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

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

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

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

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.

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.

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. Introducción

En el panorama actual de la inteligencia artificial (IA), un escenario en constante cambio y disrupción, los avances en la generación de lenguaje natural han alcanzado altos niveles de sofisticación, acercando cada vez más las características de la creación humana a la generada por máquinas. En esta línea de desarrollo se encuentran los modelos de lenguaje basados en IA, como Gemini de Google y ChatGPT de OpenAI, que son capaces de producir texto de manera fluida y coherente, incluso en áreas tan aparentemente complejas como la redacción de textos dramáticos (Anantrasirichai & Bull, 2021; Anguiano & Beckett, 2023). Este nivel de perfeccionamiento ha ocasionado la rápida irrupción de esta tecnología en diversos ámbitos de la creación literaria y de guiones audiovisuales (Chow, 2020), ocasionando multitud de interrogantes sobre su alcance, entre los que podríamos destacar, como primarios, ¿cuán presente podría estar esta tecnología sin advertirlo? y con ello ¿cuál es su potencial actual para funcionar de manera indistinguible a la de un creador humano? (Dayo et al., 2023).

La necesidad de explorar y comprender esta capacidad de discernimiento se vuelve cada vez más relevante en un mundo donde la presencia de IA en la producción de contenido es cada vez más prominente. ¿Hasta qué punto puede ser suplantable una redacción dramática humana por una generada automáticamente por tecnología actual de IA? ¿Existe una diferencia perceptible en la calidad, el estilo o la coherencia entre los textos producidos por humanos y aquellos generados por IA?

Los interrogantes sobre la capacidad de las personas para distinguir entre una obra creada por humanos y una generada gracias a estas máquinas, o lo que es lo mismo, la capacidad actual de las máquinas para engañarnos con un producto creativo (Kurt, 2018), resulta una cuestión de doble interés: por una lado evalúa la funcionalidad de la inteligencia artificial en un contexto concreto, y por otro, tantea nuestra capacidad de discernimiento, lo que puede conllevar una reflexión sobre nuestra consciencia del alcance de esta tecnología (InFocus, 2023). Pero, además, el especial interés de llevar estos análisis de identificación a las obras artísticas, como es el caso de la escritura de un texto dramático, se fundamenta en una cuestión de trasfondo: ¿pueden las máquinas producir obras expresivas y llegar a emocionarnos (Çelik, 2024; Francois, 2024)? A tenor de las convencionales premisas de distinción entre el ser humano y las máquinas, esta resulta una cuestión de calado, y todo acercamiento parcial a este campo de indagación puede colaborar, ampliando un poco más el ángulo de visión global (Li, 2022).

2. Metodología

Para la realización del estudio, se tomó una secuencia de un guion de largometraje realizada por un guionista, que desconocía la finalidad del texto y el objeto de investigación. Por otro lado, se instruyó a dos inteligencias artificiales, Gemini, de Google (Metz & Grant, 2023), y ChatGPT, de OpenAI, en su versión 3.5 (Kozachek, 2023), mediante la redacción de un prompt extenso idéntico para las dos IAs (Anexo 1). Este, contenía las instrucciones para que escribieran la escena dándoles unos grandes rasgos en forma de sinopsis dentro del mismo[1].

1 Los textos de las IA fueron obtenidos el 2 / 12 / 2023 y los cuestionarios fueron rellenados por los 24 sujetos durante febrero y marzo de 2024.

Una muestra de 24 alumnos, que compone la totalidad del alumnado de 1º, 2º y 3º del Grado en Cine de la Universidad del Atlántico Medio, participó mediante un formulario anónimo en el que cada individuo debía indicar su curso y sexo, para después distinguir, en cada una de las tres opciones de una misma escena de un guion, si estaba escrita por una persona o una IA. Se entregaron una opción escrita por una persona, otra escrita por ChatGPT y la tercera por Gemini, indicándoles que cualquiera de las tres escenas que iban a leer podría estar escrita por una IA, o por un humano (Anexo 2). También se pidió a los participantes que razonaran sobre por qué han tomado cada decisión, redactando una respuesta corta para cada caso de identificación.

Los resultados recabados se sometieron a un análisis cuantitativo para analizar la capacidad de los participantes para identificar la procedencia, buscando también variables relacionadas como el tipo de IA utilizada o el sexo de los participantes. Las justificaciones aportadas por los alumnos sirvieron para realizar un análisis cuantitativo en el que se dividieron las respuestas por categorías semánticamente afines en busca de patrones de reconocimiento. Esto se hizo de manera global, atendiendo a cada una de las tres procedencias y también distinguiendo entre las identificaciones correctas y las incorrectas de los participantes.

