary to systematically evaluate its capabilities to adapt evaluation methods appropriately in each context and moment. In this sense, this work proposes a series of recommendations regarding designing exam questions for Computer Engineering degrees. Moreover, given the non-deterministic nature of ChatGPT, it is crucial to obtain an accurate measure of its performance to analyze the distribution of the responses provided in the case of multiple-choice questions, as illustrated in this work.

As main future lines of work, we can point out the following. First, we need to develop a reference framework for evaluating the capabilities of this type of technology within the competencies of Computer Engineering, which will allow us to obtain a straightforward and quick measure of the performance of new tools or their evolutions. Additionally, there are other LLMs comparable to OpenAI's GPT-3, such as, to name a few, OPT from Meta [7] or BLOOM [8]. Second, the customization of LLMs with specific contents of computer science degrees for their innovative application in higher education contexts. It might be interesting to explore the concept of artificial stupidity, implicitly introduced by Alan Turing [21] and explicitly by Lars Liden [22]. A model adjusted with incorrect data would be used to generate essays. The student would have to analyze the work to find the inconsistencies. In both cases, the built LLMs would relieve teachers of repetitive work, such as creating quizzes adjusted to the syllabus or solving fundamental doubts. Third, the use of ChatGPT as a system for debugging exam questions to, for example, reduce the implicit context used in a question, avoid ambiguous or overlapping answer options in multiple-choice questions, or also eliminate possible distractors from the statements, as pointed out in the example of a multiple-choice question with mispriming shown at the end of Section III-B, since they can hinder its correct understanding. It would be possible to continue this line of work with a more detailed analysis of the distributions of the responses and their comparison with the most frequent mistakes made by students.

Acknowledgement

This publication is part of the R&D project ID2021-127412OB-I00, funded by MICIU/AEI/10.13039/501100011033 and "FEDER/UE".

References

[1] E. Tovar Caro, "Actas de las III Jornadas de Enseñanza Universitaria de Informática, Jenui 1997," Madrid, junio 1997.

[2] J. Divasón, F. J. Martínez de Pisón, A. Romero, and E. Sáenz de Cabezón, "Modelos de inteligencia artificial para asesorar el proceso evaluador de trabajos informáticos complejos," in Actas de las XXVII Jornadas de Enseñanza Universitaria de Informática, Jenui 2021. Asociación de Enseñantes Universitarios de la Informática (AENUI), 2021.

[3] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[4] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," Nov. 2022. [Online]. Available: https://openai.com/blog/chatgpt/.

[5] A. Vaswani et al., "Attention is All you Need," Advances in neural information processing systems, vol. 30, 2017.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, 2020.

[7] S. Zhang et al., "OPT: Open pre-trained transformer language models," arXiv:2205.01068, 2022. Available: https://arxiv.org/abs/2205.01068.

[8] B. Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," arXiv:2211.05100, 2022. Available: https://arxiv.org/abs/2211.05100.

[9] A. Gilson et al., "How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment," 2022. Available: https://www.medrxiv.org/content/10.1101/2022.12.23.22283901v1.

[10] M. J. Bommarito and D. M. Katz, "GPT Takes the Bar Exam," 2022. Available: https://ssrn.com/abstract=4314839.

[11] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet? - A systematic literature review on chatbots in education," Frontiers in Artificial Intelligence, vol. 4, 2021.

[12] X. Zhai, "ChatGPT user experience: Implications for education," 2022. Available: https://ssrn.com/abstract=4312418.

[13] M. M Alshater, "Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT," 12 2022. Available: https://ssrn.com/abstract=4312358.

[14] T. Susnjak, "ChatGPT: The End of Online Exam Integrity?" arXiv:2212.09292, 2022. Available: https://arxiv.org/abs/2212.09292.

[15] P. A. C. Debby R. E. Cotton and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," Innovations in Education and Teaching International, vol. 0, no. 0, pp. 1–12, 2023. Available: https://doi.org/10.1080/14703297.2023.2190148.

[16] Y. K. Dwivedi et al., "Opinion paper: "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," International Journal of Information Management, vol. 71, p. 102642, 2023. Available: https://www.sciencedirect.com/science/article/pii/S0268401223000233.

[17] F. J. García Peñalvo, F. Llorens-Largo, and J. Vidal, "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa," RIED-Revista Iberoamericana de Educación a Distancia, vol. 27, no. 1, p. 9–39, ene. 2024. Available: https://revistas.uned.es/index.php/ried/article/view/37716.

[18] R. Rodriguez-Echeverria, J. D. Gutiérrez, J. M. Conejero, and A. E. Prieto, "Impacto de ChatGPT en los métodos de evaluación de un grado de ingeniería informática," in Actas de las XXIV Jornadas de Enseñanza Universitaria de Informática, Jenui 2023. Asociación de Enseñantes Universitarios de la Informática (AENUI), 2023.

[19] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "Dissociating language and thought in large language models: A cognitive perspective," arXiv:2301.06627v1, 2023. Available: https://arxiv.org/abs/2301.06627.

[20] N. Kassner and H. Schütze, "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7811–7818.
[21] A. M. Turing, "I.—Computing Machinery and Intelligence," Mind, vol. LIX, no. 236, pp. 433–460, 10 1950. Available: https://doi.org/10.1093/mind/LIX.236.433.
[22] L. Lidén, "Artificial stupidity: The art of intentional mistakes," AI game programming wisdom, vol. 2, pp. 41–48, 2003.G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

Roberto Rodriguez-Echeverría is an Associate Professor in the Department of Computer and Telematic Systems Engineering at the Universidad de Extremadura (Spain), with more than 20 years of experience. He received his Ph.D. in Computer Science from Universidad de Extremadura in 2014. He is the co-author or author of more than 50 scientific papers. His research areas include Web Engineering, Big Data, Model Driven Software Development, Machine Learning, and Image Analysis. He is the Head of the Applied Information Technology Research Institute (Instituto de Tecnología Informática Aplicada, INTIA).

Juan D. Gutiérrez is an Assistant Professor in the Department of Electronics and Computing at the Universidad de Santiago de Compostela (USC). With more than twenty years of experience in the computer world, he recently presented his Ph.D. on visible LED light-based indoor positioning systems (IPS). His current research is focused on applying Artificial Intelligence to different fields of knowledge. His training includes programming in different languages, system administration, application design, and databases and the Internet.

José M. Conejero is an Associate Professor in the Department of Computer and Telematic Systems Engineering at the Universidad de Extremadura (Spain), where he has taught several courses related to Programming and Software Engineering. He received his PhD in Computer Science from Universidad de Extremadura in 2010. He is the author of more than 50 papers in journals and conference proceedings. He has also participated in different journals and conferences as a program committee member. His research areas include web engineering, big data, and model-driven development. He has also been involved in several competitive projects and contracts with entities and companies. Recently, he has co-founded MetrikaMedia, a startup company and spin-off of the Universidad de Extremadura.

Álvaro E. Prieto is an Associate Professor of Computer Languages and Systems and Head of e-administration at the University of Extremadura (Spain). He received his BSc in Computer Science from the University of Extremadura in 2000 and a Ph.D. in Computer Science in 2013. He is a member of the Quercus Software Engineering Group and is involved in various R&D&I projects. His research interests include software engineering, model-driven engineering, predictive analytics, and generative AI.