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

RELATEC
Revista Latinoamericana de Tecnología Educativa

ARTICLE / ARTÍCULO

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

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

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

Received: December 1, 2025
Reviewed: May 5, 2026
Accepted: May 8, 2026

Corresponding authors' address:

[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. (Mexico)

[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, (Mexico)

[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 (Mexico)

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

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 × 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.

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.

1. Introduction

In recent years, the expansion of Generative Artificial Intelligence (GAI), based on Large Language Models (LLMs), has transformed information-seeking practices, writing processes, and learning support in higher education (Kosmyna et al., 2025; Nasr et al., 2025; Zhai et al., 2024). Beyond the informal use of chatbots to complete assignments or draft texts, a body of research on the use of GAI in educational assessment processes is beginning to emerge: automated generation of test items and rubrics (Arif et al., 2024; Doughty et al., 2024; May et al., 2025; Mistry et al., 2024; Scaria et al., 2024), grading of essays and academic texts (Bouziane & Bouziane, 2024; Bui & Barrot, 2025; Usher, 2025), written feedback on student work (Atasoy & Arani, 2025), and evaluation of multimodal products such as images or visual compositions (Liao et al., 2025; OpenAI, 2025).

In the generation of instruments, several studies have shown that LLMs can rapidly generate item banks aligned with curricular specifications and theoretical frameworks, opening the possibility of scaling assessment design processes (Arif et al., 2024; Doughty et al., 2024; Mistry et al., 2024; Scaria et al., 2024). Nevertheless, they also document significant rates of formal errors, content misalignment and problematic distractors when an expert human editing is absent, including cases in which up to 80% of the items require a substantive correction (May et al., 2025). Similarly, in the assessment of written work, GAI has been found to align with teachers' evaluations regarding mechanical criteria—grammar, spelling, and structure—but to diverge in the assessment of thematic coherence, conceptual depth and contextual adaptation (Bouziane & Bouziane, 2024; Bui & Barrot, 2025; Usher, 2025). These findings have reinforced recommendations to adopt hybrid models in which GAI systems contribute speed and operational detail, while human judgment ensures content validity and pedagogical sensitivity (Artsi et al., 2024; Kiryakova, 2025; Kıyak & Emekli, 2024; Laupichler et al., 2024; Law et al., 2025).

Despite this progress, the available evidence has mainly focused on textual products and relatively standardizable performance dimensions, such as linguistic accuracy and achievement in cognitive content (Atasoy & Arani, 2025; Bui & Barrot, 2025). Much less is known about the performance of GAI as an evaluator in tasks involving visual inputs and complex socio-affective contexts, for example school dynamics, equity or power relations in school contexts (Fierro-Evans & Carbajal-Padilla, 2019; Rodríguez-Figueroa, 2019, 2021). In addition, most studies have been limited to calculating correlations or inter-rater agreement coefficients (κ, α) between humans and AI, without analyzing the overall stability of the scoring system or the optimal combinations of human and automated evaluators from advanced reliability frameworks such as Generalizability Theory (Brennan, 2001; Krippendorff, 2004; Shavelson & Webb, 1991).

This gap is especially significant in the field of school dynamics, which refers to the everyday interpersonal practices that shape life within educational institutions, distinct from—although closely connected to—their institutional management (Fierro-Evans, 2013; Fierro-Evans & Carbajal-Padilla, 2019). From a sociocultural perspective, school dynamics incorporate normative dimensions—justice, equity, and lasting peace—and structural dimensions—power distribution and forms of regulation—that go beyond the direct observation of specific behaviors and require the articulation of

context, structure, and social action (Giddens, 1998; Rodríguez-Figueroa, 2019, 2021; Weiss, 2015). In higher education, moreover, the literature on school dynamics is less extensive than in basic education, despite the fact that universities are key spaces for democratic formation, inclusion, and youth rights (Hirmas & Carranza, 2009; López et al., 2023).

In this sense, one way to visualize the function of metaphors is through the three-level sociocultural framework—context, structure, and social action—which organizes information and guides interpretation (see Figure 1) (Rodríguez-Figueroa, 2019). In simple terms, context situates the school within its physical and social environment, as well as the profile of its agents; structure includes "learning and educating to live together" (norms and institutional management); and social action refers to "living together" in everyday interaction.

In this context, visual metaphors created by university students constitute a valuable means of exploring how school dynamics and their tensions are represented. Different from standardized questionnaires, metaphors make it possible to simultaneously represent context, structure, and social action within a single scene, requiring the evaluator—whether human or a GAI model—to connect visible elements with interpretations related to equity, segregation, hierarchies, or practices of peace (Delors et al., 1996; Rodríguez-Figueroa, 2019; Weiss, 2015). Precisely because of their complexity, these metaphors pose challenges for GAI and provide an ideal "testing ground" for evaluating its interpretative capacities in relation to experts in school dynamics (Bertling et al., 2025; Caeiro Rodríguez et al., 2021).