3. Análisis y resultados

3.1 Resultados cuantitativos (capacidad de identificación):

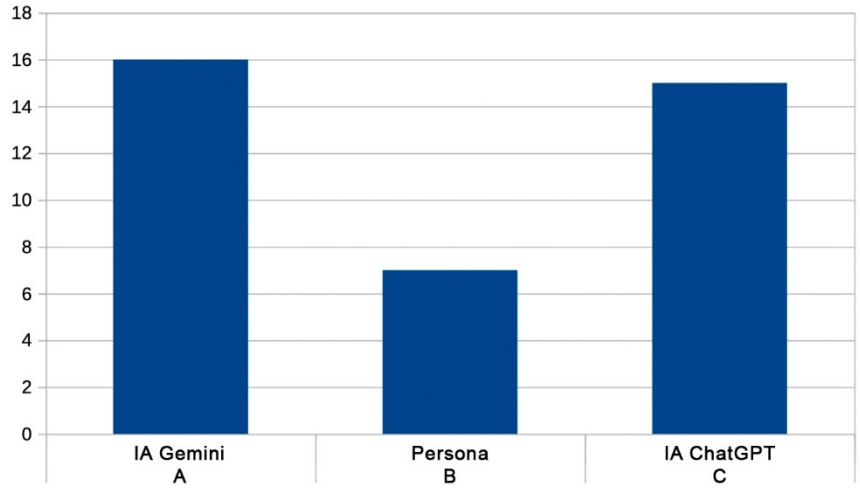
Las alternativas eran A (escena escrita por Gemini), B (escena escrita por un guionista) y C (escena escrita por ChatGPT); los estudiantes tenían que marcar si cada escena estaba escrita por una IA o una persona, resultando:

- A (Gemini): Correctas = 16, Incorrectas = 8
- B (Persona): Correctas = 7, Incorrectas = 17
- C (ChatGPT): Correctas = 15, Incorrectas = 9

En la Figura 1 se muestra un histograma con las proporciones de las respuestas correctas. A primera vista parece que la opción B, la escrita por un guionista, no fue identificada como tal. Con todo, para evaluar si los sujetos pudieron distinguir entre las opciones, es necesario comparar su desempeño con lo que podríamos esperar al azar. Si suponemos que las personas están adivinando por azar, entonces esperaríamos que acertaran aproximadamente 8 respuestas (ya que hay 24 personas y 3 opciones posibles).

Figura 1

Histograma de los aciertos, en función de las tres opciones posibles



Podemos utilizar la prueba de chi-cuadrado para determinar si existe una diferencia significativa entre la frecuencia observada y esperada de respuestas correctas (Lancaster & Seneta, 2005). Si la diferencia es significativa, se puede concluir que las personas tienen capacidad de distinguir entre las opciones por encima de la simple adivinación al azar.
Primero, calculamos los valores esperados para cada opción si las respuestas fueran al azar:

- A (Gemini): Esperadas = 8
- B (Persona): Esperadas = 8
- C (ChatGPT): Esperadas = 8

Ahora, calculamos el estadístico de chi-cuadrado:

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

Donde O es la frecuencia observada y E la frecuencia esperada.

- Para A (Gemini), $\chi^2_A = 16$
- Para B (Persona), $\chi^2_B = 0{,}25$
- Para C (ChatGPT), $\chi^2_C = 12{,}25$

Tras ello, podemos comparar los valores obtenidos de chi-cuadrado con un valor crítico para un nivel de significación dado y 2 grados de libertad (ya que tenemos 3 opciones - 1).

Para un nivel de significación estadística del 5%[2], el valor crítico de chi-cuadrado es aproximadamente 5,99. Así:

- Para A (Gemini), $\chi^2_A = 16 > 5,99$
- Para B (Persona), $\chi^2_B = 0,25 < 5,99$
- Para C (ChatGPT), $\chi^2_C = 12,25 > 5,99$

De ello podemos concluir que las personas han mostrado capacidad para distinguir como inteligencia artificial tanto la opción de Gemini como la de ChatGPT, ya que sus respuestas están significativamente alejadas de lo que se esperaría al azar (hipótesis inicial). Cabe preguntarse entonces si existe una diferencia significativa en las tasas de respuesta (IA/humano) entre los textos generados con IA y los escritos por humanos sin IA. No parece haberla. Esto implica que dos de cada tres respuestas siguen la pauta de la IA, independientemente de si el texto es generado o no con IA.

