

Can Large Language Models Substitute for Human Raters? Reliability and Validity of Urban Design Quality Assessment

Wookjae Yang

`wookjae.yang@inu.ac.kr`

Incheon National University <https://orcid.org/0000-0002-6305-3877>

Jihyun Hwang

Incheon National University <https://orcid.org/0009-0009-8268-3258>

Reid Ewing

University of Utah <https://orcid.org/0000-0002-4117-3456>

Research Article

Keywords: Urban Design Quality, Generative AI, Inter-rater Reliability, AI-Human Agreement, Multimodal Large Language Model (LLM)

Posted Date: June 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10152272/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

This study evaluates the reliability and initial validity of multimodal LLMs for urban design assessment in a low-rise neighborhood in Seoul. Three human raters and three LLMs evaluated streetscape characteristics at 40 observation points using a standardized rubric. Reliability was assessed using ICC, Fleiss' kappa, and human–LLM agreement measures. LLMs showed relatively high reliability for visually explicit design features but lower reliability for variables requiring proportional estimation, contextual interpretation, or detection of small-scale details. While some associations with perceived comfort were preserved, LLMs did not fully replicate human-rated results. Findings indicate that LLM suitability is element-specific rather than universal.

1. Introduction

The emergence of multimodal Large Language Models (LLMs) presents a fundamental methodological question for urban planning and design research: can LLMs reliably replace trained human raters in assessing urban design quality? While traditional computer vision approaches have focused on extracting and quantifying physical features from street-level imagery through semantic segmentation (Cordts et al., 2016; Chen et al., 2018), LLMs, including GPT, Gemini, and Claude, can now interpret design assessment instructions and apply structured scoring rubrics directly to visual images, positioning them as potential substitutes for human raters rather than mere feature detection tools.

Urban design quality assessment traditionally requires trained observers to evaluate fine-grained environmental attributes, such as enclosure, transparency, building material quality, signage conditions, and street furniture provision, using standardized checklists and scoring frameworks (Ewing & Clemente, 2013a; Ewing & Handy, 2009; X. Gao & Gao, 2010). These assessments require judgments about how features are configured and maintained, rather than merely identifying which objects are present. The popularity of geo-tagged street-level imagery from web mapping services and crowdsourcing platforms has created unprecedented visual datasets (Fan et al., 2024; Goodchild, 2007), yet the labor-intensive nature of human-based rating systems severely constrains their scalability, particularly in dense residential environments where conditions vary significantly over short distances.

Whether LLMs can replace human raters depends entirely on the reliability, agreement, and validity of the measurement. Methodological standards in built environment research emphasize inter-rater reliability, typically assessed using the Intraclass Correlation Coefficient (ICC) for continuous variables and Fleiss' kappa for categorical variables. If LLMs produce internally consistent ratings that align with trained human observers across diverse urban conditions, they could transform large-scale urban design assessment by processing massive volumes of imagery at substantially reduced cost and time. Conversely, systematic inconsistencies or biases would limit their utility to supportive rather than replacement roles.

This study systematically evaluates whether multimodal LLMs can reliably replace human raters by comparing assessments of urban design quality in a low-rise villa neighborhood in Hwagok-dong, Seoul.

Using street-level imagery sampled across the district, three trained human raters and three LLM platforms applied identical standardized rubrics to evaluate quantitative and categorical streetscape variables. Inter-rater reliability was assessed separately for human raters and LLMs using ICC and Fleiss' kappa, while human–LLM agreement was examined through Bland–Altman analysis and cross-group reliability statistics.

The research addresses two core questions: **(1)** Can LLMs reliably evaluate urban design quality compared with human raters? **(2)** Which urban design variables can be reliably evaluated by both humans and AI models?

By focusing strictly on measurement consistency and reliability metrics, this research provides empirical evidence on the conditions under which LLMs can serve as practical replacements for human raters in urban design assessment and identifies the current limits of AI capability in structured environmental evaluation. In this study, the human raters are treated as the gold standard in urban design assessment. If AI can reproduce human measurements, it will allow urban designers to scale up from small samples of streets subject to human assessment to large samples subject to AI assessment, at a small fraction of the processing time and without loss of accuracy in measurement.

2. Literature Review

2.1 Measuring Urban Design Quality

Traditional methods for measuring urban design quality include systematic field audits using standardized checklists, street-view-based visual assessments, expert panel evaluations, and survey-based perception studies. For example, (Kim et al., 2025) evaluated low-rise residential areas using structured analysis frameworks combined with resident surveys, enabling systematic comparison of design conditions across different housing typologies. While effective, these human-based approaches are resource-intensive and limited in geographic scalability.

The proliferation of geo-tagged street-level imagery through mobile Internet technologies, web mapping services, and crowdsourcing platforms has created unprecedented visual datasets covering urban environments at fine spatial resolution (Goodchild, 2007). These data sources, combined with advances in urban big data analytics, enable researchers to analyze physical environments, socioeconomic conditions, and human dynamics at scales previously unattainable (Fan et al., 2024). Geospatial artificial intelligence (GeoAI) technologies have enriched classical measurement approaches by integrating spatial effects such as dependency, distance decay, and scale sensitivity (Gao et al., 2024 ; Zhu et al., 2020).

Deep learning methods, particularly convolutional neural networks applied to semantic image segmentation, have enabled large-scale extraction of physical streetscape features from imagery with high accuracy. State-of-the-art approaches have extended beyond feature extraction to approximate human perceptual responses, demonstrating that computational models can replicate aggregate crowd-

sourced judgments of environmental qualities such as safety and beauty at city scale (Yao et al., 2019; Zhang et al., 2019)

However, existing AI-based approaches have predominantly focused on identifying and quantifying physical elements rather than applying structured evaluative criteria as trained human raters do. The emergence of multimodal Large Language Models capable of interpreting both textual instructions and visual inputs represents a critical methodological development that may bridge this gap between feature extraction and systematic evaluation. Unlike task-specific computer vision models, LLMs can potentially apply complex scoring rubrics and evaluation frameworks directly to street-level imagery, positioning them as potential replacements rather than supplements to human raters.

2.2. Inter-Rater Reliability in Urban Design Assessment

The fundamental methodological question for any measurement system—whether human-based or AI-generated—is reliability. Inter-rater reliability quantifies the degree to which independent evaluators produce consistent assessments when applying identical criteria to the same environments. Without adequate reliability, observed relationships between design characteristics and planning outcomes may reflect measurement error rather than genuine environmental effects.

Inter-rater reliability is typically quantified using the Intraclass Correlation Coefficient (ICC) for continuous variables and Fleiss' kappa for categorical variables. ICC values above 0.75 are generally considered indicative of good to excellent reliability, while values below 0.50 suggest poor agreement that limits analytical utility (Koo and Mae, 2016). These empirical thresholds provide objective criteria for determining which urban design elements can be measured consistently enough to support rigorous analysis of urban design effects on human outcomes such as street life, traffic safety, and physical activity generation.

