Evaluación de la fiabilidad computacional de *ChatGPT* en el cálculo de la fiabilidad intercodificadora en el análisis de contenido: evidencia a partir de datos simulados

Evaluating the computational reliability of *ChatGPT* in calculating intercoder reliability in content analysis: Evidence from simulated data

Manuel Goyanes

Citación recomendada:

Goyanes, Manuel (2026). "Evaluación de la fiabilidad computacional de ChatGPT en el cálculo de la fiabilidad intercodificadora en el análisis de contenido: evidencia a partir de datos simulados [Evaluating the computational reliability of ChatGPT in calculating intercoder reliability in content analysis: Evidence from simulated data]". *Infonomy*, 4(2) e26006. *https://doi.org/10.3145/infonomy.26.006*

Artículo recibido: 27-02-2026
Artículo aprobado: 01-04-2026

Manuel Goyanes
https://orcid.org/0000-0001-8329-0610
https://directorioexit.info/ficha3719
Universidad Carlos III de Madrid
Departamento de Comunicación
Grupo de Investigación MITCOM
Madrid, 133. 28903 Getafe (Madrid), España
manuel.goyanes@uc3m.es

Resumen

La creciente integración de los modelos de lenguaje de gran escala (LLMs) en el proceso de investigación ha suscitado importantes interrogantes sobre su fiabilidad en la realización de análisis estadísticos. Aunque estudios previos han explorado el uso de estos modelos en tareas de clasificación textual y codificación cualitativa, existe una notable falta de evidencia sobre su precisión en el cálculo de métricas estadísticas fundamentales utilizadas en el análisis de contenido. Este estudio aborda este vacío evaluando de forma sistemática el rendimiento de *ChatGPT* en el cálculo del porcentaje de acuerdo, las tablas de contingencia y la Kappa de Cohen. Mediante un conjunto de simulaciones controladas, se variaron parámetros clave como el tamaño muestral, el número de categorías, el equilibrio en la distribución y el nivel de error en las codificaciones. Los resultados generados por *ChatGPT* 5.3 *Instant* se compararon con los obtenidos mediante procedimientos estadísticos estándar, considerados como referencia (*ground truth*). Los hallazgos indican que *ChatGPT* alcanza una alta precisión únicamente en condiciones simples, especialmente en muestras pequeñas con variables binarias y distribuciones balanceadas. Sin embargo, su rendimiento se deteriora a medida que aumenta la complejidad analítica. En escenarios de complejidad moderada, el modelo presenta una precisión parcial, reproduciendo en ocasiones correctamente las tablas de contingencia pero introduciendo desviaciones en los estadísticos derivados. En condiciones más complejas, particularmente con distribuciones desbalanceadas o múltiples categorías, *ChatGPT* genera resultados sistemáticamente sesgados, tendiendo a sobreestimar el nivel de acuerdo. Asimismo, en muestras de gran tamaño, el modelo presenta limitaciones operativas que impiden la obtención de resultados. En conjunto, los resultados evidencian una falta de fiabilidad estadística por lo que no se recomienda el uso de *ChatGPT* para el cálculo de estas métricas, salvo en casos muy simples con muestras pequeñas, y siempre bajo supervisión y validación mediante software estadístico consolidado.

Palabras clave

Análisis de contenido; Fiabilidad intercodificadora; Kappa de Cohen; Inteligencia artificial; *ChatGPT*; Porcentaje de acuerdo.

Abstract

The increasing integration of large language models (LLMs) into the research workflow has raised important questions regarding their reliability in performing statistical analyses. While prior studies have explored the use of LLMs in text classification and qualitative coding, little is known about their accuracy in computing core statistical metrics used in content analysis. This study addresses this gap by systematically evaluating the performance of *ChatGPT* in calculating percentage agreement, contingency tables, and Cohen's Kappa. Using a series of controlled simulations, we varied key parameters including sample size, number of categories, distribution balance, and levels of coding error. The outputs generated by *ChatGPT* 5.3 *Instant* were benchmarked against results obtained by the author using standard statistical procedures (ground truth). Findings indicate that *ChatGPT* achieves high accuracy only under simple

conditions, particularly with small samples, binary variables, and balanced distributions. However, its performance declines as analytical complexity increases. In moderately complex scenarios, the model shows partial accuracy, often reproducing contingency tables correctly but introducing deviations in derived statistics. In more complex settings, especially with unbalanced distributions or multiple categories, *ChatGPT* produces systematically biased results, typically overestimating agreement. Additionally, in large-scale datasets, the model fails to generate outputs due to operational limitations. Overall, the results reveal a lack of consistent reliability across realistic analytical scenarios. As a bottom line, the use of *ChatGPT* for computing these statistical metrics is not recommended, except in very simple cases involving small samples, and only under strict supervision and validation using established statistical software.

Keywords
Content analysis; Intercoder reliability; Cohen's Kappa; Artificial intelligence; *ChatGPT*; Percentage agreement.

1. Introducción

En los últimos años, el desarrollo de modelos de lenguaje de gran escala (*Large Language Models*, *LLMs* por sus siglas en inglés), como *ChatGPT*, ha transformado de manera significativa múltiples etapas del proceso de investigación (**Robila**; **Robila**, 2020; **Xu** *et al.*, 2021; **Grossmann** *et al.*, 2023). Desde la generación de ideas hasta el análisis de datos (**Morgan**, 2023; **Chubb**, 2023; **Goyanes**; **De-Marcos**, 2025; **Cook** *et al.*, 2025; **Nguyen**; **Welch**, 2026), estos sistemas han comenzado a integrarse como herramientas auxiliares en tareas tradicionalmente realizadas por investigadores humanos (**Andersen** *et al.*, 2025). En el ámbito del análisis de contenido, en particular, su potencial ha despertado un creciente interés, especialmente en lo relativo a la automatización de procesos de codificación y al cálculo de métricas estadísticas asociadas a la fiabilidad intercodificadora (**Lee** *et al.*, 2020; **Fatouros** *et al.*, 2023; **Belal** *et al.*, 2023; **Gilardi** *et al.*, 2023; **Jim** *et al.*, 2024; **Lossio-Ventura** *et al.*, 2024; **Fu** *et al.*, 2024; **Törnberg**, 2025; **Goyanes**; **De-Marcos**, 2025).

