Revista Latinoamericana de Tecnología Educativa, 25(2)
https://doi.org/10.17398/1695-288X.25.2.225

RELATEC
Revista Latinoamericana de Tecnología Educativa

ARTÍCULO / ARTICLE

Evaluación de metáforas visuales sobre convivencia escolar con inteligencia artificial generativa y juicio experto

Evaluation of Visual Metaphors of School Coexistence Using Generative Artificial Intelligence and Expert Judgment

Karla Karina Ruiz-Mendoza[1], Luis Horacio Pedroza-Zúñiga[2] y Héctor Manuel Rodriguez-Figueroa[3]

Recibido: 1 de diciembre de 2025
Revisado: 5 de mayo de 2026
Aceptado: 8 de mayo de 2026

Dirección de los autores:

[1] Facultad de Humanidades y Ciencias Sociales. Universidad Autónoma de Baja California (UABC), Calzada Universidad #14418, Mesa de Otay, C.P. 22427 Tijuana, B.C. (México)

[2] Instituto de Investigación y Desarrollo Educativo. Universidad Autónoma de Baja California (UABC). Carretera Transpeninsular Ensenada-Tijuana No.3917, Fraccionamiento Playitas, C.P. 22860, Ensenada, Baja California, (México)

[3] Universidad Autónoma de Aguascalientes. DAv. Universidad n.º 940, Ciudad Universitaria, C.P. 20100 Prol. Mahatma Gandhi n.º 6601, Gigante de los Arellano, Aguascalientes (México)

E-mail / ORCID

ruiz.karla32@uabc.edu.mx

https://orcid.org/0000-0001-8978-8364

horacio.pedroza@uabc.edu.mx

https://orcid.org/0000-0002-5256-2967

hectorrdzfig@gmail.com

https://orcid.org/0000-0003-1314-4073

Resumen: Este estudio compara coincidencias y confiabilidad entre jueces humanos y modelos de IAG al evaluar metáforas visuales de convivencia escolar, entendidas como representaciones gráficas que trasladan experiencias institucionales a escenas simbólicas (rancho con corrales, mercado con puestos y países-facultades). El diseño fue comparativo, con medidas repetidas sobre tres metáforas y la explicación verbal transcrita de los estudiantes que las elaboraron en grupos de discusión. Participaron seis personas expertas (18 valoraciones) y cinco LLM (ChatGPT-5, ChatGPT-5 Pro, ChatGPT-o3, DeepSeek, Gemini 2.5), con cinco ejecuciones por modelo (75 valoraciones). La rúbrica incluyó juicio global (1-3), tres indicadores binarios y 10 ítems Likert (1-5). Se estimaron alpha ordinal, kappa de Cohen, Fleiss kappa, alpha de Krippendorff y una G-study p x r con D-study para mezclas humano-IAG. Se aclara que la entrada no fue idéntica para todos los modelos: las variantes de ChatGPT trabajaron con imagen y descripción, mientras que DeepSeek y Gemini valoraron la descripción narrativa y la transcripción; por ello, los resultados se interpretan como línea base situada en las versiones y condiciones analizadas, no como conclusión definitiva sobre todos los sistemas multimodales actuales o futuros. Los hallazgos muestran que las IAG produjeron más palabras que las personas expertas, aunque la extensión se trató como indicador descriptivo y no como medida de calidad o profundidad. El índice de Convivencia fue mayor en humanos, sobre todo en la Metáfora 2 (Delta=+0.91). El D-study indicó que 1-2 personas expertas + alrededor de 10 ejecuciones de IAG alcanzan G >= 0.95. ChatGPT 5 Pro fue el modelo más próximo al consenso humano. El estudio aporta un referente para el análisis pedagógico futuro de producciones estudiantiles con apoyo de IAG.

Palabras-clave: Inteligencia artificial, Evaluación educativa, Confiabilidad interjueces, Rúbricas de evaluación, Metáfora visual, Cultura escolar, Confiabilidad de la medición.

Abstract: This study compares agreement and reliability between human judges and GAI models when evaluating visual metaphors of school coexistence, understood as graphic representations that translate institutional experiences into symbolic scenes (a ranch with corrals, a market with stalls, and countries/faculties). The design was comparative, with repeated measures across three metaphors and the transcribed verbal explanations provided by the students who produced them in focus groups. Six human experts participated (18 ratings) and five LLMs (ChatGPT-5, ChatGPT-5 Pro, ChatGPT-o3, DeepSeek, Gemini 2.5), with five runs per model (75 ratings). The rubric included a global judgment (1-3), binary indicators, and 10 Likert items (1-5). Ordinal alpha, Cohen kappa, Fleiss kappa, Krippendorff alpha, and a p x r G-study with a D-study for human-GAI mixtures were estimated. The input was not identical across models: ChatGPT variants worked with image and description, whereas DeepSeek and Gemini evaluated the narrative description and transcript. Therefore, the results are interpreted as a baseline situated in the versions and conditions analyzed, not as a definitive conclusion about all current or future multimodal systems. Findings showed that GAI produced more words than human experts, although length was treated as a descriptive indicator rather than a measure of quality or depth. The Coexistence Index was higher for humans, especially in Metaphor 2 (Delta=+0.91). The D-study indicated that 1-2 human experts plus around 10 GAI runs reach G >= 0.95. ChatGPT-5 Pro was the model closest to human consensus. The study provides a reference point for future pedagogical analysis of student productions supported by GAI.

Keywords: Artificial Intelligence, Educational Assessment, Interrater Reliability, Scoring Rubrics, Visual metaphor, School Culture, Test Reliability.

1. Introducción

En los últimos años, la expansión de la Inteligencia Artificial Generativa (IAG) basada en Modelos de Lenguaje Grande (LLM, por sus siglas en inglés) ha transformado las prácticas de búsqueda de información, redacción y apoyo al aprendizaje en educación superior (Kosmya et al., 2025; Nasr et al., 2025; Zhai et al., 2024). Más allá del uso informal de chatbots para resolver tareas o redactar textos, comienza a consolidarse un campo de investigación sobre el empleo de la IAG en procesos de evaluación educativa: generación automatizada de ítems y rúbricas (Arif et al., 2024; Doughty et al., 2024; May et al., 2025; Mistry et al., 2024; Scaria et al., 2024), calificación de ensayos y textos académicos (Bouziane & Bouziane, 2024; Bui & Barrot, 2025; Usher, 2025), retroalimentación escrita sobre producciones estudiantiles (Atasoy & Arani, 2025) y valoración de productos multimodales como imágenes o composiciones visuales (Liao et al., 2025; OpenAI, 2025).

En la generación de instrumentos, diversos estudios muestran que los LLM pueden producir rápidamente bancos de reactivos alineados a especificaciones curriculares y marcos teóricos, lo que abre la posibilidad de escalar procesos de diseño de pruebas (Arif et al., 2024; Doughty et al., 2024; Mistry et al., 2024; Scaria et al., 2024). No obstante, también documentan tasas relevantes de errores formales, desalineación de contenidos y distractores problemáticos cuando no media una edición humana experta, e incluso casos donde hasta 80 % de los ítems requieren corrección sustantiva (May et al., 2025). De manera similar, en la evaluación de textos, se ha observado que la IAG tiende a coincidir con las y los docentes en criterios mecánicos —gramática, ortografía, estructura—, pero se distancia en la apreciación de la coherencia temática, la profundidad conceptual y la adecuación al contexto (Bouziane & Bouziane, 2024; Bui & Barrot, 2025; Usher, 2025). Estos hallazgos han reforzado la recomendación de adoptar esquemas híbridos, donde la IAG aporta velocidad y detalle operativo, mientras que el juicio humano asegura la validez de contenido y la sensibilidad pedagógica (Artsi et al., 2024; Kiryakova, 2025; Kıyak & Emekli, 2024; Laupichler et al., 2024; Law et al., 2025).

A pesar de este avance, la evidencia disponible se ha concentrado principalmente en productos textuales y dimensiones de desempeño relativamente estandarizables, como la corrección lingüística o el logro en contenidos cognitivos (Atasoy & Arani, 2025; Bui & Barrot, 2025). Mucho menos se sabe sobre el desempeño de la IAG como jueza en tareas que involucran insumos visuales y constructos socio-afectivos complejos, por ejemplo, representaciones de convivencia, equidad o relaciones de poder en contextos escolares (Fierro-Evans & Carbajal-Padilla, 2019; Rodríguez-Figueroa, 2019, 2021). Asimismo, la mayoría de los estudios se ha limitado a calcular correlaciones o coeficientes de acuerdo interjueces (κ, α) entre humanos e IA, sin analizar la estabilidad global del sistema de jueceo ni las combinaciones óptimas de jueces humanos y automatizados desde marcos avanzados de confiabilidad como la Teoría de la Generalizabilidad (Brennan, 2001; Krippendorff, 2004; Shavelson & Webb, 1991).

