PIXEL-BIT
e-ISSN: 2171-7966
REVISTA DE MEDIOS Y EDUCACIÓN
MEDIA & EDUCATION JOURNAL

Distinction between Screenplay Texts Written by Humans and Generated by AI: A Preliminary Study with Film Students

Distinción entre textos de guion escritos por humanos y generados por IA: un estudio preliminar con estudiantes de Cine

Dr. Javier Luri Rodríguez
PDI Universidad del Atlántico Medio. España
D. Elio Quiroga Rodríguez
Cineasta. PDI Universidad del Atlántico Medio. España

Recibido: 2024/06/25; **Revisado:** 2024/04/19; **Aceptado:** 2024/11/24; **Online First:** 2024/12/03; **Publicado:** 2025/01/07

ABSTRACT
Most uncertainties surrounding the development of artificial intelligence in academia or professional environments are related to a certain intrusion of technology, not only because it performs tasks traditionally done by humans, but also because it can be difficult to identify. This article studies how different students of the Bachelor's degree in Film distinguish between scripts created with and without AI. The sample was chosen under the assumption that the students possess mature judgment regarding the topic investigated. 24 students were provided with three versions of the same scene script, together with a questionnaire to identify the texts' origin, either human or from Generative Artificial Intelligence, as well as justify the reasons for their choice. Among the three texts provided, some were exclusively human-generated and others were different types of synthetic texts. A quantitative analysis shows a considerable tendency to perceive that the texts are synthetic, regardless of their actual origin, resulting in a specific inability to distinguish. On the other side, a qualitative analysis has highlighted keys on how students perceive or assume the texts' origin. Certain notions of "naturalness" described by the participants were significant.

RESUMEN
La mayoría de las incógnitas que suscita el desarrollo de la inteligencia artificial en contextos como el académico o el profesional, están relacionadas con cierta intrusión de la tecnología, no solo por realizar labores tradicionalmente humanas, sino también por hacerlo de una manera que puede resultar difícil de identificar. El presente artículo estudia cómo distintos estudiantes del Grado de Cine distinguen entre guiones creados con IA y sin ella, muestra elegida por suponérsele un juicio maduro en este campo. A 24 alumnos se les facilitaron tres opciones de una misma escena de guion, junto con un cuestionario en el que debían reconocer la procedencia, humana o de inteligencia artificial generativa, de cada uno de los textos, así como justificar las razones de su elección. Entre los tres textos facilitados había producción exclusivamente humana y texto sintético de varias procedencias. Un análisis cuantitativo muestra una destacable inclinación hacia la percepción de que se trata de textos sintéticos, independientemente de que estos realmente lo sean, resultando una particular incapacidad de discernimiento. Por otro lado, un análisis cualitativo ha podido destacar claves sobre cómo los estudiantes perciben o presuponen las diferentes procedencias, resultando destacables ciertas nociones de “naturalidad” descritas por los participantes.

PALABRAS CLAVES · KEYWORDS
Inteligencia artificial; Evaluación de la percepción; Estudiantes universitarios; Narrativa audiovisual; Guion cinematográfico; Identificación de autoría; Artificial Intelligence; User Perception Evaluation; University Students; Audiovisual Narrative; Screenwriting; Authorship Identification

1. Introduction

In today's fast-moving world of artificial intelligence (AI), constantly changing and highly disruptive, advances in natural language generation (NLG) have reached high levels of sophistication, bringing software-generated scripts ever closer to those created by human beings. Examples of this are AI-based language models such as Google's Gemini and OpenAI's ChatGPT, which are capable of producing fluent and coherent text even in areas as apparently complex as dramaturgy (Anantrasirichai & Bull, 2021; Anguiano & Beckett, 2023). This level of refinement has led to the rapid introduction of this technology into various areas of literary creation and audiovisual scriptwriting (Chow, 2020), raising a multitude of questions about its scope, of which we could highlight two as being of primary importance: how pervasive could this technology become without being identifiable as such, and to what extent could its product become indistinguishable from that of a human creator (Dayo et al., 2023)?

The need to explore and understand this capacity of recognition is becoming increasingly relevant in a world where AI's involvement in content production is becoming ever more widespread. To what extent could human dramaturgy be supplanted by the automatic generation of content by current AI technology? Is there a discernible difference in quality, style or coherence between human-produced and AI-generated scripts?