La fiabilidad intercodificadora constituye un pilar fundamental en el análisis de contenido (**Lombard** *et al.*, 2002; **Krippendorff**, 2004), ya que permite evaluar el grado de consistencia entre codificadores independientes y, por tanto, la validez de las inferencias derivadas del proceso de codificación (**Hayes**; **Krippendorff**, 2007; **Neuendorf**, 2017; **Krippendorff**, 2018; **Goyanes**; **Piñeiro-Naval**, 2024). Tradicionalmente, esta fiabilidad se ha estimado mediante una amplia variedad de indicadores, entre los que se incluyen el porcentaje de acuerdo, las tablas de contingencia y coeficientes más sofisticados como la Kappa de Cohen o el alfa de Krippendorff, entre otros (véase **Hayes** & **Krippendorff**, 2007 para una revisión). Este último ha ganado relevancia por su flexibilidad y aplicabilidad a distintos niveles de medición; sin embargo, su cálculo suele requerir procedimientos computacionales más complejos, como técnicas de remuestreo (*bootstrap*) para el cálculo de intervalo de confianza, lo que dificulta su implementación fiable en modelos de lenguaje como *ChatGPT*, donde no se recomienda su uso debido a limitaciones en la ejecución de este tipo de operaciones. En este contexto, la

Kappa de Cohen se mantiene como uno de los coeficientes más utilizados en el análisis de contenido, tanto por su solidez conceptual como por su capacidad para ajustar el acuerdo observado en función del esperado por azar, lo que la convierte en una medida especialmente adecuada para evaluar la fiabilidad intercodificadora en numerosos estudios empíricos.

En este contexto, el uso de *ChatGPT* y otros LLMs para calcular estadísticos de fiabilidad plantea una cuestión metodológica relevante: ¿hasta qué punto estos sistemas son capaces de realizar cálculos estadísticos básicos y avanzados de manera precisa y consistente? Si bien la literatura reciente ha explorado el rendimiento de los LLMs en tareas analíticas, incluyendo diferentes tareas de computación (**Amin** *et al.*, 2024) o la interpretación de resultados estadísticos (**Mondal** *et al.*, 2024; **Shukla** *et al.*, 2025; **Lu** *et al.*, 2025; **Dobler** *et al.*, 2025), existe aún una notable escasez de estudios que evalúen de forma sistemática su capacidad para computar métricas específicas del análisis de contenido. Este vacío es especialmente relevante si se considera el creciente uso académico de estas herramientas por parte de investigadores en diferentes campos científicos (**Van Noorden**; **Perkel**, 2023; **Matos** *et al.*, 2025), quienes pueden confiar en sus resultados sin una validación previa de su exactitud.

Además, la naturaleza de los LLMs introduce desafíos particulares en términos de fiabilidad computacional. A diferencia de los programas estadísticos tradicionales, diseñados para ejecutar operaciones deterministas, los modelos de lenguaje operan mediante procesos probabilísticos que pueden generar variabilidad en los resultados, especialmente cuando se enfrentan a tareas que requieren cálculos precisos o manejo de grandes volúmenes de datos. Esta característica plantea interrogantes sobre los límites de su aplicabilidad en contextos donde la exactitud es relevante. Ante este escenario, el presente estudio tiene como objetivo evaluar la fiabilidad computacional de *ChatGPT* en el cálculo de estadísticos de fiabilidad intercodificadora en análisis de contenido.

En concreto, se examina su capacidad para calcular correctamente tres elementos básicos del análisis de contenido: el porcentaje de acuerdo, las tablas de contingencia y la Kappa de Cohen. Para ello, se diseñan una serie de simulaciones que buscan aproximarse a condiciones realistas propias del análisis de contenido en comunicación, reproduciendo configuraciones habituales en términos de tamaño muestral, número de categorías, distribución de las mismas y niveles de desacuerdo entre codificadores. A diferencia de enfoques centrados en el análisis estadístico del error, este trabajo adopta una perspectiva de verificación directa, comparando los resultados generados por *ChatGPT* con valores de referencia obtenidos manualmente mediante software estadístico convencional. Este enfoque permite identificar si *ChatGPT* es capaz de realizar correctamente estos cálculos y delimitar las condiciones bajo las cuales su rendimiento puede verse comprometido.

Este estudio contribuye a la literatura metodológica ofreciendo una evaluación sistemática y replicable del uso de *ChatGPT* como herramienta para el cálculo de estadísticos de fiabilidad en análisis de contenido. Asimismo, aporta evidencia

empírica relevante para orientar a investigadores sobre las posibilidades y limitaciones de integrar modelos de lenguaje en procesos estadísticos que requieren precisión cuantitativa.