Este vacío es especialmente relevante en el campo de la convivencia escolar, entendida como las prácticas interrelacionales cotidianas que configuran la vida en las instituciones educativas, distinta aunque articulada a su gestión institucional (Fierro-Evans, 2013; Fierro-Evans & Carbajal-Padilla, 2019). Desde una perspectiva sociocultural,

la convivencia incorpora dimensiones normativas —justicia, equidad, paz duradera— y estructurales —distribución de poder, formas de regulación— que exceden la observación directa de conductas puntuales y requieren articular los planos de contexto, estructura y acción social (Giddens, 1998; Rodríguez-Figueroa, 2019, 2021; Weiss, 2015). En educación superior, además, la literatura sobre convivencia es menos abundante que en educación básica, pese a que las universidades son espacios clave para la formación democrática, la inclusión y el ejercicio de derechos en la juventud (Hirmas & Carranza, 2009; López et al., 2023).

En este sentido, una forma de visualizar la función de las metáforas es a través del esquema sociocultural de tres planos —contexto, estructura y acción social— que organiza la información y orienta la interpretación (véase Figura 1) (Rodríguez-Figueroa, 2019). En términos sencillos: el contexto ubica la escuela en su espacio físico y social y en el perfil de sus agentes; la estructura recoge el «aprender-educar para vivir juntas y juntos» (normas y gestión institucional); y la acción social remite al «vivir juntos» en la interacción cotidiana.

En este contexto, las metáforas visuales elaboradas por estudiantes universitarios constituyen un dispositivo fértil para explorar cómo se significan la convivencia y sus tensiones. A diferencia de los cuestionarios estandarizados, las metáforas permiten representar simultáneamente contexto, estructura y acción social en una sola escena, y obligan a quien juzga —humano o modelo de IAG— a articular lo visible con interpretaciones sobre equidad, segregación, jerarquías o prácticas de paz (Delors et al., 1996; Rodríguez-Figueroa, 2019; Weiss, 2015). Precisamente por su complejidad, estas metáforas plantean desafíos para la IAG y ofrecen un «campo de prueba» idóneo para evaluar sus capacidades interpretativas en relación con expertas y expertos en convivencia escolar (Bertling et al., 2025; Caeiro Rodríguez et al., 2021).

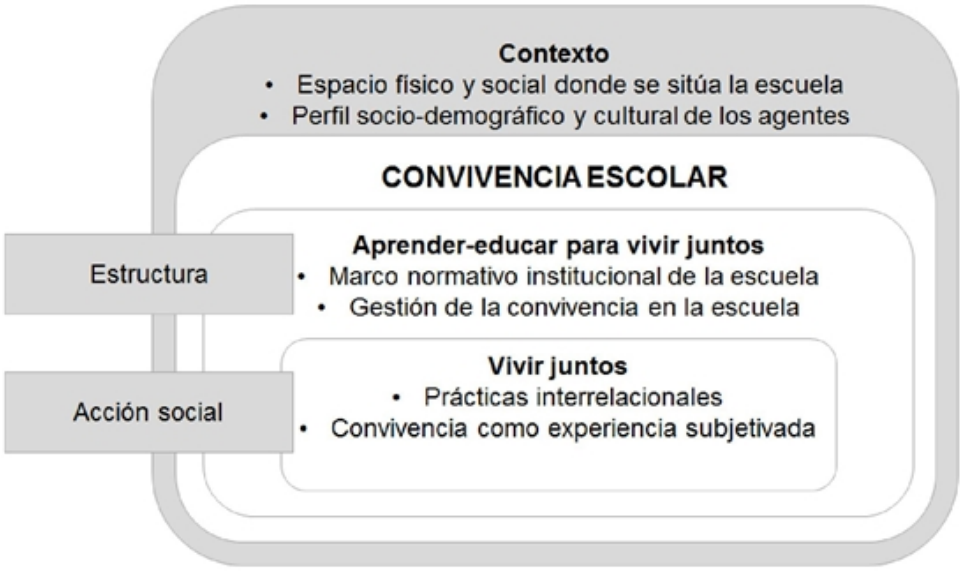

Figura 1. Convivencia escolar desde la perspectiva sociocultural. Fuente: Marco teórico para el estudio de la convivencia escolar desde la perspectiva sociocultural (Rodríguez-Figueroa, 2019, p. 35).

Para aclarar el objeto de análisis, en este artículo se entiende por metáfora visual de convivencia escolar una representación gráfica producida por estudiantes en la que un dominio de experiencia cotidiano (por ejemplo, un rancho, un mercado o una organización territorial por países/facultades) se usa para expresar cómo se perciben las relaciones, normas, conflictos, inclusión o exclusión dentro de la universidad. La evaluación no se dirige a la calidad estética de la imagen, sino a la correspondencia entre los elementos visibles o descritos y dimensiones socioculturales de la convivencia: interacción cotidiana, reglas, distribución de poder, participación, equidad y formas de resolución o ausencia de resolución de conflictos. Las tres metáforas analizadas se detallan en el Anexo 1 y se sintetizan en la Tabla 1.

Sin embargo, hasta donde se ha revisado, no se han reportado estudios que comparen de manera sistemática las valoraciones de modelos de IAG y paneles de personas expertas sobre metáforas visuales de convivencia escolar en educación superior, ni que examinen la confiabilidad global de esos sistemas de jueceo mediante Teoría de la Generalizabilidad. La mayoría de los trabajos sobre evaluación asistida por IAG se concentra en tareas textuales y se limita a reportar indicadores de acuerdo bivariado, sin estimar cuántos y qué tipos de jueces —humanos y/o automatizados— son necesarios para obtener puntajes suficientemente estables (Atasoy & Arani, 2025; Brennan, 2001; Krippendorff, 2004; Shavelson & Webb, 1991).

Este estudio busca contribuir a cubrir esa brecha. En términos generales, nuestro objetivo es comparar la confiabilidad de jueces humanos y modelos de IAG al interpretar tres metáforas visuales de convivencia escolar elaboradas por estudiantes de tres universidades mexicanas, identificar diferencias y coincidencias en sus valoraciones cerradas y estimar, mediante Teoría de la Generalizabilidad, la mezcla de evaluadores que produce puntajes más confiables (AERA, APA, & NCME, 2014; Brennan, 2001; Shavelson & Webb, 1991). En particular, nos preguntamos: (a) en qué medida coinciden humanos e IAG en la lectura de dimensiones observables y normativas de la convivencia; (b) qué tan estable es el sistema de jueceo cuando se integran paneles exclusivamente humanos, exclusivamente automatizados o mixtos; y (c) qué combinaciones de jueces humanos e IAG maximizan la confiabilidad sin perder el anclaje contextual que aporta la experiencia experta.

2. Metodología

A partir del objetivo, este es un estudio cuantitativo y comparativo, con medidas repetidas sobre tres estímulos y panel multijuez. El análisis se realizó a partir de un grupo focal previo en tres universidades de distintas regiones de México, donde se obtuvieron metáforas cercanas a la perspectiva de acción social; cada imagen fue elaborada por grupos de cuatro estudiantes. Las universidades fueron: privada en el Bajío, «rancho con corrales» (metáfora 1); pública del Centro-Occidente, «mercado con puestos» (metáfora 2); pública del noroeste, «países-facultades» (metáfora 3); véase Anexo 1 y Tabla 1. El estudio no pretende agotar todos los modos posibles de análisis visual, sino probar un procedimiento de evaluación comparada entre jueces humanos y modelos de IAG sobre metáforas estudiantiles. Para la confiabilidad del sistema se aplicó la Teoría de la Generalizabilidad en un diseño cruzado p x r (objeto de medida = Encuesta x Ítem Likert, 30 unidades; facetas: imagen (p), juez (r) e interacción p x r). Esta elección siguió guías comunes en investigación cuantitativa y evaluación educacional

(Creswell & Creswell, 2018; AERA, APA, & NCME, 2014; Shavelson & Webb, 1991; Brennan, 2001).