Assessing people's ability to distinguish between a work created by humans and one generated by these machines or, in other words, the current ability of machines to deceive us with a creative product (Kurt, 2018), is a key question in two distinct ways: on the one hand, it evaluates the functionality of artificial intelligence in a specific context and, on the other, it tests our own capacity for discernment, which can provoke some reflection on our awareness of the scope of this technology (InFocus, 2023). Moreover, applying these recognition analyses to artistic work such as scriptwriting, in this instance, is particularly interesting as it raises a fundamental question: can machines produce expressive works that move us emotionally (Çelik, 2024; Francois, 2024)? Given the conventional understanding of the distinction between humans and machines, this is a key issue, and any contribution to this field of enquiry, however partial, can usefully widen our overall understanding of it (Li, 2022).

2. Methodology

In order to carry out the study, we took a scene from a feature film written by a scriptwriter, who was unaware of the use to which the script would be put, as well as of the study's objective. Then, two generative AI models - Google's Gemini (Metz & Grant, 2023) and OpenAI's ChatGPT, version 3.5 (Kozachek, 2023) – were both inputted with an identical long prompt (see Annex 1) giving them instructions to write the scene that included a broad outline in the form of a synopsis[1].

[1] The AI texts were obtained on 2/12/2023, and the questionnaires were completed by the 24 respondents during February and March 2024.

Twenty-four students, all of the first, second and third-year undergraduates taking the Degree in Cinema Studies at the Universidad del Atlántico Medio (UNAM), participated via an anonymous questionnaire in which each recorded their course year and gender and then, after viewing the three different versions of the same scripted scene, indicated whether they judged it to have been written by a human or by AI. They were shown one version written by a human, one written by ChatGPT and one written by Gemini, and made aware that any of the three scenes that they were about to read could have been written by AI, or by a human (see Annex 2). Respondents were also asked to think about why they made each choice, and to write a brief explanation in each case.

The results obtained were subjected to quantitative analysis so as to assess the respondents' ability to identify the source, and to examine related variables such as the type of AI used and the respondents' gender. The explanations offered by the students were also quantitatively analysed and, in search of recognition patterns, their responses were sorted into semantically-related categories. This was applied globally, analysing each of the three sources and also distinguishing between correct and incorrect identifications by the respondents.

3. Analysis and results

3.1 Quantitative results (identification ability):

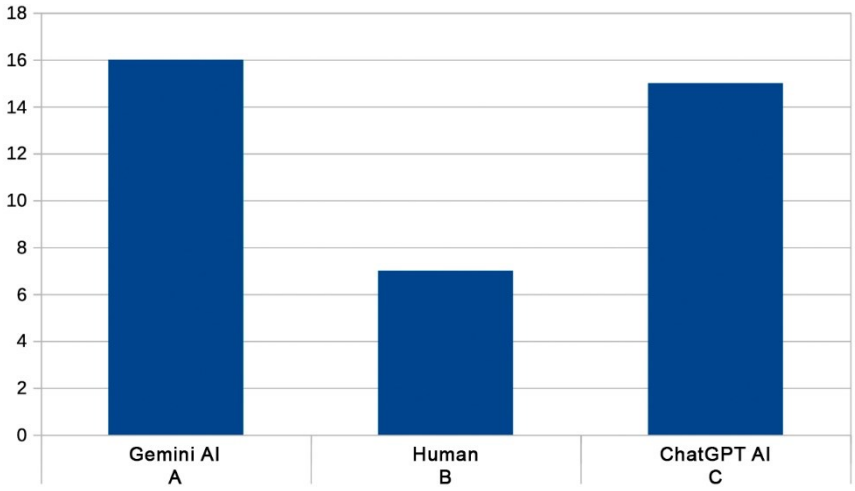
The alternatives were A (scene scripted by Gemini), B (scene scripted by a scriptwriter) and C (scene written by ChatGPT). The students had to mark whether each scene was scripted by AI or by a human, and the results were as follows:

- A (Gemini): Correct = 16 Incorrect = 8
- B (Human): Correct = 7 Incorrect = 17
- C (ChatGPT): Correct = 15 Incorrect = 9

Figure 1 is a histogram displaying the amount of correct answers for each version. At first glance, it seems clear that version B, the scene written by a scriptwriter, was the least-identified as such. But to properly assess the respondents' capacity to differentiate the versions, it is necessary to compare their performance with a randomly-generated result. If the respondents had made their judgements purely at random, given that there are 3 versions and 24 students, we would expect on average 8 correct answers.

Figure 1

Histogram of correct answers, for each of the three versions available



We can use the chi-square test to determine whether there is a significant difference between the observed and the expected frequency of correct answers (Lancaster & Seneta, 2005). If the difference is significant, we can conclude that people do have the ability to distinguish between versions over and above random guesswork.