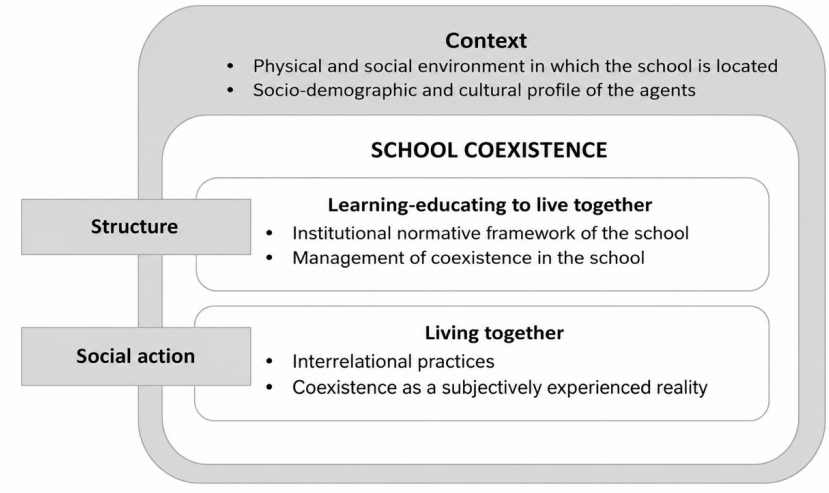

Figure 1. School dynamics from a sociocultural perspective. Theoretical framework for the study of school dynamics from a sociocultural perspective (Rodríguez-Figueroa, 2019, p. 35).

To clarify the object of analysis, this article defines a visual metaphor of school dynamics as a graphic representation produced by students in which a domain of everyday experience (for example, a ranch, a market or a territorial organization divided by countries/faculties) is used to express how relationships, norms, conflicts, inclusion or exclusion within the university are perceived. The evaluation does not focus on the aesthetic quality of the image, but rather on the correspondence between visible or described elements and sociocultural dimensions of school dynamics: everyday

interaction, rules, distribution of power, participation, equity and forms of conflict resolution or absence thereof. The three metaphors analyzed are detailed in Appendix 1 and summarized in Table 1.

However, no studies have yet systematically compared the evaluations produced by GAI models and panels of experts regarding visual metaphors of school dynamics in higher education, nor examined the overall reliability of these scoring systems through Generalizability Theory. Most studies on GAI-assisted assessment focus on textual tasks and are limited to reporting bivariate agreement indicators, without estimating how many and what types of evaluators—humans and/or automated—are necessary to obtain sufficiently stable scores (Atasoy & Arani, 2025; Brennan, 2001; Krippendorff, 2004; Shavelson & Webb, 1991).

This study seeks to help fill this gap. In general terms, our goal is to compare the reliability of human evaluators and GAI models in interpreting the three visual metaphors of school dynamics created by students from three Mexican universities, identify differences and similarities in their closed-ended evaluations, and estimate—through Generalizability Theory—the combination of evaluators that produces the most reliable scores (AERA, APA, & NCME, 2014; Brennan, 2001; Shavelson & Webb, 1991). More specifically, we ask: (a) to what extent humans and GAI agree in their interpretation of observable and normative dimensions of school dynamics; (b) how stable the scoring system is when exclusively human, exclusively automated or mixed panels are used; and (c) which combinations of human and GAI evaluators maximize reliability without losing the contextual grounding provided by expert experience.

2. Methodology

Based on the objective of the study, the research follows a quantitative repeated-measures design involving three stimuli and a multi-rater panel. The analysis was conducted following prior focus groups conducted at three universities from different regions of Mexico, where metaphors aligned with the social action perspective were obtained; each image was created by groups of four students. The universities were: a private university in the Bajío region, "ranch with corrals" (metaphor 1); a public university in the Central-West region, "market with stalls" (metaphor 2); a public university in northwestern Mexico, "countries-faculties" (metaphor 3); see Appendix 1 and Table 1. The study does not aim to exhaust all possible modes of visual analysis, but rather to test a comparative evaluation between human evaluators and GAI models using student-produced metaphors. To estimate system reliability, Generalizability Theory was applied using a crossed p × r design (object of measurement = survey × Likert item, 30 units; facets: images (p), evaluator (r) and p × r interaction). This decision followed common guidelines in quantitative research and educational assessment (Creswell & Creswell, 2018; AERA, APA, & NCME, 2014; Shavelson & Webb, 1991; Brennan, 2001).

Table 1. Visual metaphors analyzed and their relevance for rating.

Metaphor	Visual domain	Synthetic description	Relevance for rating
1. Ranch with corrals	Ranch/corrals	University dynamics are represented through differentiated spaces, social groupings and boundaries between actors.	It allows the assessment of segregation, sense of belonging, interaction among groups and symbolic or institutional boundaries.
2. Market with stalls	Campus market	The university is represented as a shared circulation space, exchange, diversity of stalls, and everyday coexistence.	It allows the analysis of cooperation, everyday social dynamics and tensions within shared spaces, participation, and integrated diversity
3. Countries-faculties	Map/territory	Each faculty or group is shown as a territory with differentiated boundaries, identities and relationships.	It allows the assessment of group identities, hierarchies, intergroup distance, distribution of power and inequality.

2.1. Participants

The study was conducted with two groups: six experts in school dynamics (one evaluation for each metaphor; 18 human evaluations) and five GAI models (ChatGPT 5, ChatGPT 5 Pro, ChatGPT o3, DeepSeek, Gemini 2.5 Flash). The ChatGPT 5 Pro version corresponded to a paid license. Each model performed five runs per metaphor (75 GAI evaluations; see Table 2). To estimate model consistency, each run was treated as if it were an independent evaluator (25 GAI evaluators per metaphor). The human evaluators were researchers with demonstrated experience in studies and/or interventions related to school dynamics: two from Chilean universities and four from Mexico. The human panel was selected using expert-judgment sampling. It included evaluators with demonstrated experience in studies and/or interventions related to school dynamics in higher education. The study was non-anonymous (name, area of expertise and role were recorded for traceability purposes) and informed consent was obtained. No missing data were recorded in the closed-ended items.

Table 2. Evaluations per rubric.