For LLMs to serve as viable replacements for human raters, their outputs must demonstrate both internal consistency across repeated evaluations and sufficient agreement with trained human observers to constitute valid measurements of the same underlying constructs. The absence of systematic reliability testing for LLM-based urban design evaluations represents a critical methodological gap. Establishing clear empirical evidence of where LLM reliability meets or falls short of established standards is essential for determining the appropriate scope of AI integration in urban design assessment workflows and for assessing whether these systems can legitimately replace human evaluators in systematic environmental analysis.

2.3. Research Gaps

Although considerable progress has been made in measuring urban design quality and examining its relationship with human experiences in cities, several gaps remain. First, most urban design evaluation methods still rely heavily on human raters to conduct visual audits or expert assessments, which can introduce subjectivity and variability among evaluators. While inter-rater reliability tests are sometimes applied in built environment research, systematic evaluations of measurement reliability remain relatively

limited in urban design studies. Second, recent advances in artificial intelligence and image-based analysis have opened new possibilities for automated or AI-assisted environmental assessment. However, existing studies primarily focus on computer vision approaches for extracting physical features from images, and comparatively little attention has been given to evaluating whether AI-based systems can produce consistent assessments of qualitative urban design attributes traditionally judged by human observers. Third, the implications of measurement reliability for downstream analyses—such as examining the relationship between urban design qualities and subjective well-being—have received limited empirical attention. If environmental variables are measured inconsistently, the robustness of observed relationships between design characteristics and human outcomes may be compromised. Addressing these issues requires research that simultaneously examines the reliability of urban design assessments and their implications for behavioral or psychological outcomes.

In response to these gaps, this study evaluates whether AI-assisted assessments can reliably measure urban design quality in residential environments and compares their consistency with human raters using the Intraclass Correlation Coefficient (ICC). Focusing on a low-rise villa neighborhood in Seoul, the study identifies design variables that can be measured consistently and examines their associations with subjective well-being. By integrating reliability testing with well-being analysis, this research contributes to methodological discussions on measuring urban design and explores the potential role of AI as a complementary tool for assessing the urban environment.

3. Data Collection

3.1 Study area

The study was conducted in a low-rise villa neighborhood in Hwagok-dong, Seoul, South Korea (Fig. 1). The area is a dense residential neighborhood dominated by small multi-family housing units of five stories or fewer, commonly referred to as "villas," a prevalent urban housing typology across Korean cities. It is also often understood as an unplanned housing type, not developed through master planning at the district level, and is contrasted with apartments. The neighborhood features narrow streets, heterogeneous building conditions, materials, and designs, significant variation in streetscape quality, and a lack of integrated maintenance. These characteristics make Hwagok-dong particularly well-suited for examining neighborhood-scale variations in urban design quality and testing the capabilities of both human raters and LLMs to assess design quality in complex, fine-grained visual environments typical of Korean residential districts. This has also limited the development of generalized design strategies to further maintain the outdoor street environment.

Observation points were selected using a systematic spatial sampling approach designed to ensure representative coverage while minimizing selection bias. Initially, candidate points were generated at 100-meter intervals along the street network within the villa residential area to provide comprehensive spatial coverage of the study district. From this systematically distributed set of candidate locations, a random sampling method was used to select the final observation points for assessment using the

processing tool, Random Extract in QGIS. This random sampling method aimed to effectively capture the diverse range of streetscape conditions, building forms, and environmental features throughout the neighborhood while reducing potential spatial clustering and selection biases that could compromise the sample's representativeness. A total of 40 observation points were selected for the final analysis.

3.2 Urban Design and Environment Variables

Urban design quality was evaluated using a comprehensive set of variables that capture key visual and environmental characteristics of residential streetscapes. These variables were selected from established urban design assessment frameworks, particularly those proposed by (Ewing & Clemente, 2013b), which emphasize measurable design qualities such as enclosure, transparency, greenery, façade conditions, and street furniture. Each variable was operationalized with standardized visual evaluation criteria, enabling both human raters and LLMs to assess the same environmental features using consistent definitions and scoring scales.

The variables were grouped into two measurement types. Continuous variables included counts and ratios, such as the counts of commercial signs, retail facilities, streetlights, and safety facilities, as well as proportional measures, including the window-to-wall ratio, tree coverage percentage, and sky view factor. Categorical variables captured the presence, absence, or type of discrete streetscape features, including lane markings, building materials, outdoor dining facilities, and infrastructure elements such as fences and utility poles. A standardized scoring rubric was applied uniformly across all 40 observation points to ensure consistency in measurement.

3.3 Descriptive Statistics and Systematic Differences

Table 1 presents descriptive statistics comparing human and LLM assessments across all variables, along with statistical tests for differences between rater types, including paired t-tests for continuous variables and chi-squared tests for categorical variables. The paired t-tests revealed significant differences between human and LLM assessments across most variables. LLMs consistently produced lower counts and proportions than human raters for most continuous measures, with particularly large differences for window-to-wall ratio (Human: 50.0 ± 19.8 ; LLM: 21.6 ± 4.83), streetlights (Human: 0.913 ± 0.712 ; LLM: 0.146 ± 0.293), and sky view factor (Human: 37.5 ± 16.4 ; LLM: 22.1 ± 6.51). Only illegal parking and benches showed no significant differences in mean ratings across rater types.

For categorical variables, LLMs showed substantially lower detection rates for certain environmental features, particularly trash cans (Human: 60.8% vs. LLM: 5.0%) and CCTV cameras (Human: 28.3% vs. LLM: 2.5%). These differences, supported by chi-squared tests, indicated that human and LLM raters do not merely apply different numerical scales to the same perceptions, but rather detect and quantify environmental features at fundamentally different rates. This pattern warranted a subsequent reliability analysis to assess whether assessments were internally consistent within each rater type and sufficiently aligned across rater types when assessing urban design quality.