Como dato adicional podemos atender a si el sexo de los sujetos influye en el índice de éxito. Primero, calculemos las proporciones de aciertos para hombres y mujeres en cada opción:

- Para A (IA-Gemini):
  - Total de aciertos = 16
  - Hombres: 8 aciertos
  - Mujeres: 8 aciertos
  - Proporción de aciertos para hombres y mujeres = 0,5

- Para B (Persona):
  - Total de aciertos = 7
  - Hombres: 4 aciertos
  - Mujeres: 3 aciertos
  - Proporción de aciertos para hombres ≈ 0,571
  - Proporción de aciertos para mujeres ≈ 0,429

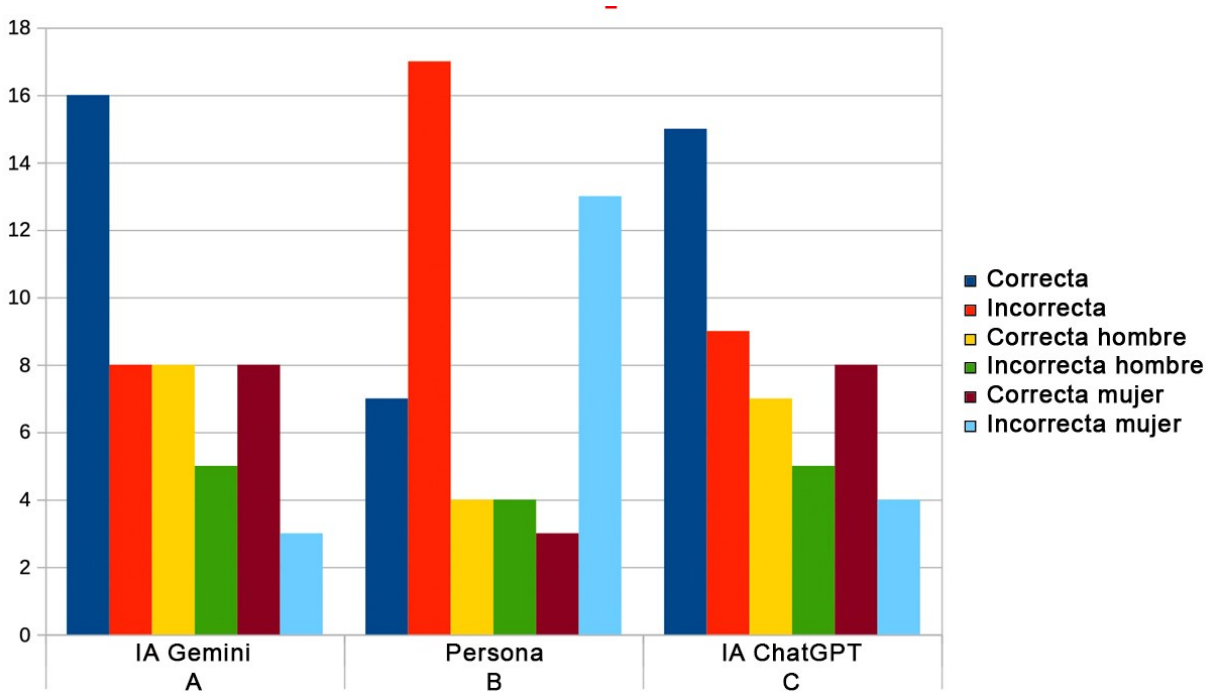
- Para C (IA-ChatGPT):
  - Total de aciertos = 15
  - Hombres: 7 aciertos
  - Mujeres: 8 aciertos
  - Proporción de aciertos para hombres ≈ 0,467
  - Proporción de aciertos para mujeres ≈ 0,533

2 El nivel de significación de 0,05 es un umbral arbitrario comúnmente utilizado en las pruebas de hipótesis estadísticas. Indica que existe una probabilidad del 5% de rechazar la hipótesis nula cuando en realidad es verdadera (Fernández Cano, 2009).

En la Figura 2 se puede observar el histograma de las respuestas generales, de hombres y de mujeres.

Figura 2

Histograma de las respuestas correctas e incorrectas, generales y por sexos



Se procede a realizar pruebas de hipótesis para cada opción para determinar si hay diferencias significativas entre las proporciones de aciertos de hombres y mujeres. Podemos usar la prueba de diferencia de proporciones para cada caso.

Para cada opción, la hipótesis nula H_0 sería que no hay diferencia entre las proporciones de aciertos de hombres y mujeres, y la hipótesis alternativa H_1 sería que sí hay una diferencia entre ambas.

Utilizando un nivel de significación de 0,05, podemos calcular el estadístico de prueba z y compararlo con el valor crítico $z_{\alpha/2}$.

