urban science

MDPI

Article

Evaluating Human Intuition, Multimodal AI, and an XGBoost Model for Estimating Ambient Population Density: A Questionnaire-Based Benchmark of Accuracy, Reasoning, and Confidence

Pasit Rojradtanasiri * and Junko Tamura

Architecture and Urbanism Program (I-AUD), Graduate School of Science and Technology, Meiji University, Tokyo 101-8301, Japan; junko@meiji.ac.jp
* Correspondence: tonch.ss77@gmail.com

Abstract

The physical environment of a neighborhood, such as its road networks and land-use distribution, appears to be closely related to its ambient population density, but this relationship is difficult to measure quantitatively. A previous study based on Japanese neighborhoods addressed this challenge by developing an XGBoost baseline model to estimate Average Hourly Ambient Population Density (AHAPD) at a neighborhood scale using 16 physical environment-related features, achieving 75.9% accuracy. This study evaluates that baseline model by comparing it with human evaluators and multimodal AI models in a spatial reasoning task based on AHAPD estimation. A questionnaire-based benchmark with 29 sets of questions was developed to compare the three evaluator types across accuracy, reasoning, and confidence. The online questionnaire was distributed from 8 June–22 September 2025. In total, 100 responses were analyzed, comprising 94 human evaluators, 5 free-tier versions of multimodal AI models, and 1 output from the baseline model. The baseline model scored 16 points, compared with a mean of 15.1 points for the human evaluators and 14.2 points for the tested AI models. Its performance was also comparable to human evaluators with domain expertise and contextual familiarity, who averaged 16.0 points. The comparison further revealed differences in feature importance, confidence patterns, and characteristic errors among the evaluator types. Selected discrepancy cases highlighted possible strengths and weaknesses of each approach. The main contribution of this study is a questionnaire-based diagnostic benchmark for comparing evaluator types with different characteristics in spatial reasoning tasks.

Keywords: neighborhood planning; machine learning; comparative study; ambient population density; GIS; AI

Check for updates

Academic Editors: Lirong Yin, Shan Liu and Kenan Li

Received: 14 July 2026
Revised: 28 August 2026
Accepted: 9 September 2026
Published: 11 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The physical environment of a neighborhood, such as its road network, land-use distribution, and building footprint, appears to be closely related to how populations are distributed and move within urban areas. This relationship can be reflected through ambient population density [1], which, unlike residential population density, also accounts for population movement and the presence of people throughout the day. Many studies have described this relationship in different ways and linked various physical environment characteristics to population distribution and urban activities. For example, LandScan [1]

uses features such as landcover, commercial centers, and transportation infrastructure as key predictors of ambient population distribution. Cevero & Kockelman [2] linked density, diversity, and design to travel demand, identifying features such as retail intensity and walking accessibility as important factors. Jacobs [3,4] highlighted that density, short blocks, history, and commercial activity contributed to vibrant urban life, which can also be related to higher levels of ambient population activity. Krugman [5,6] emphasized cluster economies, where the concentration of particular land uses can reduce transportation costs and attract similar or complementary activities to occur in the same area. Together, these studies highlight how different physical characteristics of an urban environment, although described using different concepts and terminologies, can contribute to the ambient population density of a neighborhood.

However, quantitatively measuring this relationship is challenging, as the physical environment consists of many interrelated features, and an accurate representation of ambient population density is difficult to obtain. In a previous study by Rojradtanasiri et al. [7], an alternative method for estimating ambient population density at the neighborhood scale was explored. The study captured physical environment features within an 800 m radius of train stations across Japan using Geographic Information System (GIS) [8] and the OpenStreetMap (OSM) database [9]. Basic statistical data for the municipalities in which these neighborhoods are located were also collected as supplementary information. These inputs were then paired with average hourly Mobile Spatial Statistics (MSS) [10] data available in Japan [11], which were used as a proxy of ambient population density. Together, these data were used to create individual neighborhood profiles. In total, 900 neighborhood profiles from across Japan were compiled to form the dataset used for training Machine Learning (ML) models [12]. The objective was to classify neighborhoods into categories based on their Average Hourly Ambient Population Density (AHAPD). Although the study was conducted using only Japanese neighborhood data, the methodology relied on free and widely available tools, allowing a similar approach to be replicated in other countries.

The result was a trained XGBoost model [13], referred to in this study as the baseline model. The model used 16 input features. Of these, 13 were physical environment features collected directly from OSM, a globally available open-source dataset. The remaining 3 were basic statistical features collected from e-Stat [14], Japan's official government statistics portal. The model successfully classified neighborhoods into three AHAPD classes with an accuracy of 75% (F1 score [15]: 75.9%). Feature importance was analyzed using SHapley Additive exPlanations (SHAP) [16] and Partial Dependence Plots (PDPs) [17]. Although the dataset was relatively small compared with those used to train state-of-the-art AI models [18,19], the task-specific model demonstrated promising capability. The results also provided measurable evidence of the relationship between the physical environment and AHAPD, supporting the feasibility of the proposed approach as a proof of concept.

Findings from the study include the top three most influential input features for neighborhoods with high AHAPD were LandUseDiversity, BasedDensity, and RoadCount [7]. The study also identified threshold values at which the relationship between specific features and ambient population density changed from positive to negative, pointing to new areas for investigation. The overall, by-class and predictive-instance feature importance scores provide insights into how models make each decision, or the reasoning. Additionally, the model's confidence scores for each predictive instance reveal insights about the model's confidence level and correctness of the prediction, which are also valuable. These metrics (accuracy, reasoning, and confidence) form the foundation for evaluating model performance.