First, we calculate the expected values for each version if the answers were random:

- A (Gemini) Expected = 8
- B (Human) Expected = 8
- C (ChatGPT) Expected = 8

Now, we calculate the chi-square statistic:

$$\chi^2 = \Sigma\ (O\text{-}E)^2 / E$$

Where O is the observed frequency and E is the expected frequency:

- For A (Gemini) $\chi^2_A = 16$
- For B (Human) $\chi^2_B = 0.25$
- For C (ChatGPT) $\chi^2_C = 12.25$

We can then compare the chi-square values obtained with a critical value for a given significance level and 2 degrees of freedom (since we have 3 options - 1). For a statistical significance level of 5%[2], the critical value of chi-square is approximately 5.99. Thus:

- For A (Gemini) $\chi^2_A = 16 > 5.99$.
- For B (Human) $\chi^2_B = 0.25 < 5.99$
- For C (ChatGPT) $\chi^2_C = 12.25 > 5.99$

From this we can conclude that the students have shown an ability to identify both the Gemini and ChatGPT versions as AI-generated, as their responses are significantly different from what would be expected at random (initial hypothesis). The question then arises as to whether there is a significant difference in response rates (AI/human) between AI-generated scripts and those written by humans without AI. There does not seem to be. This implies that two out of three responses follow the AI pattern, regardless of whether the script is AI or non-AI generated.

In addition, we can examine whether the gender of the respondents influences their success rate. First, we calculate the correct-answer ratios for males and females for each version:

- For A (Gemini AI):
  - Total correct answers = 16
  - Men: 8 correct answers
  - Women: 8 correct answers
  - Correct-answer ratio for men and women = 0.5

- For B (Human):
  - Total correct answers = 7
  - Men: 4 correct answers
  - Women: 3 correct answers
  - Correct-answer ratio for men ≈ 0.571
  - Correct-answer ratio for women ≈ 0.429

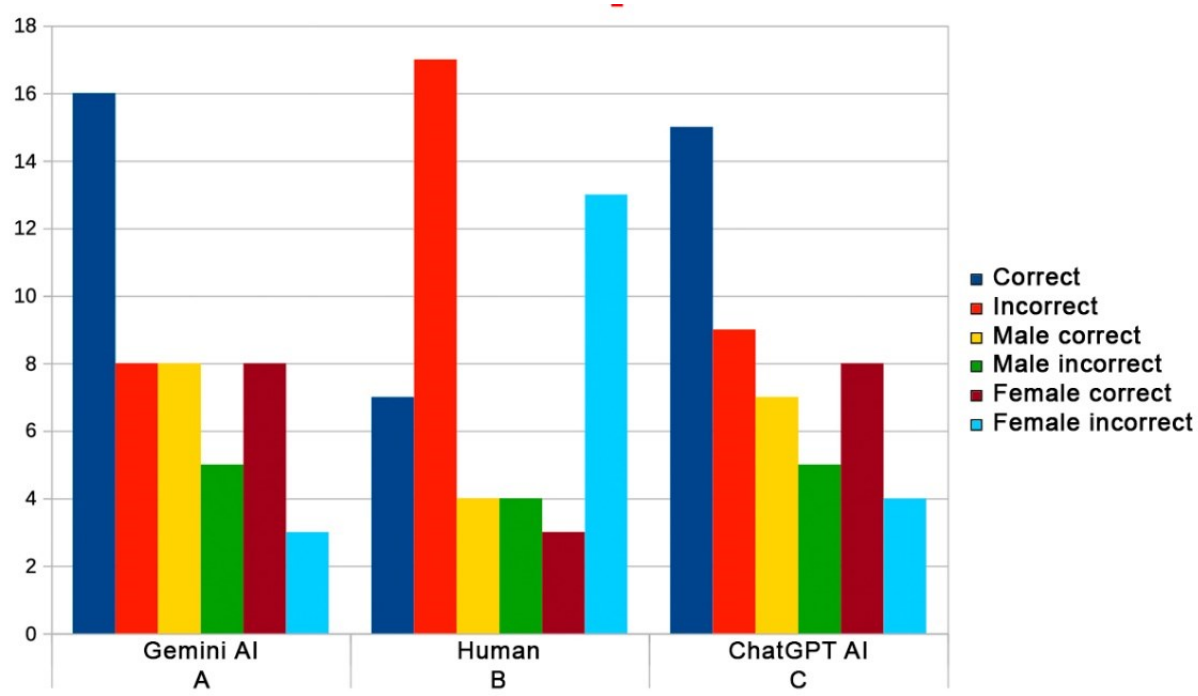
- For C (ChatGPT AI):
  - Total correct answers = 15
  - Men: 7 correct answers
  - Women: 8 correct answers
  - Correct-answer ratio for men ≈ 0.467
  - Correct-answer ratio for women ≈ 0.533

Figure 2 is a histogram showing the overall responses both of men and of women.