Tabla 1. Metáforas visuales analizadas y relevancia para la evaluación.

Metáfora	Dominio visual	Descripción sintética	Relevancia para el jueceo
1. Rancho con corrales	Rancho/corrales	La convivencia universitaria se representa mediante espacios diferenciados, agrupamientos y límites entre actores.	Permite valorar segregación, pertenencia, interacción entre grupos y fronteras simbólicas/institucionales.
2. Mercado con puestos	Mercado universitario	La universidad se presenta como un espacio de circulación, intercambio, diversidad de puestos y coexistencia cotidiana.	Permite analizar cooperación, participación, diversidad integrada, convivencia cotidiana y tensiones en espacios comunes.
3. Países-facultades	Mapa/territorios	Cada facultad o grupo aparece como territorio con límites, identidades y relaciones diferenciadas.	Permite evaluar identidad grupal, jerarquías, distancia intergrupal, distribución de poder e inequidad.

2.1. Participantes

Se trabajó con dos grupos: seis personas expertas en convivencia escolar (una valoración por cada metáfora; 18 valoraciones humanas) y cinco modelos de IAG (ChatGPT 5, ChatGPT 5 Pro, ChatGPT o3, DeepSeek, Gemini 2.5 Flash). La versión ChatGPT 5 Pro correspondió a una licencia de pago. Cada modelo realizó cinco ejecuciones por metáfora (75 valoraciones IAG; véase Tabla 2). Para estimar la consistencia de los modelos, se consideró cada ejecución como si fuera un juez independiente (25 jueces de IAG por metáfora). Las y los jueces humanos son personas investigadoras con experiencia comprobada en estudios y/o intervenciones sobre convivencia escolar: dos de universidades chilenas y cuatro de México. El muestreo del panel humano fue por juicio experto. Se incluyó a quienes contaban con experiencia comprobable en estudios o intervención de convivencia en educación superior. El estudio fue no anónimo (se registró nombre, área y rol para trazabilidad) y se obtuvo consentimiento informado. No se registraron datos faltantes en los ítems cerrados.

Tabla 2. Participantes por rúbrica.

Metáfora	Human@s (n)	IAG (n conversaciones)	Total
1	6	25	31
2	6	25	31
3	6	25	31
Total	18	75	93

2.2. Instrumento

A los jueces se les entregó una Guía para el análisis de metáforas, la cual fue estructurada por expertos en convivencia escolar y mantuvo la misma secuencia para las tres metáforas (véase Tabla 3): a) un registro descriptivo no puntuable (M0) para inventariar evidencias visuales sin interpretar; b) dimensiones puntuables (D1-D6), que combinaron un juicio global inductivo del tipo de convivencia (3.1, escala 1-3: mala-regular-buena) con indicadores binarios de procesos clave (4 subgrupos, 7.1 contención de violencia, 8.1 resolución constructiva, 9.1 paz duradera; 0 = No/1 = Sí) y un bloque de 10 proposiciones con escala Likert (1-5) sobre matices de convivencia (cooperación, antagonismo, participación, diversidad integrada, segregación, poder jerárquico, convivencia pacífica, equidad, desigualdad, violencia directa); y c) campos de apoyo no puntuables (M1) para justificar brevemente los juicios y asegurar trazabilidad.

Tabla 3. Guía de evaluación de metáforas visuales de convivencia escolar: constructo, dimensiones, indicadores y codificación.

Dimensión / módulo de la guía	**Qué evalúa (definición operativa)**	**Indicadores / Ítems (cód.)**	**Tipo de respuesta y codificación**
M0. Descripción objetiva *(no puntúa)*	Inventario de elementos visuales sin interpretación para sustentar el juicio posterior.	Listado de objetos, actores, colores, disposición espacial (1 y 2).	Texto abierto (evidencias descriptivas).
D1. Tipo de convivencia (3.1)	Valoración global inductiva de la convivencia representada.	3.1 «Tipo de convivencia»: *mala* (faltas de respeto/violencia; subgrupos separados), *regular* (mezcla de prácticas positivas y negativas), *buena* (respeto, inclusión, participación, comunicación, clima tranquilo).	Ordinal 3 puntos → 1 = mala, 2 = regular, 3 = buena.
D2. Subgrupos (4)	Presencia de separación del estudiantado en «grupitos/bolitas».	4 Subgrupos informales.	Binaria → 0 = No (incluye «No segurx»), 1 = Sí.
D3. Contención de la violencia (7.1)	Existencia de respuestas inmediatas ante situaciones que alteran la convivencia.	7.1 Escenarios de contención.	Binaria → 0 = No, 1 = Sí.
D4. Resolución constructiva (8.1)	Prácticas de diálogo, negociación, mediación o cooperación para resolver conflictos.	8.1 Resolución constructiva.	Binaria → 0 = No, 1 = Sí.

Dimensión / módulo de la guía	**Qué evalúa (definición operativa)**	**Indicadores / Ítems (cód.)**	**Tipo de respuesta y codificación**
D5. Paz duradera (9.1)	Transformaciones orientadas a equidad, inclusión y participación sostenida.	9.1 Condiciones para paz duradera.	Binaria → 0 = No, 1 = Sí.
D6. Matices de convivencia (Índice Likert 10 ítems)	Presencia de prácticas/estructuras relacionales específicas que, agregadas, estiman el Índice de Convivencia.	10.1–10.10: cooperación (+), antagonismo (–), participación activa (+), diversidad integrada (+), segregación/marginación (–), poder jerárquico/impositivo (–), convivencia pacífica (+), equidad (+), desigualdad (–), violencia directa (–).	Likert 1–5 («totalmente en desacuerdo» a «totalmente de acuerdo»).
M1. Campos de apoyo *(no puntúan)*	Sustento textual de los juicios para trazabilidad.	3.2 / 7.2 / 8.2 / 9.2 Justificaciones breves.	Texto abierto (anclajes del jueceo).

2.3. Procedimiento

Como se observa en la Figura 2, el procedimiento consistió en cinco fases. Primero se seleccionaron las tres metáforas provenientes de los grupos focales. Después se aplicó la guía a las personas expertas y se ejecutaron los cinco modelos de IAG con instrucciones idénticas. No obstante, se observó una diferencia metodológica importante: las variantes de ChatGPT trabajaron con la imagen y la descripción, mientras que DeepSeek y Gemini no observaron la imagen como tal, sino que elaboraron sus valoraciones a partir de la narrativa y la transcripción suministradas. Por tanto, la comparación no debe leerse como una competencia estrictamente equivalente entre sistemas multimodales, sino como un análisis de cómo distintos modos de entrada -visual más textual frente a texto mediador- afectan el jueceo de metáforas. Esta precisión también sitúa el alcance temporal del estudio: dada la velocidad con que evoluciona la IAG, versiones posteriores probablemente incorporen mejoras visuales y de interpretación multimodal.

En un tercer momento, se realizó la limpieza y recodificación: se integró una base única, se normalizaron encabezados y se documentaron las inversiones. Para los análisis principales se seleccionaron los ítems con respuesta cerrada (3.1, 4, 7.1, 9.1 y 10.1-10.10), y se contabilizó el total de palabras por ítem abierto (1, 2, 3.2, 5, 6.1, 6.2 y 6.4). Este conteo de palabras se mantuvo como indicador descriptivo del estilo de respuesta y de la trazabilidad de las justificaciones, no como medida de profundidad, calidad o validez del análisis. Después se calculó el índice de Convivencia por juez y metáfora a partir de los ítems 10.1 a 10.10. Finalmente, se estimó el acuerdo entre humanos e IAG por ítem y se evaluó la confiabilidad del sistema de jueceo.

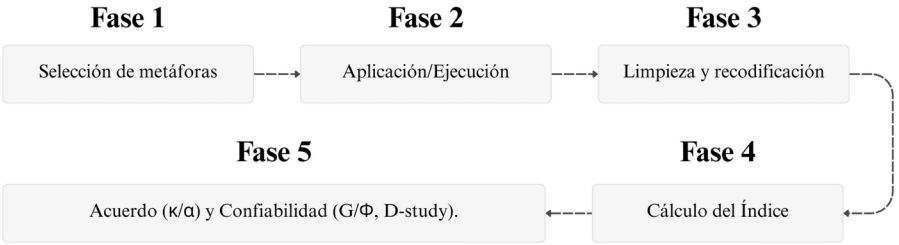

Figura 2. Procedimiento del estudio.