Metaphor	Humans (n)	GAI (n conversations)	Total
1	6	25	31
2	6	25	31
3	6	25	31
Total	18	75	93

2.2. Instrument

The judges were provided with a Metaphor Analysis Guide, structured by experts in school dynamics and using the same sequence for all three metaphors (see Table 3): a) a non-scored descriptive section (M0) used to inventory visual evidence without

interpretation; (b) scored dimensions (D1–D6), which combined an inductive global judgment of the type of coexistence (3.1, 1–3 scale: poor–fair–good) with binary indicators of key processes (4 subgroups: 7.1 violence containment, 8.1 constructive resolution, 9.1 lasting peace; 0 = No / 1 = Yes) and a set of 10 Likert-scale propositions (1–5) capturing nuances of coexistence (cooperation, antagonism, participation, integrated diversity, segregation, hierarchical power, peaceful coexistence, equity, inequality, direct violence); and (c) non-scored support fields (M1) used to briefly justify judgments and ensure traceability.

Table 3. Assessment Guide for Visual Metaphors of School Dynamics: Construct, Dimensions, Indicators and Coding.

Dimension / module of the guide	What it evaluates (operational definition)	Indicators / Items (codes)	Type of response and coding
M0. Objective description (not scored)	Inventory of visual elements without interpretation to support subsequent judgment.	List of objects, actors, colors, and spatial arrangement (1 and 2).	Open-ended text (descriptive evidence).
D1. Type of coexistence (3.1)	Global inductive assessment of the represented coexistence.	3.1 "Type of coexistence": poor (disrespect/violence, separated subgroups), fair (mix of positive and negative practices), good (respect, inclusion, participation, communication, peaceful coexistence).	Ordinal 3-point scale → 1 = poor, 2 = fair, 3 = good.
D2. Subgroups (4)	Presence of student separation into small "groups/cliques".	4 Informal subgroups.	Binary → 0 = No (including "Not sure"), 1 = Yes.
D3. Violence containment (7.1)	Presence of immediate responses to situations that disrupt coexistence.	7.1 Containment scenarios	Binary → 0 = No, 1 = Yes.
D4. Constructive resolution (8.1)	Practices of dialogue, negotiation, mediation or cooperation for conflict resolution.	8.1 Constructive resolution.	Binary → 0 = No, 1 = Yes.
D5. Lasting peace (9.1)	Transformation oriented toward equity, inclusion, and sustained participation.	9.1 Conditions for lasting peace	Binary → 0 = No, 1 = Yes.

Dimension / module of the guide	What it evaluates (operational definition)	Indicators / Items (codes)	Type of response and coding
D6. Dimensions of coexistence (10-item Likert index)	Presence of specific relational practices/structures that, when aggregated, estimate the Coexistence Index.	10.1–10.10: cooperation (+), antagonism (–), active participation (+), integrated diversity (+), segregation/marginalization (–), hierarchical/impositional power (–), peaceful coexistence (+), equity (+), inequality (–), direct violence (–).	Likert scale 1–5 ("strongly disagree" to "strongly agree").
M1. Supporting fields (not scored)	Textual support of judgments to ensure traceability.	3.2 / 7.2 / 8.2 / 9.2 Brief justifications.	Open-ended text (rater justification anchors).

2.3. Procedure

As shown in Figure 2, the procedure consisted of five phases. First, the three metaphors derived from the focus groups were selected. Next, the guide was applied to the expert evaluations and the five GAI models were executed using identical instructions. However, an important methodological difference was observed: ChatGPT variants worked with image and description, whereas DeepSeek and Gemini did not directly observe the image; instead they produced their evaluations based on the narrative and transcriptions provided. Therefore, the comparison should not be interpreted as a strictly equivalent competition between multimodal systems, but rather as an analysis of how different input modalities –visual plus textual versus text-mediated input– affect metaphor evaluation. This clarification also situates the temporal scope of the study: given the rapid evolution of GAI, later versions will likely incorporate improvements in visual processing and multimodal interpretation.

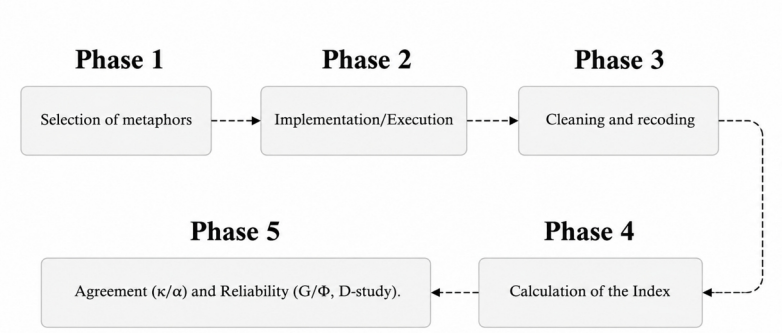

Figure 2. Study procedure

In a third stage, data cleaning and recoding were performed: a unified database was created, headers were standardized and reverse-coded items were documented. For the main analyses, closed-ended items (3.1, 4, 7.1, 9.1, and 10.1–10.10) were

selected, and the total number of words per open-ended item (1, 2, 3.2, 5, 6.1, 6.2, and 6.4) was computed. This word count was retained as a descriptive indicator of response style and traceability of justifications, not as a measure of depth, quality or validity of the analysis. Subsequently, the Coexistence Index was calculated for each rater and metaphor based on items 10.1–10.10. Finally, agreement between human and GAI was estimated per item and the reliability of the scoring system was assessed.

2.4. Data analysis

Description. Means and standard errors of the Coexistence Index were explored by metaphor and type of rater. Text length was examined only as a descriptive variable to characterize response styles (more concise or more extensive), without assuming that a higher number of words implied greater depth or quality. For the Likert-scale item block, internal consistency was estimated using ordinal alpha, which is appropriate for ordered categorical response formats (Gadermann, Guhn, & Zumbo, 2012).

