

PIXEL BIT

Nº 69 ENERO 2024
CUATRIMESTRAL

e-ISSN:2171-7966
ISSN:1133-8482

Revista de Medios y Educación





PIXEL-BIT

REVISTA DE MEDIOS Y EDUCACIÓN

Nº 69 - ENERO- 2024

<https://revistapixelbit.com>



EDITORIAL
UNIVERSIDAD DE SEVILLA

EQUIPO EDITORIAL (EDITORIAL BOARD)

EDITOR JEFE (EDITOR IN CHIEF)

Dr. Julio Cabero Almenara, Departamento de Didáctica y Organización Educativa, Facultad de CC de la Educación, Director del Grupo de Investigación Didáctica. Universidad de Sevilla (España)

EDITOR ADJUNTO (ASSISTANT EDITOR)

Dr. Juan Jesús Gutiérrez Castillo, Departamento de Didáctica y Organización Educativa. Facultad de CC de la Educación, Universidad de Sevilla (España)

Dr. Óscar M. Gallego Pérez, Grupo de Investigación Didáctica, Universidad de Sevilla (España)

EDITORES ASOCIADOS

Dra. Urtza Garay Ruiz, Universidad del País Vasco. (España)

Dra. Ivanovna Milqueya Cruz Pichardo, Pontificia Universidad Católica Madre y Maestra. (República Dominicana)

Dra. Carmen Llorente Cejudo, Universidad de Sevilla (España)

CONSEJO METODOLÓGICO

Dr. José González Such, Universidad de Valencia (España)

Dr. Antonio Matas Terrón, Universidad de Málaga (España)

Dra. Cynthia Martínez-Garrido, Universidad Autónoma de Madrid (España)

Dr. Luis Carro Sancristóbal, Universidad de Valladolid (España)

Dra. Nina Hidalgo Farran, Universidad Autónoma de Madrid (España)

CONSEJO DE REDACCIÓN

Dra. María Puig Gutiérrez, Universidad de Sevilla. (España)

Dra. Sandra Martínez Pérez, Universidad de Barcelona (España)

Dr. Selín Carrasco, Universidad de La Punta (Argentina)

Dr. Jackson Collares, Universidades Federal do Amazonas (Brasil)

Dra. Kitty Gaona, Universidad Autónoma de Asunción (Paraguay)

Dr. Vito José de Jesús Carioca. Instituto Politécnico de Beja Ciencias da Educação (Portugal)

Dra. Elvira Esther Navas, Universidad Metropolitana de Venezuela (Venezuela)

Dr. Angel Puentes Puente, Pontificia Universidad Católica Madre y Maestra. Santo Domingo (República Dominicana)

Dr. Fabrizio Manuel Sirignano, Università degli Studi Suor Orsola Benincasa (Italia)

Dra. Sonia Aguilar Gavira. Universidad de Cádiz (España)

Dra. Eloisa Reche Urbano. Universidad de Córdoba (España)

CONSEJO TÉCNICO

Dra. Raquel Barragán Sánchez, Grupo de Investigación Didáctica, Universidad de Sevilla (España)

Dr. Antonio Palacios Rodríguez, Grupo de Investigación Didáctica, Universidad de Sevilla (España)

Dr. Manuel Serrano Hidalgo, Grupo de Investigación Didáctica, Universidad de Sevilla (España)

Diseño de portada: Dña. Lucía Terrones García, Universidad de Sevilla (España)

Revisor/corrector de textos en inglés: Dra. Rubicelia Valencia Ortiz, MacMillan Education (México)

Revisores metodológicos: evaluadores asignados a cada artículo

CONSEJO CIENTÍFICO

Jordi Adell Segura, Universidad Jaume I Castellón (España)

Ignacio Aguaded Gómez, Universidad de Huelva (España)

María Victoria Aguiar Perera, Universidad de Las Palmas de Gran Canaria (España)

Olga María Alegre de la Rosa, Universidad de la Laguna Tenerife (España)

Manuel Área Moreira, Universidad de la Laguna Tenerife (España)

Patricia Ávila Muñoz, Instituto Latinoamericano de Comunicación Educativa (México)

Antonio Bartolomé Pina, Universidad de Barcelona (España)

Angel Manuel Bautista Valencia, Universidad Central de Panamá (Panamá)

Jos Beishuizen, Vrije Universiteit Amsterdam (Holanda)

Florentino Blázquez Entonado, Universidad de Extremadura (España)
 Silvana Calaprince, Università degli studi di Bari (Italia)
 Selín Carrasco, Universidad de La Punta (Argentina)
 Raimundo Carrasco Soto, Universidad de Durango (México)
 Zulma Cataldi, Universidad de Buenos Aires (Argentina)
 Luciano Cecconi, Università degli Studi di Modena (Italia)
 Jean-François Cerisier, Université de Poitiers, Francia
 Jordi Lluís Coiduras Rodríguez, Universidad de Lleida (España)
 Jackson Collares, Universidades Federal do Amazonas (Brasil)
 Enricomaria Corbi, Università degli Studi Suor Orsola Benincasa (Italia)
 Marialaura Cunzio, Università degli Studi Suor Orsola Benincasa (Italia)
 Brigitte Denis, Université de Liège (Bélgica)
 Floriana Falcinelli, Università degli Studi di Perugia (Italia)
 María Cecilia Fonseca Sardi, Universidad Metropolitana de Venezuela (Venezuela)
 Maribel Santos Miranda Pinto, Universidade do Minho (Portugal)
 Kitty Gaona, Universidad Autónoma de Asunción (Paraguay)
 María-Jesús Gallego-Arrufat, Universidad de Granada (España)
 Lorenzo García Aretio, UNED (España)
 Ana García-Valcarcel Muñoz-Repiso, Universidad de Salamanca (España)
 Antonio Bautista García-Vera, Universidad Complutense de Madrid (España)
 José Manuel Gómez y Méndez, Universidad de Sevilla (España)
 Mercedes González Sanmamed, Universidad de La Coruña (España)
 Manuel González-Sicilia Llamas, Universidad Católica San Antonio-Murcia (España)
 António José Meneses Osório, Universidade do Minho (Portugal)
 Carol Halal Orfali, Universidad Tecnológica de Chile INACAP (Chile)
 Mauricio Hernández Ramírez, Universidad Autónoma de Tamaulipas (México)
 Ana Landeta Etxeberria, Universidad a Distancia de Madrid (UDIMA)
 Linda Lavelle, Plymouth Institute of Education (Inglaterra)
 Fernando Leal Ríos, Universidad Autónoma de Tamaulipas (México)
 Paul Lefrere, Cca (UK)
 Carlos Marcelo García, Universidad de Sevilla (España)
 Francois Marchessou, Universidad de Poitiers, París (Francia)
 Francesca Marone, Università degli Studi di Napoli Federico II (Italia)
 Francisco Martínez Sánchez, Universidad de Murcia (España)
 Ivory de Lourdes Mogollón de Lugo, Universidad Central de Venezuela (Venezuela)
 Angela Muschitiello, Università degli studi di Bari (Italia)
 Margherita Musello, Università degli Studi Suor Orsola Benincasa (Italia)
 Elvira Esther Navas, Universidad Metropolitana de Venezuela (Venezuela)
 Trinidad Núñez Domínguez, Universidad de Sevilla (España)
 James O'Higgins, de la Universidad de Dublín (UK)
 José Antonio Ortega Carrillo, Universidad de Granada (España)
 Gabriela Padilla, Universidad Autónoma de Tamaulipas (México)
 Ramón Pérez Pérez, Universidad de Oviedo (España)
 Angel Puentes Puente, Pontificia Universidad Católica Madre y Maestra. Santo Domingo (República Dominicana)
 Juan Jesús Gutiérrez Castillo, Universidad de Sevilla (España)
 Julio Manuel Barroso Osuna, Universidad de Sevilla (España)
 Rosalía Romero Tena, Universidad de Sevilla (España)
 Hommy Rosario, Universidad de Carabobo (Venezuela)
 Pier Giuseppe Rossi, Università di Macerata (Italia)
 Jesús Salinas Ibáñez, Universidad Islas Baleares (España)
 Yamile Sandoval Romero, Universidad de Santiago de Cali (Colombia)
 Albert Sangrà Morer, Universidad Oberta de Catalunya (España)
 Ángel Sanmartín Alonso, Universidad de Valencia (España)
 Horacio Santángelo, Universidad Tecnológica Nacional (Argentina)
 Francisco Solá Cabrera, Universidad de Sevilla (España)
 Jan Frick, Stavanger University (Noruega)
 Karl Steffens, Universidad de Colonia (Alemania)
 Seppo Tella, Helsinki University (Finlandia)
 Hanne Wachter Kjaergaard, Aarhus University (Dinamarca)



FACTOR DE IMPACTO (IMPACT FACTOR)

SCOPUS Q1 Education: Posición 236 de 1406 (83% Percentil). CiteScore Tracker 2022: 5 - Journal Citation Indicator (JCI). Emerging Sources Citation Index (ESCI). Categoría: Education & Educational Research. Posición 257 de 739. Cuartil Q2 (Percentil: 65.29) - FECYT: Ciencias de la Educación. Cuartil 1. Posición 16. Puntuación: 35,68- DIALNET MÉTRICAS (Factor impacto 2021: 1.72. Q1 Educación. Posición 12 de 228) - REDIB Calificación Glogal: 29,102 (71/1.119) Percentil del Factor de Impacto Normalizado: 95,455- ERIH PLUS - Clasificación CIRC: B- Categoría ANEP: B - CARHUS (+2018): B - MIAR (ICDS 2020): 9,9 - Google Scholar (global): h5: 42; Mediana: 42 - Journal Scholar Metric Q2 Educación. Actualización 2016 Posición: 405a de 1,115- Criterios ANECA: 20 de 21 - INDEX COPERNICUS Puntuación ICV 2019: 95.10

Pixel-Bit, Revista de Medios y Educación está indexada entre otras bases en: SCOPUS, Fecyt, DOAJ, Iresie, ISOC (CSIC/CINDOC), DICE, MIAR, IN-RECS, RESH, Ulrich's Periodicals, Catálogo Latindex, Biné-EDUSOL, Dialnet, Redinet, OEI, DOCE, Scribd, Redalyc, Red Iberoamericana de Revistas de Comunicación y Cultura, Gage Cengage Learning, Centro de Documentación del Observatorio de la Infancia en Andalucía. Además de estar presente en portales especializados, Buscadores Científicos y Catálogos de Bibliotecas de reconocido prestigio, y pendiente de evaluación en otras bases de datos.