2.4. Análisis de datos

Descripción. Se exploraron medias (y errores estándar) del índice de convivencia por metáfora y tipo de juez. La extensión textual se examinó sólo como variable descriptiva para caracterizar estilos de respuesta (más sintéticos o más extensos), sin asumir que mayor número de palabras implicara mayor profundidad o calidad. Para el bloque de ítems Likert se estimó la consistencia interna mediante alpha ordinal, apropiado para escalas con respuestas ordenadas (Gadermann, Guhn, & Zumbo, 2012).

- Revisión cualitativa de respuestas abiertas. Para complementar el conteo de palabras, se realizó una revisión cualitativa de contenido de los campos abiertos 1, 2, 3.2, 5, 6.1, 6.2 y 6.4, entendida como una inferencia sistemática y contextualizada a partir de unidades textuales (Krippendorff, 2004). La revisión se organizó por ítem y tipo de juez, y buscó identificar la función predominante de las respuestas -descriptiva, interpretativa, justificativa o argumentativa-, su grado de síntesis, los núcleos temáticos recurrentes y la forma en que cada grupo articuló evidencia visual o textual con constructos de convivencia escolar. Este procedimiento no sustituyó los análisis estadísticos, sino que funcionó para distinguir extensión textual de profundidad interpretativa y calidad argumentativa.

- Acuerdo entre jueces. Se midió el grado de coincidencia entre humanos e IAG con κ de Cohen: ponderado cuadrático para variables ordinales (Likert) y no ponderado para variables binarias (Cohen, 1960, 1968). Para acuerdo multijuez en binarios se aplicó Fleiss' κ (Fleiss, 1971). Además, se utilizó α de Krippendorff ante niveles de medida mixtos (Krippendorff, 2004).

- Confiabilidad. Se realizó una G-study para descomponer la varianza en imagen/estímulo (p), juez (r) e interacción p×r, y obtener los coeficientes G (confiabilidad relativa) y Φ (confiabilidad absoluta) (Shavelson & Webb, 1991; Brennan, 2001). Con esos componentes se efectuó una D-study para proyectar la confiabilidad esperada al combinar distintos números de humanos e IAG.

- Ranking de modelos. Se construyó un puntaje compuesto para ordenar los modelos IAG (κ-Likert 70 %, κ-binarios 20 %, ρ de Spearman del índice 10 %) y se verificó la coherencia entre ordenamientos con la correlación de rangos (Spearman, 1904).

– Software y visualización. Se procesó en Python, específicamente Google Colab, con pandas y Matplotlib (Hunter, 2007; McKinney, 2010). Las figuras siguieron lineamientos clásicos de visualización científica y economía gráfica (Cleveland, 1993; Tufte, 2001; Wilkinson, 2005).

Es importante resaltar que los resultados estadísticos se organizaron por niveles de evidencia para evitar que un solo coeficiente sobredeterminara la interpretación: alpha ordinal examinó la consistencia del bloque Likert; kappa de Cohen comparó consensos humano-IAG por ítem; Fleiss kappa y alpha de Krippendorff abordaron acuerdo multijuez y niveles de medida mixtos; la G-study y la D-study evaluaron la estabilidad global del sistema y la mezcla de jueces; y el ranking de modelos se usó únicamente como índice de proximidad al consenso humano, no como prueba absoluta de superioridad de un modelo.

3. Resultados

El primer análisis fue descriptivo y tuvo un alcance acotado: se contabilizó el número de palabras para caracterizar el estilo de respuesta de humanos y modelos, no para inferir por sí mismo mayor profundidad, calidad o validez.

3.1. Producción de palabras por ítem según tipo de juez

En la Tabla 4 se observa este resultado según su media. Las IAG escribieron más que el panel humano en todas las secciones del instrumento. En promedio, las IAG registraron 72.4 palabras en Descripción, 85.4 en Interpretación, 46.2 en Justificación (3.2), 9.6 en Actores (5), 14.6 en Actitudes (6.1), 15.0 en Comportamientos (6.2) y 24.0 en Conflictos (6.4), frente a 38, 47, 29, 4, 5, 5 y 3 de las respuestas humanas, respectivamente. Esto implicó incrementos aproximados de 1.9x (Descripción), 1.8x (Interpretación), 1.6x (Justificación), 2.4x (Actores), 2.9x (Actitudes), 3.0x (Comportamientos) y 8.0x (Conflictos), con respecto a los humanos.

Tabla 4. Promedio de palabras por ítem según modelos de IAG y humanos.

Modelos	**ChatGPT 5**	**ChatGPT 5 Pro**	**ChatGPT o3**	**Deep-seek**	**Gemini 2.5**	**Promedio IAG**	**Humanos**
1. Descripción	78	84	71	52	77	72.4	38
2. Interpretación	85	67	83	74	118	85.4	47
3.2. Justificación	41	41	31	47	71	46.2	29
5. Actores	14	15	5	9	5	9.6	4
6.1 Actitudes	15	18	9	13	18	14.6	5
6.2 Comportam.	16	17	8	11	23	15	5
6.4 Conflictos	22	22	21	12	43	24	3

Por modelo, Gemini 2.5 fue el más extenso en la descripción de los apartados interpretativos y conductuales (Interpretación = 118; Justificación = 71; Comportamientos = 23; Conflictos = 43). Mientras que ChatGPT 5 Pro tuvo más extensión en la Descripción (84) de las metáforas y los Actores (15) que conforman las metáforas —empatando el máximo en Actitudes (18) junto con Gemini 2.5. Por su parte, DeepSeek solo destacó en Justificación (47) y ChatGPT o3 tendió a respuestas más breves en variables de contexto y conducta (Actores 5–9; Actitudes 9).

Para complementar el conteo de palabras de la Tabla 4, se revisó cualitativamente el contenido de las respuestas abiertas, conservando los mismos ítems para hacer posible la comparación. Esta lectura no asume que una mayor extensión equivalga a mayor calidad o profundidad; más bien permite distinguir si las respuestas fueron principalmente descriptivas, interpretativas, justificativas o inferenciales, y si tendieron a la síntesis, la enumeración o la argumentación.

Tabla 5. Síntesis cualitativa de las respuestas abiertas según tipo de juez.

Ítem abierto	Personas expertas	IAG/LLM	Lectura comparativa
1. Descripción	Selección de elementos centrales y rasgos con valor interpretativo: espacios, límites, agrupamientos y organización de la escena.	Inventario más amplio y ordenado de objetos, colores, posiciones, actores y relaciones espaciales.	La diferencia principal fue entre jerarquizar evidencias relevantes y enumerar con mayor exhaustividad.
2. Interpretación	Lecturas más situadas sobre convivencia, separación, pertenencia, indiferencia, integración o exclusión.	Identificación de patrones generales: diversidad, agrupamientos, convivencia parcial y tensiones entre grupos.	Hubo mayor densidad contextual en el panel experto y mayor sistematicidad categorial en los modelos.
3.2. Justificación	Argumentos más matizados; algunas respuestas reconocieron ambigüedad o límites de la escala para clasificar la convivencia.	Justificaciones más homogéneas, generalmente organizadas como balance entre señales positivas y negativas.	Las IAG justificaron con regularidad; las personas expertas introdujeron más cautela ante la ambigüedad visual.
5. Actores	Identificación breve y prudente: estudiantes, comunidad universitaria, institución y, en algunos casos, docentes o autoridades.	Listado más amplio: estudiantes, grupos, facultades, autoridades, institución y comunidad universitaria.	Los modelos ampliaron el repertorio, pero con mayor riesgo de inferir actores no explícitos.
6.1. Actitudes	Énfasis en indiferencia, pertenencia, distancia, rivalidad, aceptación o aislamiento, con sentido relacional.	Organización en categorías explícitas: orgullo grupal, separación, prejuicio, rivalidad, inclusión o necesidad de unión.	Coincidieron en pertenencia y separación; el panel experto aportó más lectura relacional y los modelos más orden categorial.
6.2. Comportamientos	Mayor cautela cuando no había acciones explícitas; se señalaron convivencia limitada, aislamiento o ausencia de evidencia suficiente.	Traducción de la metáfora en acciones posibles: permanecer en grupos, circular, competir, colaborar o compartir espacios.	Los modelos formularon más conductas; las personas expertas evitaron sobrerreaccionar ante escenas ambiguas.