- Qualitative review of open-ended responses. To complement word count analysis, a qualitative content review was conducted on open-ended fields 1, 2, 3.2, 5, 6.1, 6.2, and 6.4, understood as systematic and context-sensitive inference processes based on textual units (Krippendorff, 2004). The review was organized by item and type of rater, and aimed to identify the predominant function of the responses—descriptive, interpretive, justificatory, and argumentative—as well as their degree of conciseness, recurrent thematic cores and the way in which each group linked visual or textual evidence with constructs of school dynamics. This procedure did not replace statistical analyses; rather it served to distinguish textual length from interpretive depth and argumentative quality.

- Inter-rater agreement. The degree of agreement between human evaluators and GAI was assessed using Cohen's kappa: a quadratic weighted kappa for ordinal variables (Likert scale items) and unweighted kappa for binary variables (Cohen, 1960, 1968). For multi-rater agreement on binary variables, Fleiss' kappa was applied (Fleiss, 1971). In addition, Krippendorff's alpha was used to accommodate mixed measurement levels (Krippendorff, 2004).

- Reliability. A G-study was conducted to decompose variance into facets of images/stimulus (p), rater (r), and their interaction ($p \times r$) and to estimate generalizability coefficients (G, relative reliability) and phi coefficients (Φ, absolute reliability) (Shavelson & Webb, 1991; Brennan, 2001). Based on these variance components, a D-study was performed to project expected reliability under different combinations of human and GAI raters.

- Model ranking. A composite score was constructed to rank GAI models (70% Likert-scale κ, 20% binary κ, 10% Spearman's ρ of the index) and consistency of rankings was evaluated using rank correlation (Spearman, 1904).

- Software and visualization. Analyses were conducted in Python, specifically using Google Colab, with pandas and Matplotlib (Hunter, 2007; McKinney, 2010). Figures followed established guidelines for scientific visualization and graphical excellence (Cleveland, 1993; Tufte, 2001; Wilkinson, 2005).

It is important to emphasize that the statistical results were organized by levels of evidence to avoid letting a single coefficient dominate the interpretation. Ordinal alpha assessed the internal consistency of the Likert block; Cohen's kappa compared human-GAI agreement and mixed measurement levels; the G-study and D-study evaluated the overall stability of the scoring system and the combination of raters; and the ranking model was used solely as an indicator of proximity to human consensus, rather than as an absolute measure of model superiority.

3. Results

The first analysis was descriptive and had a limited scope: word counts were calculated to characterize response styles of both human evaluators and models, not to infer greater depth, quality or validity.

3.1. Word count per item by rater type

The first analysis was descriptive and had a limited scope: word counts were calculated to characterize response styles of both human evaluators and models, not to infer greater depth, quality or validity. Table 4 presents these results in terms of means. GAI systems produced longer responses than the human panel across all sections of the instrument. Overall, GAI models produced 72.4 words in Description, 85.4 in Interpretation, 46.2 in Justification (3.2), 9.6 in Actors (5), 14.6 in Attitudes (6.1), 15.0 in Behaviors (6.2) and 24.0 in Conflicts (6.4), compared to 38, 47, 29, 4, 5, 5 and 3 words for human responses, respectively. This represented approximate increases of 1.9x (Description), 1.8x (Interpretation), 1.6x (Justification), 2.4x (Actors), 2.9x (Attitudes), 3.0x (Behaviors) and 8.0x (Conflicts), relative to human evaluators.

Table 4. Average word count per item by GAI models and humans.

Model	**ChatGPT 5**	**ChatGPT 5 Pro**	**ChatGPT o3**	**Deep-seek**	**Gemini 2.5**	**Average IAG**	**Humans**
1. Description	78	84	71	52	77	72.4	38
2. Interpretation	85	67	83	74	118	85.4	47
3.2. Justification	41	41	31	47	71	46.2	29
5. Actors	14	15	5	9	5	9.6	4
6.1 Attitudes	15	18	9	13	18	14.6	5
6.2 Behaviors	16	17	8	11	23	15	5
6.4 Conflicts	22	22	21	12	43	24	3

Among the models, Gemini 2.5 was the most verbose in the interpretive and behavioral sections (Interpretation = 118; Justification = 71; Behaviors = 23; Conflicts = 43). In contrast, ChatGPT 5 Pro produced the most extensive responses in the Description dimension of the metaphors (84) and in the Actors dimension (15), while also tying for the highest value in Attitudes (18), alongside Gemini 2.5. As for DeepSeek, it stood out only in Justification (47), while ChatGPT o3 tended to produce shorter responses in contextual and behavioral variables (Actors = 5; Attitudes = 9).

To complement the word count analysis presented in Table 4, a qualitative review of open-ended responses was conducted, using the same items to ensure comparability. This review does not assume that greater length corresponds to higher

quality or depth; instead, it makes it possible to distinguish whether responses were primarily descriptive, interpretive, justificatory or inferential, as well as whether they tended toward synthesis, enumeration or argumentation.

Table 5. Qualitative synthesis of open-ended responses by rater type.