While an F1 score of 75.9% appears relatively high, it is difficult to assess whether this level of performance is satisfactory without comparison with other evaluator types.

Two-way comparisons between ML models and humans are commonly used to evaluate the real-world capability of ML models in specific tasks. In recent years, the capabilities of multimodal AI models [20], hereafter referred to as AI models, have improved significantly, and these models can perform some tasks at a level comparable to humans. Therefore, a three-way comparison among the task-specific baseline ML model, humans, and AI models can provide a more comprehensive evaluation.

Each evaluator type approaches the task of spatial reasoning differently. The baseline model relies on tabular numerical information. Humans rely on visual interpretation, intuition, and personal experience. AI models combine visual and textual information with knowledge acquired from large-scale training data. These differences may introduce different forms of bias. Humans and AI models may be influenced more strongly by visual features, while the baseline model may be influenced by patterns or biases present in the numerical training data. Comparing these different modes of evaluation can therefore help identify not only similarities in performance, but also differences in reasoning, confidence, strengths, and weaknesses. This allows for a more diagnostic understanding of how each evaluator type approaches spatial reasoning.

Multiple studies across various domains have compared the performance of trained ML models with humans or other models on the same tasks. The tasks that typically get compared generally fall into two categories: (1) tasks that can be evaluated using existing standardized tests as a benchmark and (2) specialized tasks without a standardized test. Examples of the standardized benchmarks include AIME [21], which contains competition mathematics problems, and Humanity's Last Exam (HLE) [22], which contains more than 2500 challenging questions from over a hundred subjects. Other examples include Physics Olympiad problems [23], the IEEExtreme coding challenge [24], and the Bar Exam [25]. Together, these benchmarks provide common, comparable metrics for comparing model capability with recorded human performance. Because AI models are versatile, they can often be adapted to existing benchmarks. Therefore, using standardized benchmarks is generally preferable.

For specialized tasks without a standardized test, customized evaluation methods are often required. For example, Salinas et al. [26] reviewed 53 studies comparing AI systems with clinicians in skin cancer classification. The included studies used different benchmark datasets and experimental settings. Shehu et al. [27] compared human judgment with ML classifiers for recognizing emotion from facial expressions under different face covering conditions. Kandul et al. [28] compared the predictability of human and AI behavior in computerized navigation tasks. Yan et al. [29] compared GPT-4 with human translators across languages, domains, and expertise levels. These studies show that researchers may need to construct task-specific datasets, procedures, or evaluation frameworks when no established benchmark fits the task.

In our case, the baseline ML model was trained for a specific task of estimating AHAPD, which has no standardized test (category 2); therefore, a novel customized benchmark is needed. The key to designing a new benchmark is fairness [30], ensuring that none of the evaluator types has an unfair advantage. For instance, the same set of information, such as the 16 features that the baseline model uses as its main basis for prediction, is also given to human evaluators and AI models. However, in this study it is very challenging to replicate the test at exactly 1:1, because each evaluator type has different strengths and weaknesses. For example, the baseline model has been trained on the AHAPD data and can interpret numbers well, but cannot process images. Humans have visual and experience advantages but may have limited capability in finding patterns within raw numerical values. Multimodal AI models, on the other hand, have the advantage of being trained on large scale dataset, can interpret text, numbers, and images very well, but face rate limits [31],

have limited context windows [32] and memory [33], and sometimes hallucinate [34]. These strengths and weaknesses must be considered when designing the benchmark.

Another important factor to consider is balancing the complexity and accessibility of the benchmark. A complex benchmark that requires time, effort, and specialized expertise may limit participation to a small group of specialists. For example, skin cancer detection generally requires evaluation by clinicians with relevant expertise. In contrast, a simpler, shorter benchmark is more suitable for the general public and can involve a larger number of participants. A more general task like emotion recognition can be evaluated by a larger pool of participants (e.g., 82 participants in Shehu et al. [27]). Estimating AHAPD lies closer to a general task than a complex one.

Although there is no official training for this specific task, there are human specialists conducting studies on similar spatial reasoning tasks. For example, Sarmadi et al. [35] compared the performance of convolutional neural networks (CNNs) [36] with 102 human experts in estimating poverty from aerial images. While the CNN outperformed the experts, the score and features identified by the experts were used for training the model. Ahn et al. [37] used a human–machine collaborative approach to estimate regional economic development from satellite imagery, in which human annotators were asked to label clusters of images for training supervised deep learning models. Panczak et al. [38] conducted a literature review of temporal population estimation methods and identified 96 relevant studies. These studies suggest that this is an active area of research and that human judgment is often relied upon for spatial reasoning. Similarly, professionals in architecture, urban design, and real estate development may have an intuitive edge in this type of visual-spatial recognition, as they are more familiar with maps, aerial images, and urban elements. Therefore, to identify where the baseline model stands relative to other groups, it is meaningful to compare people with and without such domain expertise, highlighting the need for the benchmark to be accessible to a diverse group.