Table 1
Descriptive Statistics by Rater Type

Construct	Variable	Definition	Human	LLM	Test Statistics
			Mean (SD) or Freq (%)	Mean (SD) or Freq (%)	
Subjective Well-Being	comfort_level	Subjective comfort level perceived by surveyors at the survey point [5-likert scale]	3.23 (0.89)	3.02 (0.82)	1.88 *
Street / Environment	#comercial_signs	Number of commercial signs on buildings [count]	5.58 (6.97)	1.70 (1.61)	t = 5.94 ***
	#retail	Number of retail facilities [count]	2.11 (3.32)	0.817 (1.01)	t = 4.07 ***
	donation_bin	Number of used clothing donation bins [count]	0.17 (0.29)	0 (0)	t = 6.37 ***
	%trees	Percentage of street trees in the surrounding area [ratio]	9.53 (10.30)	4.68 (3.44)	t = 4.90 ***
	window-to-wall_ratio	Ratio of window area to wall area [ratio]	50 (19.80)	21.6 (4.83)	t = 15.22 ***
	#window_security_bars	Number of window security bars or window grilles [count]	2.14 (2.60)	0.07 (0.29)	t = 8.63 ***
	#lights	Number of streetlights [count]	0.91 (0.71)	0.14 (0.29)	t = 10.91 ***
	sky view factor	Degree of sky openness visible from the location [ratio]	37.50 (16.40)	22.10 (6.51)	t = 9.56 ***
	#street_litter	Number of litter items on roads and sidewalks [count]	1.80 (4.01)	0.12 (0.48)	t = 4.54 ***
	#safety_facility	Number of safety facilities such as	1.24 (2.08)	0.29 (0.60)	t = 4.78 ***

Construct	Variable	Definition	Human	LLM	Test Statistics
			Mean (SD) or Freq (%)	Mean (SD) or Freq (%)	
		speed bumps and bollards [count]			
	# illegal_parking	Number of illegally parked vehicles [count]	1.12 (1.36)	0.85 (0.71)	t = 1.93 *
	#benches	Number of benches [count]	0.04 (0.18)	0.03 (0.17)	t = 0.18
	lane_marking	None	11 (9.2%)	4 (3.3%)	χ^2 = 30.49 ***
		One or both sides	109 (90.8%)	116 (96.7%)	
	excessive_signage (more than five)	No excessive signage	71 (59.2%)	94 (78.3%)	χ^2 = 39.17 ***
		Excessive signage on one or both sides	49 (40.8%)	26 (21.7%)	
	plaza	None	111 (92.5%)	112 (93.3%)	χ^2 = 6.96 ***
		One or both sides	9 (7.5%)	8 (6.7%)	
	trashcan	None	47 (39.2%)	114 (95%)	χ^2 = 2.44
		One or both sides	73 (60.8%)	6 (5%)	
	cctv_cam	None	86 (71.7%)	117 (97.5%)	χ^2 = 0.21
		One or both sides	34 (28.3%)	3 (2.5%)	
	building_material	Dryvit	15 (12.5%)	15 (12.5%)	χ^2 = 101.32 ***
		Brick	51 (42.5%)	39 (32.5%)	
		Stone	42 (35%)	45 (37.5%)	
		Mixed	12 (10%)	21 (17.5%)	

Construct	Variable	Definition	Human	LLM	Test Statistics
			Mean (SD) or Freq (%)	Mean (SD) or Freq (%)	
	outdoor_dining	None	106 (88.3%)	114 (95%)	χ^2 = 13.35 ***
		One or both sides	14 (11.7%)	6 (5%)	
	pilotis_parking	None	30 (25%)	81 (67.5%)	χ^2 = 13.78 ***
		One or both sides	90 (75%)	39 (32.5%)	
	wires_poles	None	1 (0.8%)	21 (17.5%)	χ^2 = 0.74
		One or both sides	119 (99.2%)	99 (82.5%)	
	fence	None	33 (27.5%)	63 (52.5%)	χ^2 = 17.35 ***
		One or both sides	87 (72.5%)	57 (47.5%)	
	shelter	None	111 (92.5%)	109 (90.8%)	χ^2 = 4.04 **
		One or both sides	9 (7.5%)	11 (9.2%)	

4. Methods

4.1 Assessment Procedure

Human rater assessments were conducted through field surveys at all sampled points. Three trained raters independently assessed streetscape conditions using standardized criteria as presented in the variable definitions in Table 1. Prior to data collection, raters completed a calibration session involving example street view images and pilot observations to ensure consistent assessments. At each location, raters evaluated visible streetscape features, including building façades, street layout, vegetation, street furniture, signage, and safety facilities. Assessments were recorded using standardized survey forms. Continuous variables were measured as counts or estimated ratios (e.g., number of signs, window-to-wall ratio), while categorical variables were coded using predefined classifications. To minimize bias, raters conducted assessments independently without discussing observations or scores. Inter-rater reliability was subsequently evaluated using the independent ratings.

LLM assessments were performed using three multimodal models: GPT 5.3 Thinking, Gemini 3.5 Flash, and Claude Sonnet 4.6. For each observation point, the same high-resolution street-level image was provided to all models to ensure consistent visual input. A standardized prompt was developed to mirror the human evaluation protocol by defining all variables, scoring criteria, and classification rules. Identical prompts were used across models. To assess output stability, each image was evaluated multiple times under the same prompt conditions, with generation parameters controlled to reduce randomness where possible.

Even if the recent studies on visual assessment actively utilize a large amount of Google Street Views extracted from the API, this study could not reach that number of images since the main focus of this study is not to conclude the best prediction model but to check the reliability and validity of LLMs over human raters.

4.2 Inter-Rater Reliability Test

To systematically evaluate whether LLMs can serve as reliable assessors in urban design, three reliability tests were conducted. This analytical framework examined: (1) consistency among human raters, (2) consistency among LLM raters, and (3) agreement between human and LLM assessments. The comprehensive and sequential approaches enable identification of variables that demonstrate sufficient measurement reliability within each rater type and adequate cross-group agreement when assessing urban design qualities.

4.2.1 Within-Group Inter-Rater Reliability Assessment

Inter-rater reliability was assessed separately for human raters and LLM models using statistical methods appropriate to each variable type. For continuous and count variables, reliability was quantified using the Intraclass Correlation Coefficient with a two-way random-effects model for absolute agreement, single measures [ICC(2,1)]. This specification treats raters as a random sample from a broader population of potential evaluators and requires absolute score agreement rather than merely rank-order consistency (Shweta Bajpai et al., 2015). ICC values were calculated separately for the three human raters and the three LLM models across all 40 observation points. The ICC estimates were interpreted using established methodological guidelines (Koo & Li, 2016). For categorical variables, inter-rater reliability was evaluated using Fleiss' kappa (κ), which quantifies agreement among multiple raters simultaneously when assigning observations to nominal categories. Unlike Cohen's kappa, which is limited to two raters, Fleiss' kappa accommodates the three-rater structure used in both human and LLM evaluations (McHugh, 2012; Tai et al., 2024). Kappa values were interpreted according to established benchmarks (Landis & Koch, 1977).