EDITA (PUBLISHED BY)

Grupo de Investigación Didáctica (HUM-390). Universidad de Sevilla (España). Facultad de Ciencias de la Educación. Departamento de Didáctica y Organización Educativa. C/ Pirotecnia s/n, 41013 Sevilla.
Dirección de correo electrónico: revistapixelbit@us.es . URL: <https://revistapixelbit.com/>
ISSN: 1133-8482; e-ISSN: 2171-7966; Depósito Legal: SE-1725-02
Formato de la revista: 16,5 x 23,0 cm

Los recursos incluidos en Pixel Bit están sujetos a una licencia Creative Commons Attribution-NonCommercial-ShareAlike 4.0 Unported (Reconocimiento-NoComercial-CompartirIgual)(CC BY-NC-SA 4.0), en consecuencia, las acciones, productos y utilidades derivadas de su utilización no podrán generar ningún tipo de lucro y la obra generada sólo podrá distribuirse bajo esta misma licencia. En las obras derivadas deberá, asimismo, hacerse referencia expresa a la fuente y al autor del recurso utilizado.

©2024 Pixel-Bit. No está permitida la reproducción total o parcial por ningún medio de la versión impresa de Pixel-Bit.

- 1.- Competencia digital en labores de Investigación: predictores que influyen en función del tipo de universidad y género del profesorado de Educación Superior** // Digital competence in research work: predictors that have an impact on it according to the type of university and gender of the Higher Education teacher. Francisco David Guillén-Gámez, Melchor Gómez-García, Julio Ruiz-Palmero **7**
- 2.- La percepción de las familias acerca de las tareas digitales** // Families' perceptions about digital homework. M^a Ángeles Hernández-Prados, José Santiago Álvarez-Muñoz, Juan Antonio Gil-Noguera **35**
- 3.- Cuestionando el papel de las tecnologías en la Educación Infantil: brechas y falsas visiones** // Questioning the role of technology in Early Childhood Education: divides and false views. Patricia Digón-Regueiro, Rosa María Méndez-García, M. Mercedes Romero-Rodrigo, Cecilia Verónica Becerra-Brito **63**
- 4.- Uso académico del smartphone en la formación de posgrado: percepción del alumnado en Ecuador** // Academic use of smartphones in postgraduate education: student perception in Ecuador. Alexander López-Padrón, Santiago Mengual-Andrés, Ernesto Andrés Hermann-Acosta **97**
- 5.- Equivalencias entre los indicadores de la herramienta SELFIE y el marco DigCompEdu a partir de la técnica Delphi** // Equivalences between SELFIE indicators and DigCompEdu framework based on Delphi technique. Jacoba Munar-Garau, Jorge Ocejja, Jesús Salinas-Ibáñez **131**
- 6.- Análisis del nivel de pensamiento computacional de los futuros maestros: una propuesta diagnóstica para el diseño de acciones formativas** // Analysis of the level of computational thinking of future teachers: a diagnostic proposal for the design of training actions. Lourdes Villalustre-Martín **169**
- 7.- Una propuesta de evaluación de Recursos Educativos Digitales a través de la metodología fsQCA longitudinal** // A proposal for the evaluation of Digital Educational Resources through the longitudinal fsQCA methodology. Cristina Mendaña-Cuervo, Nieves Remo-Diez, Enrique López-González **195**
- 8.- El uso del modelo GPT de OpenAI para el análisis de textos abiertos en investigación educativa** // Use of GPT-3 for open text analysis in qualitative research. Héctor González-Mayorga, Agustín Rodríguez-Esteban, Javier Vidal **227**
- 9.- Estrategias pedagógicas con TIC en contextos educativos vulnerables: Repercusiones en la pandemia y proyecciones de futuro** // Pedagogical strategies with ICT in vulnerable educational contexts: Impact on the pandemic and future projections. María José Alcalá del Olmo-Fernández, María Jesús Santos-Villalba, José Luis González-Sodis, Juan José Leiva-Olivencia **255**
- 10.- Perfiles de uso problemático de los videojuegos y su influencia en el rendimiento académico y los procesos de toma de decisiones en alumnado universitario** // Profiles of Undergraduates' videogames problematic use: the influence of academic performance and decision-making process. Barreto-Cabrera, Yenía del Sol, Suárez-Perdomo, Arminda, Castilla-Vallejo, José Luis **287**

El uso del modelo GPT de OpenAI para el análisis de textos abiertos en investigación educativa

Using OpenAI's GPT Model to Analyse Open Texts in Educational Research



Dr. Héctor González-Mayorga

Profesor Asociado. Universidad de León, España



Dr. Agustín Rodríguez-Esteban

Profesor Contratado Doctor. Universidad de León, España



Dr. Javier Vidal

Catedrático de Universidad. Universidad de León, España

Recibido: 2023/09/27; **Revisado:** 2023/11/07; **Aceptado:** 2023/12/11; **Online First:** 2023/12/19; **Publicado:** 2024/01/07

RESUMEN

La atribución de significado a los segmentos de información propia del análisis de textos abiertos en investigación cualitativa implica una considerable inversión de tiempo. Las herramientas de Procesamiento de Lenguaje Natural pueden convertirse en un recurso para el investigador cualitativo, en cuanto que sus algoritmos permiten interpretar los textos y codificar cualitativamente de forma mucho más rápida. Esta tarea, sin embargo, requiere probar previamente el nivel de comprensión verbal de estas herramientas. La aparición del modelo GPT-3 de OpenAI supuso un salto cualitativo respecto a modelos de Procesamiento de Lenguaje Natural anteriores. El objetivo fue analizar la capacidad de comprensión verbal de esta herramienta. Se aplicaron las pruebas del índice de comprensión verbal de la batería para la medición del cociente intelectual WAIS-IV. Los resultados de las pruebas de fiabilidad fueron satisfactorios. Respecto al nivel de comprensión verbal, las respuestas situaron a GPT-3 en un percentil superior al 99, comparado con los estándares humanos. Estos resultados demuestran que es posible utilizar este modelo como herramienta para el análisis de textos abiertos, abriendo enormes posibilidades para la investigación cualitativa, aunque su uso debe basarse en una utilización precisa y diseñada ad hoc para cada proceso de análisis.

ABSTRACT

Assigning meaning to segments of information through analysis of open texts in qualitative research requires considerable investment of time. Natural Language Processing tools can be a valuable resource for qualitative researchers, as their algorithms allow for faster, qualitative interpretation of texts. However, this requires testing these tools' levels of verbal comprehension beforehand. The introduction of OpenAI's GPT-3 model has marked a qualitative leap forward compared to previous Natural Language Processing models. The study objective was to analyse this tool's verbal comprehension ability. The tests from the verbal comprehension index of the WAIS-IV IQ battery were applied. The results of the reliability tests were satisfactory. The responses put GPT-3 higher than the 99th percentile of human standards of verbal comprehension. These results demonstrate that it is possible to use this model as a tool to analyse open texts, opening up enormous possibilities for qualitative research, although its use must be based on precise, specific utilization for each analysis process.

PALABRAS CLAVES · KEYWORDS

Inteligencia artificial, procesamiento del lenguaje natural, GPT-3, análisis textual, investigación cualitativa.
Artificial intelligence, natural language processing, GPT-3, text analysis, qualitative research.

1. Introducción

"We can only see a short distance ahead, but we can see plenty there that needs to be done." (Turing, 1950, p. 460).

La investigación cualitativa supone un acercamiento a los fenómenos sociales desde el interior, analizando las propias experiencias de los individuos y sus interacciones sociales (Aspers & Corte, 2019; Cohen et al., 2018; Hammersley, 2013). El análisis de contenido, en cuanto método habitualmente aplicado a datos cualitativos, suele requerir de un proceso previo de categorización de estos datos conforme a patrones y tendencias que se descubren tras la lectura repetida de los mismos (Grodal et al., 2021). Este proceso se lleva a cabo a través de la codificación, entendida como método de organización de los datos a través de la asignación de etiquetas descriptivas a bloques de información que emplea técnicas como la vinculación, la agrupación o la agregación para apoyar la extracción y formulación de significado (Longo, 2020, Williams & Moser, 2019). Tal y como señalan Grodal et al. (2021), a medida que el investigador codifica y categoriza, eleva el nivel conceptual del análisis desde la mera descripción hacia un nivel teórico, más abstracto. La capacidad del investigador para interpretar los datos y presentar los hallazgos con claridad hace que un estudio de investigación cualitativa sea útil. Analizar datos cualitativos supone, por tanto, un proceso complejo que conlleva mucho tiempo, siendo este uno de los motivos por el que los investigadores trabajan normalmente con muestras reducidas (Flick, 2018; Longo, 2020). Sin embargo, los investigadores han llamado la atención sobre la importancia de la cantidad y la calidad de los datos señalando que el volumen de información manejado es clave para respaldar las categorías o temas, comprender su riqueza y complejidad y fortalecer la confiabilidad de los hallazgos (Williams & Morrow, 2009).

Una línea mayoritaria de investigación cualitativa se aplica al análisis de textos abiertos. Hoy en día, una importante proporción de la interacción social humana es generada y transmitida a través de textos en redes de información digital (Cope & Kalantzis, 2015; Mills, 2019). Los datos basados en texto se acumulan en un gran volumen conformando una importante fuente de información para los investigadores cualitativos (Golder & Macy, 2011; Shlomo & Goldstein, 2015). Sin embargo, se convierte en algo imposible para estos profesionales, recopilar y analizar, de forma manual, los cientos de miles de registros que pueden generarse en estas fuentes (Longo, 2020). Las aplicaciones de software de análisis de datos cualitativos asistido por ordenador se han convertido en una herramienta sustancial del proceso de investigación. La mayoría de estas herramientas tienen funciones que permiten al investigador operar con la información, ayudando en la identificación de códigos, dividiendo los datos, almacenando fragmentos de datos por código y organizando estos códigos en una estructura, reduciendo enormemente la cantidad de tiempo que debe invertir el investigador. Además, recientemente han incorporado utilidades para el análisis, basadas en la ocurrencia de palabras (Silver & Lewins, 2014). Sin embargo, estas aplicaciones no son herramientas analíticas. La mayoría de ellas requieren, en la actualidad, que los datos se ingresen en un paquete de procesamiento de texto que el analista todavía tiene que leer y codificar. Esta tarea sigue siendo una de las fases del análisis de datos cualitativos que consume más tiempo (Bail, 2014). Los elementos de ahorro de tiempo de estas herramientas ocurren en una etapa posterior de la búsqueda y recuperación de datos (Cypress, 2019).