- Para A (IA-Gemini):
  - Proporción de aciertos para hombres $p_1 = 0,5$
  - Proporción de aciertos para mujeres $p_2 = 0,5$
  - Tamaño de la muestra para hombres $n_1 = 8$
  - Tamaño de la muestra para mujeres $n_2 = 8$

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

Donde *p* es la proporción combinada de aciertos:

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

En el caso de A, $z = 0$, ya que $p_1 = p_2$.

- Para B (Persona):
  - Proporción de aciertos para hombres $p_1 = 0{,}571$
  - Proporción de aciertos para mujeres $p_2 = 0{,}429$
  - Tamaño de la muestra para hombres $n_1 = 4$
  - Tamaño de la muestra para mujeres $n_2 = 3$
  - $z = 0{,}312$.

- Para C (IA-ChatGPT):
  - Proporción de aciertos para hombres $p_1 = 0{,}467$
  - Proporción de aciertos para mujeres $p_2 = 0{,}533$
  - Tamaño de la muestra para hombres $n_1 = 7$
  - Tamaño de la muestra para mujeres $n_2 = 8$
  - $z = -0{,}382$.

Para un nivel de significación de 0,05, $z_{\alpha/2} = \pm 1{,}96$, comparamos ahora los valores de *z* con los de $z_{\alpha/2}$.

- Para A (IA-Gemini): $z = 0$, no rechazamos H_0, por lo que no hay diferencias significativas entre hombres y mujeres en los aciertos para A.
- Para B (Persona): $z \approx 0{,}312$, no rechazamos H_0, por lo que no hay diferencias significativas entre hombres y mujeres en los aciertos para B.
- Para C (IA-ChatGPT): $z \approx -0{,}382$, no rechazamos H_0, por lo que no hay diferencias significativas entre hombres y mujeres en los aciertos para C.

En conclusión, no hay diferencias significativas entre hombres y mujeres en los aciertos para ninguna de las opciones A (IA-Gemini), B (Persona) y C (IA-ChatGPT).

Finalmente, el error muestral de este grupo de 24 sujetos, suponiendo que el nivel de confianza es del 95% (con un nivel de significación como el indicado más arriba, del 5%) y que la proporción de la población que acierta la respuesta (acierto en A, B y C) es del 16,66% (4 encuestas acertaron), tendríamos:

$$\text{Error muestral} = 1{,}96 * \sqrt{(0{,}1666(1-0{,}1666) / 24)} = 0{,}149$$

Con un nivel de confianza del 95%, el resultado de la muestra estaría dentro de un margen de error de ±0,149 del valor real de la población. Ello puede indicar una precisión moderada, pero requeriría un tamaño de muestra mayor para mejorarla si se necesitara una estimación más precisa. Esta investigación es un primer paso con una pequeña muestra, por lo que parece recomendable replicar el estudio con una muestra mayor y con mayor

diversidad (edad, formación, experiencia en tecnología, etc.), así como utilizar una variedad más amplia de textos para la prueba.

Un proceso que puede darnos datos de interés más allá de los histogramas, es el uso de una matriz de confusión, un modelo de tabla resumida inventada por Karl Pearson con el término "tabla de contingencia" (Pearson, 1904), que se utiliza actualmente para evaluar el rendimiento de modelos de clasificación, especialmente para supervisar el proceso de aprendizaje de redes neuronales en inteligencia artificial. En este caso, lo aplicaremos sobre los resultados obtenidos y colegiremos las consecuencias. En la Tabla 1 podemos ver la matriz de confusión que se ha diseñado para el análisis de este experimento.

Tabla 1

Modelo de la matriz de confusión utilizada para el experimento.

Resultado del alumno	**IA real**	**Persona real**	**Total**
IA	IA correctamente identificada (Verdaderos positivos - VP)	Persona identificada como IA (Falsos positivos - FP)	Total respuestas IA
Persona	IA identificada como persona (Falsos negativos - FN)	Persona correctamente identificada (Verdaderos negativos - VN)	Total respuestas Persona
Total	Total IA	Total Persona	Total respuestas

Por su parte, la Tabla 2 muestra la matriz de confusión con los datos del experimento introducidos.

Tabla 2

Matriz de confusión del experimento

Resultado del alumno	IA real	Persona real	Total
A (IA)	16 (VP)	8 (FP)	24
B (Persona)	17 (FN)	7 (VN)	24
C (IA)	15 (VP)	9 (FP)	24
Total Respuestas	48	24	72

Las casillas de la matriz son las siguientes:

- VP (Verdaderos positivos): Alumnos que identificaron correctamente el texto escrito por IA como tal (16 en A, 15 en C).
- FP (Falsos positivos): Alumnos que identificaron incorrectamente el texto escrito por una persona como IA (8 en A, 9 en C).
- FN (Falsos negativos): Alumnos que identificaron incorrectamente el texto escrito por IA como escrito por una persona (17 en B).
- VN (Verdaderos negativos): Alumnos que identificaron correctamente el texto escrito por una persona como tal (7 en B).