Open item	Experts	GAI/LLM	Comparative Reading
1. Description	Selection of key elements and features with interpretive value: spaces, boundaries, groupings, and the composition of the scene.	A more comprehensive and organized inventory of objects, colors, positions, actors, and spatial relationships.	The main difference was between prioritizing relevant evidence and providing a more exhaustive list.
2. Interpretation	Further reading on coexistence, separation, belonging, indifference, integration, and exclusion.	Identification of general patterns: diversity, groupings, partial coexistence, and tensions between groups.	There was greater contextual density in the expert panel and greater categorical systematicity in the models.
3.2. Justification	More nuanced arguments; some responses acknowledged the ambiguity or limitations of the scale used to classify coexistence.	More consistent justifications, generally presented as a balance between positive and negative factors.	The GAI systems provided regular justifications; the experts expressed greater caution in the face of visual ambiguity.
5. Actors	Brief and discreet identification: students, the university community, the institution, and, in some cases, faculty members or administrators.	Broader list: students, groups, faculty, administrators, the institution, and the university community.	The models expanded the repertoire, but with a higher risk of inferring non-explicit actors.
6.1. Attitudes	Emphasis on indifference, belonging, distance, rivalry, acceptance, or isolation, with a relational focus.	Organization into explicit categories: group pride, separation, prejudice, rivalry, inclusion, or the need for unity.	They agreed on membership and separation; the expert panel provided a more relational perspective, while the models offered a more categorical framework.
6.2. Behaviors	Greater caution was exercised when no explicit actions were taken; limited interaction, isolation, or a lack of sufficient evidence were noted.	Translating the metaphor into concrete actions: staying in groups, moving around, competing, collaborating, or sharing spaces.	The models inferred more behaviors; the experts avoided overinterpreting ambiguous scenes.

Open item	Experts	GAI/LLM	Comparative Reading
6.4. Conflicts	Identifying latent conflicts: segregation, indifference, inequality, lack of integration, or exclusion.	More explicit identification of conflicts: rivalry, competition, separation, tension between diversity and integration, or inequality.	The GAI systems produced additional recommendations; the expert panel tended to focus more on social and regulatory implications.

Overall, the GAI/LLM groups produced more extensive, organized, and enumerative responses, particularly regarding descriptions, actors, behaviors, and conflicts. The experts were more concise, but often more selective, cautious, and context-sensitive, particularly when the questions required justifying coexistence, interpreting attitudes, or assessing latent conflicts. Therefore, the comparison should be interpreted as a difference in the nature of the analysis: the GAI systems provide comprehensiveness and systematicity, while the expert panel offers prioritization, contextual caution, and sensitivity to social and normative dimensions.

3.2. Coexistence Index: Differences Between Humans and GAI

According to the Coexistence Index, the average reveals gaps between humans and GAI systems (Table 6; see Figure 6): the index was consistently higher in the scores assigned by the expert panel. In Metaphor 2, the difference was significant in favor of the human panel (Delta = +0.91; 4.28 vs. 3.37). In Metaphors 1 and 3, the differences were small (Delta = +0.15 and +0.21, respectively), meaning they practically coincide in their interpretation of the level of coexistence. When the image or its description offered integrative cues (e.g., shared spaces, movement, or groups), the expert panel assigned higher scores, while the models generated more conservative scores. This difference is significant because it defines the type of inferences that LLMs make regarding social constructs: it is not merely a matter of detecting objects, but of weighing their pedagogical and normative significance.

Table 6. Average coexistence rates.

Metaphor	Humans (Δ)	GAI (Δ)	Differences (H–GAI)
1	2.78	2.63	+0.15
2	4.28	3.37	+0.91
3	3.47	3.26	+0.21

Note. Scores on a scale of 1–5. A positive difference indicates a higher human score compared with the GAI.

Breaking down the GAI models, Table 7 shows that ChatGPT 5 Pro and ChatGPT o3 were the closest to human evaluations, according to the MAE (0.307 and 0.344, respectively), with small deviations in M1 and M3—and even slight overestimates (Δ<0) in some cases: o3 at M1=−0.117 and M3=−0.113, and 5 Pro at M3=−0.193—while at M2 both underestimated coexistence (0.723 and 0.803). In contrast, DeepSeek (MAE=0.758) and Gemini 2.5 Flash (MAE=0.571) showed the greatest discrepancies, with systematic underestimation across all three metaphors (e.g., M2=1.023 for both and M3=0.547/0.567), and ChatGPT 5 exhibited an intermediate pattern (MAE=0.424) but with a very large gap in M2=0.983.

In terms of the relative order of metaphors, ρ=1.00 for ChatGPT 5, DeepSeek, and Gemini 2.5 indicates that they reproduced the order observed in the human consensus (M1–M3 ranking), although they were off-scale (high deltas), whereas 5 Pro and o3 (ρ=0.50) were closer in magnitude but less stable in the ranking. Considering the index's 1–5 scale and adopting an operational threshold of Δ≥0.50 as a relevant difference for decision-making, the gaps were substantial in M2 for four out of five models and, additionally, in DeepSeek/Gemini for M3; therefore, the differences between models and the human panel were materially significant (especially in M2), even though formal statistical significance is not estimated here due to the limited number of metaphors (N=3).

Table 7. Differences in the index between the GAI model and humans.

GAI Model	M1 (Δ)	M2 (Δ)	M3 (Δ)	MAE	ρ Coexistence Index (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

Note: Positive Δ = the human score was higher than that of the model (underestimation by the AI); Negative Δ = the model overestimated compared to humans. MAE: mean absolute error between scores (average of human (H) and AI scores across the three metaphors). ρ: Spearman's rank correlation coefficient between average scores per metaphor (human compared with the GAI model).

On the other hand, as shown in Table 8, the agreement between humans and GAI on these scales was of moderate magnitude (approximate average Cohen's kappa = 0.475; approximate Krippendorff's alpha = 0.236), indicating partial alignment with some loss of detail across certain dimensions. When the cues present in the metaphors are behavioral and observable (e.g., acts of cooperation, antagonistic behaviors, or indicators of active participation), alignment increases considerably—cooperation reached kappa = 0.80 (alpha = 0.572), while antagonism and active participation had an approximate kappa of 0.67. In contrast, items requiring normative judgments or inferences about social structures (equity, inequality, hierarchical power) showed low agreement (kappa ≤ 0.25), indicating less concordance between the models and the expert panel in interpreting nuances related to justice and the distribution of power. In other words, the models detect visible groupings or inequalities, but do not reproduce the normative evaluations provided by the experts with the same fidelity.