[2] The significance level of 0.05 is an arbitrary threshold commonly used in statistical hypothesis testing. It indicates a 5% probability of rejecting the null hypothesis when it is in fact true (Fernández Cano, 2009).

Figure 2

Histogram of correct and incorrect answers, overall and by gender



Hypothesis tests are then carried out for each version to determine whether there are significant differences between the proportions of correct answers by the male and by the female respondents. In each case, we can apply the difference in proportions test.

For each version, the null hypothesis H_0 would be that there is no difference between the male and the female correct-answer ratios, and the alternative hypothesis H_1 would be that there does exist a difference between the two.

Using a significance level of 0.05, we can calculate the z-test statistic and compare it to the critical value $z_{\alpha/2}$.

- For A (Gemini AI):
  - Ratio of correct answers by men: $p_1 = 0.5$
  - Ratio of correct answers by women: $p_2 = 0.5$
  - Male sample size: $n_1 = 8$
  - Female sample size: $n_2 = 8$

$$z = \frac{(p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where p is the combined proportion of correct answers:

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

In the case of A, $z = 0$, since $p_1 = p_2$.

• For B (Human):
- Ratio of correct answers by men: $p_1 = 0.571$
- Ratio of correct answers by women: $p_2 = 0.429$
- Male sample size: $n_1 = 4$
- Female sample size: $n_2 = 3$
- $z = 0.312$.

• For C (ChatGPT AI):
- Ratio of correct answers by men: $p_1 = 0.467$
- Ratio of correct answers by women: $p_2 = 0.533$
- Male sample size: $n_1 = 7$
- Female sample size: $n_2 = 8$
- $z = - 0.382$.

For a significance level of 0.05, $z_{\alpha/2} = \pm 1{,}96$, we now compare the values of z with those of $z_{\alpha/2}$.

- For A (Gemini AI): $z = 0$, we do not reject H_0, so there is no significant difference between men and women in correct answers for A.
- For B (Person): $z \approx 0.312$, we do not reject H_0, so there is no significant difference between men and women in correct answers for B.
- For C (ChatGPT AI): $z \approx - 0.382$, we do not reject H_0, so there is no significant difference between men and women in correct answers for C.

In conclusion, there are no significant differences between men and women in giving correct answers for any of the versions A (Gemini AI), B (Human) and C (ChatGPT AI).

Finally, the sampling error for this group of 24 subjects, assuming that the confidence level is 95% (with a significance level, as indicated above, of 5%) and that the proportion of the population that gets the answer right (correct for A, B and C) is 16.66% (4 completed questionnaires got it right), giving us:

$$\text{Sampling error} = 1.96 * \sqrt{(0.1666(1-0.1666) / 24)} = 0.149$$

With a confidence level of 95%, the sample result would be within a margin of error of ±0.149 of the true population value. This may indicate moderate precision, but would require a larger sample size to improve it if a more precise estimate is required. This research study is a first step with a small sample size, so it would seem advisable to repeat it with a larger and more diverse sample (of age, education, experience with technology etc.), as well as using a wider variety of scripts for the testing.

One process that can provide us with useful data over and above that offered by the histograms is the use of a confusion matrix, a summary table model invented by Karl Pearson using the term 'contingency table' (Pearson, 1904), which is currently used to evaluate the performance of classification models, especially to monitor the learning process of neural networks in artificial intelligence. In this case, we will apply it to the data obtained and interpret the results. Table 1 (below) shows the confusion matrix that has been designed to help further analyse the outcome of this experiment.

Table 1

Model of the confusion matrix used for this experiment

Student result	AI	Human	Total
AI	AI correctly identified (true positives - TP)	Human identified as AI (false positives - FP)	Total AI answers
Human	AI identified as person (false negatives - FN)	Human correctly identified (true negatives - TN)	Total human answers
Total	AI total	Human total	Answers total

Table 2 shows the experiment's results incorporated into the confusion matrix model:

Table 2

Confusion matrix of this experiment

Student result	AI	Human	Total
A (AI)	16 (TP)	8 (FP)	24
B (Human)	17 (FN)	7 (TN)	24
C (AI)	15 (TP)	9 (FP)	24
Total answers	48	24	72

The cells of the matrix indicate the following:

- TP (True positives): students who correctly identified the script written by AI as AI-generated (16 in A, 15 in C).
- FP (False positives): students who incorrectly identified the script written by a human as AI-generated (8 in A, 9 in C).
- FN (False negatives): students who incorrectly identified the script written by AI as having been written by a human (17 in B).
- TN (True negatives): students who correctly identified the script written by a human as having been written by a human (7 in B).