Another advantage of applying AI models like ChatGPT [39] to existing tests is that they can be used to help evaluate test difficulty. For example, Suresh and Rawat [40], used ChatGPT to assess the difficulty of SAT exams [41] and re-evaluate past human performance. Similarly, the uncalibrated predicted class probabilities of the pre-trained baseline model can be used as a proxy for model confidence. These confidence scores can then be used as an operational indicator of question difficulty within the benchmark, helping to compare the strengths and weaknesses of each evaluator type.

In conclusion, this study addresses a research gap in which the previously trained baseline model had not been compared with other evaluator types performing the same spatial reasoning task. To address this gap, a novel questionnaire-based benchmark was developed to compare the baseline model with human evaluators, and multimodal AI models. The novelty of this research lies in using the benchmark as a diagnostic tool rather than focusing only on predictive accuracy. By incorporating reasoning and confidence, the framework enables deeper investigation of discrepancy cases in which different evaluator types approached the same task differently. This allows their distinct behavioral patterns, strengths, and weaknesses to be identified and analyzed.

The study had two main objectives: (1) to evaluate the performance of the previously trained baseline model by comparing it with human evaluators and multimodal AI models using a questionnaire-based benchmark and (2) to identify similarities and differences in their performance, reasoning, confidence, strengths, and weaknesses in estimating AHAPD. Through this comparison, the study aims to provide a more diagnostic understanding of the baseline model and identify areas for its future improvement.

2. Methodology

To address this challenge, we developed a questionnaire-based benchmark using Google Forms [42]. The questionnaire was designed to evaluate three aspects: accuracy, reasoning, and confidence. The benchmark was intended primarily as a diagnostic tool to better understand the behavior of the baseline model rather than as a strictly controlled comparison of accuracy. Therefore, the questionnaire cases were selected based on known patterns in the baseline model's predictions on the testing dataset, including differences in correctness and confidence.

Simple questionnaire tools were used so that the same evaluation framework could be applied across evaluator types with different capabilities. The baseline model relied only on the numerical feature values, while human evaluators and multimodal AI models received both aerial images and numerical feature information to better suit their respective modes of reasoning.

A three-way comparison among the baseline model, human evaluators, and multimodal AI models is challenging because each evaluator type has different strengths and weaknesses. Comparing classification accuracy is relatively straightforward because all evaluator types produce the same class output through multiple-choice questions. However, comparing reasoning and confidence is more difficult because these outputs are expressed in different forms. For the baseline model, feature importance was used as a proxy for reasoning, while uncalibrated predicted class probabilities were used as a proxy for confidence. For humans evaluators and AI models', reasoning was evaluated using feature-selection questions, while confidence was measured using self-rating questions. Because these outputs differ in form and scale, they were converted into comparable, although not identical, metrics before analysis. The implications of these different input conditions and measurement approaches are discussed further in the Section 4.4.

Early versions of the questionnaire were tested with peers and colleagues from the same university cohort to collect feedback. Multiple iterations were developed and tested to balance the amount of insight collected with the simplicity and accessibility of the questionnaire before external distribution. The final questionnaire consisted of four main sections: Basic Information, Training, Main Questionnaire, and Post Questionnaire. The Training and Main Questionnaire sections together contained 29 sets of questions selected based on known baseline model behavior. The response period continued until 100 responses were received. All responses were then analyzed and discussed before drawing conclusions. An overview of this process is shown in Figure 1, with further details provided in the following sections.

In this paper, generative AI (ChatGPT-5) was used as a brainstorming assistant to improve the clarity of ideas after the authors had outlined the research scope and drafted the manuscript. After the initial results were received, AI was also used as a coding assistant for data analysis. Five generative AIs (GPT-4o, GPT-5, Gemini 2.5 Flash, Claude Sonnet 4, and Grok-3) were included as an evaluator type for performance comparison between humans, AI and ML, which is the main focus of the study. Finally, ChatGPT-5 was used as a language improvement assistant at the end of the study, under the authors' close supervision at every step.

2.1. Questionnaire-Based Diagnostic Benchmark Design

To evaluate accuracy, confidence, and reasoning, we used common questionnaire tools. Classification accuracy was assessed with multiple-choice questions. Reasoning was evaluated by asking evaluators to select 5 of 16 given features they believed most influenced their decision. Confidence was measured by having evaluators rate their certainty on a scale of 1 to 10.