Las técnicas de análisis temático asistido por Inteligencia Artificial (Towler et al., 2022), especialmente aquellas basadas en Procesamiento de Lenguaje Natural (Guetterman et al.,

2018), están suponiendo un gran avance al permitir examinar frases, buscar grupos de significado más allá de las palabras, adaptarse para tener en cuenta la jerga y la gramática informal, y determinar el sentimiento (Carriere et al., 2021; Chang et al., 2021; Rezaeenour et al., 2022). Las utilidades en las que más se está desarrollando el Procesamiento de Lenguaje Natural en la actualidad son la Desambiguación del Sentido de las Palabras (Word-Sense Disambiguation), la Teoría de la Argumentación del Texto (Text Argumentation Theory), el Análisis de Sentimientos (Sentiment Analysis), el Modelado de Temas (Topic Modelling) y el Resumen Automático de Textos (Automatic Textual Summarization) (Oussalah, 2022). De especial interés, dado el objetivo de la presente investigación, son los dos últimos. El Modelado de Temas (Topic Modelling) es una técnica de aprendizaje automático no supervisado que es capaz de detectar patrones de palabras y frases dentro de un conjunto de textos y agruparlas automáticamente. Los temas y los parámetros de la mezcla son desconocidos y se infieren únicamente a partir de los datos, donde cada tema está representado por sus palabras más probables. El segundo de ellos, Resumen Automático de Textos (Automatic Textual Summarization), tiene como objetivo resumir el contenido principal de un documento bajo dos procedimientos de síntesis: a) la síntesis extractiva, que utiliza por una selección de frases del documento original considerando la similitud de las frases, la ubicación o la presencia de las palabras clave seleccionadas; y b) la síntesis abstractiva, en la que las frases del resumen pueden ser diferentes a las de los documentos originales.

Como se señaló en las líneas anteriores, el análisis de textos abiertos en la investigación cualitativa es un proceso creativo. El investigador explora y reflexiona sobre el significado de los datos (Grodal et al., 2021). Las técnicas de Resumen Automático de Textos, especialmente las que se desarrollan bajo procedimientos de síntesis abstractiva (Oussalah, 2022), parecen cumplir una función similar. ¿Replican, en este sentido, la tarea 'típicamente' humana de atribuir significado a datos objetivos? Realizar esta tarea supondría 'comprender' a nivel lingüístico y semántico el lenguaje generado por humanos.

Son escasos los estudios de índole metodológico que han comparado los resultados de los procesos de categorización realizados por humanos y la codificación y categorización automática basada en Procesamiento de Lenguaje Natural. Guetterman et al. (2018) diseñaron un estudio comparando los resultados obtenidos por máquinas y humanos, llegando a la conclusión de que, aunque los métodos basados en Procesamiento de Lenguaje Natural identificaron los temas principales de los textos que habían sido encontrados con el análisis cualitativo tradicional, fueron menos eficaces a la hora de identificar los matices. Towler et al. (2022) llevaron a cabo un proceso de análisis temático no supervisado asistido por ordenador con respuestas de texto libre de los usuarios de la intervención de control de infecciones COVID-19 que permitió identificar temas latentes en los textos. Los resultados fueron comparados con un proceso de codificación humana-manual y se observó una gran coincidencia en los temas principales con ligeras diferencias en algunos matices realizados por la codificación manual.

De entre los llamados modelos de lenguaje extenso, entrenados previamente con grandes cantidades de texto sin etiquetar con un objetivo de modelado de lenguaje, destacan los modelos Generative Pre-trained Transformer. La tercera versión de estos modelos (GPT-3), presentada en mayo de 2020 en su versión beta y en julio de ese año en su versión final, fue particularmente popular debido a su flexibilidad y facilidad de uso desde la API de OpenAI (Brown et al., 2020; Si et al., 2022). Los modelos GPT son capaces de

resumir textos, responder a preguntas, escribir ensayos, traducir idiomas e incluso generar código como Python o HTML, lo que ha supuesto un gran avance en las técnicas Procesamiento de Lenguaje Natural al incorporar la Inteligencia Artificial Generativa al análisis del lenguaje natural y ser capaz de crear contenido, en este caso de tipo textual, a través de la aplicación de algoritmos de aprendizaje automático (Bajaj et al., 2022). Utilizan un modelo de lenguaje generativo y autorregresivo de tipo no supervisado, que toma una entrada y, utilizando el campo de la semántica para entender el texto, intenta ofrecer una respuesta de texto coherente y con sentido (Bajaj et al., 2022; Binz & Schulz, 2023). Esta herramienta es 10 veces mayor que otros modelos lingüísticos previos. El término preentrenado indica que estos modelos ya han sido alimentados con todos los datos necesarios para llevar a cabo su tarea generativa (Binz & Schulz, 2023). En concreto, GPT-3 se ha alimentado con unos 570 GB de información textual recopilada rastreando Internet, WebText2. A finales del año 2022, OpenAI lanzó una herramienta específica de este modelo: ChatGPT (<https://openai.com/blog/chatgpt/>), que incorpora una sencilla interfaz web que sigue el modelo imperativo, donde el usuario ingresa una solicitud y el chatbot devuelve resultados, todo en un contexto multilingüe (García-Peñalvo, 2023).

Tal y como apuntan Xie et al. (2022), en la actualidad no existe una métrica establecida para medir la calidad de los resultados generados por los diversos métodos de interpretación del lenguaje natural existentes. Este es el hecho que motiva la presente investigación, que tiene como objetivo analizar la capacidad de comprensión verbal (escrita) de estas máquinas y su uso para tareas de análisis de textos abiertos en investigación cualitativa.

2. Metodología

De forma similar a lo realizado en experimentos previos (Dobrev, 2019), hemos utilizado una prueba psicológica de comprensión verbal para humanos con el fin de determinar el nivel del modelo text-davinci-003 de GPT-3 (a partir de aquí, nos referiremos a él como GPT-3) comparado con un estándar en humanos. Tal y como señala Dobrev (2019), todo lo que se necesita para identificar los programas informáticos que son IA es someterlos a una prueba y otorgarles esta denominación a aquellos que la superan. Es importante que aclaremos que no nos planteamos aquí la pregunta de Turing: ¿pueden pensar las máquinas? (Turing, 1950). Nuestra pregunta es diferente: ¿puede una máquina entender una pregunta expresada con lenguaje natural humano? Utilizaremos la estrategia de Turing afirmando que, si la respuesta de la máquina a nuestra pregunta es similar a la de un ser humano, entonces podremos afirmar que la máquina tiene la función de la comprensión verbal similar a la función cognitiva de los humanos. En el presente estudio cuestionamos a la máquina con un conjunto de preguntas preestablecidas que conforman el índice de Comprensión Verbal de la conocida batería para la medición del cociente intelectual (CI): WAIS-IV. La utilización de instrumentos de medición de la inteligencia o, más concretamente, de medición CI para evaluar el rendimiento de las máquinas de IA no es nueva (Dowe & Hernández-Orallo, 2012). En el año 1998 se publicó el primer artículo que introdujo el test de CI para evaluar la IA (Hernández-Orallo & Minaya-Collado, 1998). Detterman (2011) propuso poner a prueba los programas informáticos mediante test de inteligencia para personas. Wang et al. (2016) se plantearon cómo crear un programa que pueda resolver test de inteligencia diseñados para humanos. Recientemente, Binz & Schulz

(2023) simularon el comportamiento de la mente humana en GPT-3 con distintas tareas propias de la psicología cognitiva: a) experimentos basados en tareas, y b) experimentos basados en viñetas. En el segundo caso, se demostró que GPT-3 pudo resolver problemas desafiantes basados en viñetas, aunque su comportamiento dependía en gran medida de cómo se presenten las viñetas.

2.1. Instrumento de evaluación: WAIS-IV

La Escala de Inteligencia de Wechsler para Adultos (WAIS) es un instrumento estandarizado diseñado para medir el cociente intelectual de personas de edades comprendidas entre los 16 y los 89 años. La cuarta y, hasta la fecha, última edición del instrumento (WAIS-IV) está compuesta de quince pruebas (diez principales y cinco opcionales) mediante las cuales se pueden calcular, además del cociente intelectual, cuatro índices: comprensión verbal, razonamiento perceptivo, memoria de trabajo y velocidad de procesamiento. Para el presente trabajo se ha empleado el índice de comprensión verbal (ICV) de la versión española del WAIS-IV (Wechsler, 2012a, 2012b), el cual está formado por tres pruebas principales: la prueba de semejanzas (S), la prueba de vocabulario (V) y la prueba de información (I); y una opcional, la prueba de comprensión (CO). Las respuestas a los ítems de las pruebas de semejanzas, vocabulario y comprensión se valoran con 2 puntos, 1 punto o 0 puntos, en función de las respuestas del individuo; mientras que las respuestas en la prueba de conocimientos son valoradas con una puntuación de 1 o 0. En algunas respuestas imprecisas o poco claras, cabe la posibilidad de solicitar más información al sujeto para poder considerarla correcta o incorrecta de una forma más precisa (Wechsler, 2012a). Se han seguido las normas de aplicación de cada una de las cuatro pruebas respecto al punto de comienzo, regla de retorno y regla de terminación.

Una vez puntuados los ítems aplicados, se ha calculado la puntuación directa (PD) de cada prueba y las correspondientes puntuaciones escalares (PE) para cada rango de edad de acuerdo con el manual de aplicación y corrección (Wechsler, 2012a, pp. 192-201). Con ellas, se obtiene el valor del ICV (Wechsler, 2012a, p. 203). Además, este baremo proporciona información sobre el percentil en el que se sitúa el sujeto.

En el presente trabajo se ha calculado el ICV con todas las combinaciones posibles de las pruebas para cada uno de los rangos de edad.

2.2. Instrumento de análisis: el modelo text-davinci-003 de GPT-3

Se utiliza OpenAI Playground, una interfaz basada en web de GPT-3. Se ha elegido el ajuste preestablecido de «Preguntas y respuestas» (Q&A preset). Este sistema mantiene el contexto de la conversación, lo que hace que esta interfaz sea mucho más conversacional que otras (Bajaj et al., 2022).

El prompt se ha definido con el texto literal que recoge el WAIS para cada una de las pruebas. No se proporciona entrenamiento adicional, al margen de los ítems de entrenamiento que describe el WAIS. El modelo permite decidir sobre varios parámetros. Respecto los parámetros de response duration, best of, frequency penalty y presence penalty, se han utilizado los valores por defecto (Bajaj et al., 2022; Binz & Schulz, 2023).

Los dos más relevantes para el objetivo del estudio son temperature y Top_P. Ambos están relacionados con la creatividad de la respuesta. Sus valores posibles oscilan entre 0 y 1. Dado que nuestro objetivo es utilizar la máquina como herramienta de diagnóstico, es imprescindible que la máquina sea fiable, es decir, que de un resultado idéntico cada vez que se realice la misma pregunta. Por ello, se decidió optar por la estrategia más conservadora fijando ambos parámetros en 0, limitando todo lo que la máquina permite su posible creatividad (Binz & Schulz, 2023).

2.3. Procedimiento

El procedimiento de análisis se llevó a cabo en dos fases.