4.2.2 Human-LLM Agreement Analysis

To determine whether LLM assessments can serve as valid replacements for human evaluations, agreement between rater types was examined using methods designed specifically for method comparison studies. For continuous variables, Bland–Altman analysis was conducted to quantify

agreement between human and LLM mean scores at each observation point. This approach calculates the mean difference between rater types (bias) and establishes 95% limits of agreement (LOA) using the formula:

$$LOA = d \pm 1.96 \times SD_d$$

, where d represents the mean difference between human and LLM scores and SD_d represents the standard deviation of those differences. Bland–Altman analysis is particularly appropriate for method comparison studies because it quantifies both systematic bias and the range of individual-level discrepancies between measurement approaches (Martin Bland & Altman, 1986). Narrow limits of agreement, combined with minimal bias, indicate that LLM and human assessments yield interchangeable results for a given variable, supporting the viability of AI replacement. For categorical variables, cross-group agreement was assessed using Fleiss' kappa between two groups. This approach directly quantifies categorical agreement between human and AI assessments, enabling comparison with within-group reliability values and assessing whether human–LLM agreement meets the same methodological standards used for human inter-rater reliability.

4.3 Construct Validity Test

To provide an initial assessment of whether the rated variables captured significant urban design constructs, the study estimated regression models linking rated design variables to subjective well-being and comfort level perceived by human and LLM raters. This analysis was intended as a minimal validity test, not as a definitive validation exercise. Following established measurement principles, expected relationships with an external outcome were treated as evidence of construct-related validity (Boateng et al., 2018; Mokkink et al., 2010).

The same model specification was estimated twice: once using the urban design elements rated by human raters and once by LLMs. The dependent variable was perceived comfort level, measured as the comfort level reported by survey participants at each survey location on a five-point Likert scale. Independent variables included design variables and environmental controls, such as temperature, humidity, and noise. If variables rated by LLMs showed coefficient directions and broad substantive patterns similar to those obtained from the human raters, this was interpreted as initial evidence that the LLM measures captured at least part of the same urban design constructs measured by human observers. Conversely, divergence between the models was interpreted more cautiously as suggesting that some variables may not have been operationalized equivalently across evaluator types.

5. Results

5.1. Inter-Rater Reliability

Table 2 shows that human raters generally achieved stronger inter-rater reliability than LLMs across most continuous variables. Human agreement was highest for clearly countable and visually explicit

features, such as #commercial_signs, #retail, #window_security_bars, and #donation_bin, with ICC values exceeding 0.6. By contrast, reliability was significantly lower for more ambiguous or infrequent features, including #illegal_parking and #benches.

The LLMs showed a more uneven pattern. Reliability remained relatively strong for # commercial_signs, #retail, and #window_security_bars, but dropped substantially for variables requiring proportional estimation or broader spatial judgment, such as %trees, sky_view_factor, #lights, and #safety_facility across LLMs.

Table 2
Intraclass Correlation Coefficients by Rater Type

Variable	Human Raters			LLM Raters		
	ICC [95% CI]	F	p	ICC [95% CI]	F	p
#comercial_signs	0.919 [0.868, 0.953]	34.3	< 0.01***	0.635 [0.473, 0.77]	6.28	< 0.01***
#retail	0.844 [0.755, 0.908]	17.4	< 0.01***	0.812 [0.708, 0.888]	14.5	< 0.01***
#donation_bin	0.785 [0.67, 0.871]	11.9	< 0.01***	N/A	N/A	N/A
%trees	0.564 [0.232, 0.766]	8.19	< 0.01***	0.380 [0.0994, 0.615]	4.47	< 0.01***
window-to-wall_ratio	0.593 [0.415, 0.741]	5.88	< 0.01***	0.407 [0.213, 0.596]	3.39	< 0.01***
#window_security_bars	0.626 [0.463, 0.764]	6.2	< 0.01***	0.692 [0.545, 0.809]	7.85	< 0.01***
#lights	0.481 [0.291, 0.656]	3.73	< 0.01***	0.179 [0.00786, 0.382]	1.74	< 0.05**
sky_view_factor	0.47 [0.0463, 0.74]	10.5	< 0.05**	0.260 [0.0429, 0.484]	2.92	< 0.01***
#street_litter	0.436 [0.246, 0.619]	3.38	< 0.01***	0.491 [0.303, 0.663]	3.88	< 0.01***
#safety_facility	0.462 [0.274, 0.64]	3.67	< 0.01***	0.093 [-0.0664, 0.294]	1.34	< 0.05**
# illegal_parking	0.371 [0.104, 0.601]	4.17	< 0.01***	0.430 [0.164, 0.647]	4.69	< 0.01***
#benches	0.222 [0.0387, 0.43]	1.90	< 0.01***	0.266 [0.072, 0.474]	2.08	< 0.01***

Note. Two-way random effects model, absolute agreement, single measures ICC(2,1). CI = Confidence Interval. N/A indicates variables not available in the AI dataset. *** $p < 0.01$, ** $p < 0.05$.

A similar pattern emerged for the categorical variables (Table 3). Human raters showed consistently stronger agreement across most variables, with especially high reliability for lane marking, plaza, and excessive signage. Moderate agreement was also observed for trash cans, CCTV cameras, outdoor dining, pilotis parking, and wires/poles, whereas agreement was lower for fences, shelters, and building materials.

Among the LLMs, reliability was strongest for building material and was moderate for shelter, fence, pilotis parking, and excessive signage. By contrast, agreement was substantially weaker for lane marking, plaza, outdoor dining, trash cans, CCTV cameras, and wires/poles. These results suggest that LLMs classify stable architectural or façade-level attributes more consistently than small-scale, partially occluded, or street-level elements.

Taken together, the reliability analysis suggests that LLMs do not underperform uniformly across all variables. Rather, their performance is selective. Human raters were more consistent for variables requiring contextual interpretation, proportional judgment, or detection of small environmental details, whereas LLMs performed comparatively well on a narrower subset of visually explicit and structurally stable features, such as #retail, #window_security_bars, and building_material. These findings indicate that current multimodal LLMs are not yet reliable substitutes for human raters across the full range of urban design elements, although they show clear promise for specific classes of visually legible attributes.

Table 3
Fleiss' Kappa by Rater Type (Categorical Type of Variables)

Variable	Human Raters			LLM Raters		
	κ	Z	p	κ	Z	P
lane_marking	0.860	11.7	< 0.01***	0.134	1.71	0.088*
excessive_signage (more than five)	0.725	9.99	< 0.01***	0.447	6.32	< 0.01***
plaza	0.760	8.32	< 0.01***	0.276	3.79	< 0.01***
trash_can	0.562	8.08	< 0.01***	0.303	3.75	< 0.01***
building_material	0.397	6.85	< 0.01***	1.000	17.8	< 0.01***
cctv_cam	0.547	6.70	< 0.01***	0.316	3.46	< 0.01***
fence	0.280	4.32	< 0.01***	0.463	6.90	< 0.01***
shelter	0.404	4.82	< 0.01***	0.504	5.93	< 0.01***
wires_poles	0.479	5.56	< 0.01***	0.287	4.25	< 0.01***
outdoor_dining	0.599	6.96	< 0.01***	0.298	3.27	< 0.01***
pilotis_parking	0.554	8.49	< 0.01***	0.459	5.71	< 0.01***
Note. κ = Fleiss' kappa; CI = Confidence Interval. Z-statistics test the significance of kappa values. *** $p < 0.01$, ** $p < 0.05$.						