Table 8. Results for items using a Likert scale.

Item (summary)	Cohen's κ (H vs GAI)	Krippendorff's α
Cooperation	0.80	0.572
Antagonism	0.67	0.238
Active participation	0.67	0.409
Inclusive Diversity	0.55	0.354

Item (summary)	Cohen's κ (H vs GAI)	Krippendorff's α
Segregation/marginalization	0.25	0.390
Hierarchical authority	0.25	0.128
Peaceful coexistence	1.00†	0.105
Equity	0.10	0.021
Inequality	0.00	0.103
Direct violence	NaN	0.042
Average (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».

On the other hand, Table 9 shows the overall results by model, including all three metaphors. Using Cohen's κ as the primary criterion and Krippendorff's α as a tiebreaker, the best model by item was: for Cooperation, GPT-5 Pro stood out (κ=.80; α=.613, higher than GPT-3 in α); in Antagonism, Gemini 2.5 Flash (κ=.67; α=.359); in Active Participation, DeepSeek (tie at κ=.67, tiebreaker by α=.302 versus GPT-5's α=.209); in Integrated Diversity, DeepSeek (κ=.64); in Segregation/Marginalization, GPT-5 Pro (κ=.27); in Hierarchical Power, GPT-5 Pro (tie in κ=.25, higher α=.254); in Peaceful Coexistence, DeepSeek (κ=1.00; probable ceiling effect with N=3); in Equity, GPT-5 Pro (tie in κ=.31, higher α=.094); in Inequality, GPT-5 Pro (κ=.21); and in Direct Violence, DeepSeek (only one with an estimate; κ≈0, α=.169). On average, DeepSeek obtained the highest κ (0.403) and GPT-5 Pro the highest α (0.282), but with a κ (0.388) similar to DeepSeek.

Table 9. Results of the agreement index by item, model vs. humans (Cohen's κ and 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
Cooperation	0.571	0.549	0.800	0.613	0.800	0.464	0.727	0.600	0.640	0.565
Antagonism	0.000	0.076	0.667	0.264	0.000	0.077	0.400	0.300	0.667	0.359
Active participation	0.667	0.209	0.625	0.296	0.625	0.246	0.667	0.302	0.595	0.355
Inclusive Diversity	0.545	0.315	0.571	0.495	0.571	0.401	0.640	0.172	0.400	0.332
Segregation/ marginalization	0.250	0.270	0.267	0.475	0.250	0.285	0.250	0.320	0.250	0.293
Hierarchical authority	0.250	0.140	0.250	0.254	0.250	0.179	0.250	0.150	0.000	-0.011
Peaceful coexistence	0.000	0.128	0.000	0.199	0.000	0.085	1.000†	0.306	0.400	0.208
Equity	0.308	0.067	0.100	0.064	0.308	0.094	0.100	0.072	-0.200	-0.049
Inequality	0.000	0.013	0.211	0.171	0.100	0.128	0.000	0.012	-0.200	-0.055
Direct violence	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

Note. "GPT 5 κ/α" = Cohen's κ / Krippendorff's α (ordinal) for ChatGPT 5; "GPT 5 Pro κ/α" = κ/α for ChatGPT 5 Pro; "GPT o3 κ/α" = κ/α for ChatGPT o3; "DS κ/α" = κ/α for DeepSeek; "Gf κ/α" = κ/α of Gemini 2.5 Flash. κ reflects the model's agreement with the human consensus on a per-item basis (weighted for ordinal scales where applicable); α was estimated using all model runs against the human panel. In cases of tied κ values,

the model with the higher α was considered "better." † A κ=1.00 with a low N may indicate a ceiling effect rather than robust agreement.

In the binary items designed to capture specific processes—scenes of violence containment (7.1), constructive conflict resolution (8.1), and conditions for lasting peace (9.1)—both the experts and the GAI models mostly responded with the "No" option (0). That is, in the three metaphors reviewed, there were insufficient indications in either the descriptions or the guided readings to affirmatively confirm the presence of containment mechanisms, resolution practices, or transformations toward peace (see Table 10). First, the low kappa values on these questions do not necessarily invalidate the instrument: rather, they indicate a lack of signal in the stimuli under the test conditions. Second, when attempting to assess reliability in the detection of infrequent processes (e.g., explicit containment actions), it is advisable to increase the number and diversity of stimuli or rebalance the categories to produce information interpretable by the agreement metrics. Furthermore, the closed-ended items in Table 10 were reviewed individually according to the models; there were no significant differences to report.

Table 10. Results for closed items.

Ítem	Escala	Fleiss' κ (multi-rater)	Cohen's κ (Consensus on H vs. GAI)	Krippendorff's α
3.1 Perception	ordinal	—	0.00	0.023
4. Division into subgroups	binary	0.014	0.00	0.024
7.1 Scenarios for violence containment	binary	−0.029	NaN	−0.018
8.1 Constructive resolution	binary	−0.001	NaN	0.010
9.1 Conditions for lasting peace	binary	0.095	0.00	0.105

3.3. Reliability of the judgment

As shown in Table 11, generalizability estimates provide a complementary and more stable perspective than Kappa coefficients when the number of stimuli is small or the sample is unbalanced. In the G-study, we calculated variance components and reliability coefficients for three panels: human (n=6), GAI (n=25), and mixed (n=31). The human panel showed moderate reliability (G ≈ 0.73; Phi ≈ 0.68) accompanied by a very high image × judge interaction (sigma2_pxr = 0.908). This interaction variance indicates that experts differ from one another in how they evaluate each metaphor; that is, the differences are not concentrated so much in the stimuli as in each judge's individual response to each image, a characteristic expected in interpretive tasks where professional subjectivity influences judgment.