Metrics drawn from the confusion matrix:

- Overall accuracy: (TP + TN) / Total responses = (16 + 15 + 7) / 72 = 0.403 (40.3%).
- Accuracy with AI-written script: TP / AI = (16 + 15) / 2 = 15.5 (77.5%)
- Accuracy with human-written script: TN / human = 7 / 1 = 7 (70%)
- Identifiability of AI-written script: TP / (TP + FN) = (16 + 15) / (16 + 15 + 17) = 0.545 (54.5%)
- Identifiability of human-written script: TN / (TN + FP) = 7 / (7 + 8 + 9) = 0.318 (31.8%).

The overall accuracy of the students is relatively low (40.3%), which would suggest that they found it hard to successfully distinguish scripts written by AI from those written by a person. However, their accuracy with AI-generated scripts is considerably higher (77.5%) than their accuracy with human-written scripts (70%), suggesting that students were better at identifying AI-generated writing than at recognising human-written writing. Identifiability for AI is moderate (54.5%), indicating that students correctly identified a decent proportion of AI-generated scripts. Identifiability for script written by a human is low (31.8%), indicating that students tended to incorrectly identify human-written text as being written by AI.

The results of this confusion matrix analysis suggest that the students did have difficulties in accurately distinguishing between AI-generated and human-written scripts. While they were better at identifying AI-generated scripts, they also made a considerable number of mistakes when classifying human-written scripts. This is consistent with previous chi-square test findings.

3.2 Qualitative results (stated criteria)

Apart from having to identify scripts as AI-generated or written by humans, respondents were also asked for a brief written explanation of each choice they made, in order to examine descriptive judgement patterns in this context. Below, we discuss what respondents recorded as their reasons for ascribing authorship, artificial or human, to each of the three scripts under consideration.

Firstly, in the case of the first script that was produced with artificial intelligence (Version A, Gemini AI), it was found that:

a. Of the correct answers, three sets of answers can be identified that stand out as the three main indicators for correct identification of the artificial provenance of the text.

   In the first set, the one with the highest level of accuracy (two thirds of those that made the right choice), we find explanations that relate to sensing an *excess functionality,* using terms such as 'schematic', 'direct', 'flat', 'methodical' and 'literal', or comments such as 'it goes straight to the point', 'it says exactly what the synopsis says', or 'it's as if it nails all the plot points but without any emotional depth'.

   The next set of answers (half of those that got it right) indicates awareness of a certain lack of *naturalness.* This set groups answers using terms such as 'artificial', 'robotic', 'automated', 'automatic' or expressions such as 'not natural', 'not very personal', or 'sounds like a dictionary'. In a few instances, this reading of the script also focused on the issue of *emotion,* using comments such as 'lacking the corresponding emotion', 'cold', 'apathetic' and 'not very expressive'.

   The third notable set (also half of those that got it right), gives readings that refer to *simplicity,* grouping answers that use terms such as 'simplistic', 'brief', 'empty', 'lame' or comments such as 'character descriptions are thin' or 'very short on details'.

b. At the same time, among the incorrect answers that judged the script to be human-written, with a smaller representative sample, the main justifications (half of those that failed) also related to *naturalness,* with comments such as 'quite human', 'quite natural', 'credible', or even more explicitly: 'I think that the delivery of the speeches and the way that the stage directions and the dialogue are laid out is just like what I see in my colleagues' work and in my own when I'm writing screenplays'. Here there is no mention of *emotion.*

Secondly, with the second AI-generated script (Version C, ChatGPT AI), we found that:

a. Of the correct answers, two identifiable sets of responses stand out, both with the same number of mentions (just over half of those that got it right).

   One of these two sets coincides with the category also identified in the first script and described above as a lack of *naturalness.* This grouping includes comments such as 'lacks naturalness', 'sounds very forced', 'very robotic questions and answers', 'not very credible', 'mechanical', 'unnatural dialogue', or 'hackneyed and clichéd expressions'. Here too there are no references to *emotional* aspects.

   The other significant set of responses focuses on the identification of errors of consistency in the *content,* including expressions such as 'incongruities', 'inconsistencies', 'coherence errors', 'problems of coherence', and various comments such as 'people's replies don't relate to what the other person is asking'.

b. In the case of respondents' justifications when incorrectly identifying the AI-generated script as written by a human, as with the previous AI script there is a smaller representative sample (just over half of those that failed), and in the first instance they give reasons related to a lack of *simplicity,* with comments such as 'more intricate', 'a lot of detail', 'more complete' and 'developed'. The set previously identified as focusing on *naturalness* also appears (with only a third of those that failed), with observations such as 'fluid and natural dialogue', 'normal conversation between people', or 'shows how each person reacts to a particular situation'. Another justification is related to *emotion,* and another small set (half of those that failed) that argues that the *format* leads them (wrongly) to believe that it is a human creation: 'technical terminology more accurately utilised', 'effective stage directions', and 'shows 100% respect for the rules of the format'.

