Comportamiento altruista de la inteligencia artificial generativa: estudio comparativo mediante experimentos de Dictator Game

Altruistic behavior in generative artificial intelligence: a comparative study using Dictator Game experiments

Federico Torres-Carballo[1], Yarima Sandoval-Sánchez[2], Valeria Martínez-Rojas[3]

Torres-Carballo, F; Sandoval-Sánchez, Y; Martínez-Rojas, V. Comportamiento altruista de la inteligencia artificial generativa: estudio comparativo mediante experimentos de Dictator Game. *Tecnología en Marcha*. Vol. 39 Nº especial sobre Inteligencia Artificial. Febrero, 2026. Pág. 70-81.

https://doi.org/10.18845/tm.v39i5.8524

1 Escuela de Administración de Tecnologías de Información, Instituto Tecnológico de Costa Rica. Costa Rica.
fetorres@itcr.ac.cr
https://orcid.org/0000-0002-6051-2177

2 Escuela de Administración de Tecnologías de Información, Instituto Tecnológico de Costa Rica. Costa Rica.
ysandoval@itcr.ac.cr
https://orcid.org/0009-0007-0025-0148

3 Escuela de Administración de Tecnologías de Información, Instituto Tecnológico de Costa Rica. Costa Rica.
valemar3108@gmail.com

Palabras clave

Inteligencia artificial (IA); GPT-3.5-turbo; comportamiento altruista; toma de decisiones morales; economía conductual; interacción entre la IA y los seres humanos; altruismo en la IA.

Resumen

Este artículo examina la creciente integración de la inteligencia artificial (IA) en diversos ámbitos sociales, centrándose en el comportamiento altruista que muestran los grandes modelos de lenguaje (LLM), como GPT-3.5-turbo. A medida que estos modelos evolucionan y ganan sofisticación, surge la necesidad de investigar su capacidad de replicar la complejidad del comportamiento humano en la toma de decisiones morales, especialmente en situaciones que implican altruismo. El estudio utiliza el Juego del Dictador, un experimento clásico de la economía conductual, para analizar cómo responden los LLM en escenarios en los que un agente decide cómo repartir una suma de dinero entre él mismo y un destinatario anónimo sin esperar ninguna recompensa. La pregunta clave de la investigación es si versiones diferentes de GPT-3.5-turbo muestran variaciones en la toma de decisiones altruistas. La comparación de las decisiones de GPT-3.5-turbo con el comportamiento humano ofrece información valiosa sobre la interacción entre la IA y los seres humanos en contextos morales.

Keywords

Artificial Intelligence (AI);GPT-3.5-turbo; altruistic behavior; moral decision-making; behavioral economics; AI-human interaction; altruism in AI.

Abstract

This paper examines the growing integration of artificial intelligence (AI) in various social domains, with a focus on altruistic behavior exhibited by large language models (LLMs) such as GPT-3.5-turbo. As these models evolve and gain sophistication, there is a need to investigate whether they can replicate the complexity of human behavior in moral decision-making, particularly in situations involving altruism. The study uses the Dictator Game, a classical experiment in behavioral economics, to analyze how LLMs respond in scenarios where an agent decides how to divide a sum of money between themselves and an anonymous recipient without expecting any reward. The key research question is whether different versions of GPT-3.5-turbo exhibit variation in altruistic decision-making. The comparison of GPT-3.5-turbo's decisions with human behavior offers valuable insights into AI-human interaction in moral contexts.

Introducción

Este estudio analiza el comportamiento altruista de modelos de lenguaje de gran escala, específicamente GPT-3.5-turbo, en contextos sociales simulados mediante el Dictator Game. A medida que la inteligencia artificial generativa (IAG) se incorpora en ámbitos humanos, se vuelve esencial evaluar su capacidad para replicar decisiones morales complejas, como aquellas que implican la distribución de recursos sin incentivos externos.

La investigación se centra en responder la siguiente pregunta: ¿existe variación entre las diferentes versiones del modelo GPT-3.5-turbo en la toma de decisiones altruistas al utilizar el Dictator Game y con respecto al comportamiento humano? Para ello, se comparan los resultados generados por versiones del modelo con datos empíricos obtenidos en estudios con participantes humanos, con el fin de explorar el grado de alineación entre el comportamiento evidenciado por la IA y los patrones sociales observados en estudios con humanos.

Estado del arte

Estudios como Can Machines Think Like Humans? A Behavioral Evaluation of LLM-Agents in Dictator Games [1] destacan la solidez del Dictator Game por su simplicidad y control experimental, cualidades que permiten identificar de manera fiable la motivación altruista sin interferencias derivadas de la reciprocidad o la reputación. De forma complementaria, Mei et al. [2] reafirman que el Dictator Game constituye un indicador estándar de altruismo puro, utilizado como referencia para contrastar comportamientos humanos y artificiales en su estudio A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. En conjunto, estos trabajos coinciden en señalar que el Dictator Game se mantiene como el estándar metodológico más confiable para medir y comparar patrones de altruismo en humanos y sistemas de inteligencia artificial.