En la primera fase, se escribió en el playground la consigna que establece el manual de aplicación del WAIS-IV para la prueba de semejanzas (S) («Ahora le voy a decir dos palabras y a preguntarle en qué se parecen») seguida de la pregunta correspondiente al ítem de ejemplo: «¿En qué se parecen dos y siete?». Seguidamente, se introdujo un salto de línea o intro. Finalmente, se hizo clic en el botón submit, obteniendo el resultado correspondiente al primer ítem. Según las instrucciones del WAIS-IV, en función de las respuestas obtenidas, se deben realizar preguntas de aclaración o profundización (P) mediante la consigna «¿Puede decirme algo más?» para otorgar la puntuación correspondiente. Es importante señalar que el modelo text-davinci-003 contesta a la segunda pregunta teniendo en cuenta el contexto de la primera y su propia respuesta, por lo que el procedimiento es idéntico al previsto en la prueba para humanos.

Para describir la fiabilidad de las respuestas, se replicó tres veces la misma acción, analizando el grado de variación en las respuestas. Asimismo, se comprobó que los resultados eran idénticos si esta prueba se realizaba desde equipos e IPs diferentes.

En la segunda fase, se procedió a replicar el proceso de la primera fase con las siguientes variaciones en el prompt del playground:

- Dos saltos de línea después del ítem y de la pregunta de aclaración.
- Dos saltos de línea después del ítem y de la pregunta de aclaración y escribiendo en letras mayúsculas los conceptos clave a comparar.
- Dos saltos de línea después del ítem y de la pregunta de aclaración y escribiendo la totalidad del texto en letras mayúsculas.

Estas pruebas nos han permitido determinar que el procedimiento que obtuvo una mayor puntuación según lo establecido en el manual de aplicación y corrección del WAIS fue aplicar dos saltos de línea después de la pregunta inicial y de la pregunta de aclaración, escribiendo la totalidad del texto en letras mayúsculas. Con este formato se realizaron el resto de las pruebas del índice de comprensión verbal del WAIS-IV de las que se ofrecen los resultados. No obstante, las escasas diferencias encontradas en las puntuaciones permiten afirmar que escribir las preguntas en minúsculas e incluir un único salto de línea ofrece resultados de fiabilidad y validez suficientemente altos.

3. Análisis y resultados

3.1. Fiabilidad

Se puede afirmar que la fiabilidad del modelo text-davinci-003, utilizando el procedimiento y parámetros descritos en el apartado anterior es del 100%. Aunque con los resultados iniciales puedan parecer menores, tal y como se puede comprobar, se producen pequeñas variaciones de expresión en los resultados que no afectan al significado.

En el 92.59% de los casos, la respuesta dada por el modelo fue idéntica para cada opción probada y sus posteriores réplicas. En las pruebas de información (I) y comprensión (CO), la fiabilidad fue del 100%. En la prueba de semejanzas, el porcentaje de fiabilidad fue del 86.66% (13 de 15 ítems aplicados), mientras que en la prueba de vocabulario (V) se situó en 84.61% (22 de 26 ítems aplicados). En la Figura 1 se presenta un ejemplo de ítem de la prueba de vocabulario (V) en la que, siguiendo el mismo procedimiento, se obtienen respuestas diferentes. En este ejemplo, se puede comprobar cómo son expresiones no idénticas con idéntico significado, lo que nos lleva a la afirmación sobre la fiabilidad total del modelo.

Figura 1

Aplicación y respuesta (y posterior réplica) del ítem 13 de vocabulario

The figure displays two screenshots of the OpenAI API interface, illustrating the model's responses to a specific prompt. The prompt is: "DÍGAME QUÉ QUIERE DECIR CADA PALABRA. MEDITAR".

Top Screenshot (Initial Response):

- Prompt:** DÍGAME QUÉ QUIERE DECIR CADA PALABRA. MEDITAR
- Response:** Meditar es una práctica mental que implica una concentración profunda y una atención consciente a los pensamientos, sentimientos y sensaciones. Esta práctica se utiliza para relajarse, reducir el estrés, mejorar la concentración y aumentar la conciencia de uno mismo.

Bottom Screenshot (Replica Response):

- Prompt:** DÍGAME QUÉ QUIERE DECIR CADA PALABRA. MEDITAR
- Response:** Meditar es una práctica mental que implica una concentración profunda y una atención consciente a los pensamientos, sentimientos y sensaciones. Se trata de una forma de relajación profunda que puede ayudar a reducir el estrés, mejorar la concentración y aumentar la claridad mental.

Both screenshots show the same settings on the right side of the interface:

- Mode:** ChatGPT
- Model:** text-davinci-003
- Temperature:** 0
- Maximum length:** 256
- Stop sequences:** Enter sequence and press Tab
- Top P:** 0

3.2. Nivel de comprensión verbal

En la Tabla 1 se pueden observar las PD y las PE obtenidas por el modelo text-davinci-003 en cada una de las cuatro pruebas del ICV.

Tabla 1

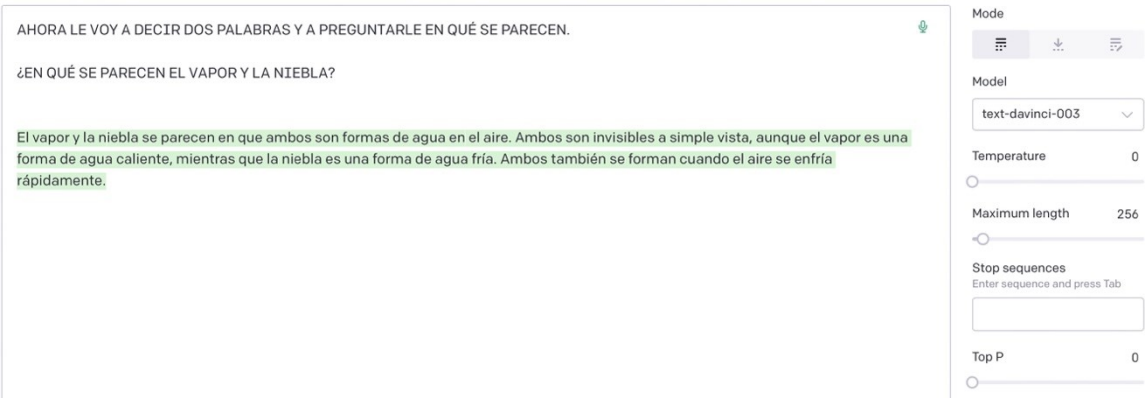
PD y PE por rangos de edad

Prueba	PD	PE por rangos de edad									
		16-17	18-19	20-24	25-34	35-44	45-54	55-69	70-79	80-84	85-89
Semejanzas (S)	30	18	18	17	17	17	17	19	19	19	19
Vocabulario (V)	52	19	19	19	19	19	19	19	19	19	19
Información (I)	24	19	19	18	17	16	16	17	19	19	19
Comprensión (CO)	36	19	19	19	19	19	19	19	19	19	19

La PD obtenida en la prueba de semejanzas (S) es de 30 puntos sobre 36 posibles. El modelo alcanza la máxima puntuación (2) en catorce de los dieciocho ítems puntuables, mientras que en dos ítems la puntuación es de 1 y en dos ítems es de 0. A modo de ejemplo, se puede observar en la Figura 2 la aplicación del ítem 10 de la prueba, a cuya respuesta se le otorga la puntuación máxima de 2 puntos.

Figura 2

Aplicación y respuesta del ítem 10 de semejanzas



En la prueba de vocabulario (V), obtiene una puntuación de 52 sobre 57 puntos, alcanzando la máxima puntuación en 27 de los 30 ítems; mientras que en un ítem obtiene una puntuación de 1 y en dos ítems una puntuación de 0. En la Figura 3 se puede observar el ítem 11, en el que se obtiene la puntuación intermedia (1) y cuya respuesta requirió de una pregunta de aclaración.

Figura 3

Aplicación y respuesta del ítem 11 de vocabulario

DÍGAME QUÉ QUIERE DECIR CADA PALABRA.

COMPASIÓN

Compasión es un sentimiento de profunda empatía y comprensión hacia el sufrimiento de otra persona. Se trata de una emoción que nos lleva a querer ayudar a alguien que está pasando por un momento difícil.

¿PUEDE EXPLICÁRMELO UN POCO MÁS?

La compasión es una emoción que nos lleva a sentir empatía y comprensión hacia el sufrimiento de otra persona. Esto significa que nos preocupamos por el bienestar de los demás y queremos ayudarles a superar sus dificultades. La compasión nos motiva a actuar de manera amable y compasiva hacia los demás, incluso cuando no estamos de acuerdo con sus opiniones o acciones. Esto nos ayuda a construir relaciones más fuertes y saludables con los demás.

Mode

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 0

En la prueba de información (I), el modelo obtuvo una puntuación de 24 sobre 26 puntos posibles, fallando en los ítems 14 («¿Por qué es famoso Fleming?») y 23 («¿Cuál es el órgano del cuerpo humano más grande?»). Sin embargo, en ambos casos, los autores del presente trabajo, consideramos que las respuestas deben considerarse válidas, lo que elevaría la puntuación a 26 sobre 26. En el caso del ítem 14, no se especifica que la pregunta se refiera a Alexander Fleming, descubridor de la penicilina, por lo que la respuesta del modelo hace referencia a Ian Fleming, creador de la saga de novelas de James Bond (ver Figura 4). En el caso del ítem 23, la respuesta del modelo interpretó como “grande” el órgano más pesado del cuerpo, el hígado, y no el más extenso, la piel, que es la respuesta correcta según el manual de aplicación y corrección del instrumento. En ambos casos, se realizó una prueba posterior añadiendo el nombre de Alexander en el ítem 14 («¿Por qué es famoso Alexander Fleming?») y sustituyendo el adjetivo ‘grande’ por ‘extenso’ en el ítem 23 («¿Cuál es el órgano del cuerpo humano más extenso?»), obteniendo el modelo la máxima puntuación en ambos casos.

Figura 4

Aplicación y respuesta del ítem 14 de información

RESPONDA A LAS SIGUIENTES PREGUNTAS.

¿POR QUÉ ES FAMOSO FLEMING?

Fleming es famoso por ser el creador de la famosa serie de novelas de espionaje James Bond. Estas novelas han sido llevadas a la pantalla grande en varias ocasiones, lo que ha contribuido a la fama de Fleming.

Mode

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 0

Por último, en la prueba optativa de comprensión (CO), el modelo alcanzó la máxima puntuación en todos los ítems, si bien en el ítem 16 (¿Por qué cuesta más un terreno en la ciudad que en el campo?) se necesitó la pregunta de aclaración para obtener los 2 puntos.