Métricas a partir de la matriz de confusión:

- Precisión general: (VP + VN) / Total respuestas = (16 + 15 + 7) / 72 = 0.403 (40.3%)
- Precisión para texto escrito por IA: VP / IA real = (16 + 15) / 2 = 15.5 (77.5%)
- Precisión para texto escrito por persona: VN / Persona real = 7 / 1 = 7 (70%)
- Identificabilidad para texto escrito por IA: VP / (VP + FN) = (16 + 15) / (16 + 15 + 17) = 0.545 (54.5%)
- Identificabilidad para texto escrito por persona: VN / (VN + FP) = 7 / (7 + 8 + 9) = 0.318 (31.8%)

La precisión general de los alumnos es relativamente baja (40.3%), lo que revelaría que tuvieron dificultades para distinguir con precisión entre los textos escritos por una IA y los escritos por una persona. Sin embargo, la precisión para IA es considerablemente más alta (77.5%) que la precisión para los textos escritos por personas (70%), lo que sugiere que los alumnos fueron mejores para identificar las escrituras generadas por IA, que para reconocer aquellas realizadas por personas. La sensibilidad para IA es moderada (54.5%), indicando que los alumnos identificaron correctamente una proporción decente de textos escritos por IA. La especificidad para el texto escrito por una persona es baja (31.8%), lo

que significa que los alumnos identificaron incorrectamente el texto escrito por una persona como si fuera escrito por una IA.

Los resultados del análisis mediante matriz de confusión sugieren que los alumnos tuvieron dificultades para distinguir con precisión entre los textos escritos por IA y los escritos por personas; si bien fueron mejores para identificar textos escritos por IA, también cometieron una cantidad considerable de errores al clasificar textos escritos por personas. Todo ello es coherente con lo obtenido anteriormente mediante chi-cuadrado.

3.2 Resultados cualitativos (criterios manifestados):

Aparte de la identificación entre texto con IA y puramente humano, a los participantes se les pidió una breve justificación escrita de cada elección, con la intención de buscar patrones descriptivos de los juicios al respecto. A continuación, se comentará lo que los participantes han manifestado sobre las razones para deducir la autoría, artificial o humana, de cada uno de los tres textos.

- En primer lugar, sobre el primer texto realizado con inteligencia artificial (Opción A, IA Gemini), se ha encontrado que:

  a) De entre las respuestas correctas, se pueden identificar tres bloques de respuestas que destacan como las tres razones principales para identificar la procedencia artificial del texto.
  En el primer bloque, el de mayor nivel de coincidencias (dos tercios de los que acertaron), encontramos justificaciones que pueden relacionarse con una *funcionalidad excesiva*. En él encontramos explicaciones basadas en los términos “esquemático”, “directo”, “plano”, “metódico”, “literalidad” o expresiones como “va al grano”, “dicen exactamente lo que dice la sinopsis”, o “es como si dieran en el clavo en todos los puntos dramáticos pero sin la emoción necesaria”.
  El siguiente bloque de respuestas (la mitad de los que acertaron), tienen que ver con la apreciación de falta de *naturalidad*, este bloque agrupa respuestas en torno a términos como “artificial”, “robotizado”, “robótico”, “automático” o expresiones como “no natural”, “poco personal”, o “de diccionario”. En unos pocos casos estos motivos además se asociaban con la *emoción*, añadiendo expresiones como “sin la emoción necesaria”, “frío”, “desganado” o “poco expresivo”.
  El tercer bloque destacable (también de la mitad de los que acertaron), está compuesto con explicaciones relacionadas con la *simplicidad*, agrupando frases en torno a los términos “simple”, breve”, “vacío”, “pobre” o expresiones como “descripciones de los personajes son escasas” o “tiene muy pocos detalles”.

  b) Por otro lado, con menor muestra representativa, de entre las respuestas que no acertaron, creyendo que era humano, las justificaciones principales (la mitad de los que fallaron) se relacionan también con la *naturalidad*, con respuestas en torno a ideas como “bastante humano”, “bastante natural”, “creíble” o incluso,

más explícitamente: “Creo que la expresión de las oraciones y el modo de ordenar las acotaciones e intervenciones son como las que veo en mis compañeros y hago yo al escribir guiones”. En este caso, sin hacer alusión a la *emoción*.