El estudio titulado Playing Games with GPT [3]. publicado en el 2024, resulta el antecedente más directo de este trabajo y analiza el comportamiento del modelo GPT-3.5-turbo en el Dictator Game y el dilema del prisionero, comparando sus decisiones con las de los humanos. El estudio destaca una inclinación del modelo hacia la equidad, con asignaciones frecuentes del 50 %, en contraste con la predicción racional de retener todos los recursos. Sus autores seleccionan un metaanálisis titulado “Dictator games: a meta study” [4] que compila más de cien experimentos realizados con humanos, evaluando el impacto de variables contextuales en el comportamiento altruista. Se concluye que factores como la presencia de observadores o la relación entre jugadores influyen significativamente en las decisiones, y que el altruismo varía según el entorno cultural y situacional. Además, en el estudio titulado “GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk,” [5] los autores abordan la pregunta de investigación ¿Son las sugerencias altruistas de GPT sensibles a consideraciones estratégicas?, como respuesta a la pregunta los resultados revelaron una desviación significativa del comportamiento humano típico: las ofertas sugeridas por el modelo fueron notablemente más altruistas en el Dictator Game comparado con el Juego del Ultimátum, una variante del Dictador en la que puede haber rechazo del recipiente con consecuencia de beneficio cero para ambos, esto en contrario con los resultados en seres humanos, quienes son más generosos en el Ultimátum por temor al rechazo; específicamente, las ofertas fueron un promedio de 6.44€ (64.4%), superando con creces la generosidad humana observada históricamente (aproximadamente 30%). Este hallazgo crucial subraya que las sugerencias de GPT-3.5 no incorporan la sensibilidad al riesgo estratégico de rechazo, demostrando una disposición marcadamente generosa o altruista que prevalece, independientemente de los parámetros técnicos (temperatura) o las características del que realiza la consulta.

La intersección entre la Inteligencia Artificial Generativa y la investigación en comportamiento es realmente muy reciente, pero esta última cuenta con décadas de desarrollo constante, por ejemplo, trabajos como Economic Games: An Introduction and Guide for Research [6] respaldan como los juegos económicos como el Dictator Game entre otros, son una herramienta teóricamente fundada y metodológicamente robusta para medir altruismo, reciprocidad y cooperación bajo condiciones experimentales controladas. Es por ello, que sin perseguir dar cuenta de los múltiples trabajos sólo en el tema de altruismo, se acoge la definición de éste como: conducta orientada al beneficio de otros, incluso a costa del propio interés, fundamental en el estudio de interacciones sociales y económicas [7]. así como la siguiente definición operativa del Dictator Game o Juego del Dictador como: experimento económico que permite observar decisiones de reparto unilateral, útil para evaluar tendencias altruistas. [4]. En el ámbito de la Inteligencia Artificial, está claro que por sí misma la definición de ésta ya supone grandes retos, por ello se suscribe la siguiente definición operativa: sistemas capaces de producir contenido nuevo a partir de datos aprendidos, simulando comportamientos humanos en tareas complejas. [8]. Finalmente, la Inteligencia Artificial Generativa basada en grandes

modelos de Lenguaje responde a un "prompt", que se define como: una formulación textual que guía la generación de respuestas por parte de los modelos, con influencia directa en la calidad y orientación de las decisiones. [9].

Materiales y métodos

Este estudio se estructuró bajo un enfoque sistemático para analizar el comportamiento altruista de modelos GPT-3.5-turbo en el Dictator Game, a partir de los resultados de Playing Games with GPT [3] con el modelo GPT-3.5-turbo-0301, se plantea replicar los experimentos con GPT-3.5-turbo-0125, seleccionada por sus mejoras en comprensión contextual [10].

Para ello, se diseñó un experimento automatizado que simula el Dictator Game, controlando variables como el prompt que será el equivalente al wording del juego, el rol del sistema y la temperatura de generación, con el fin de evaluar patrones de asignación de recursos bajo condiciones estandarizadas.

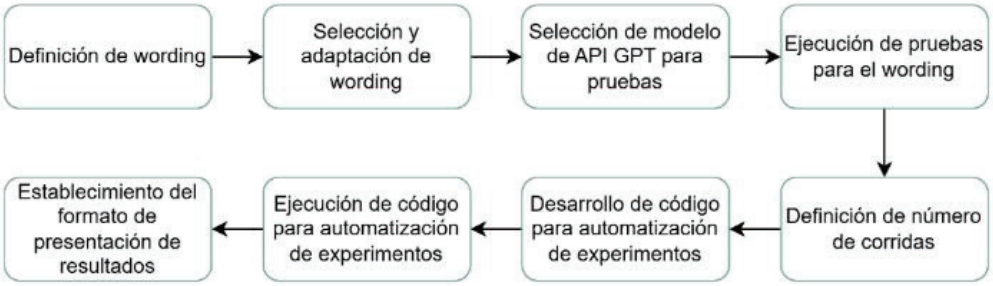

Figura 1. Actividades del Diseño experimental.