Thirdly, in relation to the script produced by a human scriptwriter (Version B), we should point out that:

a. Of the correct responses, we can highlight only a relatively modest set of answers (just over half of those who got it right) whose analysis focused on the perceived *naturalness* of the script*,* as mentioned above, and whose comments included 'clearly human conversation', 'does not seem forced', 'fluent', and 'colloquial and natural'.

b. Of the incorrect responses, which judged the script to be artificially generated, there is a set of respondents in a fairly large majority (just over half of those who failed), who explained their choice by reference to a perceived lack of *naturalness.* Their comments included terms such as 'not very organic', 'unnatural', 'stilted expressions', 'robotic dialogue', or 'not at all convincing', which once again emerge as the main criteria for their final identification choice. In this instance, these assessments of naturalness are not explicitly linked to *emotion,* which appears only in the comments of one sole respondent.

No pattern of any statistical significance has been found that relates the response type to the course year or the gender of the respondents.

In summary, it can be said that there are several sets of recurring responses, of which the most significant one focuses on what we have called '*naturalness*', being the most referenced in both correct and incorrect responses about Version B (Human), the most referenced in correct responses about Version C (ChatGPT AI), and the most referenced in incorrect responses about Version A (Gemini AI). Another relevant response set was found to focus on excessive *functionality,* being most common with Version A (Gemini AI). Other response sets also seem to be significant, such as the one focusing on *simplicity,* that is the most frequently referenced in the incorrect responses about Version C (ChatGPT AI), and is also present in the correct responses about Version A (Gemini AI).

4. Discussion

Despite the limited scope of the present study, the results obtained suggest that university students are currently unable to distinguish between drama scripts written by humans and those written by generative artificial intelligence. The fact that these data were obtained from students in a university film faculty reinforces the significance of the results, given that here it is people with a more than usual interest in and familiarity with the language of film who have demonstrated an inability to recognise the human origin of a scripted scene.

The reasoning behind respondents' identification of the scripts' human or AI origin seems to have been based on typifying human and robotic characteristics, which are then broken down into different attributes such as excess functionality, the presence or absence of naturalness and of simplicity, coherence or incoherence both in content and in format, and the presence or absence of emotion. While recognising the limited sample size, this discriminatory approach has allowed for a fairly detailed description of the different elective criteria of the respondents.

If the data presented here on the respondents' inability to discern the human origin of dramatic texts could be confirmed by further research studies with a broader scope, this would lead to more profound reflections and conclusions that are beyond the reach of this study. In addition, the extension and the broadening of the present study could help to explain in more detail and with greater reliability the characteristics attributed to human and to artificial dramatists. The implications of such comparative analyses would clearly be far-reaching, both for the humans involved in scriptwriting and for the developers of artificial intelligence.

At this point, it should be pointed out that this study is not an in-depth comparison of the particular form of human creation with what a machine is capable of: it does examine the capacity to generate literary text, but on the basis of specific guidelines previously introduced by a human being. Therefore, rather than confront the human directly with the machine, what is being compared here is human production *without* the help of an automated writing tool with human production *with* the help of such a tool. It is this automation of the literary expression of a pre-determined concept which is under consideration here, a field that was once the exclusive preserve of human skill and ingenuity. Determining the final authorship of an artistic product generated with AI tools is an analytical question well beyond the scope of this study.

With the accelerated development and increasing relevance of artificial intelligence nowadays, there is unparalleled interest in keeping a close eye on the constantly moving boundaries between artificial intelligence and human capacities and skills. As others have highlighted, 'ChatGPT is a digital life form constantly seeking evolution (...) it might blur the lines of its tool-like nature' (Luchen & Zhongwei, 2023).

At the same time, the idea that tests such as the one applied in this present study, or even the Turing test itself, are manifestly insufficient to reliably identify synthetic text seems to be gaining strength. As long ago as 1950, Alan Turing had already proposed a test in which an interrogator communicated in writing with a person and with a machine; if the machine could not be identified as such, it would pass the test (French, 2000). In current research, it is interesting to consider whether what we are investigating is the machine or the interrogator, in other words whether we are evaluating the sophistication of the technology or whether, on the contrary, we are evaluating the judgement, skills and capacities of the interrogator. The fact that in the present study a significant proportion of

the respondents saw what they felt were artificial and ‘robotic’ qualities in the human script could possibly be interpreted as a heightened mistrust provoked by a prideful reluctance to be fooled by a machine.