In contrast, the GAI panel showed very high reliability (G ≈ 0.98; Phi ≈ 0.98), with low variance (σ^2_r ≈ 0.083) and lower residual variance; this means that the models reproduced criteria very consistently across runs. The mixed panel combined both properties, yielding an approximate G = 0.976 and an approximate Phi = 0.970: the combination of GAIs and expert humans produces very high reliability, suggesting that

GAIs stabilize the overall score while human participation preserves the critical anchor for content validity. As for the model with the highest reliability under these conditions, ChatGPT 5 Pro achieved an approximate G of 0.98. In practical terms, the G-study shows that the inclusion of automated models reduces inter-rater variability and increases the reliability of the evaluation system, although the interpretation of specific cases still requires human oversight.

Table 11. G-study Results.

Panel/Model	n_judges	σ^2_p	σ^2_r	σ^2_p×r	G	Φ
Humans	6	0.404	0.214	0.908	0.728	0.684
GAI (global)	25	0.896	0.083	0.366	0.984	0.980
Mixed (humans and GAI systems)	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

Note. G = relative reliability (consistent ranking of objects); Φ = absolute reliability (absolute decisions); σ^2_p = variance among objects; σ^2_r = variance among judges; σ^2_p×r = object×judge interaction (and residual).

3.4. Ideal mix for human judging plus GAI

The D-study translates variance components into recommendations regarding how many human judges and GAI runs to include in order to achieve desired reliability thresholds. The results show that the largest increase in G occurs when incorporating GAI models: increasing the number of runs to 5–10 produces substantial increases in G, while adding more than 10–15 runs yields diminishing returns. Specifically, small combinations such as 1 human evaluator + 10 GAI runs already achieve an approximate G of 0.95 (approximate Phi of 0.94), and 1–2 humans + 15 GAI runs raise G above 0.96; on the other hand, panels with multiple humans but no GAI exhibit lower G values (for example, 2–4 human judges without AI do not reach those thresholds).

Table 12. G-theory D-study: Expected G and Φ values for H + GAI mixtures

H - GAI	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

Note: Cells = G / Φ. Bold text indicates G ≥ 0.94. Combinations not shown were estimated but are omitted for brevity.

Furthermore, when the number of GAI runs is 10 or more, adding more than 1–2 human evaluators does not increase reliability and may even slightly reduce it due to the increased inter-rater variance introduced by humans. The operational

interpretation under these conditions is as follows: for institutional monitoring and scaling tasks, a compact mix—1–2 supervising human judges alongside approximately 10 GAI runs—offers a balance between cost and reliability.

Figure 3 depicts, via a heatmap, the expected G coefficient for different mixtures of human judges (H) and GAI models (A) in the panel. The horizontal axis shows the number of GAI runs (0, 5, 10, 15, 20, 25) and the vertical axis the number of humans (1–6). Lighter shades indicate higher reliability (G). With A=0, G increases with H but remains moderate: ≈0.45 (2H), ≈0.56 (3H), ≈0.63 (4H), ≈0.68 (5H), and ≈0.73 (6H). Case 1H+0A appears without an estimate (blank cell) because with a single judge, G is not defined for relative decisions.

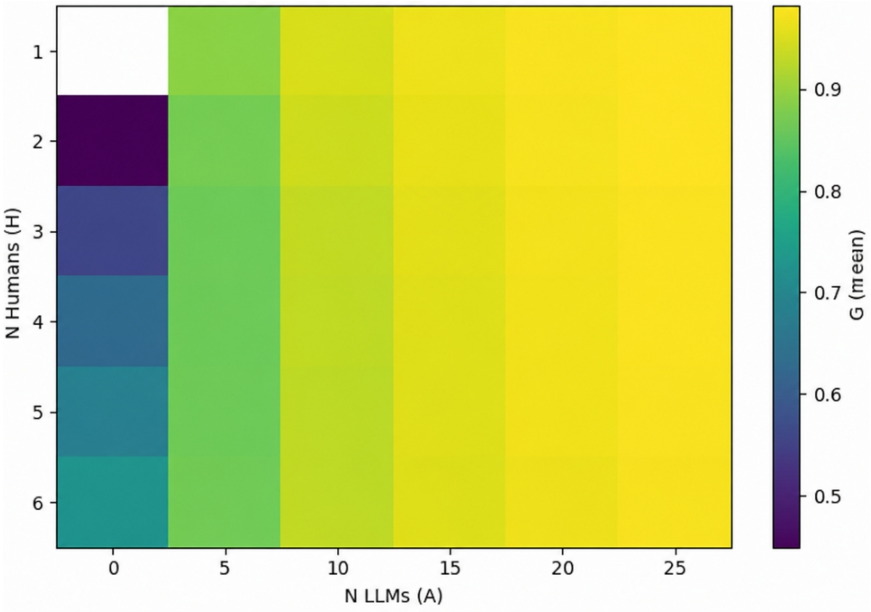

Figure 3. D-study results.