2. Método

2.1. Diseño experimental

El presente estudio adopta un diseño basado en simulación con el objetivo de evaluar la fiabilidad computacional de *ChatGPT* 5.3 *Instant* en el cálculo de estadísticos de fiabilidad intercodificadora. En concreto, el enfoque se centra en la verificación directa de la exactitud de los resultados generados por el modelo, comparándolos con valores de referencia obtenidos mediante software estadístico convencional y realizados por el autor del estudio. Para ello, se diseñó un conjunto de escenarios simulados que varían en complejidad, permitiendo evaluar el rendimiento de la inteligencia artificial bajo condiciones progresivamente más exigentes. Estos escenarios se construyeron manipulando cuatro dimensiones clave: tamaño muestral, número de categorías, distribución de las categorías y grado de desacuerdo entre codificadores. Estas simulaciones permiten capturar tanto situaciones típicas en análisis de contenido como casos límite que pueden comprometer la precisión computacional. En la Tabla 1, se describen de manera detallada las diez simulaciones consideradas.

Tabla 1. Resumen de escenarios de simulación

Escenario	N	Categorías	Distribución	Error (%)	Tipo de complejidad
1	100	2	Balanceada	12.0	Baja
2	1,000	2	Balanceada	10.4	Baja-media
3	10,000	2	Balanceada	9.47	Media
4	50,000	2	Balanceada	9.75	Alta
5	1,000	3	Balanceada	9.2	Media
6	1,000	5	Balanceada	9.2	Media-alta
7	1,000	2	Desbalanceada	8.5	Media
8	1,000	3	Desbalanceada	9.2	Media-alta
9	1,000	3	Balanceada	27.3	Alta
10	10,000	5	Desbalanceada	28.79	Muy alta

Escenario 1: Condición base (baja complejidad)

Parámetro	Valor
N	100
Categorías	2
Distribución	Balanceada (≈ 50% cada categoría)
Error entre codificadores	12%
Nivel de complejidad	Baja

Este escenario constituye una condición de referencia con baja complejidad y un tamaño muestral reducido. Su propósito es verificar que *ChatGPT* es capaz de realizar correctamente los cálculos en condiciones estándar, donde no se esperan dificultades ni en la construcción de la tabla de contingencia ni en el cálculo de la Kappa de Cohen. Este tipo de configuración es habitual en estudios piloto o en análisis exploratorios.

Escenario 2: Escalabilidad inicial

Parámetro	Valor
N	1,000
Categorías	2
Distribución	Balanceada (≈ 50% cada categoría)
Error entre codificadores	10.4%
Nivel de complejidad	Baja-media

Este escenario introduce un incremento significativo en el tamaño muestral, manteniendo aproximadamente constantes el resto de los parámetros. Su objetivo es evaluar si el modelo mantiene la precisión cuando aumenta el volumen de datos, una condición frecuente en análisis de contenido automatizados o estudios con grandes corpus textuales.

Escenario 3: Escalabilidad intermedia

Parámetro	Valor
N	10,000
Categorías	2
Distribución	Balanceada (≈ 50% cada categoría)
Error entre codificadores	9.47%
Nivel de complejidad	Media

En este escenario se evalúa la capacidad de *ChatGPT* para procesar volúmenes de datos considerablemente mayores. Este nivel de N permite identificar posibles limitaciones relacionadas con la gestión de información extensa o problemas derivados de la longitud del *input*.

Escenario 4: Escalabilidad extrema (límite)

Parámetro	Valor
N	50,000
Categorías	2
Distribución	Balanceada (≈ 50% cada categoría)
Error entre codificadores	9.75%
Nivel de complejidad	Alta

Este escenario representa una condición de estrés computacional en términos de tamaño muestral. Su inclusión responde a la necesidad de explorar los límites operativos del modelo, especialmente en contextos de *big data*, donde el volumen de observaciones puede comprometer la precisión de los cálculos.

Escenario 5: Complejidad categorial moderada

Parámetro	Valor
N	1,000
Categorías	3
Distribución	Balanceada (≈ 33% cada categoría)
Error entre codificadores	9.2%
Nivel de complejidad	Media

Este escenario introduce un incremento en el número de categorías, lo que implica una mayor complejidad en la construcción de la tabla de contingencia y en el cálculo de probabilidades esperadas en la Kappa. Este tipo de configuración es habitual en estudios que emplean escalas nominales con múltiples categorías (por ejemplo, positivo, negativo, neutral).

Escenario 6: Alta complejidad categorial

Parámetro	**Valor**
N	1,000
Categorías	5
Distribución	Balanceada (≈ 20% cada categoría)
Error entre codificadores	9.2%
Nivel de complejidad	Media-alta

Este escenario amplía aún más la complejidad categorial, generando una tabla de contingencia de mayor dimensión (5×5). Su inclusión permite evaluar si el aumento en el número de celdas y combinaciones posibles afecta la precisión del modelo, especialmente en el cálculo de la Kappa.

Escenario 7: Distribución desbalanceada (binaria)

Parámetro	**Valor**
N	1,000
Categorías	2
Distribución	Desbalanceada (≈ 90% / ≈ 10%)
Error entre codificadores	8.5%
Nivel de complejidad	Media

Este escenario introduce una distribución altamente asimétrica, condición conocida por generar discrepancias entre el porcentaje de acuerdo y la Kappa de Cohen. Su inclusión permite evaluar si *ChatGPT* es capaz de manejar correctamente estos casos, donde el acuerdo esperado por azar adquiere un papel determinante.

Escenario 8: Distribución desbalanceada (multicategórica)

Parámetro	**Valor**
N	1,000
Categorías	3
Distribución	Desbalanceada (≈ 70% / ≈ 20% / ≈ 10% cada categoría)
Error entre codificadores	9.2%
Nivel de complejidad	Media

Este escenario combina complejidad categorial y desbalance en la distribución, lo que incrementa la dificultad del cálculo, especialmente en la estimación de probabilidades marginales. Este tipo de distribución es frecuente en análisis de contenido donde ciertas categorías son claramente dominantes.