Especially in the field of education, there is urgent concern about the capacity to identify falsification and combat the intrusion associated with these synthetic productions in the student body. This is leading to an analysis and software-development race, in which the synthetic-creation side is currently ahead of the human AI-detection side. However, in both the academic and the professional worlds, the ranks of deserters from this particular battle are growing, seeing it as futile or unnecessary, so as to focus instead on designing ways to integrate artificial intelligence seamlessly and effectively into the processes of both education and creation.

5. Conclusions

The aim of this paper is to evaluate the ability of students to distinguish between scripts written by humans and those generated by artificial intelligence (AI), specifically Google's Gemini and OpenAI's ChatGPT, as well as to analyse the reasons given by respondents when distinguishing between these sources.

A quantitative analysis of the results indicates that respondents were able to distinguish between the versions generated by Gemini and ChatGPT, but were not significantly capable of distinguishing between human-written scripts and those generated by AI. Analysis of the results using the chi-square test showed that respondents' answers about the Gemini and ChatGPT scripts were significantly different from what would be expected on a purely random basis, and therefore suggested that they were better able to distinguish between these versions than by simply choosing at random. In addition, a gender analysis was conducted to determine whether there were differences between men and women in giving the right answers, and the results indicated that there were no significant differences between males and females in this regard for any of the versions evaluated.

In any case, success or otherwise in identifying the origin of the scripts can be considered a separate issue from the analysis of the respondents' identification criteria. The characteristics used and highlighted by the respondents have a significant value of their own in describing the image, prejudices and expectations that they have of this technology and its use on literary texts. This subjective image of AI has been described in other research studies with other types of university students, such as Computer Science (Singh et al., 2023) and Science (Yilmaz et al., 2023), as well as with faculty (Iqbal et al., 2022) with more generic impressions of the use of this technology, of which some highlight scepticism about its impact on learning (Lozano et al., 2021), especially with regard to critical thinking (González, 2023), and some pay particular attention to the trust derived from the perceived credibility of these tools.

A qualitative analysis of the answers written by the respondents has highlighted different sets of recurring answers, of which the main one focuses on what has been described as ‘naturalness’. This set, of which we have cited various examples of different description-types, has been grouped separately from other related criteria such as simplicity, errors of format or content, or explicit reference to issues relating to emotion. So, in this response category, there are judgements more related to mode of expression, as is reflected in comments like ‘the scene feels too forced, the dialogue comes across as cartoon-like’, ‘very

“robotic” questions and answers’, and even analyses with a more psychological slant such as ‘I don't see past the words that are used. There is no background depth’, and with judgements such as ’this script is poor, it’s lame, it's cold and flat. If a human being wrote it, it must have been a young boy or girl, or some soulless grown-up who’s lost the will to live’.

All of these statements were made while accurately identifying an AI-generated script, but similar expressions are used when respondents mistakenly identify the human script as AI-generated: ‘the dialogue is not very organic’, it is ‘lacking in substance and not very human’, ‘sometimes the dialogue strikes me as stilted and unnatural’, ‘robotic dialogue’ and ‘if this was done by a human, they’re clearly short of inspiration and understanding’. When respondents incorrectly identify an AI-generated script as being of human origin, they give similar naturalness-related explanations: ‘The dialogues are credible, they’re not over-embellished’, and one goes so far as to say that ‘the delivery of the speeches and the way that the stage directions and the dialogue are laid out is just like what I see in my colleagues’ work and in my own when I’m writing screenplays’.

Although the categorisation of the respondents' observations has made it possible to recognise particular features and to evaluate their relevance according to their relative frequency and weight, a larger and more diverse sample would undoubtedly provide the basis for a more concrete and well-founded analysis. This study is a modest first step in understanding how people perceive and distinguish between AI-generated texts and those written by humans, and offers a functional model for qualitative and quantitative research. However, in terms of the results of such a survey, it is acknowledged that the sample size is small, giving a sample error that needs improvement. Our goal, therefore, must be to repeat this study with a larger and more diverse sample, and to use a wider variety of scripts for testing, in order to obtain results which are both more robust and more generalisable.

Authors' contributions

Conceptualisation: J.L.R.and E.Q.R.. Data processing: J.L.R and E.Q.R.. Formal analysis: J.L.R and E.Q.R.. Research: J.L.R and E.Q.R.. Methodology: J.L.R and E.Q.R.. Writing - original draft: J.L.R and E.Q.R.. Writing - revising and editing: J.L.R and E.Q.R..