Ítem abierto	Personas expertas	IAG/LLM	Lectura comparativa
6.4. Conflictos	Lectura de conflictos latentes: segregación, indiferencia, desigualdad, falta de integración o exclusión.	Nombramiento más explícito de conflictos: rivalidad, competencia, separación, tensión diversidad-integración o desigualdad.	Las IAG produjeron más formulaciones; el panel experto tendió a profundizar en implicaciones sociales y normativas.

En conjunto, las IAG/LLM produjeron respuestas más extensas, ordenadas y enumerativas, especialmente en descripción, actores, comportamientos y conflictos. Las personas expertas fueron más sintéticas, pero con frecuencia más selectivas, prudentes y contextualizadas, sobre todo cuando las preguntas exigían justificar la convivencia, interpretar actitudes o valorar conflictos latentes. Por ello, la comparación debe leerse como una diferencia en el tipo de elaboración: las IAG aportan exhaustividad y sistematicidad, mientras que el panel experto aporta jerarquización, cautela contextual y sensibilidad ante dimensiones sociales y normativas.

3.2. Índice de convivencia: diferencias entre Humanos e IAG

Según el índice de Convivencia, el promedio muestra brechas entre humanos e IAG (Tabla 6; véase la Figura 6): el índice fue consistentemente mayor en las puntuaciones realizadas por las personas expertas. En la Metáfora 2 la diferencia fue amplia a favor del panel humano (Delta=+0.91; 4.28 vs. 3.37). En las Metáforas 1 y 3 las diferencias fueron pequeñas (Delta=+0.15 y +0.21, respectivamente), por lo que prácticamente coinciden en la lectura del nivel de convivencia. Cuando la imagen o su descripción ofrece señales integradoras (por ejemplo, espacios compartidos, circulación o grupos), las personas expertas asignaron puntajes más altos, mientras que los modelos generaron puntuaciones más conservadoras. Esta diferencia es relevante porque delimita el tipo de inferencias que los LLM sostienen frente a constructos sociales: no se trata sólo de detectar objetos, sino de ponderar su significado pedagógico y normativo.

Tabla 6. Promedios en los índices de convivencia.

Metáfora	Humanos (Δ)	IAG (Δ)	Diferencia (H–IAG)
1	2.78	2.63	+0.15
2	4.28	3.37	+0.91
3	3.47	3.26	+0.21

Nota. Valores en escala 1–5. Diferencia positiva indica mayor puntaje humano respecto de la IAG.

Desagregando los modelos de IAG, la Tabla 7 muestra que ChatGPT 5 Pro y ChatGPT o3 fueron los más cercanos a las evaluaciones humanas, según el MAE (0.307 y 0.344, respectivamente) y con deltas pequeños en M1 y M3 —incluso con leves sobreestimaciones (Δ<0) en algún caso: o3 en M1=−0.117 y M3=−0.113, y 5 Pro en M3=−0.193—, mientras que en M2 ambos subestimaron la convivencia (0.723 y 0.803). En contraste, DeepSeek (MAE=0.758) y Gemini 2.5 Flash (MAE=0.571) mostraron las mayores discrepancias, con subestimación sistemática en las tres metáforas (p. ej., M2=1.023 en ambos y M3=0.547/0.567), y ChatGPT 5 exhibió un patrón intermedio (MAE=0.424) pero con una brecha muy amplia en M2=0.983.

En términos de orden relativo entre metáforas, ρ=1.00 en ChatGPT 5, DeepSeek y Gemini 2.5 indica que reprodujeron el orden observado en el consenso humano (ranking M1–M3) aunque desfasaron la escala (deltas altos), mientras que 5 Pro y o3 (ρ=0.50) fueron más próximos en magnitud pero menos estables en el ranking. Considerando la escala 1–5 del índice, y adoptando un umbral operativo de Δ≥0.50 como diferencia relevante para decisión, las brechas fueron sustantivas en M2 para cuatro de cinco modelos y, adicionalmente, en DeepSeek/Gemini para M3; por tanto, las diferencias entre modelos y el panel humano fueron materialmente importantes (especialmente en M2), aun cuando la significación estadística formal no se estime aquí por el número limitado de metáforas (N=3).

Tabla 7. Diferencias del índice por modelo de IAG y humanos.

Modelo de IAG	M1 (Δ)	M2 (Δ)	M3 (Δ)	MAE	ρ Índice (vs. H)
ChatGPT 5	0.043	0.983	0.247	0.424	1.00
ChatGPT 5 Pro	0.003	0.723	−0.193	0.307	0.50
ChatGPT o3	−0.117	0.803	−0.113	0.344	0.50
DeepSeek	0.703	1.023	0.547	0.758	1.00
Gemini 2.5 Flash	0.123	1.023	0.567	0.571	1.00

Nota. Δ positivo = el índice humano fue mayor que el del modelo (subestimación por la IA); Δ negativo = el modelo sobreestimó respecto de humanas/os. MAE: error medio absoluto entre índices (promedio entre Humanos (H) e IA en las tres metáforas). ρ: correlación de rangos de Spearman entre índices promedio por metáfora (humano comparado con el modelo de IAG).

Por otro lado, como se observa en la Tabla 8, el acuerdo entre humanos e IAG en estas escalas tuvo una magnitud media (kappa de Cohen promedio aproximado = 0.475; Krippendorff alpha aproximado = 0.236), lo que indica alineación parcial con pérdida de detalle en algunas dimensiones. Cuando las señales presentes en las metáforas son conductuales y observables (por ejemplo, actos de cooperación, comportamientos antagónicos o indicadores de participación activa), el alineamiento sube considerablemente -cooperación alcanzó kappa=0.80 (alpha=0.572), antagonismo y participación activa kappa aproximado=0.67-. En contraste, los ítems que requieren juicios normativos o inferencias sobre estructuras sociales (equidad, desigualdad, poder jerárquico) mostraron acuerdos bajos (kappa <= 0.25), lo que indica menor concordancia de los modelos con el panel experto en la lectura de matices relacionados con justicia y distribución de poder. En otras palabras, los modelos detectan agrupamientos o desigualdades visibles, pero no reproducen con la misma fidelidad las evaluaciones normativas que aportan las personas expertas.

Tabla 8. Resultados de ítems con escala Likert.

Ítem (resumen)	Cohen's κ (H vs IAG)	Krippendorff's α
Cooperación	0.80	0.572
Antagonismo	0.67	0.238
Participación activa	0.67	0.409
Diversidad integrada	0.55	0.354

Ítem (resumen)	Cohen's κ (H vs IAG)	Krippendorff's α
Segregación/marginación	0.25	0.390
Poder jerárquico	0.25	0.128
Convivencia pacífica	1.00†	0.105
Equidad	0.10	0.021
Desigualdad	0.00	0.103
Violencia directa	NaN	0.042
Promedio (10)	0.475	0.236

Nota. † κ=1.00 sugiere que en las 3 encuestas el consenso humano e IA coincidieron exactamente en «convivencia pacífica». Con tan pocas unidades puede ser techo; léase como «coinciden en el patrón general» más que como «acuerdo perfecto estable».

Por otro lado, en la Tabla 9 se pueden observar los resultados por modelo de forma global; con las tres metáforas. Tomando Cohen's κ como criterio principal y Krippendorff's α como desempate, el mejor modelo por ítem fue: en Cooperación destacó GPT-5 Pro (κ=.80; α=.613, por encima de GPT-o3 en α); en Antagonismo, Gemini 2.5 Flash (κ=.67; α=.359); en Participación activa, DeepSeek (empate en κ=.67, desempata por α=.302 frente a GPT-5 α=.209); en Diversidad integrada, DeepSeek (κ=.64); en Segregación/marginación, GPT-5 Pro (κ=.27); en Poder jerárquico, GPT-5 Pro (empate en κ=.25, mayor α=.254); en Convivencia pacífica, DeepSeek (κ=1.00; probable efecto techo con N=3); en Equidad, GPT-o3 (empate en κ=.31, mayor α=.094); en Desigualdad, GPT-5 Pro (κ=.21); y en Violencia directa, DeepSeek (único con estimación; κ≈0, α=.169). En promedio, DeepSeek obtuvo el κ más alto (0.403) y GPT-5 Pro el α más alto (0.282), pero con un κ (0.388) similar a DeepSeek.