Escenario 9: Alto desacuerdo entre codificadores

Parámetro	**Valor**
N	1,000

Categorías	3
Distribución	Balanceada (≈ 33% cada categoría)
Error entre codificadores	27.3%
Nivel de complejidad	Alta

Este escenario incrementa el nivel de discrepancia entre codificadores, reduciendo el acuerdo observado y generando una estructura más compleja en la tabla de contingencia. Su objetivo es evaluar si *ChatGPT* mantiene la precisión en contextos donde los datos presentan mayor variabilidad y menor consistencia.

Escenario 10: Condición extrema (máxima complejidad)

Parámetro	**Valor**
N	10,000
Categorías	5
Distribución	Desbalanceada (≈ 50%, ≈ 20%, ≈ 15%, ≈ 10%, ≈ 5%)
Error entre codificadores	28.79%
Nivel de complejidad	Muy alta

Este escenario combina simultáneamente todas las fuentes de complejidad: gran tamaño muestral, múltiples categorías, distribución asimétrica y alto nivel de desacuerdo. Su inclusión responde a la necesidad de evaluar el rendimiento del modelo en condiciones de máxima exigencia, donde la probabilidad de errores en el cálculo es significativamente mayor.

2.2. Generación de datos y cálculo de los valores de referencia

Para cada escenario, se generaron simulaciones que representan el proceso de codificación de dos codificadores independientes. En primer lugar, se creó una variable de referencia ("verdad") siguiendo tanto el número de categorías como la distribución específica de estas definida en cada escenario (ya fuera balanceada o desbalanceada). A partir de esta, se generó la codificación del primer codificador, que replica la distribución original establecida. Posteriormente, se generó la codificación del segundo codificador introduciendo un nivel controlado de error (del ≈ 10% al ≈ 30%), asignando de manera aleatoria discrepancias respecto a la codificación original. Este procedimiento permite simular distintos niveles de acuerdo y desacuerdo entre codificadores de forma controlada y reproducible, al tiempo que incorpora variaciones sistemáticas en la estructura categorial y en la distribución de las categorías, dos elementos clave que pueden afectar al rendimiento computacional del LLM.

Los valores correctos de cada métrica se calcularon manualmente utilizando software estadístico convencional, concretamente *IBM SPSS Statistics*, que se empleó tanto para la generación de las simulaciones como para el cálculo de los estadísticos de referencia. Para cada simulación se obtuvieron: 1) Porcentaje de acuerdo, 2) Tabla de contingencia (matriz de frecuencias), 3) Kappa de Cohen. Estos resultados se utilizaron como criterio de comparación para evaluar la exactitud de los cálculos realizados por *ChatGPT*, permitiendo contrastar directamente sus *outputs* con los obtenidos mediante un entorno estadístico consolidado y ampliamente validado en la investigación empírica.

2.3. Prompt y evaluación de la exactitud

Cada simulación fue introducida en *ChatGPT* con un *prompt* en el que se solicitaba explícitamente el cálculo de las tres métricas de interés. Para garantizar la consistencia del procedimiento, se utilizó un mismo formato de *prompt* en todos las simulaciones, evitando variaciones que pudieran influir en el rendimiento del modelo. El *prompt* es el siguiente:

> *Actúa como un software estadístico.*
> *Calcula a partir de las variables coder1 y coder2:*
> *Porcentaje de acuerdo*
> *Tabla de contingencia (frecuencias)*
> *Kappa de Cohen*
> *Devuelve solo los resultados finales, sin explicación ni pasos intermedios.*
> *Usa 4 decimales en todos los resultados.*
> *Datos: [BASE DE DATOS PEGADA SEGÚN PAQUETE ESTADÍSTICO]*

La evaluación de la fiabilidad computacional se realizó mediante comparación directa entre los resultados generados por *ChatGPT* y los valores de referencia. Para cada métrica y escenario, los resultados se clasificaron de manera dicotómica: 1) Correcto: cuando el valor coincide exactamente con el resultado de referencia, 2) Incorrecto: cuando el valor difiere del resultado de referencia

3. Resultados

Escenario 1: Condición base (baja complejidad)

El primer escenario, caracterizado por un tamaño muestral reducido (N = 100), dos categorías, distribución balanceada (50% / 50%) y un nivel de desacuerdo aproximado del 10%, los resultados obtenidos muestran una coincidencia total entre los valores calculados por *ChatGPT* y los generados con el valor de referencia humano. En concreto, tanto el porcentaje de acuerdo como la tabla de contingencia y la Kappa de Cohen fueron reproducidos con exactitud por el modelo.

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	0.880	0.880	Correcta
Tabla de contingencia	Correcta	Correcta	Correcta
Kappa de Cohen	0.760	0.760	Correcta

Escenario 2: Escalabilidad inicial

El segundo escenario se diseñó para evaluar la capacidad de *ChatGPT* de mantener la precisión computacional cuando aumenta el tamaño muestral, manteniendo constantes el resto de parámetros estructurales. En este caso, se generó un dataset con un total de 1,000 observaciones, dos categorías y una distribución balanceada (50% en cada categoría).