Acknowledgements

The authors would like to thank Marta Núñez Zamorano, Director of Academic Affairs of the Universidad del Atlántico Medio, for her generous help with the logistical practicalities of this research project, as well as the students of the first three years of the University's Degree in Cinema Studies, for their kind and generous collaboration. Thanks also for the support of Ignacio Luri Rodríguez, of DePaul University, Chicago.

Approval by the Ethics CommitteeAprobación por Comité Ético

This research work has been conducted with the approval of the Research Ethics Committee of the Universidad del Atlántico Medio (reference code: CEI/03-002).

References

Anantrasirichai, N., & Bull, D. (2021). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review, 55*(4), 589-656. https://doi.org/10.1007/s10462-021-10039-7

Anguiano, D., & Beckett, L. (2023, October 1). *How Hollywood writers triumphed over AI – and why it matters*. The Guardian.

Chow, P.-S. (2020). Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry. *NECSUS_European Journal of Media Studies, 9*(1), 193–214. https://doi.org/10.25969/mediarep/14307

Çelik, K. AI vs. Human in Screenwriting: Is AI the Future Screenwriter?. (2024) *Sakarya İletişim, 4*(1), 1-22.

Dayo, F., Memon, A. A., & Dharejo, N. (2023). Scriptwriting in the Age of AI: Revolutionizing Storytelling with Artificial Intelligence. *Journal of Media & Communication, 4*(1), 24-38.

Fernández Cano, A. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*. Ed. La Muralla.

Francois, S. (2024). *AI in Scriptwriting: Can a Computer Capture Human Emotion?*. Claremont McKenna College.

French, R. M. (2000). The Turing Test: The first 50 years. *Trends in Cognitive Sciences, 4*(3):115–122. ISSN 1364-6613, 1879-307X. https://doi.org/10.1016/S1364-6613(00)01453-4

González, M. A. M. (2023). Uso responsable de la inteligencia artificial en estudiantes universitarios: Una mirada recnoética. *Revista Boletín Redipe, 12*(9), 172-178.

InFocus Film School. (2023, 8 agosto). *Will AI replace screenwriters?: 8 reasons AI can't write good scripts.* Retrieved from https://infocusfilmschool.com/will-ai-replace-screenwriters/

Kozachek, D. (2023, June). Investigating the Perception of the Future in GPT-3,-3.5 and GPT-4. In *Proceedings of the 15th Conference on Creativity and Cognition*, 282-287. https://doi.org/10.1145/3591196.3596827

Kurt, D. E. (2018). *Artistic creativity in artificial* intelligence (Master's thesis). Radboud University.

Iqbal, N., Ahmed, H., & Azhar, K. A. (2022). Exploring teachers' attitudes towards using chatgpt. *Global Journal for Management and Administrative Sciences, 3*(4), 97–111. https://doi.org/10.46568/gjmas.v3i4.163

Lancaster, H. O., & Seneta, E. (2005). *Chi square distribution*. Encyclopedia of biostatistics, 2. https://doi.org/10.1002/0470011815.b2a15018

Li, Y. (2022). Research on the application of artificial intelligence in the film industry. In *SHS Web of Conferences* (Vol. 144, p. 03002). EDP Sciences.

Lozano, I. A., Molina, J. M., & Gijón, C. (2021). Perception of Artificial Intelligence in Spain. *Telematics and Informatics, 63*, 101672. https://doi.org/10.1016/j.tele.2021.101672

Luchen, F., & Zhongwei, L. (2023). ChatGPT begins: A reflection on the involvement of AI in the creation of film and television scripts. *Frontiers in Art Research, 5*(17). https://doi.org/10.25236/FAR.2023.051701

Metz, C., & Grant, N. (2023, December 8). Google Updates Bard Chatbot With'Gemini'AI as It Chases ChatGPT. *International New York Time*. https://www.nytimes.com/2023/12/06/technology/google-ai-bard-chatbot-gemini.html

Pearson, K. (1904). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london, 60*(359-367), 489-498. https://api.semanticscholar.org/CorpusID:119589729

Singh, H., Tayarani-Najaran, M. H., & Yaqoob, M. (2023). Exploring computer science students' perception of ChatGPT in higher education: A descriptive and correlation study. *Education Sciences, 13*(9), 924. https://doi.org/10.3390/educsci13090924

Yilmaz, H., Maxutov, S., Baitekov, A., & Balta, N. (2023). Student attitudes towards chat GPT: A technology acceptance Model survey. *International Educational Review, 1*(1), 57-83. https://doi.org/10.58693/ier.114

Annexes

https://doi.org/10.5281/zenodo.14176306