5.2 Human-LLM Agreement

5.2.1 Bland–Altman Analysis (Continuous variables)

Bland–Altman analysis was used to evaluate whether human and LLM ratings for the continuous variables were sufficiently similar to be considered interchangeable (Table 4). Across all variables, the estimated biases were positive, indicating that, under the present coding of the paired differences, LLMs systematically assigned lower values than human raters on average. However, the magnitude of disagreement varied substantially across variables (Fig. 2).

The largest discrepancies were observed for variables requiring proportional judgment or broader spatial estimation. window-to-wall_ratio showed the widest limits of agreement (bias = 2.349; LOA = -57.461 to 0.786), followed by sky_view_factor (bias = 1.609; LOA = -35.354 to 4.537), %trees (bias = 1.058; LOA = -17.969 to 8.278), and #commercial_signs (bias = 0.892; LOA = -14.938 to 7.179). These wide intervals indicate substantial observation-level disagreement between human and LLM ratings, even where the average bias was moderate. In practical terms, these variables cannot be regarded as interchangeable across the two evaluator groups.

By contrast, agreement was relatively closer for low-frequency or low-range variables such as donation_bin, #benches, and #lights. # illegal_parking also showed comparatively narrower limits of agreement than most other variables. However, the apparent closeness of agreement among these variables should be interpreted with caution, because their limited variation and low prevalence reduce the possible range of disagreement.

Overall, the Bland–Altman results indicate that human and LLM ratings for continuous variables were not interchangeable for several substantively important urban design elements, particularly those involving proportional estimation, visual openness, or cumulative streetscape complexity. Rather than reflecting a simple rescaling of the same observations, the pattern of bias and limits of agreement suggests that human and LLM raters often quantified the same environments in systematically different ways.

Table 4
Bland-Altman Analysis Results

Variable	Bias	Lower LOA	Upper LOA
#commercial_signs	0.892	-14.938	7.179
#retail	0.372	-5.913	3.330
donation_bin	0.043	-0.706	0.365
%trees	1.058	-17.969	8.278
window-to-wall_ratio	2.349	-57.461	0.786
#window_security_bars	0.348	-6.385	2.260
#lights	0.087	-1.850	0.317
sky_view_factor	1.609	-35.354	4.537
#street_litter	0.454	-7.315	3.965
#safety_facility	0.234	-3.847	1.955
# illegal_parking	0.157	-2.227	1.685
#benches	0.013	-0.177	0.169

5.2.2 Fleiss' Kappa (Categorical variables)

Cross-group agreement between human and LLM ratings for the categorical variables was generally modest and highly variable across items (Table 5). The strongest human–LLM agreement was observed for excessive banners ($\kappa = 0.670$), indicating that this was the only categorical variable for which the two evaluator groups approached substantial agreement. Lane marking and plaza showed weaker but still positive agreement (both $\kappa = 0.360$), while fence ($\kappa = 0.297$) and outdoor_dining ($\kappa = 0.279$) also fell in the low-to-moderate range.

For several variables, however, agreement was limited or absent. shelter ($\kappa = 0.217$) and pilotis_parking ($\kappa = 0.044$) showed little correspondence between human and LLM classifications, despite moderate within-group reliability reported earlier for at least some of these variables. More notably, agreement was negative for trash can ($\kappa = -0.254$), cctv_cam ($\kappa = -0.176$), and wires_poles ($\kappa = -0.060$), indicating that human and LLM raters were not merely inconsistent but often classified the same scenes in systematically different ways.

Taken together, these results suggest that human–LLM agreement for categorical variables was substantially weaker than within-group reliability, particularly for small-scale elements, such as trash can, cctv_cam, wires_poles, and items found at low-frequency levels, such as plaza, fence, and outdoor_dining. In contrast, agreement was relatively stronger for more visually prominent and threshold-based variables such as excessive banners. This pattern supports the broader conclusion that current multimodal LLMs do not yet provide categorical judgments based on the operationally predefined criteria that are consistently interchangeable with those of trained human observers across the full range of urban design elements.

Table 5
Human-LLM Agreement Based on Fleiss' Kappa

Variable	Human Versus LLM		
	K	Z	p
lane_marking	0.360	2.28	< 0.01***
excessive_banners	0.670	4.24	0.022**
plaza	0.360	2.28	0.022**
trash_can	-0.254	-1.60	0.109
cctv_cam	-0.176	-1.12	0.264
fence	0.297	1.88	0.060*
shelter	0.217	1.37	0.169
wires_poles	-0.06	-0.42	0.673
outdoor_dining	0.279	1.77	0.077*
pilotis_parking	0.044	0.28	0.778
Note. κ = Fleiss' kappa; CI = Confidence Interval. Z-statistics test the significance of kappa values. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.			

5.3 Validity Test

Table 6 presents the regression models linking the human-rated and LLM-rated urban design elements to comfort level. As previously defined, the dependent variable was perceived comfort level, measured as

the comfort level reported by survey participants at each survey location on a five-point Likert scale. At the overall model level, the two specifications performed very similarly. The human-rated model explained 43.0% of the variance in comfort level, while the LLM-rated model explained 42.6%, and both models were statistically significant. This similarity in overall fit suggests that, at an aggregate level, LLM-based ratings retained a broadly comparable amount of explanatory information to human ratings.

At the variable level, however, the results were more selective. Three variables showed consistent directional associations across both models. Tree coverage was positively associated with comfort in both humans and LLMs, with coefficients of 0.028 and 0.065, respectively. Window-to-wall ratio was also positively related to comfort in both models, although the effect was stronger in the LLM-based specification, where the coefficient reached 0.047 compared with 0.008 in the human-rated model. Fence presence was negatively associated with comfort in both models, with coefficients of -0.446 for the human ratings and -0.275 for the LLM ratings. These shared patterns provide the clearest evidence that the two rating systems captured at least some common urban design dimensions in substantively similar ways. From the perspective of Ewing and Handy's urban design qualities, these variables are visually salient, occupy a substantial share of the streetscape, and shape how observers perceive greenery, openness, boundary conditions, and façade engagement—particularly through the perceptual dimensions of transparency, enclosure, human scale, imageability, and coherence.