Tabla 9. Resultados del índice de convivencia por ítem, modelo vs humanos (Cohen's κ y Krippendorff's α, ordinal).

Ítem	GPT 5 κ	GPT 5 α	GPT 5 Pro κ	GPT 5 Pro α	GPT o3 κ	GPT o3 α	DS κ	DS α	Gf κ	Gf α
Cooperación	0.571	0.549	0.800	0.613	0.800	0.464	0.727	0.600	0.640	0.565
Antagonismo	0.000	0.076	0.667	0.264	0.000	0.077	0.400	0.300	0.667	0.359
Participación activa	0.667	0.209	0.625	0.296	0.625	0.246	0.667	0.302	0.595	0.355
Diversidad integrada	0.545	0.315	0.571	0.495	0.571	0.401	0.640	0.172	0.400	0.332
Segregación/ marginación	0.250	0.270	0.267	0.475	0.250	0.285	0.250	0.320	0.250	0.293
Poder jerárquico	0.250	0.140	0.250	0.254	0.250	0.179	0.250	0.150	0.000	-0.011
Convivencia pacífica	0.000	0.128	0.000	0.199	0.000	0.085	1.000†	0.306	0.400	0.208
Equidad	0.308	0.067	0.100	0.064	0.308	0.094	0.100	0.072	-0.200	-0.049
Desigualdad	0.000	0.013	0.211	0.171	0.100	0.128	0.000	0.012	-0.200	-0.055
Violencia directa	NaN	0.015	NaN	-0.009	NaN	-0.006	0.000	0.169	NaN	0.010
Promedio	0.288	0.178	0.388	0.282	0.323	0.195	0.403	0.240	0.283	0.201

Nota. «GPT 5 κ/α» = Cohen's κ / Krippendorff's α (ordinal) de ChatGPT 5; «GPT 5 Pro κ/α» = κ/α de ChatGPT 5 Pro; «GPT o3 κ/α» = κ/α de ChatGPT o3; «DS κ/α» = κ/α de DeepSeek; «Gf κ/α» = κ/α de Gemini 2.5 Flash. El κ refleja el acuerdo del modelo con el consenso humano por ítem (ponderado para escalas ordinales cuando aplica); el α se estimó con todas las ejecuciones del modelo frente al panel humano. En empates de κ se consideró «mejor» el modelo con α más alto. † Un κ=1.00 con N bajo puede indicar efecto techo más que acuerdo robusto.

En los ítems binarios diseñados para captar procesos concretos -escenas de contención de violencia (7.1), resolución constructiva de conflictos (8.1) y condiciones para una paz duradera (9.1)- tanto las personas expertas como los modelos de IAG respondieron mayormente con la opción «No» (0). Es decir, en las tres metáforas revisadas no hubo, en las descripciones ni en las lecturas guiadas, indicios suficientes para calificar afirmativamente la presencia de mecanismos de contención, prácticas de resolución ni transformaciones hacia la paz (véase la Tabla 10). Primero, los bajos kappa en estas preguntas no invalidan necesariamente el instrumento: señalan más bien ausencia de señal en los estímulos bajo las condiciones de la prueba. Segundo, cuando se pretende evaluar fidelidad en la detección de procesos poco frecuentes (por ejemplo, acciones de contención explícita), conviene aumentar el número y la diversidad de estímulos o reequilibrar las clases para producir información interpretable por las métricas de acuerdo. Además, los ítems cerrados de la Tabla 10 se revisaron de forma individual conforme a los modelos; no hubo diferencias significativas que reportar.

Tabla 10. Resultados de ítems cerrados.

Ítem	Escala	Fleiss' κ (multi-juez)	Cohen's κ (consenso H vs IAG)	Krippendorff's α
3.1 Percepción	ordinal	—	0.00	0.023
4. Separación en subgrupos	binario	0.014	0.00	0.024
7.1 Escenarios de contención de violencia	binario	−0.029	NaN	−0.018
8.1 Resolución constructiva	binario	−0.001	NaN	0.010
9.1 Condiciones de paz duradera	binario	0.095	0.00	0.105

3.3. Confiabilidad del sistema de evaluación

Como se muestra en la Tabla 11, las estimaciones de generalizabilidad aportan una perspectiva complementaria y más estable que los kappa cuando los estímulos son pocos o desbalanceados. En la G-study calculamos componentes de varianza y coeficientes de confiabilidad para tres paneles: humanos (n=6), IAG (n=25) y mixto (n=31). El panel humano presentó una confiabilidad moderada (G aproximado=0.73; Phi aproximado=0.68) acompañada de una muy alta interacción imagen x juez (sigma2_pxr=0.908). Esta varianza de interacción indica que las personas expertas difieren entre sí en la forma en que valoran cada metáfora; es decir, las diferencias no se concentran tanto en los estímulos como en la respuesta individual de cada juez frente a cada imagen, característica esperable en tareas interpretativas donde la subjetividad profesional influye en el juicio.

Por contraste, el panel de IAG mostró una confiabilidad muy alta (G aproximado=0.98; Phi aproximado=0.98), con escasa varianza (sigma2_r aproximado=0.083) y menor varianza residual; esto significa que los modelos reprodujeron criterios de forma muy consistente entre ejecuciones. El panel mixto combinó ambas propiedades, ofreciendo G aproximado=0.976 y Phi aproximado=0.970: la combinación de IAG y personas expertas produce una confiabilidad muy alta, lo que sugiere que las IAG estabilizan la puntuación global mientras que la participación humana conserva el anclaje crítico para la validez de contenido. En cuanto al modelo con mayor confiabilidad en estas condiciones, ChatGPT 5 Pro obtuvo G aproximado=0.98. En términos prácticos, la G-study muestra que la inclusión de modelos automatizados reduce la variabilidad entre jueces y eleva la confiabilidad del sistema de evaluación, aunque la interpretación de casos concretos sigue requiriendo supervisión humana.

Tabla 11. Resultados de G-study.

Panel/Modelo	n_jueces	σ^2_p	σ^2_r	σ^2_p×r	G	Φ
Humanos	6	0.404	0.214	0.908	0.728	0.684
IAG (global)	25	0.896	0.083	0.366	0.984	0.980
Mixto (humanos e IAG)	31	0.719	0.128	0.553	0.976	0.970
ChatGPT 5	5	0.917	0.043	0.242	0.974	0.970
ChatGPT 5 Pro	5	1.160	0.010	0.120	0.980	0.978
ChatGPT o3	5	0.767	0.092	0.271	0.934	0.913
Deepseek	5	0.949	0.057	0.480	0.908	0.898
Gemini 2.5 Flash	5	1.216	0.000	0.310	0.951	0.951

Nota. G = fiabilidad relativa (ordenamiento consistente de los objetos); Φ = fiabilidad absoluta (decisiones absolutas). σ^2_p = varianza entre objetos; σ^2_r = varianza entre jueces; σ^2_p×r = interacción objeto×juez (y residual).

3.4. Mezcla idónea para evaluaciones humanas más IAG

El D-study traduce los componentes de varianza en recomendaciones sobre cuántos jueces humanos y ejecuciones de IAG incluir para lograr umbrales de confiabilidad deseados. Los resultados muestran que el mayor salto en G ocurre al incorporar modelos de IAG: pasar a 5-10 ejecuciones produce incrementos sustantivos en G, mientras que añadir más allá de 10-15 arroja rendimientos decrecientes. Concretamente, combinaciones pequeñas como 1 persona + 10 ejecuciones de IAG ya alcanzan G aproximado=0.95 (Phi aproximado=0.94), y 1-2 humanos + 15 IAG elevan G por encima de 0.96; por su parte, paneles con múltiples humanos pero sin IAG presentan G menores (por ejemplo, 2-4 humanos sin IA no alcanzan esos umbrales).

Además, cuando el número de IAG es igual o superior a 10, sumar más de 1–2 humanos no aumenta la confiabilidad e incluso puede reducirla levemente debido a la mayor varianza interjuez que introducen los humanos. La interpretación operativa, bajo estas condiciones, es la siguiente: para monitoreo institucional y tareas de escalado, una mezcla compacta —1–2 jueces humanos supervisores junto con ~10 modelos de IAG— ofrece un equilibrio entre costo y confiabilidad.

Tabla 12. TG – D-study: G y Φ esperados para mezclas H + IAG.