- En segundo lugar, sobre el segundo texto realizado con inteligencia (Opción C, IA ChatGPT), se ha encontrado que:

  a) De entre las respuestas correctas, se pueden identificar dos bloques de respuestas destacables, ambos con la misma cantidad de mención (poco más de la mitad de los que acertaron).
  Uno de estos dos bloques coincide con la categoría identificada también en el primer texto, nombrada anteriormente como falta de *naturalidad.* En ella se agrupan esta vez expresiones como “falta naturalidad”, “suena muy forzado”, “preguntas y respuestas muy robóticas”, “poco creíble”, “mecánico”, “diálogos no naturales”, o “expresiones rehechas y precocinadas”. En este caso, tampoco se relaciona con aspectos relacionados con la *emoción*.
  El otro bloque relevante de respuestas se centra en la identificación de errores de congruencia en el *contenido*, agrupando expresiones como “incongruencias”, “incoherencias”, “errores de coherencia”, “problemas de coherencia”, y varios ejemplos como “respuestas no tienen relación con lo que el otro pregunta”.

  b) Del otro lado, el de las justificaciones cuando son respuestas erróneas, creyendo que era texto humano, como en la IA anterior hay menor muestra representativa (poco más de la mitad de los que fallaron), e indican en primer lugar razones relacionadas con la no *simplicidad*, con frases en torno a “más especificado”, “mucho detalle”, “más completo” o “desarrollado”. El bloque antes identificado como *naturalidad* también aparece (con solo un tercio de los que fallaron), con ideas como “diálogos fluidos y naturales”, “conversación normal entre personas”, o “se muestra cómo reacciona cada persona ante alguna situación”; a esto se le podría añadir otra razón relacionada con la *emoción*, y otro pequeño bloque (la mitad de los que fallaron) que aduce que el *formato* le lleva (erróneamente) a creer que es humano: “nomenclatura mejor presentada”, “suficientes acotaciones”, “respeta al cien por cien las reglas del formato”.

- En tercer lugar, en relación con el texto realizado por un guionista humano (Opción B), se puede destacar que:

  a) De entre las respuestas correctas, solo se puede destacar un bloque modesto de respuestas (poco más de la mitad de los que acertaron) que corresponde con cuestiones relacionadas con la mencionada *naturalidad* percibida,

recogiendo explicaciones basadas en expresiones como “conversación humana”, “no se ve forzado”, “fluido”, “coloquiales y naturales”.

b) En lo referente a las respuestas incorrectas, creyendo que el texto no es humano, destaca de nuevo el bloque, bastante mayoritario (poco más de la mitad de los que fallaron) que podemos agrupar con las justificaciones relacionadas con la *naturalidad* percibida. Respuestas basadas en términos como “poco orgánicos”, “sin naturalidad”, “expresiones rebuscadas”, “diálogo robótico”, o “no son del todo convincentes” vuelven a concurrir como principales criterios de identificación. Estas respuestas sobre la naturalidad, en este caso carecen de vinculación explícita con la *emoción*, que aparece solo en la respuesta de otro participante aislado.

No se ha encontrado ningún patrón significativo que relacione el tipo de respuestas con el curso ni con el sexo de los participantes.

Con todo ello, se puede sintetizar que enquisten varios bloques de respuestas recurrentes, entre los que destaca en primer lugar el que hemos denominado “*naturalidad*”, siendo el más presente tanto en las respuestas correctas como en las incorrectas de la Opción B (Humana), también el más presente en las respuestas correctas de la Opción C (IA ChatGPT), y en las incorrectas de la Opción A (IA Gemini). Otro bloque de respuestas relevante resultó ser el que se relaciona con un exceso de *funcionalidad*, siendo el de mayor presencia en la Opción A (IA Gemini). Otros bloques de respuestas también parecen destacables, como el relacionado con la *simplicidad*, el mayoritario entre las respuestas incorrectas de la Opción C (IA ChatGPT), también presente entre las respuestas correctas de la Opción A (IA Gemini).

4. Discusión

A pesar del limitado alcance del presente estudio, los resultados arrojados apuntan hacia la idea de que alumnos universitarios no pueden distinguir ahora mismo los textos dramáticos escritos por personas de los escritos con inteligencias artificiales generativas. El hecho de que los datos se hayan obtenido de alumnado de una Facultad de Cine, refuerza la significación de los resultados, ya que son estudiantes supuestamente interesados y familiarizados con el lenguaje cinematográfico los que han mostrado incapacidad a la hora de distinguir la procedencia humana de una secuencia de guion.