Divergence was concentrated in a smaller set of variables. Pilotis parking showed a strong positive association with comfort only in the human-rated model, whereas lane marking and building material were significant only in the LLM-rated model. Window security bars also differed in direction across the two models. These contrasts suggest that human and LLM raters did not assign equivalent environmental meaning to all streetscape features, even when applying the same formal rubric.

Among the control variables, only noise was significant, and only in the human-rated model, where higher noise levels were associated with lower comfort. Temperature and humidity were not significant in either model.

Table 6
Ordinary Least Squares Regression Models Predicting Comfort Level

Variable	Model 1: Human		Model 2: LLM	
	Coef.	Std. Err.	Coef.	Std. Err.
# illegal_parking	-0.063	0.053	0.065	0.106
%trees	0.028***	0.009	0.065***	0.023
#lights	0.055	0.114	0.314	0.258
#benches	0.292	0.455	0.560	0.435
#commercial_signs	0.028	0.024	-0.182*	0.103
#retail	-0.049	0.046	0.266*	0.160
window-to-wall_ratio	0.008*	0.004	0.047**	0.019
#window_security_bars	0.064**	0.032	-0.123	0.262
sky_view_factor	-0.008*	0.005	-0.018*	0.011
lane_marking1	-0.185	0.296	0.967**	0.400
pilotis_parking1	0.633***	0.225	-0.221	0.151
fence1	-0.446**	0.187	-0.275*	0.151
building_material (Brick)	-0.133	0.241	-0.554**	0.241
building_material (Stone)	-0.086	0.254	-0.869***	0.261
building_material (Mixed)	-0.096	0.318	-0.732**	0.286
temp	0.017	0.057	-0.058	0.055
humid	-0.012	0.024	-0.023	0.023
noise	-0.023*	0.012	-0.005	0.010
Constant	4.453	3.113	5.194*	3.035
Note. Unstandardized coefficients reported with standard errors in parentheses. Binary variables coded as 1 = present, 0.5 =, 0 = absent. building_material reference category = Dryvit. *p < 0.1; **p < 0.05; ***p < 0.01.				

6. Discussion

This study examined whether multimodal LLMs can substitute for trained human raters in assessing urban design quality in a low-rise residential neighborhood in Seoul. Three findings are central. First, LLM performance was selective rather than general. The models showed stronger reliability for visually explicit, structurally stable attributes but weaker reliability for variables requiring proportional estimation,

contextual interpretation, or detection of small-scale, low-frequency environmental elements. Second, within-group consistency did not translate into cross-group equivalence. Agreement between human and LLM ratings was consistently lower than reliability within either group, indicating that internally consistent AI ratings do not necessarily reproduce human judgments. Third, the validity analysis provided partial rather than complete support for LLM-based measures. Some variables exhibited substantively similar associations with comfort across the human and LLM models, but others diverged sharply, indicating that the two evaluator types did not operationalize all urban design constructs in equivalent ways.

These findings matter because they suggest that LLMs are not simply weaker or noisier versions of human raters. Instead, they appear to follow a different perceptual logic. LLMs rated variables such as retail presence, commercial signs, window security bars, and building materials more consistently than variables such as tree coverage, sky view factor, lighting, and safety facilities. The difference is not trivial. The former are visually bounded, object-like, and often architecturally salient, whereas the latter require proportional estimation, integration of multiple visual cues, or interpretation of spatial relationships across the scene. In other words, the challenge is not only one of model accuracy but also of construct validity. Overall, the results indicate that LLM performance varies systematically by urban design quality: the models are more reliable for visually explicit architectural and activity-related cues, but less reliable for qualities tied to openness, greenery, and fine-grained public-realm conditions.

This distinction is especially consequential in low-rise villa neighborhoods in Korea. These environments are visually dense, incrementally modified, and often marked by small details, partial occlusion, and heterogeneous building conditions. In such settings, urban design quality rarely emerges from a single dominant feature. Rather, it arises from the cumulative interplay among façade treatment, vegetation, barriers, signage, parking, and maintenance. Human raters can synthesize these conditions in context. LLMs, by contrast, appear to perform better when the target attribute is directly visible and less reliant on contextual inference. The findings therefore underscore a broader point for urban design theory: not all dimensions of design quality are equally amenable to standardized machine assessment, even when they can be formally defined in a rubric.

The study also makes a methodological contribution by distinguishing reliability, agreement, and initial validity, rather than collapsing them into a single notion of model performance. This distinction is important because AI-based urban assessment is often evaluated primarily through predictive performance or internal consistency. The present results show why that is insufficient when the question is substitution rather than automation. Some variables exhibited moderate reliability within the LLM group but weak agreement with human ratings. This means that an AI system may apply its own scoring logic consistently while still producing measures that are not directly comparable to those of trained observers. The Bland–Altman results reinforce this point: for several key variables, including window-to-wall ratio, sky view factor, tree coverage, and commercial signs, the limits of agreement were wide, indicating substantial observation-level discrepancies. These were not simply scale differences. Human and LLM raters were often quantifying the same scenes in different ways.

The validity results extend this argument. At the model level, the human- and LLM-rated specifications explained a similar share of the variance in comfort, suggesting that LLM-based ratings retained meaningful environmental information. At the variable level, however, the overlap was selective. Tree coverage, window-to-wall ratio, and fence presence showed consistent directional associations across the two models. From the perspective of Ewing and Handy's urban design qualities, these variables are visually salient, occupy a substantial share of the streetscape, and directly shape perceived greenery, openness, boundary conditions, and façade engagement—particularly through the perceptual dimensions of transparency, enclosure, human scale, imageability, and coherence. Their consistency across models therefore suggests that LLMs can capture some higher-order spatial qualities in ways that remain substantively meaningful. At the same time, divergence in pilotis parking, window security bars, lane markings, and building materials indicates that not all streetscape attributes carry equivalent meaning across human and LLM assessments. The more important implication is therefore not whether LLMs can replace humans in general, but which urban design constructs can be measured by AI without losing substantive meaning.

These findings have several implications for urban planning and design research. First, LLMs appear better suited as selective measurement tools than as full replacements for human observers. A more realistic application is a hybrid workflow in which AI is used for visually explicit and stable variables, while human raters remain necessary for context-sensitive or low-frequency features. Second, AI-assisted urban assessment should be developed on an indicator-specific basis. Rather than assuming uniform performance across variables, researchers should identify which measures can be automated with acceptable reliability and agreement and which still require human judgment. Third, the study shows that evaluating AI as a measurement instrument requires more than a single performance metric. By combining ICC, Fleiss' kappa, and Bland–Altman analysis, this study demonstrates that reliability, agreement, and interchangeability are analytically distinct and should be assessed separately. Finally, for planning practice, the results suggest that LLMs are not yet suitable as fully autonomous evaluators in large-scale neighborhood audits. At present, they are better understood as screening or support tools that can reduce labor for some variables while still requiring human oversight, particularly in dense residential environments where subtle environmental cues matter for design assessment.