Una vez recogidas las PD y convertidas a PE, se obtuvo la suma de las PE para cada uno de los rangos de edad, teniendo en cuenta las cuatro combinaciones de pruebas posibles para calcular la puntuación compuesta del ICV, así como el percentil en el que se sitúa el modelo. En la Tabla 2, se puede observar que el ICV del modelo text-davinci-003 se sitúa entre 143 (puntuación únicamente obtenida en los rangos de edad de 35-44 años y 45-54 años) y los 150 (máxima puntuación posible), por lo que el modelo obtiene la clasificación considerada de ‘muy superior’ (Wechsler, 2012b). En todos los casos, el modelo se sitúa en el percentil 99.8 o superior, alcanzando un percentil mayor de 99.9 en la totalidad de las combinaciones de pruebas en los intervalos de edad de 16 a 19 años y de 55 a 89 años, lo que quiere decir que tiene un nivel de comprensión verbal superior al 99.9% de los humanos. Asimismo, el modelo obtiene las máximas puntuaciones compuestas de ICV (y, por lo tanto, se sitúa en un percentil superior a 99.9) en la combinación de las pruebas de semejanzas (S), vocabulario (V) y comprensión (CO) en todos los rangos de edad.

Tabla 2

Conversión de la suma de las PE a ICV y sus percentiles

Rangos de edad	Pruebas	Suma PE	ICV	Percentiles
16-17 años 18-19 años	S, V, I	56	150	>99.9
	S, V, CO	56	150	>99.9
	S, I, CO	56	150	>99.9
	V, I, CO	57	150	>99.9
20-24 años	S, V, I	54	147	99.9
	S, V, CO	55	150	>99.9
	S, I, CO	54	147	99.9
	V, I, CO	56	150	>99.9
25-34 años	S, V, I	53	145	99.9
	S, V, CO	55	150	>99.9
	S, I, CO	53	145	99.9
	V, I, CO	55	150	>99.9
35-44 años 45-54 años	S, V, I	52	143	99.8
	S, V, CO	55	150	>99.9
	S, I, CO	52	143	99.8
	V, I, CO	54	147	99.9
55-69 años	S, V, I	55	150	>99.9
	S, V, CO	57	150	>99.9
	S, I, CO	55	150	>99.9
	V, I, CO	55	150	>99.9
70-79 años 80-84 años 85-89 años	S, V, I	57	150	>99.9
	S, V, CO	57	150	>99.9
	S, I, CO	57	150	>99.9
	V, I, CO	57	150	>99.9

4. Discusión y conclusiones

Parafraseando a Turing (1950), el modelo text-davinci-003 del GPT-3 entiende las preguntas expresadas en lenguaje natural y lo sabemos por el análisis de las respuestas que nos da. De acuerdo con las pruebas utilizadas, ha demostrado tener un nivel de comprensión verbal máximo comparado con los estándares humanos. La fiabilidad y validez de los resultados obtenidos demuestran que es posible utilizar este modelo de IA como herramienta para el análisis de textos. Es decir, es posible codificar textos de forma válida y fiable. Sin embargo, su flexibilidad, deseable en su función más creativa, debe ser limitada al máximo en su función analítica. Para ello es imprescindible tener en cuenta los siguientes resultados.

En primer lugar, GPT-3 analiza todo el texto que le proporcionas cada vez y mantiene en su memoria el contexto de toda la sesión de preguntas. Por ello, para el análisis de textos independientes deben introducirse uno a uno para obtener el resultado sin riesgo de sesgos de contexto. Si le haces dos preguntas a la vez, la respuesta a cada una de ellas estará afectada por la otra pregunta que se ha hecho.

En segundo lugar, pequeñas variaciones en la forma de preguntar provocan diferentes respuestas. Lo hemos comprobado al usar mayúsculas y minúsculas, o introducir diferentes saltos de línea después de la pregunta. En los resultados ofrecemos la opción que mejores puntuaciones han obtenido, pero lo que debe inferirse aquí no es mantener estas opciones, sino que siempre deben hacerse pruebas hasta encontrar el prompt adecuado al tipo de análisis que quiera hacerse y mantenerlo en todo el análisis para que ofrezca la fiabilidad y validez exigible en cada estudio. Se ha comprobado que, si se utiliza un mismo prompt con los parámetros de temperature y Top P a cero, en línea con lo realizado en estudios similares (Binz & Schultz, 2023), la máquina ofrece una fiabilidad del 100%.

En tercer lugar, el nivel de comprensión verbal es superior al de los humanos. Sin embargo, su contexto se limita a lo aprendido en los entrenamientos y al prompt que se utilice. Esto puede provocar lo que podemos considerar fallos de interpretación. Hemos dado el ejemplo de su respuesta a quién es Fleming. Por ello, es necesario ofrecer un contexto suficiente para que la máquina identifique el ámbito en el que se debe encontrar su respuesta buscada (Guetterman et al., 2018).

En cuarto lugar, la máquina ha sido entrenada para tener una funcionalidad de conversación, véase al respecto el estudio comparativo realizado por Tack y Piech (2022) sobre la habilidad conversacional de esta máquina en comparación con un maestro, y no para tener conocimiento preciso de diversos campos. Sin embargo, los textos utilizados han proporcionado conocimiento suficiente para obtener máximas puntuaciones en las pruebas de información. Esto implica que su grado de cultura general, por así llamarlo, también es superior a la mayoría de los humanos. Una vez más, debe insistirse en que debe ser controlada esta funcionalidad dado que, al igual que con los humanos, puede ser una posible fuente de sesgo.

En quinto lugar, sabemos que el nivel de comprensión verbal es diferente en función de la edad. Sin embargo, hemos comprobado que dicho nivel es máximo para un amplio rango de edad. En nuestro caso, ofrecemos los datos para edades entre 16 y 89 años (los establecidos por la prueba utilizada). Esto permite utilizar la máquina independientemente de la edad de los autores de los textos. Debe ampliarse este estudio para comprobar si mantiene su nivel de comprensión en textos elaborados por sujetos hasta los 16 años.

Por último, mencionaremos que no disponemos de evidencias que permitan realizar afirmaciones sobre los resultados en otras lenguas, pero es muy plausible la hipótesis de que aquellas lenguas de las que se hayan utilizado una cantidad similar de textos a la española tengan resultados similares.

En síntesis, el objetivo de este estudio es analizar la capacidad de comprensión verbal (escrita) del modelo text-davinci-003 de GPT-3 y su uso para tareas de análisis de textos abiertos en investigación cualitativa. Como conclusión, podemos afirmar que el nivel de comprensión verbal establecido en las pruebas utilizadas y su fiabilidad permite su uso como herramienta para el análisis de textos abiertos y abre enormes posibilidades a su uso en este tipo de investigación.

Este uso debe basarse en una utilización precisa y diseñada ad hoc para cada proceso de análisis: hay que elegir el corpus a analizar y los procedimientos de análisis de la misma manera que lo hacemos para procedimientos no asistidos por máquinas. Son posibles estrategias deductivas (con sistemas de categorías previamente definido) o inductivas (estableciendo categorías sobre las agrupaciones de textos realizadas por las máquinas), pero todas deben cumplir las mismas garantías metodológicas que en análisis no asistidos.

El hecho de que la máquina ofrezca sus respuestas utilizando elementos de contexto hace, al igual que sucede con un analista humano, que estos elementos actúen como aquellos que aumenten la precisión y también como elementos que provoquen sesgos. La forma de preguntar condiciona la respuesta: pero este problema no es específico de las IA. Por ello, todos los sistemas de control de calidad, fiabilidad y validez (Williams & Morrow, 2009) de los análisis habitualmente utilizados en investigación cualitativa deben seguir formando parte del método de estudios asistidos por las máquinas.

Mientras que, hasta ahora, en el análisis de textos abiertos se ha trabajado con corpus con unos pocos decenas o cientos de documentos, actualmente podemos plantearnos miles o cientos de miles (Longo, 2020; Mills, 2019). Es un salto muy relevante, dado que una de las principales críticas a los estudios cualitativos siempre ha sido la calidad de sus muestras reducidas y sus limitaciones para generalizar las conclusiones.

Las aplicaciones de esta herramienta de análisis para las ciencias sociales son enormes, desde el análisis de respuestas abiertas en encuestas o entrevistas, la identificación de temas tratados en redes sociales o la utilización como un juez más para analizar la consistencia interna de procesos de codificación.

Somos conscientes de los riesgos de utilizar estas IA para la toma de decisiones. Por ello, nuestro enfoque es utilizarlas como una herramienta más de análisis, pero manteniendo todas las garantías que la investigación en ciencias sociales ha ido depurando a lo largo de años. Los riesgos no deben frenar su uso, pero deben maximizar los sistemas de control de calidad de sus resultados. Esta tecnología nos obliga, como afirmaba Turing, a recorrer múltiples opciones, que deben ser probadas, no asumidas sin crítica (García-Peñalvo et al., 2023).

Si superamos la fascinación inicial de lo creativas que son estas máquinas y las cosas tan asombrosas que nos responden, podremos centrarnos en usar una de sus funcionalidades específicas (la comprensión de textos) para aumentar nuestra capacidad de análisis y de explicación de los fenómenos sociales, lo que permitirá encontrar, de forma más ajustada y rápida, las soluciones útiles que necesitamos.

5. Financiación

Proyecto titulado «El uso de la IA en el análisis de redes informales sociales para la orientación en Educación Superior.» (Ref. PID2021-125405NB-I00), dentro de la convocatoria de Proyectos de Generación de Conocimiento (BOE, 13 de noviembre de 2022) en el marco del Programa Estatal para Impulsar la Investigación Científico-Técnica y su Transferencia del Plan Estatal de Investigación Científica, Técnica y de Innovación 2021-2023 (Identificador de la convocatoria en la Base de Datos Nacional de Subvenciones ID BDNS: 590214).

Using OpenAI's GPT Model to Analyse Open Texts in Educational Research

1. Introduction

"We can only see a short distance ahead, but we can see plenty there that needs to be done." (Turing, 1950, p. 460).

Qualitative research involves an approach to social phenomena from the inside, analysing individuals' experiences and social interactions (Aspers & Corte, 2019; Cohen et al., 2018; Hammersley, 2013). Content analysis, a method commonly applied to qualitative data, typically requires preliminary categorisation of data according to patterns and trends discovered after repeated reading (Grodal et al., 2021). This process is done via coding—which is a method of organising data by assigning descriptive labels to blocks of information—using techniques such as linking, grouping, or aggregating to support the extraction and formulation of meaning (Longo, 2020; Williams & Moser, 2019). As Grodal et al. (2021) noted, as the researcher codes and categorizes, they elevate the conceptual level of the analysis from mere description to a more abstract or theoretical level. The researcher's ability to interpret the data and present their findings clearly makes qualitative research helpful. Analysing qualitative data is a time-consuming, complex process, which is one reason why researchers usually work with small samples (Flick, 2018; Longo, 2020). However, researchers have highlighted the importance of the quantity and quality of data, noting that the volume of information handled is critical to supporting categories or themes, understanding their richness and complexity, and strengthening the reliability of the findings (Williams & Morrow, 2009).