Las razones de esta identificación aportadas por los participantes parecen describir componentes humanos y robóticos, que se han podido desglosar en distintas características como la excesiva funcionalidad, la falta o presencia de naturalidad, así como de simplicidad, las incoherencias o coherencias tanto de contenido como de formato y la presencia o carencia de emoción. Teniendo en cuenta la limitada muestra usada, esta discriminación ha permitido una descripción bastante detallada de los distintos criterios electivos de los sujetos.

Con el refuerzo de otras investigaciones con mayor alcance, se podrían confirmar los datos aquí expuestos sobre la incapacidad de discernir la procedencia humana de textos dramáticos, lo que podría mover a una reflexión que escapa al alcance de este trabajo. Además, este refuerzo o la ampliación del presente estudio, puede ayudar a explicar con más detalle y fiabilidad las características que se atribuyen a los creadores dramáticos humanos y a los artificiales. Las implicaciones de tales descripciones comparativas se pueden intuir trascendentes, tanto para los humanos implicados en la literatura cinematográfica, como para los desarrolladores de inteligencia artificial.

En este punto, resulta pertinente matizar que el presente estudio no recoge una comparación radical entre la forma de creación humana y la que es capaz de hacer una máquina; se ha valorado la capacidad de generar texto literario, pero en base a unas directrices concretas introducidas previamente por un humano. Por tanto, más que confrontar al humano con la máquina, lo que se ha comparado es la producción humana, sin ayuda de una herramienta de redacción automatizada, con la producción con la ayuda de esta herramienta. Es esa automatización, la de la redacción literaria de una idea prefijada, a la que aquí se ha atendido, como un campo antes exclusivo de la maña humana. La autoría final del producto artístico que haya sido generado con herramientas de IA, es una cuestión de análisis ajena a las pretensiones de este trabajo.

Con el acelerado desarrollo y la creciente relevancia de la inteligencia artificial hoy, es indiscutible el interés por seguir atendiendo a esta movilidad de fronteras entre las competencias de la inteligencia artificial y las humanas. Como otros han puesto de relieve, “ChatGPT es una forma de vida digital en constante evolución (…) podría desdibujar las líneas de su naturaleza de herramienta” (Luchen & Zhongwei, 2023).

Por otro lado, parece coger fuerza la idea de que pruebas como la del presente trabajo o incluso el mismo test de Turing resultan claramente insuficientes para identificar texto sintético. Turing, ya en los años cincuenta, propuso una prueba en la que un interlocutor se comunicaba por escrito con un humano y con una máquina, de no distinguirse a la máquina, esta estaría pasando la prueba (French, 2000). En las investigaciones actuales, resulta interesante atender a si se investiga a la máquina o al humano, es decir, si se trata de valorar la sofisticación de la tecnología o si por el contrario se está evaluando la capacidad de discernimiento del usuario, sus capacidades y recursos para ello. Que en el presente trabajo los participantes hayan visto aspectos artificiales y “robóticos” de manera masiva en el texto humano, podría interpretarse como una desconfianza impostada por el orgullo del que no quiere ver cuestionado su criterio. Y es que, especialmente en el ámbito de la educación, existe un interés acuciante por identificar la falsificación y la intrusión asociada a estas producciones sintéticas entre el alumnado. Esto está llevando a una carrera de perspicacia y desarrollo de software, en la que el bando de la creación sintética va por delante del de la identificación humana, pero tanto en el mundo académico como en el profesional, crecen las filas de los que desertan de esta batalla, por imposible o innecesaria, para centrarse en diseñar formas en la que integrar la inteligencia artificial de manera transparente y efectiva en los procesos de educación y de creación.

5. Conclusiones

Este trabajo busca evaluar la capacidad de estudiantes para distinguir entre textos escritos por personas y aquellos generados por inteligencias artificiales (IA),

específicamente Gemini de Google y ChatGPT de OpenAI, así como de hacer un análisis de las razones aducidas por los participantes a la hora de distinguir esta procedencia.

Los resultados, en un análisis cuantitativo, indican que los participantes lograron distinguir entre las opciones proporcionadas por Gemini y ChatGPT, pero no lograron distinguir de manera significativa entre textos escritos por personas y aquellos generados por las IA. El análisis de los resultados a través de la prueba de chi-cuadrado mostró que las respuestas de los participantes estaban significativamente alejadas de lo que se esperaría al azar para las opciones de Gemini y ChatGPT, lo que sugiere que los participantes pudieron distinguir entre estas opciones mejor que simplemente eligiendo al azar. Además, se llevó a cabo un análisis por género para determinar si había diferencias en los aciertos entre hombres y mujeres. Los resultados indicaron que no hubo diferencias significativas entre hombres y mujeres en los aciertos para ninguna de las opciones evaluadas.