Several limitations should also be noted. The study was conducted in a single low-rise villa neighborhood in Seoul, and the findings may not generalize directly to other urban contexts. Although the sample was adequate for the present reliability analysis, it remains limited for broader claims about generalizability. Image inputs were standardized for the LLMs, but factors such as lighting, weather, seasonal vegetation, and camera perspective could not be fully controlled and may have influenced both human and AI evaluations. In addition, the validity analysis should be interpreted as initial rather than definitive. Similarities in coefficient direction and model fit are informative, but they do not establish full equivalence between human and AI measurement. More robust validation would require additional external criteria, known-group comparisons, and direct tests of differences in coefficients across evaluator types. Finally, urban design quality was necessarily operationalized through a limited set of

observable variables. Broader dimensions such as imageability, coherence, and social meaning may not be fully captured through single-image assessment, whether by humans or LLMs.

Overall, this study clarifies the current scope of AI substitution in urban design assessment. Current multimodal LLMs are not yet general substitutes for trained human raters. Their performance is strongest where urban design variables are visually explicit, spatially bounded, and structurally legible, and weakest where evaluation depends on contextual reading or cumulative scene interpretation. The most productive path forward is therefore not full replacement, but carefully validated hybrid assessment, in which AI expands the scale of urban design measurement while human judgment remains central to the interpretation of more ambiguous and context-dependent qualities.

7. Conclusion

In summary, this study shows that multimodal LLMs are not yet general substitutes for trained human raters in urban design quality assessment. Their performance is selective: they show relatively strong reliability for some visually explicit and architecturally legible variables, but weaker reliability and limited agreement for variables that depend on contextual interpretation, proportional estimation, or detection of small environmental details. The validity results point in the same direction. LLM-based ratings preserve some meaningful associations with comfort, but they do not reproduce the full pattern of relationships observed in human-rated measures. The broader implication is that AI substitution in urban design assessment should be understood as quality-specific rather than universal. In that sense, the present study contributes less to a simple yes-or-no answer than to a more useful conclusion: the future role of LLMs in urban design assessment is likely to be complementary, selective, and hybrid, rather than fully replacement-based.

It should be kept in mind that urban design assessment is just a means to an end. In several published studies, qualities such as transparency and imageability (familiar to urban designers) have been shown to relate to street life and pedestrian activity. These measures of urban livability are the ends that ultimately matter to residents, designers and planners, and public officials. Only to the extent that AI helps us build an evidentiary base that urban design qualities matter, does the present study contribute to human well-being.

References

1. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quinonez HR, Young SL (2018) Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Front Public Health* 6:366616. <https://doi.org/10.3389/FPUBH.2018.00149/BIBTEX>
2. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49/TABLES/7

3. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December*, 3213–3223. <https://doi.org/10.1109/CVPR.2016.350>
4. Ewing R, Clemente O (2013a) Measuring urban design: Metrics for livable places. *Measuring Urban Design: Metrics for Livable Places*, 1–183. <https://doi.org/10.5822/978-1-61091-209-9/SAVE-RESEARCH>
5. Ewing R, Clemente O (2013b) Measuring urban design: Metrics for livable places. *Measuring Urban Design: Metrics for Livable Places*, 1–183. <https://doi.org/10.5822/978-1-61091-209-9/SAVE-RESEARCH>
6. Ewing R, Handy S (2009) Measuring the Unmeasurable: Urban Design Qualities Related to Walkability. *J Urban Des* 14(1):65–84. <https://doi.org/10.1080/13574800802451155>
7. Fan Y, Tang L, Le H, Shen K, Tan S, Zhao Y, Shen, YL X., Dragan, Gašević (2024) |. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *Br J Educ Technol*, 00, 1–42. <https://doi.org/10.1111/bjet.13544>
8. Gao S, Hu, Yingjie., Li W (2024) *Handbook of geospatial artificial intelligence* (1st Edition). CRC Press
9. Gao X, Gao X (2010) Incentives of Environmental Design and Management in Urban Neighborhoods. *Environ Manage*. <https://doi.org/10.5772/10102>
10. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 2007 69:4, 69(4), 211–221. <https://doi.org/10.1007/S10708-007-9111-Y>
11. Kim YJ, Kim MJ, Jun HJ (2025) Revitalizing old low-rise residential environments by focusing on communities: case studies in Seoul, Korea. *Archnet-IJAR: Int J Architectural Res*. <https://doi.org/10.1108/ARCH-08-2024-0373>
12. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 15(2):155–163. <https://doi.org/10.1016/J.JCM.2016.02.012>
13. Landis JR, Koch GG (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159. <https://doi.org/10.2307/2529310>
14. Martin Bland J, Altman DG, FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT (1986) *Lancet* 327(8476):307–310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8). STATISTICAL METHODS
15. McHugh ML (2012) Interrater reliability: The kappa statistic. *Biochemia Med* 22(3):276–282. <https://doi.org/10.11613/BM.2012.031/FULLARTICLE>
16. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, De Vet HCW (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Research: Int J Qual Life Aspects Treat Care Rehabilitation* 19(4):539–549. <https://doi.org/10.1007/S11136-010-9606-8>

17. Shweta Bajpai R Bajpai, Himanshu K, Chaturvedi (2015) An Overview of Concepts and Methods. J Indian Acad Appl Psychol 41(3):20–27.
https://www.researchgate.net/publication/273451591_Evaluation_of_Inter-Rater_Agreement_and_Inter-Rater_Reliability_for_Observational_Data_An_Overview_of_Concepts_and_Methods Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data:
18. Tai RH, Bentley LR, Xia X, Sitt JM, Fankhauser SC, Chicas-Mosier AM, Monteith BG (2024) An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods*, 23.
<https://doi.org/10.1177/16094069241231168;PAGE:STRING:ARTICLE/CHAPTER>
19. Yao Y, Du X, Diao Y, Zhu H (2019) An integration of deep learning with feature embedding for protein–protein interaction prediction. PeerJ 2019(6):e7126.
<https://doi.org/10.7717/PEERJ.7126/SUPP-1>
20. Zhang F, Wu L, Zhu D, Liu Y (2019) Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS J Photogrammetry Remote Sens 153:48–58.
<https://doi.org/10.1016/J.ISPRSJPRS.2019.04.017>
21. Zhu D, Zhang F, Wang S, Wang Y, Cheng X, Huang Z, Liu Y (2020) Understanding Place Characteristics in Geographic Contexts through Graph Convolutional Neural Networks. Annals Am Association Geographers 110(2):408–420. <https://doi.org/10.1080/24694452.2019.1694403>