One common strategy in qualitative research is analysing open texts. Nowadays, a significant proportion of human social interaction is generated and transmitted through texts in digital information networks (Cope & Kalantzis, 2015; Mills, 2019). Text-based data accumulates in large volumes, forming a significant source of information for qualitative researchers (Golder & Macy, 2011; Shlomo & Goldstein, 2015). However, it is impossible for qualitative researchers to manually collect and analyse the hundreds of thousands of records that can be produced from these sources (Longo, 2020). Computer-assisted qualitative data analysis software has become a valuable tool in the research process. Most of these tools have functions that allow researchers to operate on information, aiding in identifying codes, dividing the data, storing data snippets by code, and organising these codes into a structure, significantly reducing the time researchers need to spend. Moreover, they have recently incorporated utilities for analysis based on word occurrence (Silver & Lewins, 2014). However, these applications are not analytical tools. Most require the data to be entered into a word processing package that the analyst must still read and code. This task remains one of the most time-consuming phases of qualitative data analysis (Bail, 2014) and the time-saving elements these tools offer come later, in data search and retrieval (Cypress, 2019).

Artificial Intelligence-assisted thematic analysis techniques (Towler et al., 2022), especially those based on Natural Language Processing (Guetterman et al., 2018), are

making significant advances by enabling examination of phrases, searching for meaning groups beyond words, adapting to account for slang and informal grammar, and determining sentiment (Carriere et al., 2021; Chang et al., 2021; Rezaeenour et al., 2022). The utilities where Natural Language Processing is currently most developed are Word-Sense Disambiguation, Text Argumentation Theory, Sentiment Analysis, Topic Modelling, and Automatic Textual Summarization (Oussalah, 2022). Given the goal of our research, the last two are of particular interest. Topic Modelling is an unsupervised machine learning technique capable of detecting patterns of words and phrases within a set of texts and automatically grouping them. The mix's topics and parameters are unknown and inferred solely from the data, where each topic is represented by its most probable words. The second, Automatic Textual Summarization, aims to summarise the main content of a document using two synthesis procedures: a) extractive synthesis, which uses a selection of phrases from the original document considering the similarity of the phrases, location, or presence of selected keywords; and b) abstractive synthesis, where the summary phrases may be different from those in the original documents.

As mentioned above, analysing open texts in qualitative research is a creative process. The researcher explores and reflects on the meaning of the data (Grodal et al., 2021). Automatic Textual Summarisation techniques fulfil a similar function, especially those which use abstractive synthesis procedures (Oussalah, 2022). In this sense, do they replicate the 'typically' human task of attributing meaning to objective data? To perform this task would mean to 'understand' the language generated by humans at a linguistic and semantic level.

Few methodological studies have compared the results of human categorisation processes and automatic coding and categorisation based on Natural Language Processing. Guetterman et al. (2018) designed a study comparing the results obtained by machines and humans, concluding that although Natural Language Processing-based methods identified the main themes of the texts found with traditional qualitative analysis, they were less effective in identifying nuances. Towler et al. (2022) conducted a computer-assisted unsupervised thematic analysis process with free-text responses from users of the COVID-19 infection control intervention, which allowed identification of latent themes in the texts. The results were compared with a human-manual coding process, and there was good agreement in the main themes, with slight differences in some nuances made by manual coding.

Generative Pre-trained Transformer models stand out among many language models—pre-trained with large amounts of unlabelled text for language modelling purposes. The third version of these models (GPT-3), the beta version of which was presented in May 2020 and the final version in July the same year, was prevalent due to its flexibility and ease of use through OpenAI's API (Brown et al., 2020; Si et al., 2022). GPT models can summarise texts, answer questions, write essays, translate languages, and even generate code like Python or HTML. This represents a significant advance in Natural Language Processing techniques by incorporating Generative Artificial Intelligence into the analysis of natural language and being able to create content—in this case textual—through the application of machine learning algorithms (Bajaj et al., 2022). They use an unsupervised, generative, autoregressive language model, which takes input and, using the field of semantics to understand the text, tries to offer a coherent, meaningful text response (Bajaj et al., 2022; Binz & Schulz, 2023). This tool is ten times larger than previous linguistic models. The term pre-trained indicates that these models have already been fed all the necessary data for

their generative task (Binz & Schulz, 2023). Specifically, GPT-3 has been fed about 570 GB of textual information collected by scanning the Internet, WebText2. At the end of 2022, OpenAI launched a specific tool for this model: ChatGPT (<https://openai.com/blog/chatgpt/>), which incorporates a simple web interface that follows the imperative model, where the user enters a request, and the chatbot returns results, all in a multilingual context (García-Peñalvo, 2023).

As Xie et al. (2022) pointed out, there is currently no established metric to measure the quality of the results generated by the various existing methods of natural language interpretation. That was the motivation behind the present study, which aims to analyse these machines' verbal (written) comprehension capacity and their use for analysing open texts in qualitative research.

2. Methodology

Similarly to previous experiments (Dobrev, 2019), we used a psychological test of verbal comprehension for humans in order to determine the level of the GPT-3's text-davinci-003 model (from now on referred to as GPT-3) compared to a human standard. As Dobrev (2019) indicated, all that is needed to identify computer programs that are AI is to subject them to a test and grant this designation to those that pass it. It is essential to clarify that we are not posing Turing's question: can machines think? (Turing, 1950). Our question is different: Can a machine understand a question expressed in a human's natural language? We will use Turing's strategy, stating that if the machine's response to our question is similar to that of a human, then we can say that the machine has the function of verbal comprehension similar to the cognitive function of humans. In the present study, we asked the machine the pre-established questions that make up the Verbal Comprehension Index in the well-known battery for measuring IQ: WAIS-IV. The use of intelligence measurement instruments or, more specifically, IQ measurement to assess the performance of AI machines is not new (Dowe & Hernández-Orallo, 2012): the first study to use the IQ test to evaluate AI was in 1998 (Hernández-Orallo & Minaya-Collado, 1998). Detterman (2011) proposed testing computer programs with intelligence tests for people. Wang et al. (2016) considered how to create a program that can solve intelligence tests designed for humans. Recently, Binz & Schulz (2023) simulated the behaviour of the human mind in GPT-3 with different tasks from cognitive psychology: a) experiments based on tasks and b) experiments based on vignettes. The second case demonstrated that GPT-3 could solve challenging problems based on vignettes, although its behaviour depended greatly on how they were presented.

2.1. Assessment instrument: WAIS-IV

The Wechsler Adult Intelligence Scale (WAIS) is a standardised instrument designed to measure the intelligence quotient of individuals aged 16 to 89 years old. The fourth, and to date latest, edition of the instrument (WAIS-IV) consists of fifteen tests (ten primary and five optional) that allows four indices to be calculated in addition to the intelligence quotient: verbal comprehension, perceptual reasoning, working memory, and processing speed. For the present study, we used the Verbal Comprehension Index (VCI) from the Spanish version of the WAIS-IV (Wechsler, 2012a, 2012b), which is composed of three main tests: the

Similarities test (S), the Vocabulary test (V), and the Information test (I); and one optional test, the Comprehension test (CO). Responses to the items in the Similarities, Vocabulary, and Comprehension tests are scored with 2 points, 1 point, or 0 points, depending on the individual's responses, while the responses in the Information test are scored with 1 or 0. In some imprecise or unclear responses, subjects can be asked for more information before scoring their response more accurately (Wechsler, 2012a). The norms of application for each of the four tests in terms of starting point, return rule, and termination rule were followed.

Once the items were scored, the direct score (DS) for each test and the corresponding scaled scores (SS) for each age range were calculated according to the application and correction manual (Wechsler, 2012a, pp. 192-201). These scores produce the VCI (Wechsler, 2012a, p. 203). In addition, the scale provides information about the percentile the subject is in.

In the present study, the VCI was calculated with all possible combinations of tests for each age range.

2.2. Analysis instrument: the text-davinci-003 model of GPT-3

The study used OpenAI Playground, a web-based interface of GPT-3, with the "Questions and Answers" (Q&A) pre-set setting chosen. This maintains the conversation context, making this interface more conversational than others (Bajaj et al., 2022).

The prompt was defined with the literal text that the WAIS provides for each test. No additional training was provided apart from the training items described by the WAIS. The model allows decisions on various parameters. The default values were used for the response duration, best of, frequency penalty, and presence penalty parameters (Bajaj et al., 2022; Binz & Schulz, 2023).

The most important two parameters for the study objective are temperature and Top_P. Both are related to the creativity of the response. Possible values range from 0 to 1. Given that our objective is to use the machine as a diagnostic tool, it must be reliable, giving an identical result each time the same question is asked. Therefore, the most conservative strategy was chosen, setting both parameters to 0, limiting the machine's potential creativity (Binz & Schulz, 2023).

2.3. Procedure

The analysis procedure had two phases. In the first phase, the instruction from the WAIS-IV application manual for the Similarities test (S) ("Now I am going to tell you two words and ask you how they are similar") was written in the playground, followed by the question corresponding to the example item: "How are two and seven similar?". Then, a line break or return was entered. Finally, the submit button was clicked, producing the result corresponding to the first item. According to the WAIS-IV instructions, depending on the responses, clarification or expansion questions must be asked using the instruction "Can you tell me anything more?" to award the corresponding score. It is important to note that the text-davinci-003 model answers the second question considering the context of the first and its response, so the procedure is identical to that envisaged in the test for humans.

To describe the reliability of the responses, the same action was replicated three times, analysing the degree of variation in the responses. We also checked that the results were identical if the test was done using different computers and IP addresses.

In the second phase, the same process as the first phase was replicated, with the following variations in the playground prompt:

- Two line breaks after the item and the clarification question.
- Two line breaks after the item and the clarification question and writing the key concepts to be compared in capital letters.
- Two line breaks after the item and the clarification question and writing the entire text in capital letters.

These tests allowed us to determine that the procedure that achieved the highest score according to the WAIS application and correction manual was to apply two line breaks after the initial question and the clarification question and write the entire text in capital letters. The rest of the WAIS-IV verbal comprehension index tests were conducted using this format, producing the results described below. However, the small differences in the scores mean that we can say that writing the questions in lowercase and including a single line break offers high reliability and validity.