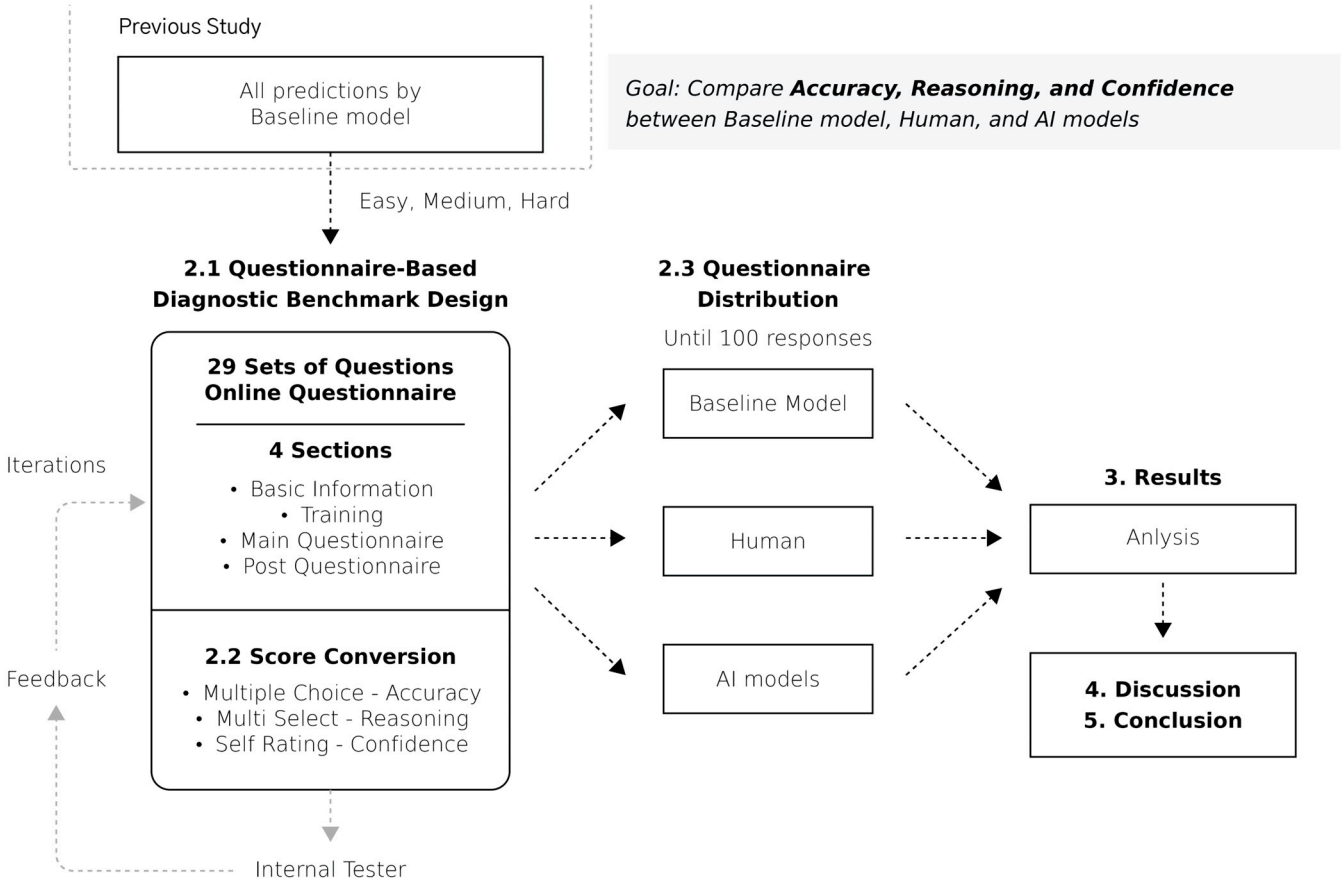

Figure 1. Methodology Overview.

The first version of the questionnaire included 20 sets of questions, each with three in-depth questions. Early feedback indicated that the questionnaire was too long, overly technical, and difficult to complete, even among those with domain expertise. Evaluators reported that reasoning-type questions were particularly time-consuming and difficult, especially in the pre-questionnaire section where no examples were provided before answering. Some early testers without domain expertise reported struggling with domain-specific terms such as "neighborhood", "features" and "traffic". Several also noted that without a training section it was difficult to complete the questionnaire because there was no feedback or chance to recalibrate their responses. Additionally, evaluators were curious to know their scores after completing, which were not shown in the first version.

In response, we simplified the language, reduced the number of reasoning questions, made some questions optional, and introduced a training section where evaluators could check the accuracy of their first three answers and recalibrate their decisions. A final score display was also added to satisfy evaluator curiosity about their performance. The final version of the questionnaire consisted of four sections: Basic Information, Training, Main Questionnaire, and Post-Questionnaire. Because the number of reasoning questions was reduced, we were able to include more multiple-choice and rating questions without increasing the overall time required to complete the questionnaire. The final version of the questionnaire is described in detail below.

2.1.1. Basic Information

This section consists of six demographic questions: name (optional), age, nationality, country of residence, occupation, and experience in Japan (in years). A seventh optional question asked evaluators to select 5 of the 16 features that they believed most strongly

influenced the ambient population density of a neighborhood. This question was intended to capture evaluators' overall preconceptions and was planned to be compared with their selections in the post-questionnaire section.

Names were initially excluded but were later added as an optional field to identify which invitees had already submitted a response. These questions were intended to capture evaluators' experience, familiarity with Japan, and level of domain expertise because these factors may influence performance and reasoning.

2.1.2. Training

The training section included 3 sets of questions (S1–S3), each with three questions (Q1–Q3).

- Q1: a multiple-choice classification question aiming to evaluate accuracy, asking each evaluator to select A, B or C (A = Class 0, B = Class 1, C = Class 2), based on an aerial image and a bar chart showing the value of 16 features associated with the neighborhood.
- Q2: a multi-select question to evaluate reasoning, asking each evaluator to choose 5 of 16 features that most influenced their decision.
- Q3: a rating question asking each evaluator to rate their confidence on a scale of 1 to 10. The results are used to identify which questions the evaluators found easier or harder.

For the training questions, we chose one example from each class in which the model made a correct prediction with high confidence, representing easier questions (Table 1). After completing the three sets, the correct answers were shown, allowing evaluators to recalibrate their decisions. Figure 2 shows an example of the three question types used in the distributed questionnaire.