Figures

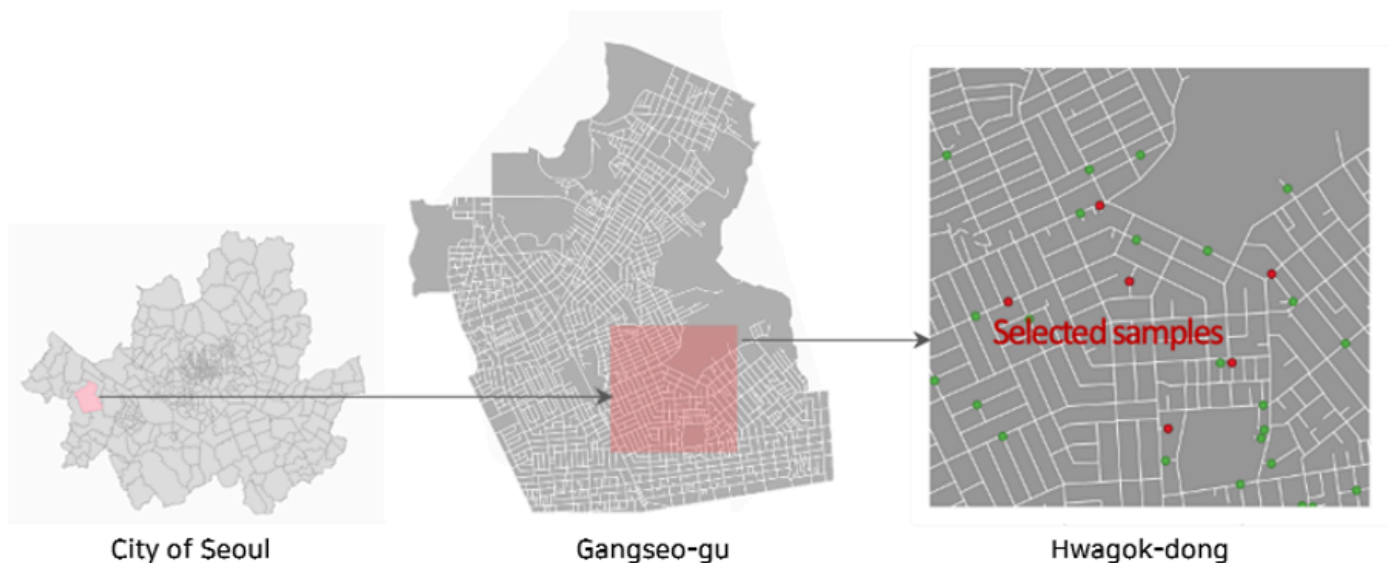


Figure 1

Study Area

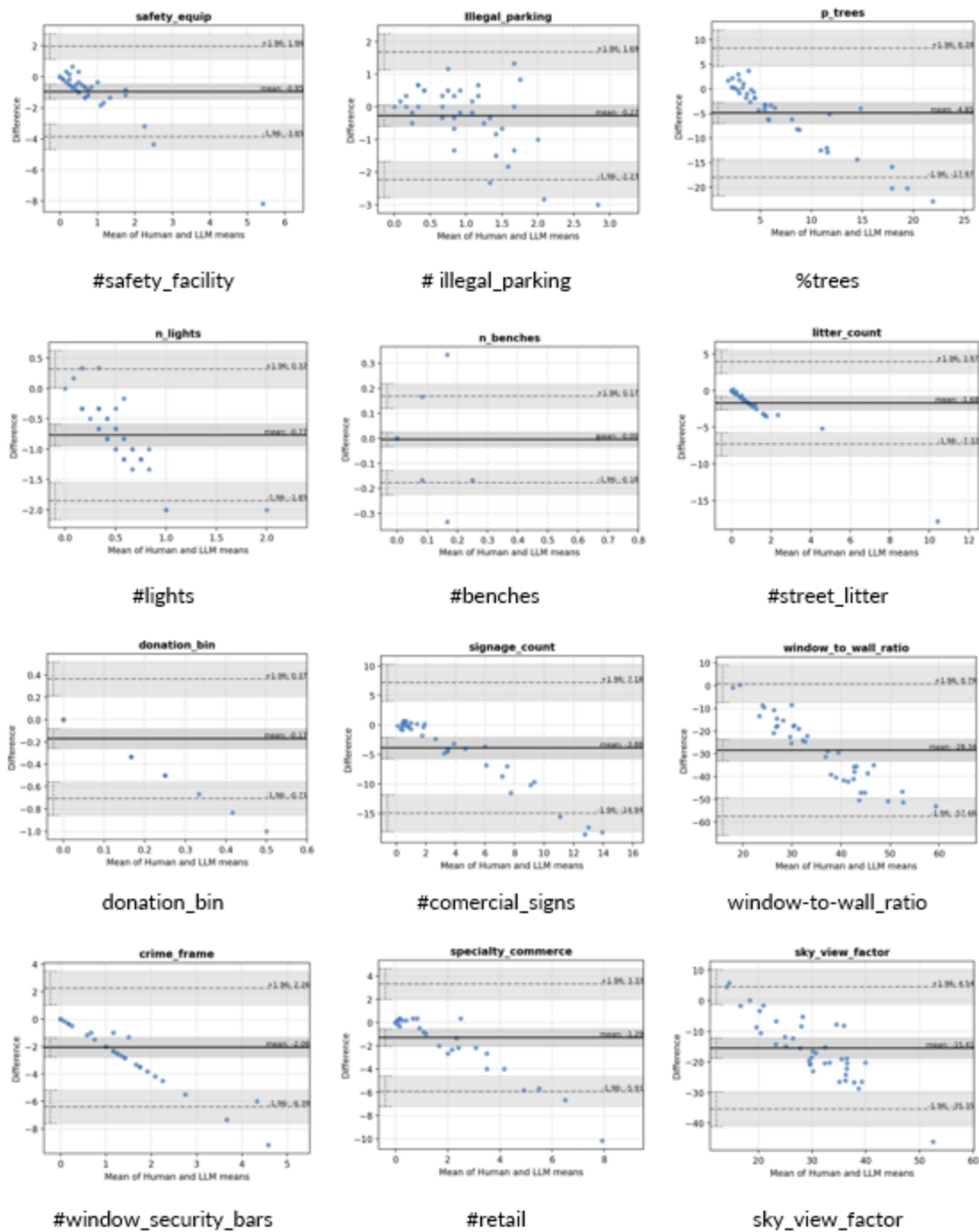


Figure 2

Bland-Altman Analysis Identifying Agreement Between Human and LLM Raters