3. Analysis and Results

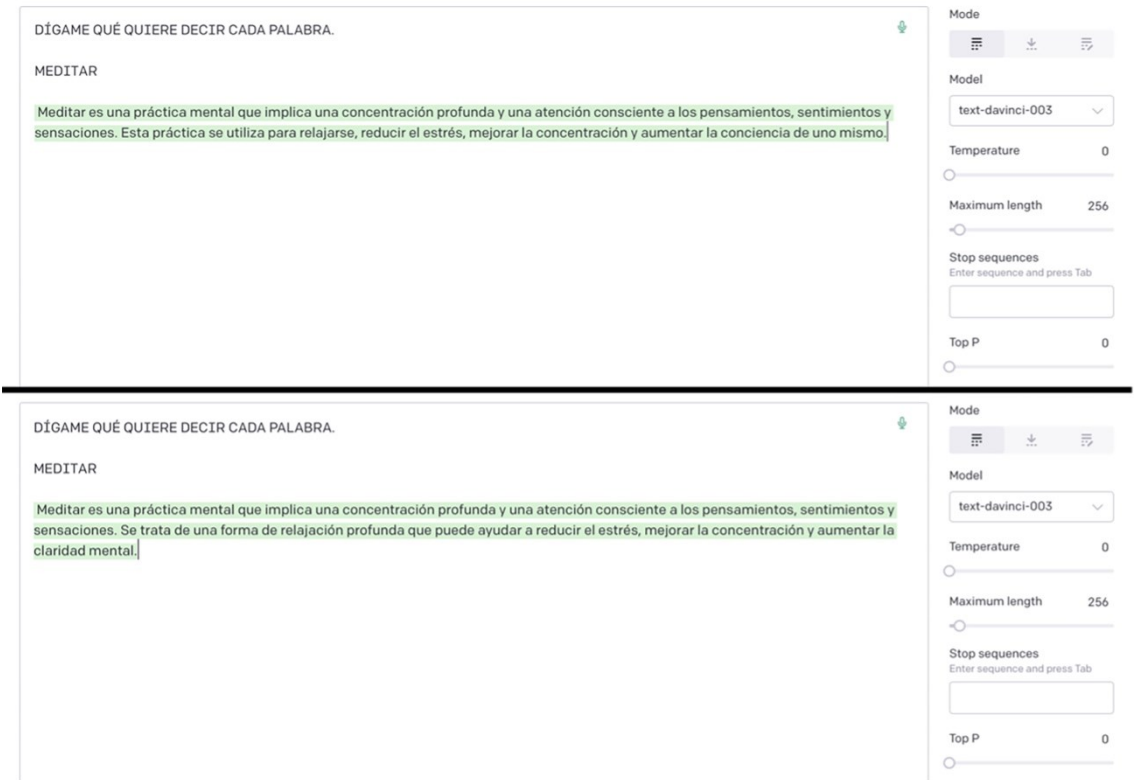
3.1. Reliability

The reliability of the text-davinci-003 model—using the procedure and parameters described in the previous section—was 100%. Although the initial results may seem lower, minor variations in expression did not affect the meaning.

In 92.59% of cases, the model's response was identical for each option tested and its subsequent replicas. The reliability of the information (I) and comprehension (CO) tests was 100%. In the similarities (S) test, the reliability percentage was 86.66% (13 out of 15 items), while in the Vocabulary (V) test, it was 84.61% (22 out of 26 items applied). Figure 1 presents an example of a vocabulary test item which produced different responses following the same procedure. In this example, they are non-identical expressions with identical meanings, leading us to state the model's complete reliability.

Figure 1

Application and response (and subsequent replica) to item 13 in the vocabulary test



3.2. Validity: Level of verbal comprehension

Table 1 shows the DS and SS obtained by the text-davinci-003 model in the four VCI tests.

Table 1

DS y SS by age ranges

Test	DS	SS by age ranges									
		16-	18-	20-	25-	35-	45-	55-	70-	80-	85-
		17	19	24	34	44	54	69	79	84	89
Similarities (S)	30	18	18	17	17	17	17	19	19	19	19
Vocabulary (V)	52	19	19	19	19	19	19	19	19	19	19
Information (I)	24	19	19	18	17	16	16	17	19	19	19
Comprehension (CO)	36	19	19	19	19	19	19	19	19	19	19

The DS in the Similarities (S) test was 30 points out of 36. The model achieved the maximum score (2) in fourteen of the eighteen scoreable items, while the score was 1 in two items and 0 in two. For example, Figure 2 shows the application of item 10, which achieved the maximum score of 2 points.

Figure 2

Application and response to item 10 in the similarities test

Mode

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 0

AHORA LE VOY A DECIR DOS PALABRAS Y A PREGUNTARLE EN QUÉ SE PARECEN.

¿EN QUÉ SE PARECEN EL VAPOR Y LA NIEBLA?

El vapor y la niebla se parecen en que ambos son formas de agua en el aire. Ambos son invisibles a simple vista, aunque el vapor es una forma de agua caliente, mientras que la niebla es una forma de agua fría. Ambos también se forman cuando el aire se enfría rápidamente.

In the vocabulary (V) test, the model scored 52 out of 57 points, achieving the maximum in 27 of the 30 items. In one item it scored 1, while in two items it scored 0. Figure 3 shows item 11, where it had an intermediate score (1) and where the response required a clarification question.

Figure 3

Application and response to item 11 in the vocabulary test

Mode

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 0

DÍGAME QUÉ QUIERE DECIR CADA PALABRA.

COMPASIÓN

Compasión es un sentimiento de profunda empatía y comprensión hacia el sufrimiento de otra persona. Se trata de una emoción que nos lleva a querer ayudar a alguien que está pasando por un momento difícil.

¿PUEDE EXPLICÁRMELO UN POCO MÁS?

La compasión es una emoción que nos lleva a sentir empatía y comprensión hacia el sufrimiento de otra persona. Esto significa que nos preocupamos por el bienestar de los demás y queremos ayudarles a superar sus dificultades. La compasión nos motiva a actuar de manera amable y compasiva hacia los demás, incluso cuando no estamos de acuerdo con sus opiniones o acciones. Esto nos ayuda a construir relaciones más fuertes y saludables con los demás.

In the information (I) test, the model scored 24 out of 26 possible points, failing in items 14 ("Why is Fleming famous?") and 23 ("What is the largest organ of the human body?"). However, in both cases, we believe that the responses should be considered valid, which

would raise the score to 26 out of 26. In the case of item 14, it is not specified that the question refers to Alexander Fleming, discoverer of penicillin, and the model's response refers to Ian Fleming, creator of James Bond (see Figure 4). In the case of item 23, the model's response interpreted "largest" as the heaviest organ of the body, the liver, and not the most extensive, the skin, which is the correct answer according to the instrument application and correction manual. In both cases, a subsequent test was conducted by adding the name Alexander in item 14 ("Why is Alexander Fleming famous?") and replacing the adjective 'large' with 'extensive' in item 23 ("What is the most extensive organ of the human body?"), with the model scoring the maximum in both cases.

Figure 4

Application and response to item 14 in the information test

The screenshot shows a chat window with a light blue header. On the left, a question is asked in Spanish: "¿POR QUÉ ES FAMOSO FLEMING?". On the right, the model's response is displayed in a green bubble: "Fleming es famoso por ser el creador de la famosa serie de novelas de espionaje James Bond. Estas novelas han sido llevadas a la pantalla grande en varias ocasiones, lo que ha contribuido a la fama de Fleming." To the right of the chat window is a settings panel. It includes a 'Mode' section with icons for chat, download, and share. Below that is the 'Model' section with a dropdown menu set to 'text-davinci-003'. The 'Temperature' is set to 0. The 'Maximum length' is set to 256. There is a 'Stop sequences' section with a text input field and a 'Top P' slider set to 0.

Finally, in the optional comprehension (CO) test, the model achieved the maximum score in all items, although in item 16 ("Why is land more expensive in the city than in the country?"), a clarification question was needed to obtain the 2 points.

Once the DS were collected and converted to SS, the sum of the SS for each age range was calculated, considering the four possible combinations of tests to calculate the composite VCI score and the percentile the model was in. Table 2 shows that the VCI of the text-davinci-003 model ranged between 143 (a score only obtained in age ranges 35-44 years and 45-54 years old) and 150 (maximum possible score), meaning that the model was classified as 'very superior' (Wechsler, 2012b). In all cases, the model was in the 99.8 percentile or higher, reaching higher than the 99.9 percentile in all test combinations in the age ranges 16 to 19 and 55 to 89 years old. This means it has a higher level of verbal comprehension than 99.9% of humans. Likewise, the model achieved the maximum composite VCI scores (and, therefore, is in a percentile above 99.9), and a 'very superior' classification (Wechsler, 2012b). In all cases, the model was placed in the 99.8 percentile or higher, reaching higher than a 99.9 percentile in all combinations of tests in the age ranges 16 to 19 and 55 to 89 years old, indicating that it has a higher level of verbal comprehension than 99.9% of humans. Furthermore, the model achieved the maximum composite VCI scores (and therefore is in a percentile above 99.9) in the combination of similarities (S), vocabulary (V), and comprehension (CO) tests across all age ranges.

Table 2*Conversion of the sum from SS to VCI and its percentiles*

Age ranges	Test	SS sum	VCI	Percentiles
16-17 years 18-19 years	S, V, I	56	150	>99.9
	S, V, CO	56	150	>99.9
	S, I, CO	56	150	>99.9
	V, I, CO	57	150	>99.9
20-24 years	S, V, I	54	147	99.9
	S, V, CO	55	150	>99.9
	S, I, CO	54	147	99.9
	V, I, CO	56	150	>99.9
25-34 years	S, V, I	53	145	99.9
	S, V, CO	55	150	>99.9
	S, I, CO	53	145	99.9
	V, I, CO	55	150	>99.9
35-44 years 45-54 years	S, V, I	52	143	99.8
	S, V, CO	55	150	>99.9
	S, I, CO	52	143	99.8
	V, I, CO	54	147	99.9
55-69 years	S, V, I	55	150	>99.9
	S, V, CO	57	150	>99.9
	S, I, CO	55	150	>99.9
	V, I, CO	55	150	>99.9
70-79 years 80-84 years 85-89 years	S, V, I	57	150	>99.9
	S, V, CO	57	150	>99.9
	S, I, CO	57	150	>99.9
	V, I, CO	57	150	>99.9

4. Discussion y conclusions

Paraphrasing Turing (1950), the text-davinci-003 model of GPT-3 understands questions expressed in natural language, and we know this by analysing the responses it gave. According to the tests used in the study, it has demonstrated a very high level of verbal comprehension compared to human standards. The reliability and validity of the results demonstrate that it is possible to use this AI model as a tool for text analysis. That is, it is possible to code texts validly and reliably. However, its flexibility, desirable in its more creative functions, must be maximally limited in its analytical function. For this, it is essential to consider the following results.

Firstly, GPT-3 analyses all the text provided each time and kept the context of the entire question session in memory. Therefore, for analysis of independent texts, they must be entered separately to obtain a result without the risk of context biases. If two questions are asked simultaneously, the answer to one will be affected by the other.

Secondly, minor variations in the way questions are asked cause different responses. We verified this by using uppercase and lowercase letters or entering different line breaks after the question. In the results, we offer the option that produced the best scores, but the takeaway here is not to just use these options, but rather always conduct tests until the

appropriate prompt is found for the type of analysis to be done and to maintain it throughout the analysis so that it offers the reliability and validity required in each study. We verified that if the same prompt is used with the temperature and Top P parameters set to zero, in line with what similar studies have done (Binz & Schultz, 2023), the machine offers 100% reliability.

Thirdly, the level of verbal comprehension is greater than that of humans. However, its context is limited to what it learned in training and the prompt used. This limitation can lead to what we might consider interpretation errors. We gave the example of its response to who Fleming is. Sufficient context is needed for the machine to identify what field the desired response is in (Guetterman et al., 2018).