H - IA	0	5	10	15
1	—	0.888 / 0.866	0.950 / 0.939	0.968 / 0.962
2	0.449 / 0.416	0.870 / 0.845	0.942 / 0.930	0.964 / 0.957
3	0.560 / 0.514	0.864 / 0.838	0.936 / 0.922	0.961 / 0.953
4	0.632 / 0.585	0.863 / 0.836	0.933 / 0.919	0.959 / 0.950
6	0.728 / 0.684	0.870 / 0.845	0.931 / 0.916	0.955 / 0.946

Nota. Celdas = G / Φ. Negritas marcan G ≥ 0.94. Las combinaciones no mostradas fueron estimadas pero se omiten por brevedad.

La Figura 3 representa, mediante un mapa de calor, el coeficiente G esperado para distintas mezclas de jueces humanos (H) y modelos IAG (A) en el panel. El eje horizontal muestra el número de IAG (0, 5, 10, 15, 20, 25) y el eje vertical el número de humanos (1–6). Los tonos más claros indican mayor confiabilidad (G). Con A=0, G aumenta con H pero se mantiene moderado: ≈0.45 (2H), ≈0.56 (3H), ≈0.63 (4H), ≈0.68 (5H) y ≈0.73 (6H). El caso 1H+0A aparece sin estimación (celda en blanco) porque con un solo juez no se define G para decisiones relativas.

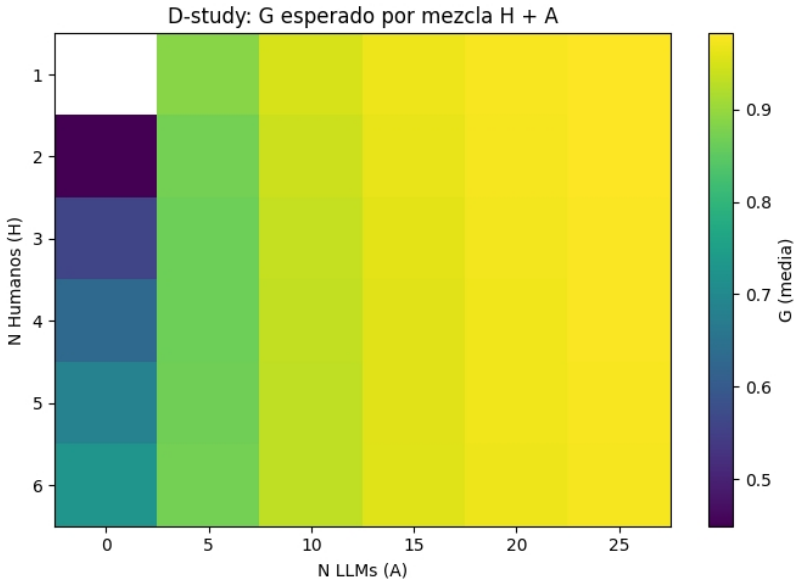

Figura 3. Resultado D-study.

3.5. LLM con mejor desempeño

El ranking de modelos se construyó mediante un puntaje compuesto que pondera kappa-Likert (70 %), kappa-binario (20 %) y correlación del índice global (10 %). Bajo ese criterio, ChatGPT 5 Pro fue el modelo con mayor proximidad al consenso humano (score aproximado=0.975; kappa-Likert aproximado=0.648; kappa-binario aproximado=0.250), seguido por ChatGPT o3 (score aproximado=0.725). Este ordenamiento no debe interpretarse como «mejor desempeño» universal, sino como

mayor cercanía al consenso experto en las tres metáforas, con las versiones y condiciones de entrada empleadas en el estudio. Los otros modelos evaluados (ChatGPT 5, DeepSeek, Gemini 2.5 Flash) reprodujeron con fidelidad el orden global de las metáforas en el índice (correlación rho=1.00), aunque exhibieron menor kappa-Likert; es decir, afinaron menos en el detalle ítem a ítem.

Tabla 13. Ranking de LLM's.

LLM	κ-Likert	κ-Bin	ρ Índice	Score
ChatGPT 5 Pro	0.648	0.250	0.50	0.975
ChatGPT o3	0.629	0.111	0.50	0.725
ChatGPT 5	0.476	0.111	1.00	0.345
Deepseek	0.392	0.111	1.00	0.121
Gemini 2.5 Flash	0.384	0.111	1.00	0.100

4. Discusión y conclusiones

Desde el marco sociocultural de la convivencia -que distingue entre prácticas interrelacionales cotidianas y gestión institucional (Fierro-Evans, 2013; Fierro-Evans & Carbajal-Padilla, 2019; Giddens, 1998; Weiss, 2015; Rodríguez-Figueroa, 2019, 2021)- las diferencias entre personas expertas e IAG se interpretan como formas distintas de operar con la evidencia disponible. La lectura humana fue más situada y holística; la de los modelos, más estandarizable y dependiente de las señales explícitas de la imagen o de su descripción textual. Cuando las metáforas ofrecieron claves integradoras -espacios comunes, circulación o vínculos visibles entre grupos-, el panel humano asignó valores más altos de convivencia, mientras que los modelos produjeron puntuaciones más conservadoras. Este contraste es coherente con la literatura sobre uso educativo de IAG: sin guía, los LLM pueden generar respuestas homogéneas o apoyarse en patrones superficiales; con una dirección clara, pueden apoyar procesos de análisis y retroalimentación (Zhai et al., 2024; Kosmya et al., 2025; Nasr et al., 2025).

Al comparar modelos, ChatGPT 5 Pro presentó la mayor alineación global con el consenso humano. Esta ventaja debe leerse con cautela porque las variantes de ChatGPT trabajaron con imagen y descripción, mientras que DeepSeek y Gemini valoraron una mediación textual. Por ello, la diferencia no permite afirmar una superioridad general del modelo, sino una mayor proximidad en las condiciones de este diseño. En el caso de DeepSeek, las coincidencias fueron más visibles en ítems puntuales, pero no siempre en la magnitud global del índice. Esta calibración diferencial coincide con estudios sobre valoración de textos donde los modelos replican patrones, pero requieren anclas pedagógicas explícitas, teoría de apoyo y revisión humana para sostener juicios de conjunto (Bui & Barrot, 2025; Bouziane & Bouziane, 2024; Usher, 2025; Atasoy & Arani, 2025).

Las diferencias entre ítems de la escala Likert confirmaron que los modelos se alinean mejor con el panel humano cuando las señales son observables. El acuerdo fue alto ante conductas visibles (por ejemplo, cooperación, antagonismo y participación). En cambio, la alineación disminuyó cuando entraron en juego constructos normativos

y subjetivos (equidad, desigualdad, poder jerárquico): allí, el juicio humano aportó contextos, historias institucionales y relaciones de poder que los modelos no reprodujeron con la misma precisión. Por tanto, los LLM funcionan mejor con señales claras y estandarizables, pero requieren blueprinting, grounding específico y revisión posterior para acercarse a valoraciones que dependen de nociones como justicia y distribución de oportunidades (Doughty et al., 2024; Mistry et al., 2024; Scaria et al., 2024; Arif et al., 2024; May et al., 2025).

En cuanto a producción de texto, todas las IAG escribieron más que las personas expertas, especialmente en interpretación y conflicto. Este hallazgo no se interpreta como evidencia de mayor profundidad, sino como una característica de generación textual ante consignas abiertas: los modelos suelen cubrir más puntos, añadir redundancias y ofrecer explicaciones extensas. Las personas expertas, en cambio, fueron más sintéticas y concentraron la escritura donde tenía sentido justificar. La extensión puede ser útil si se canaliza con guías específicas, rúbricas claras y límites de longitud; de lo contrario, puede aumentar el ruido analítico. Esta precisión responde directamente a la necesidad de diferenciar cantidad de palabras, profundidad, síntesis, tipo textual y calidad argumentativa (Kiryakova, 2025; Nasr et al., 2025).

Sobre la confiabilidad, los hallazgos mostraron un patrón: las IAG fueron muy consistentes entre ejecuciones; el panel humano fue más variable -como se espera en tareas interpretativas-; y la mezcla de ambas fuentes elevó la estabilidad sin eliminar el anclaje contextual del juicio experto. En términos prácticos, para monitoreo, depuración y estudios exploratorios se recomienda una mezcla pequeña de personas con un grupo moderado de ejecuciones de IAG (por ejemplo, una a dos personas expertas con alrededor de diez ejecuciones de uno o varios modelos). Para decisiones sensibles, en especial cuando están en juego equidad, paz o relaciones de poder, conviene elevar el peso del juicio humano (Shavelson & Webb, 1991; Brennan, 2001; AERA, APA, & NCME, 2014).