Table 1. List of 29 neighborhoods used in the questionnaire (3 Training + 26 Main Questionnaires) along with the model's confidence scores.

C0 P0 (Correct)	C0 P1 (Over Est.)	C0 P2 (Over Est.)	C1 P1 (Correct)	C1 P0 (Under Est.)	C1 P2 (Over Est.)	C2 P2 (Correct)	C2 P1 (Under Est.)	C2 P0 (Under Est.)
Iwatenuma-kunai, Iwate (0.999) ET	**Kameyama, Mie (0.983) H**	-	**Kokubu, Kagoshima (0.997) ET**	**Hino, Shiga (0.998) H**	**Sakuramachi, Nagasaki (0.996) H**	**Gojo, Kyoto (0.999) ET**	**Fukaya, Saitama, (0.988) H**	**Sasabaru, Fukuoka (0.488) H**
Daishaka, Aomori (0.998) E	**Yodoe, Tottori (0.936) H**	-	**Shimodate, Ibaraki (0.997) E**	**Tsubata, Ishikawa (0.997) H**	**Miebashi, Okinawa (0.989) H**	**Fushimi, Aichi (0.999) E**	**SakuraHom-machi, Gifu (0.983) H**	-
Nahari, Kochi (0.997) E	**Kesennuma, Miyagi (0.821) H**	-	**Shirakawa, Fukushima (0.994) E**	**Yasu, Kochi (0.997) H**	**Kofu, Yamanashi (0.957) H**	**Kamimaezu, Aichi (0.998) E**	**Yokkaichi, Mie (0.953) H**	-
30 more	4 more	-	48 more	5 more	2 more	39 more	9 more	-
Urasa, Niigata (0.601) M	Chuden, Tokushima/ (0.602)	-	**Kiyotake, Miyazaki (0.532) M**	Sembokucho, Iwate (0.831)	Tottori, Tottori (0.697)	**Kashiwa, Chiba (0.544) M**	Akita, Akita, (0.584)	-
Unazuki-Onsen, Toyama (0.566) M	Obama, Fukui (0.555)	-	Kyozuka, Okinawa (0.519)	Nirasaki, Yamanashi (0.768)	Minami Kofu, Yamanashi (0.604)	Miyazaki, Miyazaki (0.536)	Maebashi, Gumma, (0.561)	-
Yuasa, Wakayama (0.397) M	Tamura, Fukushima/ (0.516)	-	**Inarimachi, Toyama (0.511) M**	Hizen Kashima, Saga (0.553)	Saidaiji, Okayama (0.507)	**DaigakuByoin -mae, Nagasaki (0.525) M**	Awa-Tomida, Tokushima (0.522)	-

E = Easy, M = Medium, H = Hard, ET = Easy questions used for training (underlined), Main Questionnaire = Bold, C = Correct Class, P = Predicted Class.

Figure 2. Example of Q1 (Accuracy), Q2 (Reasoning), and Q3 (Confidence).

2.1.3. Main Questionnaire

This section contained 26 sets of questions (S4–S29), following the same structure as the training section but without immediate feedback. The first 23 sets (S4–S26) included only Q1 (Accuracy) and Q3 (Confidence) to reduce time requirements and complexity. The final 3 sets (S27–S29) included Q2 (Reasoning). Because the earlier reasoning questions were optional, these final reasoning questions were prioritized for analysis.

The neighborhoods used in each set were selected from predictions made by the baseline model on the testing dataset. The full list of predictions was extracted from the confusion matrix [43] (Figure A1) and reorganized based on prediction confidence and correctness. The uncalibrated predicted class probabilities of the baseline model were used as a proxy for model confidence and as an indicator of question difficulty. Questions that the baseline model answered correctly with high confidence were classified as Easy. Questions where the model had low confidence were classified as Medium. Questions the model answered incorrectly were classified as Hard. The final 26 questions in this section consisted of 6 Easy, 7 Medium, and 13 Hard questions.

A random selection of questions, or the use of the entire testing dataset, would provide a less biased estimate of overall accuracy. However, due to time constraints, only a limited number of questions could be included without making the questionnaire too long and potentially reducing human participation. A random sample might also include too few low-confidence or high-confidence incorrect cases, which were important for the diagnostic purpose of this study. The selected cases therefore allowed the benchmark to examine not only accuracy, but also differences in confidence and reasoning across different patterns of baseline-model behavior. This design supports a more diagnostic comparison in which the difficulty and response patterns of each evaluator type can be examined separately.

Among the 26 main questionnaire cases, the baseline model answered 13 correctly and 13 incorrectly, resulting in a constructed accuracy of 50% within the main questionnaire. Including the 3 training questions, evaluators needed to score more than 16 points to outperform the baseline model. Table 1 lists the neighborhoods selected for the questionnaire. The locations of these 29 neighborhoods across different regions of Japan are shown in Figure 3. The blue dots represent cities across Japan, while the yellow dots represent neighborhoods included in the training dataset. The red dots represent neighborhoods selected for the questionnaire, and the accompanying numbers indicate their corresponding questionnaire items.