In addition, the machine has been trained for conversational functionality. See the comparative study by Tack and Piech (2022) on this machine's conversational ability compared to a teacher and not having precise knowledge of various fields. However, the texts used provided sufficient knowledge to achieve maximum scores in the information tests. This good result implies that its degree of general culture is also greater than most humans. Once again, it must be emphasised that this functionality must be controlled, as, like humans, it can be a potential source of bias.

Furthermore, we know that verbal comprehension levels vary according to age. However, we found that the model demonstrated the maximum level for a wide age range. Our study provides data for ages between 16 and 89 years (the range established by the test used). This means that the machine could be used regardless of the age of the authors of texts. The study should be extended to verify if it maintains its level of comprehension in texts written by subjects under 16 years of age.

Finally, we do not have evidence to make assertions about results in other languages, but it is very plausible to hypothesise that those languages from which a similar number of texts to Spanish have been used will have similar results.

In summary, this study aimed to analyse the (written) verbal comprehension capacity of GPT-3's text-davinci-003 model and its use for analysing open texts in qualitative research. In conclusion, the level of verbal comprehension established in the tests used and the reliability allow it to be used as a tool for analysing open texts and opens enormous possibilities for its use in this type of research.

This use must be based on a precise design that is specific to each analysis process: the corpus to be analysed and the analysis procedures must be chosen in the same way as for non-machine-assisted procedures. Deductive strategies (with previously defined category systems) or inductive strategies (establishing categories on the groupings of texts made by machines) are possible, but must all have the same methodological assurances as non-assisted analysis.

The fact that the machine offers its responses using contextual elements means that these elements—just as with a human analyst—act as elements that increase precision and as elements that cause biases. The way questions are asked affects the response, but this problem is not specific to AI. Therefore, all of the usual quality control, reliability, and validity systems (Williams & Morrow, 2009) used in qualitative research must continue to be part of the method in machine-assisted studies.

While up to now, in analysis of open texts, work has been done with corpora of a few dozens or hundreds of documents, we can currently consider thousands or hundreds of thousands (Longo, 2020; Mills, 2019). This capacity is a significant improvement, given that one of the main criticisms of qualitative studies has always been the quality of their small samples and their limitations in generalising conclusions.

The applications of this analysis tool for social sciences are enormous, from analysis of open-ended responses in surveys or interviews, identification of themes in social networks, or use as an additional judge to analyse the internal consistency of coding processes.

We are aware of the risks of using these AIs for decision-making. Therefore, our approach is to use them as another analysis tool while maintaining all the safeguards that social science research has refined over the years. The risks should not prevent them from being used but should encourage us to maximise the quality control systems for their results. As Turing stated, this technology forces us to explore multiple options, which must be tested, not accepted without criticism (García-Peñalvo et al., 2023).

If we overcome the initial fascination with how creative these machines are and the astonishing responses they give us, we can focus on using one of their specific functionalities (text comprehension) to increase our capacity for analysis and explanation of social phenomena, which will allow us to more accurately and quickly find the solutions we need.

5. Funding

Project titled "The use of AI in the analysis of informal social networks for guidance in Higher Education" (Ref. PID2021-125405NB-I00), within the Subprogram for Knowledge Generation (BOE, November 13th, 2022) in the framework of the State Program to Promote Scientific-Technical Research and its Transfer of the State Plan for Scientific, Technical and Innovation Research 2021-2023 (Call identifier in the National Grant Database ID BDNS: 590214).

References

- Aspers, P., & Corte, U. (2019). What is qualitative in qualitative research. *Qualitative Sociology*, 42(2), 139–160. <https://doi.org/10.1007/s11133-019-9413-7>
- Bail, C. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, 43(3), 465–482. <https://doi.org/10.1007/s11186-014-9216-5>
- Bajaj, D., Goel, A., Gupta, S.C., & Batra, H. (2022). MUCE: A multilingual use case model extractor using GPT-3. *International Journal of Information Technology*, 14, 1543–1554. <https://doi.org/10.1007/s41870-022-00884-2>
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *PNAS. Proceedings of the National Academy of Sciences*, 120(6), e2218523120. <https://doi.org/10.1073/pnas.2218523120>

- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M.A. Ranzato, R. Hadsell, M.F. Balcan, & H.T. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. <https://bit.ly/3ZmRPQG>
- Carriere, J., Shafi, H., Brehon, K., Pohar, M.K., Churchill, K., Ho, C., & Tavakoli, M. (2021). Case report: Utilizing AI and NLP to assist with healthcare and rehabilitation during the COVID-19 pandemic. *Frontiers in Artificial Intelligence*, 4, 613637. <https://doi.org/10.3389/frai.2021.613637>
- Chang, T., DeJonckheere, M., Vydiswaran, V.G.V., Li, J., Buis, L.R., & Guetterman, T.C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research*, 15(3), 398–412. <https://doi.org/10.1177/15586898211021196>
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education*. Routledge. <https://doi.org/10.4324/9780203224342>
- Cope, B., & Kalantzis, M. (2015). Sources of evidence-of-learning: Learning and assessment in the era of big data. *Open Review of Educational Research*, 2(1), 194–217. <https://doi.org/10.1080/23265507.2015.1074869>
- Cypress, B.S. (2019). Data analysis software in qualitative research. Preconceptions, expectations, and adoption. *Dimensions of Critical Care Nursing*, 38(4), 213-220. <https://doi.org/10.1097/DCC.0000000000000363>
- Detterman, D. (2011). A challenge to Watson. *Intelligence*, 39(2-3), 77-78. <https://doi.org/10.1016/j.intell.2011.02.006>
- Dobrev, D. (2019). The IQ of artificial intelligence. *Serdica Journal of Computing*, 13(1-2), 41-70. <https://doi.org/10.55630/sjc.2019.13.41-70>
- Dowe, D.L., & Hernández-Orallo, J. (2012). IQ tests are not for machines, yet. *Intelligence*, 40(2), 77–81. <https://doi.org/10.1016/j.intell.2011.12.001>
- Flick, U. (2018). *Designing qualitative research*. SAGE.
- García-Peñalvo, F.J. (2023). La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: Disrupción o pánico. *Education in the Knowledge Society*, 24, e31279. <https://doi.org/10.14201/eks.31279>
- García Peñalvo, F.J., Llorens-Largo, F., & Vidal, J. (2023). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1). <https://doi.org/10.5944/ried.27.1.37716>
- Golder, S.A., & Macy, M.W. (2011). Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. *Science*, 333(6051), 1878–1881. <https://doi.org/10.1126/science.1202775>
- Grodal, S., Anteby, M., & Holm, A.L. (2021). Achieving rigor in qualitative analysis: The role of active categorization in theory building. *Academy of Management Review*, 46(3), 591-612. <https://doi.org/10.5465/amr.2018.0482>

- Guetterman, T.C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, 20(6), e231. <https://doi.org/10.2196/jmir.9702>
- Hammersley, M. (2013). *What is qualitative research?* Bloomsbury Academic.
- Hernández-Orallo, J., & Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the international symposium of engineering of intelligent systems (EIS'98)* (pp. 146–163). ICSC Press. <https://bit.ly/3JdQTbK>
- Longo, L. (2020). Empowering qualitative research methods in education with artificial intelligence. In A.P. Costa, L.P. Reis, & A. Moreira (Eds.), *Computer supported qualitative research. New trends on qualitative research (WCQR2019)*. (pp. 1-21). Springer. <https://doi.org/10.1007/978-3-030-31787-4>
- Mills, K. A. (2019). *Big data for qualitative research*. Routledge Focus. <https://doi.org/10.4324/9780429056413>
- Oussalah, M. (2022). AI explainability. A bridge between machine vision and natural language processing. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G.M. Farinella, T. Mei, M. Bertini, H.J. Escalante, & R. Vezzani (Eds.), *Pattern recognition. ICPR International workshops and challenges* (pp. 257-273). Springer. <https://doi.org/10.1007/978-3-030-68790-8>
- Rezaeenour, J., Ahmadi, M., Jelodar, H., & Shahrooei, R. (2022). Systematic review of content analysis algorithms based on deep neural networks. *Multimedia Tools & Applications*. <https://doi.org/10.1007/s11042-022-14043-z>
- Shlomo, N., & Goldstein, H. (2015). Big data in social research. *Journal of the Royal Statistical Society*, 178(4), 787–790. <https://doi.org/10.1111/rssa.12144>
- Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2022). *Prompting GPT-3 to be reliable*. arXiv. <https://doi.org/10.48550/arXiv.2210.09150>
- Silver, C., & Lewins, A. (2014). *Using software in qualitative research: A step-by-step guide*. SAGE.
- Tack, A. & Piech, C. (2022). *The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues*. arXiv. <https://doi.org/10.48550/arXiv.2205.07540>
- Towler, L., Bondaronek, P., Papakonstantinou, T., Amlôt, R., Chadborn, T., Ainsworth, B., & Yardley, L. (2022). *Applying machine-learning to rapidly analyse large qualitative text datasets to inform the COVID-19 pandemic response: Comparing human and machine-assisted topic analysis techniques*. MedRxiv. <https://doi.org/10.1101/2022.05.12.22274993>
- Turing, A.M.I. (1950). Computing machinery and intelligence. *Mind*, 236(LIX), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Wang, H., Tian, F., Gao, B., Bian, J., & Liu, T.Y. (2016). *Solving verbal comprehension questions in IQ test by knowledge-powered word embedding*. arXiv. <https://doi.org/10.48550/arXiv.1505.07909>
- Wechsler, D. (2012a). WAIS-IV. *Escala de inteligencia de Wechsler para adultos-IV. Manual de aplicación y corrección*. Pearson.
- Wechsler, D. (2012b). WAIS-IV. *Escala de inteligencia de Wechsler para adultos-IV. Manual técnico y de interpretación*. Pearson.

- Williams, E.N., & Morrow, S.L. (2009). Achieving trustworthiness in qualitative research: A pan-paradigmatic perspective. *Psychotherapy Research*, 19(4-5), 576-582. <https://doi.org/10.1080/10503300802702113>
- Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. *International Management Review*, 15(1), 45-55. <https://bit.ly/3mt4BPh>
- Xie, Y., Vosoughi, S., & Hassanpour, S. (2022). *Interpretation quality score for measuring the quality of interpretability methods*. arXiv. <https://doi.org/10.48550/arXiv.2205.12254>

Cómo citar:

- González-Mayorga, H., Rodríguez-Esteban, A., & Vidal, J. (2024). El uso del modelo GPT de OpenAI para el análisis de textos abiertos en investigación educativa [Using OpenAI's GPT Model to Analyse Open Texts in Educational Research]. *Pixel-Bit. Revista de Medios y Educación*, 69, 227-253. <https://doi.org/10.12795/pixelbit.102032>