4.1. Alcances, limitaciones y proyecciones

Este estudio debe leerse como un referente situado y no como una evaluación definitiva de la IAG. La tecnología evoluciona con gran velocidad y es probable que versiones más actuales o futuras mejoren la lectura directa de imágenes, el razonamiento multimodal y la calibración pedagógica. Precisamente por ello, el valor del trabajo consiste en ofrecer una línea base documentada: muestra qué ocurre cuando se comparan personas expertas y modelos bajo condiciones explícitas de entrada, cómo cambia el acuerdo entre dimensiones observables y normativas, y qué mezclas de jueces producen mayor confiabilidad. Estos resultados forman parte de un conjunto posible de investigaciones; en el futuro se podrían realizar análisis pedagógicos más profundos de las producciones estudiantiles, incorporar modelos multimodales actualizados, comparar rúbricas más específicas y evaluar no sólo extensión, sino calidad argumentativa, profundidad interpretativa y utilidad para la retroalimentación educativa.

En conclusión, y respondiendo al objetivo de investigación, humanos e IAG coincidieron principalmente en dimensiones observables y se diferenciaron más en dimensiones normativas y subjetivas. Si bien este patrón podría parecer esperable, su relevancia radica en haberlo documentado empíricamente en una tarea específica de evaluación de metáforas visuales sobre convivencia escolar. Así, lo aparentemente

obvio adquiere valor científico cuando se precisa, se mide y se sitúa en un campo empírico concreto. El panel humano mostró mayor capacidad para contextualizar nociones como justicia, equidad o poder; los modelos produjeron puntuaciones más conservadoras cuando estos constructos no estaban explícitos en la imagen o en la descripción. ChatGPT 5 Pro fue el sistema más cercano al consenso humano en estas condiciones, mientras que DeepSeek destacó en algunos ítems puntuales. Todas las IAG generaron más texto que el panel humano, pero esa abundancia debe entenderse como rasgo de estilo y no como medida directa de calidad. La principal contribución del estudio es metodológica y pedagógica: ofrece un referente para combinar jueces humanos e IAG, identificar en qué dimensiones la automatización puede apoyar, y delimitar dónde sigue siendo imprescindible la interpretación experta, especialmente para futuros análisis de producciones estudiantiles sobre convivencia escolar.

5. Referencias

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arif, T., Asthana, S., & Collins-Thompson, K. (2024). Generation and assessment of multiple-choice questions from video transcripts using large language models. *Proceedings of the 11th ACM Conference on Learning at Scale (L@S '24)*, 530–534. https://doi.org/10.1145/3657604.3664714

Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G., & Klang, E. (2024). Large language models for generating medical examinations: Systematic review. *BMC Medical Education, 24*, 354. https://doi.org/10.1186/s12909-024-05239-y

Atasoy, A., & Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts? *Education and Information Technologies*. https://doi.org/10.1007/s10639-025-13553-1

Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education, 3*, 201. https://doi.org/10.1007/s44217-024-00320-6

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://link.springer.com/book/10.1007/978-1-4757-3456-0

Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2025.2521003

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE. https://us.sagepub.com/en-us/nam/research-design/book246125

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. *Avances en Medición, 6*, 27–36. https://www.researchgate.net/publication/302438451_Validez_de_contenido_y_juicio_de_expertos_Una_aproximacion_a_su_utilizacion

Fierro-Evans, C., & Carbajal-Padilla, P. (2019). Convivencia escolar: una revisión del concepto. *Psicoperspectivas, 18*(1), 1–14. https://doi.org/10.5027/psicoperspectivas-vol18-issue1-

Fierro-Evans, M. C. (2013). Convivencia inclusiva y democrática: una perspectiva para gestionar la seguridad escolar. *Sinéctica: Revista Electrónica de Educación*, (40), 1–18. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-109X2013000100005&nrm=iso

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data. *Practical Assessment, Research, and Evaluation,* 17(1). https://doi.org/10.7275/n560-j767

Giddens, A. (1998). *La constitución de la sociedad: bases para la teoría de la estructuración* (2da reimpr.). Amorrortu Editores.

Gemini Team Google. (2025). Multi-task performance of LLMs: A Google perspective. *Google Research Papers, 12*(4), 76–89.

Hirmas, C. y Carranza, G. (2009). Matriz de indicadores sobre convivencia democrática y cultura de paz en la escuela. En *III Jornadas de Cooperación Iberoamericana sobre Educación para la Paz, la Convivencia Democrática y los Derechos Humanos* (pp. 56-136). Salesianos Impresores.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kiryakova, G. (2025). ChatGPT as a supportive tool for creating assessment resources. *TEM Journal, 14*(2), 1014–1023. https://doi.org/10.18421/TEM142-04

Kıyak, Y. S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal, 100*(1189), 858–865. https://doi.org/10.1093/postmj/qgae065

Kosmyna, N., et al. (2023). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (arXiv:2506.08872). https://arxiv.org/abs/2506.08872

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1093/hcr/30.3.411

Laupichler, M. C., Rother, J. F., Grunwald Kadow, I. C., Ahmadi, S., & Raupach, T. (2024). Large language models in medical education: Comparing ChatGPT- to human-generated exam questions. *Academic Medicine, 99*(5), 508–512. https://doi.org/10.1097/ACM.0000000000005626

Law, A. K. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Hung, K. K.-C., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: A cohort study in a high-stakes examination. *BMC Medical Education, 25*, 208. https://doi.org/10.1186/s12909-025-06796-6

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Liao, Z., Liu, X., Qin, W., Li, Q., Wang, Q., Wan, P., Zhang, D., Zeng, L., & Feng, P. (2025). HumanAesExpert: Advancing a multi-modality foundation model for human image aesthetic assessment (arXiv:2503.23907). https://arxiv.org/abs/2503.23907

Martínez-Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Síntesis.

May, T. A., Fan, Y. K., Stone, G. E., Koskey, K. L. K., Sondergeld, C. J., Folger, T. D., Archer, J. N., Provinzano, K., & Johnson, C. C. (2025). An effectiveness study of generative AI tools used to develop multiple-choice test items. *Education Sciences, 15*(2), 144. https://doi.org/10.3390/educsci15020144

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

Mistry, N. P., Saeed, H., Rafique, S., Le, T., Obaid, H., & Adams, S. J. (2024). Large language models as tools to generate radiology board-style multiple-choice questions. *Academic Radiology, 31*(9), 3872–3878. https://doi.org/10.1016/j.acra.2024.06.046

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7–16. https://doi.org/10.7334/psicothema2018.291

Nasr, N. R., Tu, C.-H., Sujo-Montes, L., Yen, C.-J., *et al.* (2025, abril). *Generative artificial intelligence impact on critical thinking in higher education: A comprehensive bibliometric analysis and systematic review.* Preprint. ResearchGate. https://doi.org/10.13140/RG.2.2.18979.16168

OpenAI. (2025). Introducing GPT-5. https://openai.com/index/introducing-gpt-5

Art of Problem Solving. (2025). *AMC historical results*. https://wiki.artofproblemsolving.com/wiki/index.php?title=AMC_historical_results

Rodríguez-Figueroa, H. M. (2019). *La convivencia escolar desde la perspectiva sociocultural* [tesis doctoral]. Universidad Autónoma de Aguascalientes.

Rodríguez-Figueroa, H. M. (2021). Convivencia escolar: Revisión del concepto a partir de dos estudios de caso. *Sinéctica, (57)*, e1272. https://doi.org/10.31391/S2007-7033(2021)0057-003

Scaria, N., Chenna, S. D., & Subramani, D. (2024). How good are modern LLMs in generating relevant and high-quality questions at different Bloom's skill levels for Indian high school social science curriculum? *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. https://aclanthology.org/2024.bea-1.1

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.1093/ije/dyq191

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press. https://www.edwardtufte.com/book/the-visual-display-of-quantitative-information

Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2025.2487495

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). Springer. https://link.springer.com/book/10.1007/0-387-28695-0

Weiss, E. (2015). Más allá de la socialización y de la sociabilidad: jóvenes y ba-chillerato en México. *Educ. Pesqui.*, 41, número especial, pp. 1257-1272. https://doi.org/10.1590/S1517-9702201508144889