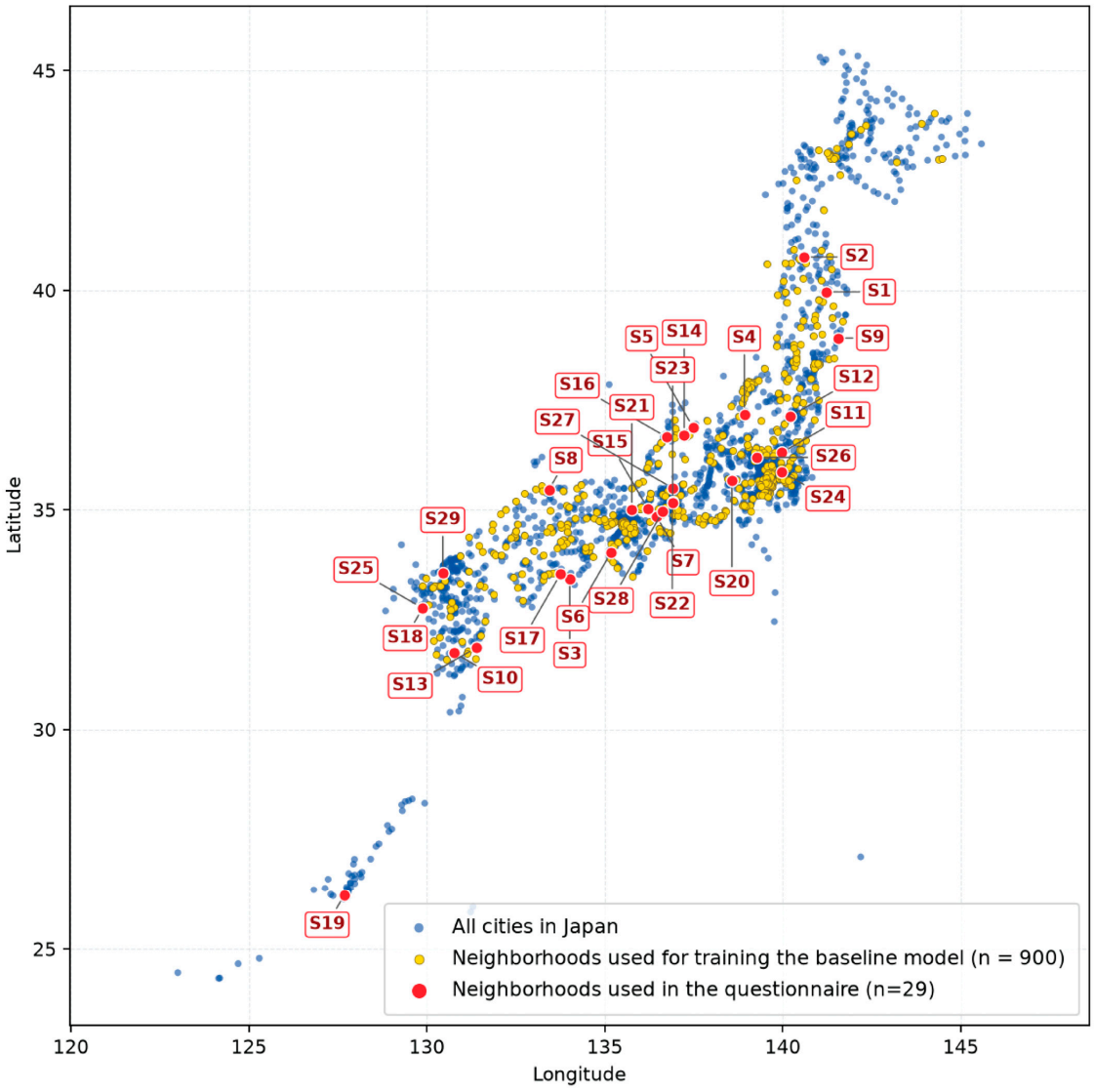

Figure 3. Locations of the neighborhoods selected for the questionnaire across Japan.

2.1.4. Post-Questionnaire

This section consisted of 6 reflective questions, asking evaluators to rate their overall confidence and explain their reasoning when choosing A, B or C in free-form text format. This was intended to capture reasoning that may not be fully captured by the multi-selection Q2 questions. For one last time, evaluators were asked to select the 5 most influential features for their overall reasoning. The final question asked evaluators for any final comments on the questionnaire (optional).

2.2. Score Conversion

Because the outputs from the three evaluator types are expressed in different forms and scales, they need to be converted into comparable metrics before analysis. Accuracy is relatively straightforward because it can be calculated directly from the questionnaire score. Reasoning and confidence, however, require additional processing because they are represented differently by the baseline model and the other evaluator types.

For the baseline model, reasoning was represented by feature importance values, while human and AI reasoning was represented by the five features selected in the questionnaire. Confidence is represented by uncalibrated predicted class probabilities for the baseline model and by rating-scale responses for humans and AI models. In addition, the reasoning and confidence score of the baseline model represent outputs from a single model, whereas the human and AI scores represent collective responses from each evaluator group. Therefore, the human and AI responses were aggregated and their weights were distributed collectively to create comparable group-level metrics.

Table 2 summarizes which questionnaire items were used to evaluate each metric and where they appear in the questionnaire structure. After all responses were collected, the raw answers were processed to calculate the final Accuracy, Reasoning and Confidence metrics. The detailed conversion method for each metric is explained in the following sections.

Table 2. The overall structure and which questionnaire item evaluates which metric.