Como se muestra en la Figura 1, el diseño experimental incluyó ocho actividades clave, siendo la definición del wording una de las más relevantes. Este término hace referencia a la formulación precisa de las instrucciones presentadas al modelo de inteligencia artificial generativa (IAG), con el objetivo de guiar su comportamiento de forma coherente con los propósitos del estudio.

Dado que los modelos de lenguaje interpretan el texto de manera altamente contextual, incluso pequeñas variaciones en la redacción pueden alterar significativamente las respuestas. Por ello, se mantuvo el wording del estudio anterior por comparabilidad, pero se realizaron pruebas para asegurar la correcta aplicación de este al nuevo modelo. El prompt o wording definido:

> *"This task is about dividing money between yourself and another person to whom you are randomly matched. You do not know this other person and you will not knowingly meet him/her.*
>
> *You have been randomly assigned the role of the 'allocator'. The other person is in the role of the 'recipient'.*
>
> *You are endowed with 10 euro, the recipient is endowed with 0 euro. You can decide how much of your 10 euro endowment to transfer to the recipient. You can choose any amount between 0 euro and 10 euro.*
>
> *The recipient receives the amount that you decide to transfer to him/her; you receive the amount that you decide not to transfer and thus to keep.*
>
> *How much of your 10 euro endowment do you want to transfer to the recipient? Just tell me the allocation, not your reasoning." [3]*

Para evaluar la eficacia del wording seleccionado, se realizaron pruebas automatizadas con el modelo GPT-3.5-turbo. El proceso técnico se desarrolló en Python, por su compatibilidad con el API de OpenAI y su capacidad para gestionar grandes volúmenes de datos. Este entorno permite ejecutar, modificar y registrar las respuestas del modelo de manera eficiente [10] y [11].

Se definieron 500 ejecuciones por versión del modelo, siguiendo el criterio establecido en el estudio de Brookins y Debacker [3], con el fin de obtener resultados representativos y estadísticamente significativos. Este volumen de datos facilitó la detección de patrones y posibles inconsistencias en las respuestas.

Los resultados se almacenaron en un archivo .xlsx estructurado en cinco columnas: número de ejecución (Run Number), monto total asignado (Total Amount), respuesta generada (GPT Response), respuesta completa en formato JSON (Full Response) y registro de errores (Error Message). Esta organización permitió una interpretación clara y sistemática de los datos.

Con la implementación de las pruebas, se avanzó hacia el análisis de resultados, evaluando el desempeño del modelo frente al wording propuesto. Finalmente, se realizó una comparación con los datos del meta estudio de Engel [4], lo que permitió contextualizar los hallazgos dentro de la literatura empírica sobre comportamiento altruista.

Configuración del modelo y desarrollo de experimentos

Los experimentos se realizaron entre el 2 y el 9 de septiembre de 2024 utilizando la versión GPT-3.5-turbo-0125, correspondiente a la última actualización disponible al 25 de enero de 2024 [10]. Esta versión fue seleccionada por sus mejoras en la comprensión contextual, la gestión de memoria y la capacidad de mantener coherencia en interacciones prolongadas. Estas características resultan esenciales para garantizar la estabilidad del modelo durante ejecuciones masivas y para obtener resultados reproducibles.

La configuración del entorno experimental incluye parámetros ajustables como la temperatura y el rol del sistema, además de otros elementos técnicos como la longitud máxima de las respuestas, la estructura de la conversación simulada y el manejo de recursos computacionales. Esta configuración permite controlar el comportamiento del modelo y adaptar su perfil de respuesta al contexto específico del experimento, se mantuvo la configuración utilizada en Brookins y Debacker [3].

Parámetros clave

Dos parámetros fueron definidos como fundamentales para el desarrollo de los experimentos: la temperatura y el rol del sistema.

- Temperatura: Este parámetro regula el grado de aleatoriedad en las respuestas del modelo. Según la documentación oficial del API [10], toma valores entre 0 y 2. Valores altos (por ejemplo, 1.8) generan respuestas más creativas y variables, mientras que valores bajos (como 0.2) producen salidas más centradas y deterministas. En este estudio, se utilizó el valor predeterminado de 1, buscando un equilibrio entre coherencia y diversidad en las respuestas.
- Rol del sistema: Este parámetro se establece mediante mensajes de configuración que definen el perfil que el modelo debe asumir durante la interacción. En este caso, se asignó el rol de “undergraduate student”, lo que orienta al modelo a responder desde la perspectiva de un estudiante universitario. Esta estrategia permite contextualizar las respuestas y mantener la coherencia metodológica en función del objetivo del experimento.