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	0.900	0.896	Incorrecta
Tabla de contingencia	Incorrecta	Incorrecta	Incorrecta
Kappa de Cohen	0.800	0.792	Incorrecta

Escenario 3: Escalabilidad intermedia

En el tercer escenario (N = 10,000, dos categorías, distribución balanceada y un nivel de desacuerdo del 9.47%), los resultados humanos obtenidos mediante *SPSS* mostraron un correcto funcionamiento del diseño de simulación. En concreto, la tabla de contingencia reflejó una distribución consistente con el error introducido, y el valor de la Kappa de Cohen alcanzó 0.811, indicando un nivel elevado de acuerdo entre codificadores. Sin embargo, *ChatGPT* no fue capaz de procesar correctamente el la base de datos en este escenario. En lugar de generar los estadísticos solicitados, el modelo devolvió un mensaje de error ("mm… something seems to have gone wrong"), lo que impidió la obtención de resultados para el porcentaje de acuerdo, la tabla de contingencia y la Kappa de Cohen.

Estadístico	**Resultado *ChatGPT***	**Resultado *SPSS***	**Coincidencia**
Porcentaje de acuerdo	No calculado	0.905	No calculado
Tabla de contingencia	No calculada	Correcta	No calculada
Kappa de Cohen	No calculada	0.811	No calculada

Escenario 4: Escalabilidad extrema (límite)

En el cuarto escenario (N = 50,000, dos categorías, distribución balanceada y un nivel de desacuerdo del 9.75%), los resultados obtenidos mediante *SPSS Statistics* confirmaron la estabilidad del diseño de simulación en condiciones de gran tamaño muestral. La tabla de contingencia reflejó una distribución coherente con el nivel de error introducido, y la Kappa de Cohen alcanzó un valor de 0.805, indicando un alto grado de acuerdo entre codificadores. En contraste, *ChatGPT* no fue capaz de procesar la base de datos. En este caso, el modelo devolvió un mensaje indicando que la entrada era demasiado extensa ("El mensaje que enviaste es demasiado largo, edítalo y envíalo de nuevo"), lo que impidió la generación de cualquier estadístico.

Estadístico	**Resultado *ChatGPT***	**Resultado *SPSS***	**Coincidencia**
Porcentaje de acuerdo	No calculado	0.902	No calculado
Tabla de contingencia	No calculada	Correcta	No calculada
Kappa de Cohen	No calculada	0.805	No calculada

Escenario 5: Complejidad categorial moderada

En el escenario multicategórico balanceado (1,000 casos, tres categorías con distribución aproximadamente equivalente y un nivel de desacuerdo cercano al 10%), los resultados obtenidos mediante *IBM SPSS Statistics* mostraron una estructura de frecuencias consistente con el diseño de simulación. En concreto, la tabla de contingencia reflejó una distribución equilibrada entre categorías, y el valor de la Kappa de Cohen alcanzó 0.862, indicando un alto nivel de acuerdo entre codificadores. *ChatGPT* reprodujo correctamente la estructura de la tabla de contingencia, coincidiendo exactamente con las frecuencias observadas en *SPSS*. Sin embargo, se identificaron discrepancias en los cálculos derivados. El porcentaje de acuerdo fue estimado por *ChatGPT* en 0.905, mientras que el valor correcto obtenido en *SPSS* fue 0.908. De forma similar, la Kappa de Cohen calculada por el modelo (0.857) fue ligeramente inferior al valor de referencia (0.862).

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	0.905	0.908	Incorrecto
Tabla de contingencia	Correcta	Correcta	Correcta
Kappa de Cohen	0.857	0.862	Incorrecta

Escenario 6: Alta complejidad categorial

En el escenario con cinco categorías (N = 1,000) y distribución balanceada, los resultados obtenidos mediante *SPSS* mostraron una estructura de frecuencias consistente con el diseño de simulación, así como un alto nivel de acuerdo entre codificadores. En concreto, la Kappa de Cohen alcanzó un valor de 0.885, indicando una elevada fiabilidad intercodificadora. *ChatGPT* reprodujo correctamente la tabla de contingencia, coincidiendo exactamente con las frecuencias observadas en *SPSS* en todas las celdas. Sin embargo, se identificaron discrepancias en los cálculos derivados. El porcentaje de acuerdo fue estimado por *ChatGPT* en 0.892, mientras que el valor correcto obtenido en *SPSS* fue 0.908. Asimismo, la Kappa de Cohen calculada por el modelo (0.865) fue inferior al valor de referencia (0.885).

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	0.892	0.908	Incorrecto
Tabla de contingencia	Correcta	Correcta	Correcta
Kappa de Cohen	0.865	0.885	Incorrecta

Escenario 7: Distribución desbalanceada (binaria)

En el escenario con dos categorías y distribución desbalanceada (90% vs. 10%), los resultados obtenidos mediante *SPSS Statistics* reflejaron la asimetría esperada en la distribución de frecuencias, con una clara predominancia de la categoría mayoritaria. En este contexto, la Kappa de Cohen alcanzó un valor de 0.636, indicando un nivel moderado de acuerdo entre codificadores, inferior al observado en escenarios balanceados. *ChatGPT* no logró reproducir correctamente los resultados en este escenario. En particular, el modelo generó una tabla de contingencia incorrecta, asumiendo implícitamente una distribución balanceada entre categorías, lo que derivó en errores sustanciales tanto en el porcentaje de acuerdo como en el valor de la Kappa de Cohen.

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	0.904	0.915	Incorrecto
Tabla de contingencia	Incorrecta	Correcta	Incorrecta
Kappa de Cohen	0.808	0.636	Incorrecta

Escenario 8: Distribución desbalanceada (multicategórica)

El escenario 8 se diseñó para evaluar el rendimiento de *ChatGPT* en un contexto caracterizado por una distribución desbalanceada de las categorías. La muestra estuvo compuesta por N = 1,000 unidades de análisis, clasificadas en tres categorías con una distribución aproximada del 70%, 20% y 10%, respectivamente. Asimismo, se introdujo un nivel de desacuerdo entre codificadores del 9.2%, con el objetivo de simular condiciones de codificación imperfecta pero plausibles en estudios reales.