Metric	Basic Info Section	Training Section	Main Section	Post Section
Accuracy	-	S1Q1–S3Q1 (n = 3)	S4Q1–S29Q1 (n = 26)	-
Reasoning	Pre-Q6 (Opt., n = 1)	S1Q2–S3Q2 (Opt., n = 3)	S27Q2–S29Q2 (n = 3)	Post-Q4 (n = 1)
Confidence	-	S1Q3–S3Q3 (n = 3)	S4Q3–S29Q3 (n = 26)	Post-Q1 (n = 1)

S = set; Q = question; Opt. = optional question; n = total number of items.

2.2.1. Accuracy—Multiple Choice (Q1)

For the accuracy score, the raw data were evaluated using several statistics, such as mean, minimum, maximum, median, and mode, across all evaluators and demographic subgroups. This analysis helped identify behavioral patterns and allowed performance comparison between evaluator types.

2.2.2. Reasoning—Select 5 Features (Q2)

By analyzing the baseline model using SHAP, feature importance scores can be extracted for each context (t), either overall or class-specific (Classes 0, 1 or 2). These scores represent the relative contribution of each feature to the model's prediction. The feature importance values are converted into percentages, with the combined contribution of all features equal to 100%. This shows how much weight each feature contributes to the prediction and is used to represent the baseline model's reasoning score for each feature.

To capture human- and AI-perceived feature importance collectively, the evaluators were asked to select 5 of the 16 features that most influenced their decision in up to eight instances: four at the beginning of the questionnaire (optional) and four at the end. Each set of four selections corresponded to one overall feature importance context (FI_{oa}) and three class-specific contexts (FI_{c0} for Class 0, FI_{c1} for Class 1, and FI_{c2} for Class 2). Because the initial selections were optional and may be incomplete, only the four end-of-questionnaire selections were used in the analysis.

For each context t in $\{oa, c0, c1, c2\}$, each evaluator distributed a total weight of 1 equally among the five selected features, giving each selected feature a weight of 0.2 (1/5). Let N be the number of evaluators, and let $x_{e,t,p}$ indicate whether the evaluator e selected features p in context t where $x_{e,t,p} = 1$ if selected and 0 otherwise. The Derived Feature Importance (DFI) for features p in context t is calculated as

$$DFI_t(p) = \frac{1}{5N} \sum_{e=1}^{N} x_{e,t,p} \quad (1)$$

where

- N is the total number of evaluators.
- e represents an individual evaluator.
- t represents the evaluation context (overall, Class 0, Class 1, or Class 2).
- p represents one of the 16 input features.
- $x_{e,t,p}$ equals 1 if evaluator e selected feature p in context t, and 0 otherwise.

This DFI score represents the average feature importance assigned to each feature across all evaluators within a given context. This metric enables direct comparison between human-, AI-, and baseline model-perceived feature importance, allowing similarities and differences in reasoning to be identified across all evaluator types and contexts.

2.2.3. Confidence—Rating (Q3)

The uncalibrated predicted class probabilities of the baseline model were used as its confidence scores. These scores ranged from 0 to 1, with higher values representing greater confidence in the predicted class. The confidence scores were used to classify question difficulty and to analyze the behavior of the baseline model. For human evaluators and AI models, confidence level was collected using self-rating questions. Because raw confidence scores may vary according to response style, (e.g., conservative versus bold evaluators) each evaluator's scores were normalized using that evaluator's observed range.

For each evaluator e, the minimum (m_e) and maximum (M_e) confidence scores across all questions were identified. For a given question (Q), the normalized confidence score is calculated as

$$\frac{C_{Q,e} - m_e}{M_e - m_e} \quad (2)$$

where $C_{Q,e}$ is the raw confidence score provided by evaluator e for question Q. This normalization scales each evaluator's confidence scores to a range between 0 and 1, with their minimum confidence mapped to 0 and maximum confidence mapped to 1.

The Average Confidence for question Q (ACQ) is then calculated as

$$ACQ = \frac{1}{N} \sum_{e=1}^{N} \frac{C_{Q,e} - m_e}{M_e - m_e} \quad (3)$$

where

- N is the total number of evaluators.
- $C_{Q,e}$ is the raw confidence score provided by evaluator e for question Q.

- m_e and M_e are the minimum and maximum confidence scores reported by evaluator e respectively.

This ACQ metric represents the average normalized confidence of all evaluators for a given question. Higher ACQ values indicate that evaluators collectively found a question easier, whereas lower values indicate that it was perceived as more difficult. Human-derived ACQ values were compared with the baseline model's confidence scores to identify similarities and discrepancies, including instances where predictions were confidently correct or confidently incorrect. For the AI models, ACQ was calculated using the same procedure as for the human evaluators.

2.3. Questionnaire Distribution

The questionnaire was distributed in two phases. The internal phase involved early testers and was used to collect feedback and refine the questionnaire. The external, invitation-only phase targeted a diverse group of peers, colleagues, and family members. For evaluators who are unfamiliar with online forms, one-on-one support was provided via online meetings. The author presented the questions, translated where necessary, and submitted answers on their behalf with minimal interference to preserve objectivity.

The questionnaire was submitted to free-tier versions of five commercially available multimodal AI models: GPT-4o [44], GPT-5 [45], Gemini 2.5 Flash [46], Claude Sonnet 4 [47], and Grok-3 [48], on 21 August 2025. These free-tier versions were subjected to lower rate limits, restricted context windows, and possible performance degradation during the long session. The result should therefore be interpreted as performance under these specific testing conditions and not as the upper-bound capability of each model.