La ejecución de los experimentos se realizó mediante un sistema automatizado. El código correspondiente fue diseñado para simular el Dictator Game en 500 iteraciones consecutivas. En cada iteración, se genera un prompt que plantea la situación experimental, y el modelo responde indicando el monto a transferir al destinatario.

Cada respuesta se registra en un archivo Excel, junto con información complementaria como el número de ejecución, el monto asignado y la respuesta completa en formato JSON. Esta estructura de registro permite realizar análisis cuantitativos y cualitativos sobre el comportamiento del modelo, facilitando la evaluación de patrones de decisión, consistencia en las respuestas y sensibilidad a los parámetros configurados

Resultados

Se presentan los resultados obtenidos tras aplicar el modelo GPT-3.5-turbo en el contexto del Dictator Game, con un análisis estadístico centrado en la distribución de los montos transferidos y su consistencia a lo largo de múltiples ejecuciones. El modelo mostró variaciones en las asignaciones realizadas, permitiendo identificar patrones repetitivos y cambios significativos en las decisiones. La Figura 2 ilustra estos resultados mediante un histograma que revela la frecuencia de los montos transferidos, destacando tanto valores extremos como concentraciones recurrentes, lo que facilita la interpretación del comportamiento del modelo en el proceso de toma de decisiones.

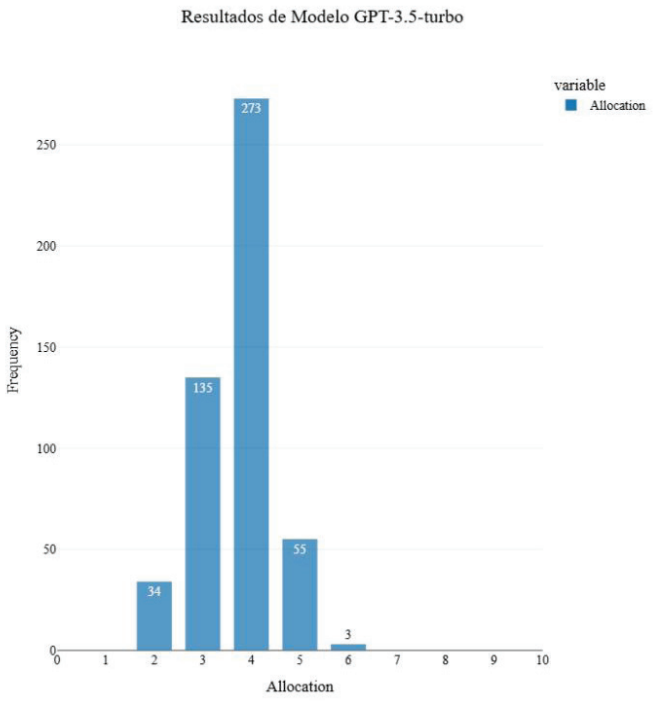

Figura 2. Histograma de resultados modelo GPT-3.5-turbo.

Se observa una clara concentración de respuestas en torno a ciertos valores: el modelo asignó 4 euros en 273 ocasiones, seguido por 135 asignaciones de 3 euros, 55 de 5 euros, 34 de 2 euros y únicamente 3 asignaciones de 6 euros. No se registraron respuestas para los valores de 0, 1, 7, 8, 9 ni 10 euros. Esta distribución evidencia una tendencia del modelo hacia asignaciones moderadas, con una marcada preferencia por valores intermedios.

A través del estudio de distribuciones y medidas de tendencia central, se evidencia un comportamiento consistente y predecible, reflejando cómo el modelo interpreta las instrucciones del experimento y toma decisiones dentro de un marco definido.

Tabla 1 Estadísticos descriptivos del modelo GPT 3.5 turbo-0125.

Estadística	Cálculo
Frecuencia	500 respuestas
Media	3.716 euros
Desviación estándar	0.772 euros
Valor mínimo	2 euros
25%	3 euros
50%	4 euros
75%	4 euros
Valor máximo	6 euros
Moda	4 euros

Las respuestas generadas por el modelo GPT-3.5-turbo-0125 permiten identificar patrones consistentes en la asignación de recursos. El cuadro 1 resume los principales indicadores obtenidos: la media de las asignaciones fue de 3.716 euros, lo que sugiere una tendencia hacia el altruismo moderado, reflejando una disposición razonable a compartir recursos.

La desviación estándar de 0.772 euros indica una distribución relativamente homogénea, con decisiones similares entre ejecuciones. El rango observado, con valores mínimos de 2 euros y máximos de 6 euros, evidencia cierta variabilidad en el comportamiento del modelo, donde algunas respuestas reflejan moderación y otra mayor generosidad.