Los resultados reflejan la asimetría esperada en la distribución de frecuencias, con una clara predominancia de la categoría mayoritaria cuando el análisis se hace manualmente mediante *SPSS*. En este contexto, la Kappa de Cohen alcanzó un valor de 0.813, indicando un nivel sustancial de acuerdo entre codificadores, aunque condicionado por el desbalance en la distribución de las categorías. *ChatGPT* no logró reproducir correctamente los resultados en este escenario. En particular, el modelo generó una tabla de contingencia incorrecta, asumiendo implícitamente una distribución más equilibrada entre categorías, lo que derivó en errores tanto en el porcentaje de acuerdo como en el valor de la Kappa de Cohen.

Estadístico	**Resultado *ChatGPT***	**Resultado *SPSS***	**Coincidencia**
Porcentaje de acuerdo	0.904	0.908	Incorrecto
Tabla de contingencia	Incorrecta	Correcta	Incorrecta
Kappa de Cohen	0.808	0.813	Incorrecta

Escenario 9: Alto desacuerdo entre codificadores

El escenario 9 se diseñó para evaluar el rendimiento de *ChatGPT* en un contexto multicategórico con distribución equilibrada de las categorías. La muestra estuvo compuesta por N = 1,000 unidades de análisis, clasificadas en tres categorías con una distribución aproximada del 33% cada una. Asimismo, se introdujo un nivel de desacuerdo entre codificadores del 9.2%, con el objetivo de simular condiciones de codificación imperfecta pero realistas en estudios de análisis de contenido. Los resultados reflejan una distribución homogénea de las categorías cuando el análisis se realiza correctamente en *SPSS*. En este contexto, la Kappa de Cohen alcanzó un valor de 0.591, indicando un nivel moderado de acuerdo entre codificadores. *ChatGPT* no logró reproducir correctamente los resultados en este escenario. En particular, el modelo generó una tabla de contingencia incorrecta, inflando artificialmente el nivel de acuerdo observado y produciendo estimaciones sesgadas. Como consecuencia, sobreestimó tanto el porcentaje de acuerdo como la Kappa de Cohen.

Estadístico	**Resultado *ChatGPT***	**Resultado *SPSS***	**Coincidencia**
Porcentaje de acuerdo	0.884	0.727	Incorrecto
Tabla de contingencia	Incorrecta	Correcta	Incorrecta
Kappa de Cohen	0.809	0.591	Incorrecta

Escenario 10: Condición extrema (máxima complejidad)

El escenario 10 se diseñó para evaluar el rendimiento de *ChatGPT* en un contexto de alta complejidad caracterizado por un gran tamaño muestral y una distribución fuertemente desbalanceada de las categorías. La muestra estuvo compuesta por N = 10,000 unidades de análisis, clasificadas en cinco categorías con una distribución aproximada del 50%, 20%, 15%, 10% y 5%, respectivamente. Asimismo, se introdujo un nivel elevado de desacuerdo entre codificadores del 28.79%. Los resultados manuales reflejan claramente la asimetría en la distribución de frecuencias, con un claro predominio de la categoría mayoritaria. En este contexto, el porcentaje de acuerdo alcanzó el 71.21%, mientras que la Kappa de

Cohen se situó en 0.601, indicando un nivel moderado de acuerdo entre codificadores, a pesar del elevado nivel de error introducido. *ChatGPT* no logró reproducir los resultados en este escenario. En lugar de generar estimaciones, el modelo directamente falló en la ejecución del cálculo, devolviendo un error ("Hmm…something seems to have gone wrong"). Esto evidencia limitaciones importantes en el manejo de grandes volúmenes de datos y escenarios de mayor complejidad.

Estadístico	Resultado *ChatGPT*	Resultado *SPSS*	Coincidencia
Porcentaje de acuerdo	No calculado	0.712	No calculado
Tabla de contingencia	No calculada	Correcta	No calculada
Kappa de Cohen	No calculada	0.601	No calculada

4. Discusión de los resultados

Tabla 2. Resumen de escenarios de simulación

Escenario	% Acuerdo	Tabla contingencia	Kappa	Evaluación global
1	Correcto	Correcto	Correcto	Correcto
2	Incorrecto	Incorrecto	Incorrecto	Incorrecto
3	No calculado	No calculado	No calculado	No calculado
4	No calculado	No calculado	No calculado	No calculado
5	Incorrecto	Incorrecto	Incorrecto	Incorrecto
6	Incorrecto	Incorrecto	Incorrecto	Incorrecto
7	Incorrecto	Incorrecto	Incorrecto	Incorrecto
8	Incorrecto	Incorrecto	Incorrecto	Incorrecto
9	Incorrecto	Incorrecto	Incorrecto	Incorrecto
10	No calculado	No calculado	No calculado	No calculado

Los resultados obtenidos permiten identificar un patrón claro en el rendimiento de *ChatGPT* en tareas de cálculo de fiabilidad intercodificadora. En condiciones simples (caracterizadas por tamaños muestrales reducidos, variables binarias y distribuciones balanceadas) el modelo es capaz de reproducir con precisión los estadísticos clave, alcanzando coincidencias exactas con los resultados obtenidos mediante *SPSS*. Sin embargo, este nivel de precisión no se mantiene cuando aumenta la complejidad del diseño.