3.5. Top-performing LLM

The model ranking was constructed using a composite score that weights Kappa-Likert (70%), Kappa-binary (20%), and global index correlation (10%). Under this criterion, ChatGPT 5 Pro was the model closest to human consensus (approximate score = 0.975; approximate Kappa-Likert = 0.648; approximate Kappa-binary = 0.250), followed by ChatGPT o3 (approximate score = 0.725). This ranking should not be interpreted as universal "best performance," but rather as greater proximity to expert consensus on the three metaphors, given the versions and input conditions used in the study. The other models evaluated (ChatGPT 5, DeepSeek, Gemini 2.5 Flash) faithfully reproduced the overall order of the metaphors in the index (rho correlation = 1.00), although they exhibited lower Likert kappa; that is, they were less precise in the item-by-item details.

Table 13. LLM Rankings

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. Discussion and Conclusions

From the sociocultural framework of coexistence—which distinguishes between everyday interpersonal practices and institutional management (Fierro-Evans, 2013; Fierro-Evans & Carbajal-Padilla, 2019; Giddens, 1998; Weiss, 2015; Rodríguez-Figueroa, 2019, 2021)—the differences between experts and models are interpreted as distinct ways of working with the available evidence. The human interpretation was more contextual and holistic; that of the models, more standardized and dependent on explicit cues from the image or its textual description. When metaphors provided integrative clues—shared spaces, circulation, or visible links between groups—the human panel assigned higher coexistence scores, while the models produced more conservative scores. This contrast is consistent with the literature on the educational use of GAI: without guidance, LLMs can generate homogeneous responses or rely on superficial patterns; with clear direction, they can support processes of analysis and feedback (Zhai et al., 2024; Kosmyna et al., 2025; Nasr et al., 2025).

When comparing models, ChatGPT 5 Pro showed the highest overall alignment with human consensus. This advantage should be interpreted with caution because the ChatGPT variants used images and descriptions, while DeepSeek and Gemini relied on textual input. Therefore, the difference does not allow us to conclude that the model is generally superior, but rather that it is more closely aligned under the conditions of this study. In the case of DeepSeek, the matches were more visible in specific items, but not always in the overall magnitude of the index. This differential calibration aligns with studies on text evaluation where models replicate patterns but require explicit pedagogical anchors, supporting theory, and human review to sustain overall judgments (Bui & Barrot, 2025; Bouziane & Bouziane, 2024; Usher, 2025; Atasoy & Arani, 2025).

Differences among items on the Likert scale confirmed that the models align better with human behavior when the cues are observable. Agreement was high regarding visible behaviors (e.g., cooperation, antagonism, and participation). In contrast, alignment decreased when normative and subjective constructs (equity, inequality, hierarchical power) came into play: in those cases, human judgment provided contextual understanding, institutional histories, and attention to power relations that the models did not reproduce with the same precision. Therefore, LLMs perform best with clear, standardizable signals, but require blueprinting, specific grounding, and subsequent review to approximate evaluations that depend on notions

such as justice and the distribution of opportunities (Doughty et al., 2024; Mistry et al., 2024; Scaria et al., 2024; Arif et al., 2024; May et al., 2025).

In terms of text production, all GAI systems wrote more than the experts, especially regarding interpretation and conflict. This finding is not interpreted as evidence of greater depth, but rather as a characteristic of text generation in response to open-ended prompts: the models tend to cover more points, add redundancies, and offer lengthy explanations. The experts, on the other hand, were more concise and focused their writing where it made sense to elaborate. Length can be useful if guided by specific prompts, clear rubrics, and length limits; otherwise, it can increase analytical noise. This precision directly addresses the need to distinguish between word count, depth, conciseness, text type, and argumentative quality (Kiryakova, 2025; Nasr et al., 2025).

Regarding reliability, the findings revealed a pattern: the GAI systems were highly consistent across runs; the human panel was more variable—as expected in interpretive tasks—and combining both sources increased stability without eliminating the contextual anchor of expert judgment. In practical terms, for monitoring, debugging, and exploratory studies, a small mix of people with a moderate number of GAI runs is recommended (for example, one to two experts with around ten runs of one or more models). For sensitive decisions, especially when equity, peace, or power relations are at stake, it is advisable to increase the weight of human judgment (Shavelson & Webb, 1991; Brennan, 2001; AERA, APA, & NCME, 2014).

4.1. Scope, limitations, and projections

This study should be viewed as a preliminary reference rather than a definitive assessment of GAI. Technology is evolving rapidly, and it is likely that more recent or future versions will improve direct image reading, multimodal reasoning, and pedagogical calibration. Precisely for this reason, the value of this work lies in providing a documented baseline: it shows what happens when experts and models are compared under explicit input conditions, how agreement varies across observable and normative dimensions, and which combinations of evaluators yield the highest reliability. These results are part of a potential set of investigations; in the future, deeper pedagogical analyses of student work could be conducted, updated multimodal models could be incorporated, more specific rubrics could be compared, and not only text length but also argumentative quality, interpretive depth, and usefulness for educational feedback could be evaluated.

In conclusion, and in response to the research objective, humans and GAI systems largely agreed on observable dimensions and differed more on normative and subjective dimensions. While this pattern might seem to be expected, its significance lies in having documented it empirically in a specific task involving the evaluation of visual metaphors related to school coexistence. Thus, what is seemingly obvious acquires scientific value when it is specified, measured, and situated within a specific empirical field. The human panel demonstrated a greater ability to contextualize concepts such as justice, equity, or power; the models produced more conservative scores when these constructs were not explicitly present in the image or description. ChatGPT 5 Pro was the system closest to human consensus under these conditions, while DeepSeek stood out on specific items. All the AI systems generated more text than the human panel, but this abundance should be understood as a stylistic trait

rather than a direct measure of quality. The study's main contribution is methodological and pedagogical: it offers a framework for combining human judges and AIs, identifying the areas where automation can be helpful, and defining where expert interpretation remains essential—especially for future analyses of student work on school coexistence.

5. References

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