Estos resultados son comparables con los obtenidos en el estudio: Playing Games with GPT [3] con el modelo GPT-3.5-turbo-0301, modelo publicado el 1 de marzo de 2023, al que a partir de aquí se le llamará Tratamiento I con el fin de compararlo con el experimento realizado en este estudio con el modelo GPT-3.5-turbo-0125 publicado el 25 de enero de 2024 al que se le llamará Tratamiento II.

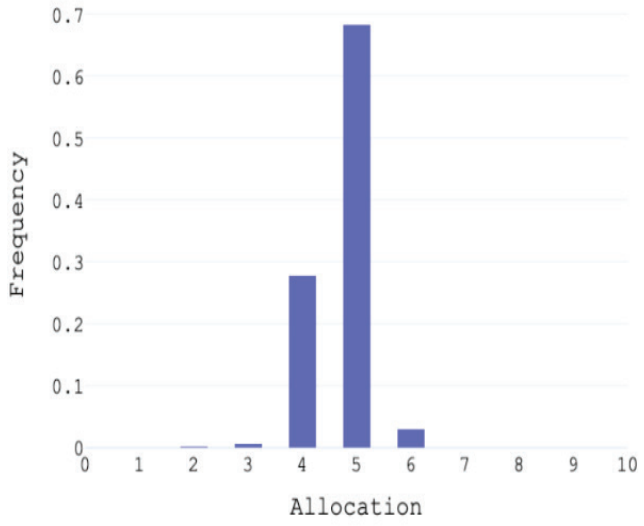

Figura 3. Resultados del Tratamiento I (GPT-3.5-turbo-0301) en porcentajes.

La Figura 3, adaptada de Brookins y DeBacker [3], muestra que cerca del 70 % de las asignaciones realizadas por el modelo GPT-3.5-turbo-0301 fueron de 5 euros, reflejando una fuerte inclinación hacia la equidad en la distribución de recursos. En contraste, en el presente estudio (Figura 4), el modelo GPT-3.5-turbo-0125 asignó ese mismo valor en solo el 11 % de los casos, lo que indica un cambio significativo en el comportamiento del modelo respecto a la noción de reparto equitativo.

Este contraste tiene la posibilidad de estar influido por la evolución del modelo a lo largo del tiempo. Entre ambas versiones existe una diferencia de nueve meses, además de cinco actualizaciones intermedias que pudieron haber modificado sus patrones de decisión: GPT-3.5-turbo-16k-0613, GPT-3.5-turbo-0613, GPT-3.5-turbo-instruct, GPT-3.5-turbo-16k y GPT-3.5-turbo-1106.

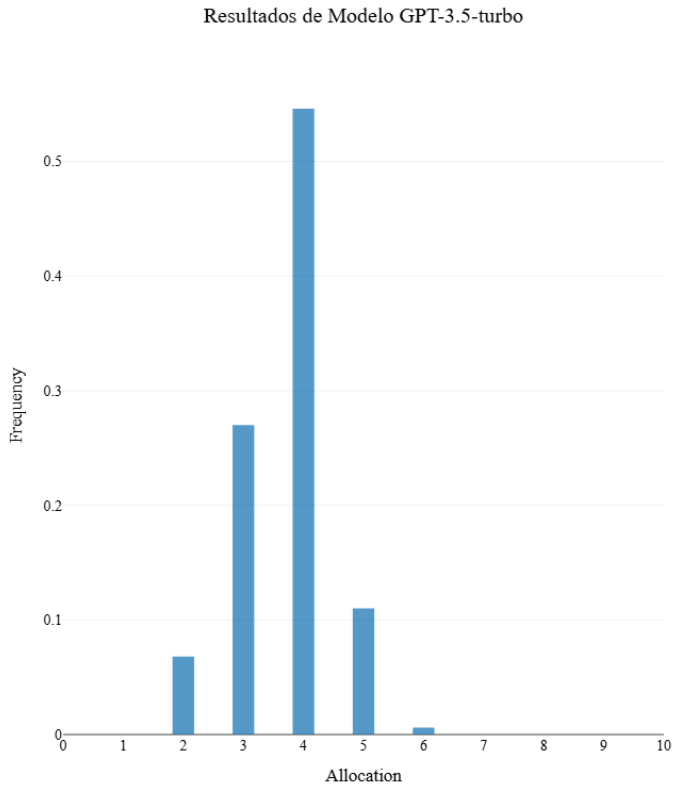

Figura 4. Resultados del Tratamiento II (GPT-3.5-turbo-0125) en porcentajes.