En cualquier caso, el acierto o no a la hora de identificar la procedencia de los textos, puede considerarse una cuestión al margen del análisis de los criterios de identificación de los participantes. Las características atendidas y reseñadas por los participantes contienen un valor significativo propio, al describir la imagen, prejuicios y expectativas, que se tiene de esta tecnología, y su uso en textos literarios. Esta imagen subjetiva de la IA se está describiendo en otras investigaciones con otros tipos de alumnado universitario, como el de Informática (Singh et al., 2023) o el de Ciencias (Yilmaz et al., 2023), como en el profesorado (Iqbal et al., 2022) con impresiones más genéricas sobre el uso de esta tecnología, entre las que destaca el escepticismo sobre el impacto en el aprendizaje (Lozano et al., 2021), especialmente en lo referente al pensamiento crítico (González, 2023), y en las que se manifiesta una atención especial a la confianza derivada de la credibilidad percibida en estas herramientas.

Un análisis cualitativo sobre las respuestas redactadas por los participantes, ha destacado distintos bloques de respuestas recurrentes, entre los que destaca en primer lugar el que se ha denominado como “naturalidad”. Este bloque, sobre el que hemos descrito diversos ejemplos en tipos de frases, se ha aislado de otros aspectos relacionados, como la simplicidad, los errores de formato o contenido o la explícita referencia a cuestiones relacionadas con la emoción. Queda por tanto en esta categoría de respuestas, las ideas más relacionadas con un modo de expresión que se ha descrito como “la escena se siente demasiado forzada, las expresiones se sienten como dibujos animados”, “respuestas y preguntas muy “robóticas”, incluso con análisis más relacionados con connotaciones psicológicas como “no veo más allá de las palabras que se utilizan. No hay transfondo”, o incluso con juicios como “este guión es pobre y desganado, frío y plano. Si fue realizado por una persona, se trataría de un niño o una niña, o al menos de un adulto sin alma y sin ganas de vivir”. Todas estas manifestaciones son acertando en identificar una IA, pero expresiones similares se encuentran cuando erran diciendo que es de IA el texto humano: “diálogos poco orgánicos”, “sin chicha y poco humano”, “expresiones a veces me parecen rebuscadas y poco naturales” o incluso “diálogo robótico” o “si lo ha hecho una persona resultaría evidente su falta de inspiración y carencias en la comprensión”. También cuando fallan identificando como humano lo que es IA se dan razones similares relacionadas con la naturalidad: “Los diálogos son creíbles, no están excesivamente adornados”, un estilo que confunde hasta hacer manifestar que “la expresión de las oraciones y el modo de ordenar las acotaciones e intervenciones son como las que veo en mis compañeros y hago yo al escribir guiones”.

Aunque la categorización de las observaciones de los participantes ha permitido visualizar distintas características y valorar su relevancia según su presencia relativa, sin duda con una muestra mayor y más diversa podrían alcanzarse características más concretas y fiables. Este estudio constituye un modesto primer paso en la comprensión de cómo las personas perciben y distinguen entre textos generados por IA y aquellos escritos por humanos, aportando para ello un modelo de investigación cualitativa y cuantitativa funcional. Sin embargo, en cuanto a los resultados de tal diseño, se reconoce que el tamaño de la muestra es pequeño, dando un error muestral mejorable. Por ello, se anima a replicar el estudio con una muestra más grande y diversa, así como utilizar una variedad más amplia de textos para la prueba, con el fin de obtener resultados más robustos y generalizables.

Contribuciones de autores

Conceptualización – J.L.R. y E.Q.R.. Curación de datos – J.L.R. y E.Q.R.. Análisis formal – J.L.R. y E.Q.R.. Investigación – J.L.R. y E.Q.R.. Metodología – J.L.R. y E.Q.R.. Redacción – borrador original – J.L.R. y E.Q.R.. Redacción – revisión y edición – J.L.R. y E.Q.R..

Agradecimientos

Los autores quieren expresar su agradecimiento a Marta Núñez Zamorano, Directora de Secretaría Académica de la Universidad del Atlántico Medio, por su ayuda generosa en la logística de esta investigación, así como a las alumnas y alumnos de los tres primeros cursos del Grado en Cine de la Universidad, por su amable y desinteresada colaboración. También el apoyo de Ignacio Luri Rodríguez, de la DePaul University, Chicago.

Aprobación por Comité Ético

Este trabajo de investigación ha sido realizado con la aprobación del Comité Ético de Investigación de la Universidad del Atlántico Medio, con código N.º: CEI/03-002.

Referencias

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

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