A medida que se incrementa el tamaño muestral, se introduce mayor número de categorías o se altera la distribución de las mismas, el rendimiento de *ChatGPT* se deteriora de forma sistemática. En escenarios de complejidad moderada, el modelo puede reproducir correctamente la estructura de la tabla de contingencia, pero presenta desviaciones en los cálculos derivados, como el porcentaje de acuerdo o la Kappa de Cohen. Estas discrepancias, aunque en algunos casos son pequeñas, evidencian limitaciones en la precisión computacional incluso cuando la representación de los datos es correcta.

En escenarios más exigentes (especialmente aquellos con distribuciones desbalanceadas o mayor número de categorías) los errores se vuelven estructurales. *ChatGPT* tiende a asumir distribuciones más equilibradas de lo que realmente

son, lo que conduce a una sobreestimación sistemática del acuerdo entre codificadores. Este patrón se observa claramente en los escenarios 7, 8 y 9, donde tanto la tabla de contingencia como los estadísticos derivados son incorrectos, reflejando una dificultad para captar la estructura empírica de los datos. Finalmente, en condiciones de alta carga computacional, como grandes tamaños muestrales (≥10,000 casos), el modelo directamente falla en la ejecución de la tarea, devolviendo errores o indicando limitaciones en la longitud de entrada. Esto sugiere que, más allá de problemas de precisión, existen restricciones operativas relevantes para el uso de *ChatGPT* como herramienta de análisis estadístico en contextos de gran escala.

Estos resultados indican que el uso de *ChatGPT* para el cálculo de métricas de fiabilidad en análisis de contenido debe abordarse con cautela. Si bien puede ser útil en contextos simples o como herramienta exploratoria, su rendimiento es inconsistente y se degrada notablemente en escenarios más realistas y complejos. Esto refuerza la necesidad de validar sistemáticamente los resultados obtenidos con modelos de lenguaje frente a software estadístico consolidado antes de su uso en investigación empírica.

Este estudio presenta varias limitaciones que deben ser consideradas al interpretar los resultados. En primer lugar, el rendimiento observado de *ChatGPT* podría estar condicionado por el diseño del *prompt* utilizado. La evidencia previa sugiere que los modelos de lenguaje son altamente sensibles a variaciones en la formulación de instrucciones, por lo que modificaciones en el *prompt* podrían potencialmente alterar la precisión de los resultados obtenidos. En este sentido, futuros estudios deberían explorar de manera sistemática el efecto de diferentes estrategias de *prompting* sobre el rendimiento en tareas estadísticas.

En segundo lugar, el análisis se ha centrado en una única versión de *ChatGPT*, lo que limita la generalización de los hallazgos. Dado el rápido desarrollo de los modelos de lenguaje, sería pertinente comparar el rendimiento entre diferentes versiones dentro de la misma familia (e.g., distintas versiones de *GPT*), así como con otros modelos disponibles, con el fin de evaluar la consistencia y robustez de los resultados. Por último, este trabajo debe entenderse como un estudio inicial que ofrece una primera aproximación empírica al rendimiento de los modelos de lenguaje en el cálculo de métricas de fiabilidad intercodificadora. Investigaciones futuras podrían ampliar este enfoque incorporando diseños experimentales más complejos, mayor variedad de métricas estadísticas y condiciones adicionales que permitan establecer conclusiones más generales sobre el uso de inteligencia artificial en la computación de estadísticos relevantes del análisis de contenido.

5. Referencias

Amin, M. M.; **Mao, R.**; **Cambria, E.**; **Schuller, B. W.** (2024). A wide evaluation of *ChatGPT* on affective computing tasks. *IEEE Transactions on Affective Computing*.
https://doi.org/10.1109/TAFFC.2024.3419593

Andersen, J. P.; **Degn, L.**; **Fishberg, R.**; **Graversen, E. K.**; **Horbach, S. P.**; **Schmidt, E. K.**; **Schneider, J. W.**; **Sørensen, M. P.** (2025). Generative artificial intelligence (GenAI) in the research process: A survey of researchers' practices and perceptions. *Technology in Society, 81*, 102813.
https://doi.org/10.1016/j.techsoc.2025.102813

Belal, M.; **She, J.**; **Wong, S.** (2023). Leveraging *ChatGPT* as a text annotation tool for sentiment analysis. *arXiv*.
https://arxiv.org/abs/2306.17177

Chubb, L. A. (2023). Me and the machines: Possibilities and pitfalls of using artificial intelligence for qualitative data analysis. *International Journal of Qualitative Methods, 22*, 16094069231193593.
https://doi.org/10.1177/16094069231193593

Cook, D. A.; **Ginsburg, S.**; **Sawatsky, A. P.**; **Kuper, A.**; **D'Angelo, J. D.** (2025). Artificial intelligence to support qualitative data analysis: Promises, approaches, pitfalls. *Academic Medicine, 100*(10), 1134–1149.
https://doi.org/10.1097/ACM.0000000000006134

Dobler, D.; **Binder, H.**; **Boulesteix, A. L.**; **Igelmann, J. B.**; **Köhler, D.**; **Mansmann, U.**; **Pauly, M.**; **Scherag, A.**; **Schmid, M.**; **Tawil, A. A.**; **Weber, S.** (2025). *ChatGPT* as a tool for biostatisticians: A tutorial on applications, opportunities, and limitations. *Statistics in Medicine, 44*(23–24), e70263.
https://doi.org/10.1002/sim.70263

Fatouros, G.; **Soldatos, J.**; **Kouroumali, K.**; **Makridis, G.**; **Kyriazis, D.** (2023). Transforming sentiment analysis in the financial domain with *ChatGPT*. *Machine Learning with Applications, 14*, 100508.
https://doi.org/10.1016/j.mlwa.2023.100508