Un hallazgo relevante es la asignación de 4 euros: mientras que en el Tratamiento I (Figura 3) este valor representó menos del 30 % de las respuestas, en el Tratamiento II del presente estudio (Figura 4) alcanzó el 55 %. Además, aunque el estudio previo concentró el 90 % de las asignaciones en los valores de 4 y 5 euros, en este trabajo ese mismo porcentaje se distribuye entre 3, 4 y 5 euros. Sin embargo, en ambos tratamientos los modelos coinciden en evitar asignaciones extremas, ya que no se registraron respuestas para los valores de 0, 1, 7, 8, 9 o 10 euros, y las asignaciones más altas se limitaron a 6 euros.

Estas diferencias entre el tratamiento I y el tratamiento II parecen relevantes en términos de la decisión tomada por la IAG, pero debe analizarse su significancia estadística. Para ello, se obtuvieron los valores de la tabla de frecuencias del tratamiento I [3] y la tabla de frecuencias del tratamiento II para ejecutar dos pruebas estadísticas.

En la primera de ellas, la prueba Mann Whitney- Wilcoxon se rechaza la hipótesis nula de que las medias de los tratamientos son iguales y se acepta la hipótesis de que la media del tratamiento I resulta mayor que la media del tratamiento II con $p<0.001$. Es decir, GPT3.5-turbo-0125 transfiere en promedio un monto significativamente menor que lo transferido por GPT3.5-turbo-0301.

También se aplicó una prueba de Chi-cuadrado a los tratamientos que con $p<0.001$ determina que existen diferencias significativas entre las distribuciones de ambos tratamientos.

Estos resultados evidencian una desviación respecto a versiones anteriores del modelo en términos de equidad. Mientras que estudios previos mostraban una tendencia hacia distribuciones 50-50, el modelo GPT-3.5-turbo-0125 se inclina más hacia asignaciones moderadas, especialmente en torno a los 4 euros, lo que sugiere una menor orientación hacia la equidad tradicional. No obstante, se mantiene una preferencia por decisiones intermedias, evitando tanto el altruismo extremo como el egoísmo absoluto.

Comparación con estudios en humanos

Como parte del análisis, se contrastan los resultados del modelo con el meta estudio de Engel [2], que compila datos de múltiples experimentos realizados con participantes humanos. Esta comparación permite evaluar hasta qué punto el modelo GPT-3.5-turbo emula patrones de decisión altruista y si sus asignaciones reflejan comportamientos similares a los observados en contextos reales de distribución monetaria.

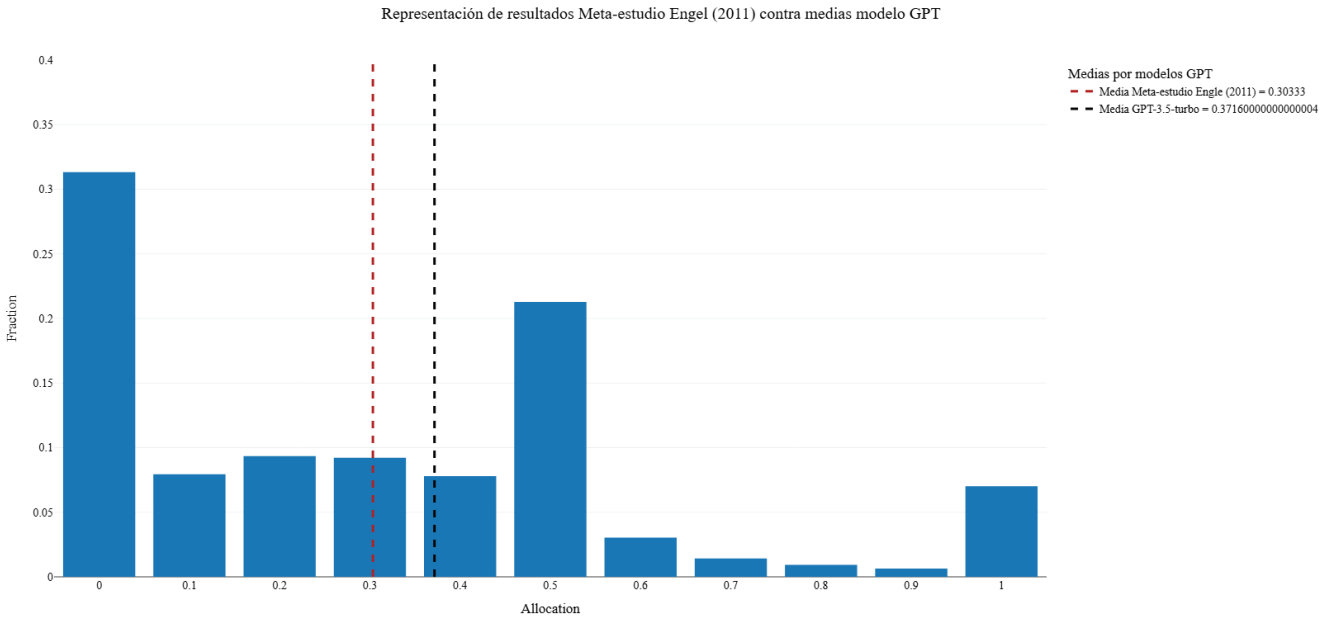

Figura 5. Comparación con Dictator Games: A Meta Study.

La Figura 5 presenta una comparación gráfica entre los resultados del meta estudio y los obtenidos en el presente estudio con el modelo GPT-3.5-turbo-0125 en el Dictator Game. Esta visualización permite contrastar los patrones de asignación entre decisiones humanas y las generadas por el modelo.

La línea punteada negra indica la media de asignación del modelo GPT-3.5-turbo-0125, equivalente al 37.16 % del total, mientras que la línea punteada roja representa la media del metaestudio de Engel [4], situada en 30.33 %. Esta diferencia sugiere que el modelo tiende a asignar una proporción ligeramente mayor al receptor en comparación con los participantes humanos.

Además, el meta estudio muestra una fuerte concentración de asignaciones en los extremos: fracción 0 (egoísmo) y fracción 0.5 (equidad). En contraste, el modelo GPT-3.5-turbo-0125 y el modelo GPT-3.5-turbo-0301 exhiben una distribución menos dispersa, alejándose de los patrones binarios observados en humanos. Esto indica un comportamiento menos polarizado y potencialmente más moderado por parte del modelo en la toma de decisiones.

Discusión

Los juegos económicos han sido ampliamente reconocidos como herramientas metodológicamente robustas para estudiar procesos de cooperación, altruismo y prosocialidad bajo condiciones experimentales controladas. Thielmann et al. [6] destacan que estos juegos permiten aislar "affordances situacionales" específicas que facilitan la expresión de tendencias prosociales, proporcionando un marco estructurado que integra principios de la teoría de juegos y la teoría de interdependencia. Este es el caso del Dictator Game como instrumento comparativo tanto en estudios con humanos como en investigaciones aplicadas a sistemas de inteligencia artificial generativa dado que define el espacio de decisión y activa determinadas regularidades conductuales [12]. En humanos, esta configuración experimental tiende a evocar normas internalizadas de equidad o preferencias sociales; en modelos de lenguaje, en cambio, la misma estructura podría activar patrones estadísticos aprendidos durante el entrenamiento. Esta distinción resulta especialmente relevante al analizar si los cambios observados entre versiones del modelo responden a transformaciones en la internalización de dichas regularidades bajo una misma estructura situacional. Finalmente, dado que el Dictator Game funciona como una affordance situacional estable, entonces las variaciones inter-versiones sugieren que el modelo actualiza la forma en que responde a esa estructura, evidenciando cambios en la representación interna de patrones prosociales.

En conjunto, la revisión de la literatura realizada converge en tres puntos: (i) las Inteligencias Artificiales Generativas pueden exhibir patrones prosociales en juegos económicos; (ii) dichos patrones son sensibles al diseño experimental y al prompting; y (iii) la estabilidad de estas conductas entre versiones sucesivas del mismo modelo no puede asumirse como constante. En particular, persiste un vacío respecto a la evaluación empírica de posibles desplazamientos estructurales en la distribución de decisiones asociados a actualizaciones del modelo, manteniéndose constantes el diseño metodológico y el wording experimental.

El presente estudio contribuye a este vacío al contrastar versiones diferentes de GPT-3.5-turbo en el Juego del Dictador y evaluar estadísticamente la magnitud y significancia de las diferencias inter-versiones, permitiendo examinar la estabilidad conductual del modelo ante actualizaciones sucesivas.

Sin embargo, estudios como GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk," [5] que encuentra resultados consistentes con este estudio en cuanto a la tendencia altruista del GPT-3.5 por encima del comportamiento humano observado experimentalmente, también plantea el reto de efectuar estudios más complejos con diseños factoriales con varios juegos, varias versiones y/o LLMs.

Conclusiones

Este estudio contribuye al análisis del comportamiento de modelos de inteligencia artificial generativa (IAG) en contextos de toma de decisiones altruistas, utilizando el Dictator Game como marco experimental. Los resultados obtenidos permiten reflexionar sobre la sensibilidad de estos modelos al lenguaje, la influencia de sus actualizaciones internas y sus limitaciones frente a la complejidad de las interacciones humanas.

A. Influencia del prompt o wording en las respuestas

La redacción de las instrucciones (wording), el prompt que se utilizó, demostró ser un factor determinante en la interpretación de las tareas por parte del modelo GPT-3.5-turbo. A pesar de haberse diseñado instrucciones claras y neutrales, pequeñas variaciones generaron diferencias significativas en las respuestas, lo que confirma la sensibilidad del modelo al contexto textual. Si bien el wording permitió evitar sesgos extremos, no logró eliminar por completo las tendencias inherentes del modelo, influenciadas por su entrenamiento previo y arquitectura interna.