Fu, Z.; **Hsu, Y. C.**; **Chan, C. S.**; **Lau, C. M.**; **Liu, J.**; **Yip, P. S. F.** (2024). Efficacy of *ChatGPT* in Cantonese sentiment analysis: Comparative study. *Journal of Medical Internet Research, 26*, e51069.
https://doi.org/10.2196/51069

Gilardi, F.; **Alizadeh, M.**; **Kubli, M.** (2023). *ChatGPT* outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*.
https://doi.org/10.1073/pnas.2305016120

Goyanes, M.; **De-Marcos, L.** (2025). Protocolo metodológico para el desarrollo de análisis de contenido asistido por inteligencia artificial fiable y válido: Guía práctica con *ChatGPT*. *Anuario ThinkEPI, 19*.
https://doi.org/10.3145/thinkepi.2025.e19a07

Goyanes, M.; **Lopezosa, C.**; **Jordá, B.** (2025). Thematic analysis of interview data with *ChatGPT*: Designing and testing a reliable research protocol for qualitative research. *Quality & Quantity*.
https://doi.org/10.1007/s11135-025-02199-3

Goyanes, M.; **Piñeiro Naval, V.** (2024). Análisis de contenido en SPSS y KALPHA: Procedimiento para un análisis cuantitativo fiable con la Kappa de Cohen y el alpha de Krippendorff. *Estudios sobre el Mensaje Periodístico, 30*(1), 123–140.
https://doi.org/10.5209/esmp.92732

Grossmann, I.; **Feinberg, M.**; **Parker, D. C.**; **Christakis, N. A.**; **Tetlock, P. E.**; **Cunningham, W. A.** (2023). AI and the transformation of social science research. *Science, 380*(6650), 1108–1109.
https://doi.org/10.1126/science.adi1778

Hayes, A. F.; **Krippendorff, K.** (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.
https://doi.org/10.1080/19312450709336664

Jim, J. R.; **Talukder, M. A. R.**; **Malakar, P.**; **Kabir, M. M.**; **Nur, K.**; **Mridha, M. F.** (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 100059.
https://doi.org/10.1016/j.nlp.2024.100059

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433.
https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage.
https://doi.org/10.4135/9781071878781

Lee, L. W.; **Dabirian, A.**; **McCarthy, I. P.**; **Kietzmann, J.** (2020). Making sense of text: Artificial intelligence-enabled content analysis. *European Journal of Marketing, 54*(3), 615–644.
https://doi.org/10.1108/EJM-02-2019-0219

Lombard, M.; **Snyder-Duch, J.**; **Bracken, C. C.** (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587–604.
https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Lossio-Ventura, J. A.; **Weger, R.**; **Lee, A. Y.**; **Guinee, E. P.**; **Chung, J.**; **Atlas, L.**; **Linos, E.**; **Pereira, F.** (2024). A comparison of *ChatGPT* and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: Sentiment analysis of COVID-19 survey data. *JMIR Mental Health, 11*, e50150.
https://doi.org/10.2196/50150

Lu, Y.; **Yang, R.**; **Zhang, Y.**; **Yu, S.**; **Dai, R.**; **Wang, Z.**; **… Zhou, F.** (2025). Stateval: A comprehensive benchmark for large language models in statistics. *arXiv*. *https://arxiv.org/abs/2510.09517*

Matos, T.; **Santos, W.**; **Zdravevski, E.**; **Coelho, P. J.**; **Pires, I. M.**; **Madeira, F.** (2025). A systematic review of artificial intelligence applications in education: Emerging trends and challenges. *Decision Analytics Journal, 15*, 100571. *https://doi.org/10.1016/j.dajour.2025.100571*

Mondal, H.; **Mondal, S.**; **Mittal, P.** (2024). Evaluating large language models for selection of statistical test for research: A pilot study. *Perspectives in Clinical Research, 15*(4), 178–182. *https://doi.org/10.4103/picr.picr_275_23*

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of *ChatGPT*. *International Journal of Qualitative Methods, 22*, 16094069231211248. *https://doi.org/10.1177/16094069231211248*

Neuendorf, K. A. (2017). *The content analysis guidebook*. SAGE. ISBN: 978 1 412979474

Nguyen, D. C.; **Welch, C.** (2026). Generative artificial intelligence in qualitative data analysis: Analyzing—or just chatting? *Organizational Research Methods, 29*(1), 3–39. *https://doi.org/10.1177/10944281251377154*

Robila, M.; **Robila, S. A.** (2020). Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies, 29*(10), 2954–2966. *https://doi.org/10.1007/s10826-019-01689-x*

Shukla, M.; **Pandey, D.**; **Kaur, S.**; **Agarwal, M.**; **Goyal, A.**; **Sharma, H.**; **Sharma, H., Jr.** (2025). Evaluating the accuracy and explanatory quality of large language models *ChatGPT*, Claude, DeepSeek, Gemini, Grok, and Le Chat in statistical test selection for hypothesis testing decisions. *Cureus, 17*(10), eXXXXX. *https://doi.org/10.7759/cureus.94949*

Törnberg, P. (2025). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review, 43*(6), 1181–1195. *https://doi.org/10.1177/08944393241286471*

Van Noorden, R.; **Perkel, J. M.** (2023). AI and science: What 1,600 researchers think. *Nature, 621*(7980), 672–675. *https://doi.org/10.1038/d41586-023-02980-0*

Xu, Y.; **Liu, X.**; **Cao, X.**; **Huang, C.**; **Liu, E.**; **Qian, S.**; **… Zhang, J.** (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(4). *https://doi.org/10.1016/j.xinn.2021.100179*