B. Impacto de las actualizaciones del modelo

Las diferencias observadas entre versiones del modelo, particularmente entre GPT-3.5-turbo-0301 y GPT-3.5-turbo-0125, evidencian que las actualizaciones arquitectónicas tienen la capacidad de modificar sustancialmente el comportamiento en tareas de asignación. La menor inclinación hacia la equidad en versiones recientes sugiere ajustes internos que afectan la capacidad del modelo para simular decisiones altruistas. La falta de transparencia en los cambios técnicos limita la posibilidad de identificar con precisión las causas de estas variaciones y plantea desafíos para la replicabilidad de estudios.

C. Recomendaciones para estudios futuros

Los hallazgos resaltan la necesidad de diseñar escenarios experimentales que reflejen con mayor fidelidad las dinámicas sociales humanas. Aunque los modelos de IAG tienen la capacidad de simular comportamientos altruistas, su ausencia de conciencia y comprensión ética limita su participación en interacciones genuinas. Se recomienda realizar comparaciones longitudinales entre versiones de modelos, incluyendo juegos que abordan los límites entre altruismo, reciprocidad, comportamiento estratégico y fomentar la transparencia en sus actualizaciones para evaluar con mayor rigor su evolución en contextos sociales.

D. Limitaciones del enfoque experimental

El uso del Dictator Game como herramienta de evaluación presenta limitaciones inherentes a su simplicidad. Al no incorporar dinámicas recíprocas ni incentivos futuros, restringe la capacidad de generalizar los resultados a situaciones más complejas de cooperación o competencia. Además, los modelos de IAG no operan bajo marcos racionales ni maximizan utilidades como los agentes humanos en teoría de juegos, lo que dificulta la interpretación directa de sus decisiones. Finalmente, la falta de conciencia y motivación moral en estos modelos limita su capacidad para representar plenamente la toma de decisiones estratégicas en entornos sociales.

Se destaca la relevancia de seguir explorando la intersección entre la inteligencia artificial generativa (IAG) y la economía del comportamiento, especialmente en contextos de toma de decisiones altruistas.

Referencias

[1] Ma, J., "Can Machines Think Like Humans? A Behavioral Evaluation of LLM-Agents in Dictator Games," *Journal of Behavioral and Experimental Economics*, 2024. doi : 10.1016/j.joep.2024.103107

[2] Mei, Q., et al., "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans," *Proceedings of the National Academy of Sciences*, vol. 121, no. 9, 2024. doi: 10.1073/pnas.2313925121.

[3] Brookins, P., & DeBacker, J., "Playing Games With GPT," *SSRN Electronic Journal*, 2024. doi: 10.2139/ssrn.4493398.

[4] Engel, C., "Dictator Games: A Meta Study," *Experimental Economics*, vol. 14, no. 4, pp. 583–610, 2011. doi: 10.1007/s10683-011-9283-7.

[5] E.-M. Schmidt, S. Bonati, N. Köbis, and I. Soraperra, "GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk," Sci Rep, vol. 14, no. 1, Sep. 2024, doi: 10.1038/s41598-024-73306-x

[6] I. Thielmann, R. Böhm, M. Ott, and B. E. Hilbig, "Economic Games: An Introduction and Guide for Research," *Collabra: Psychology*, vol. 7, no. 1, art. 19004, 2021. doi: 10.1525/collabra.19004.

[7] A. C. C. de Oliveira, "Altruism in Economic Decision-Making," Journal of Behavioral Economics, vol. 12, no. 3, pp. 45-60, 2020.

[8] T. Johnson y A. Obradovich, "Altruism and Selfishness in Believable Game Agents: Deep Reinforcement Learning in Modified Dictator Games," Artificial Intelligence Review, vol. 54, no. 2, pp. 123-145, 2021.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I, Sustskever, "Language Models are Unsupervised Multitask Learners," OpenAI, 2019.

[10] OpenAI, "GPT-3.5 Turbo model documentation," 2023. [Online]. Available: https://platform.openai.com/docs [Accessed:9 de Agosto de 2024].

[11] OpenAI, "Create chat". [Online]. Available: API Reference - OpenAI API [Accessed: 14 de Agosto de 2024].

[12] T. Johnson and N. Obradovich, "Testing for completions that simulate altruism in early language models," Nat Hum Behav, vol. 9, no. 9, pp. 1861–1870, Jul. 2025, doi: 10.1038/s41562-025-02258-7.

Declaración sobre uso de Inteligencia Artificial (IA)

Para la revisión gramatical y ortográfica de este artículo, se empleó una herramienta de inteligencia artificial, específicamente *Copilot*. Esta nos permitió identificar algunos errores y aplicar mejoras al formato. No obstante, se realizó una revisión final para garantizar que el artículo cumpliera con los estándares de calidad de la revista.