Screen-captured images of each questionnaire page were saved in .jpeg format and submitted to each AI model one image at a time, following the same sequence and within a single chat session. When a model reached its rate limit, evaluation was paused until the cooldown period ended and then continued in the same conversation. Each AI model completed the questionnaire only once, consistent with the single-attempt condition given to each human evaluator. However, because AI-generated responses can be stochastic, the result from a single session may vary across repeated runs and should not be interpreted as the best possible performance of each model. To maintain consistency, each conversation began with the same prompt:

> *"Today, I would like you to participate in my research by answering a set of questionnaires. I will provide you with screenshots of the questions, and you should respond with how you personally would answer them. I will then input your answers into the questionnaire on your behalf. Please do not look up the answers on the internet. Simply answer based on your own knowledge, experience, or opinion."*

Their responses were then manually entered by the author into the same Google Form that was distributed to other evaluators. The AI model results were recorded and analyzed alongside human results.

3. Results

In total, it took 107 days, from 8 June to 22 September 2025, to receive all 100 responses. Responses from one minor tester and early testers were excluded from the analysis. The final response sets used for analysis are from 94 human evaluators, 5 AI models, and 1 recorded output from the baseline model. The ages of the human evaluators ranged from 18 to 78 years, with a mean age of 35 years. The top nationalities by percentage were Thai (41.5%), Japanese (26.6%), Chinese (9.6%), and others (22.3%), which included American, Austrian, Australian, Bhutanese, French, Irish, Italian, Indian, Macedonian, Singaporean,

and more. Among the human evaluators, 59.5% had experience living in Japan (34.0% lived in Japan; 25.5% were native Japanese) and 40.5% had no experience living in Japan (9.6% had never been to Japan; 30.9% had visited only as tourists).

There were 41 unique professions among human evaluators. University students formed the largest occupational group (29.8%), followed by architects (9.6%). The remaining 60.6% were a combination of other professions such as banker, investor, animator, engineer, designer, marketer, and more. Occupations such as architects, engineers, real estate agents, and landscape designers were classified as having domain expertise because their work may involve maps, urban planning, or related technical concepts. In contrast, roles such as sales, business managers, and bankers were classified as without domain expertise. Overall, 47.9% of human evaluators were from professions with domain expertise and 52.1% were from professions without domain expertise. AI models and the baseline model are reported as separate categories and are excluded from the human domain expertise counts. Table 3 summarizes the demographic of the evaluators and their average, minimum, and maximum scores.

Table 3. The overall number (n) of evaluators in each demographic group, their average score (s), and their score range (r).

Total Human Evaluators (n = 94, s = 15.1)/r = 9–22		
Experience in Japan	Without Domain Expertise (n = 49, s = 14.5, r = 9–20)	With Domain Expertise (n = 45, s = 15.8, r = 11–22)
Native Japanese (n = 24, s = 15.5, r = 11–21)	n = 9, s = 15.2, r = 12–18	n = 15, s = 15.8, r = 11–21
Experienced Living in Japan (n = 32, s = 15.7, r = 11–22)	n = 13, s = 15.3, r = 11–20	n = 19, s = 16.0, r = 11–22
Visited Japan as Tourist (n = 29, s = 14.1, r = 9–20)	n = 25, s = 13.8, r = 9–18	n = 4, s = 16.2, r = 11–20
Never been to Japan Before (n = 9, s = 15.0, r = 11–20)	n = 2, s = 15.5, r = 11–20	n = 7, s = 14.9, r = 12–17
Baseline model (s = 16)/Average AI models (s = 14.2, r = 10–18)		
GPT-4o (s = 18), GPT-5 (s = 17), Grok 3 (s = 10), Gemini 2.5 Flash (s = 12), Claude Sonnet 4 (s = 14)		

3.1. Accuracy Score

The average score across all evaluators was 15.1 (training questions included), with a range of 9 to 22 points (Table 3). 33% of human evaluators outperformed the baseline model (s = 16), and 67% scored less than or equal to the baseline (Figure 4). Domain expertise appears to be an indicator associated with the performance (Figure 5). Within the same subgroup, those with domain expertise outperformed those without by an average of 1.3 points. Among those who had previously visited Japan as tourists, evaluators with domain expertise outperformed those without domain expertise by 2.4 points, representing the largest difference among the subgroups. Experience in Japan showed mixed results among those with domain expertise. The best-performing subgroups were those who had visited Japan as a tourist with domain expertise; however, this may be due to the small sample size in this subgroup (n = 4). Among those without domain expertise, the highest average score was from evaluators who had never been to Japan, although this group was also very small (n = 2). The AI models scored between 10 and 18 points, with an average of 14.2 points, similar to the average of evaluators without domain expertise.

The highest score among all human evaluators was 22 points, achieved by an individual with both domain expertise and contextual familiarity. In contrast, the lowest score was 9 points, achieved by an individual with limited contextual familiarity and no domain expertise. Although the highest-performing evaluator in each subgroup exceeded the baseline model, the baseline model outperformed the lowest-performing evaluator in every subgroup. Overall, the baseline model performed 5.96% higher than the average human evaluator, 10.3% higher than the average of human evaluators without domain expertise, and 1.27% higher than the average of human evaluators with domain expertise.

The best-performing AI model scored 18, two points more than the baseline model. The baseline scored 1.8 points more than the AI-model average of 14.2 points.

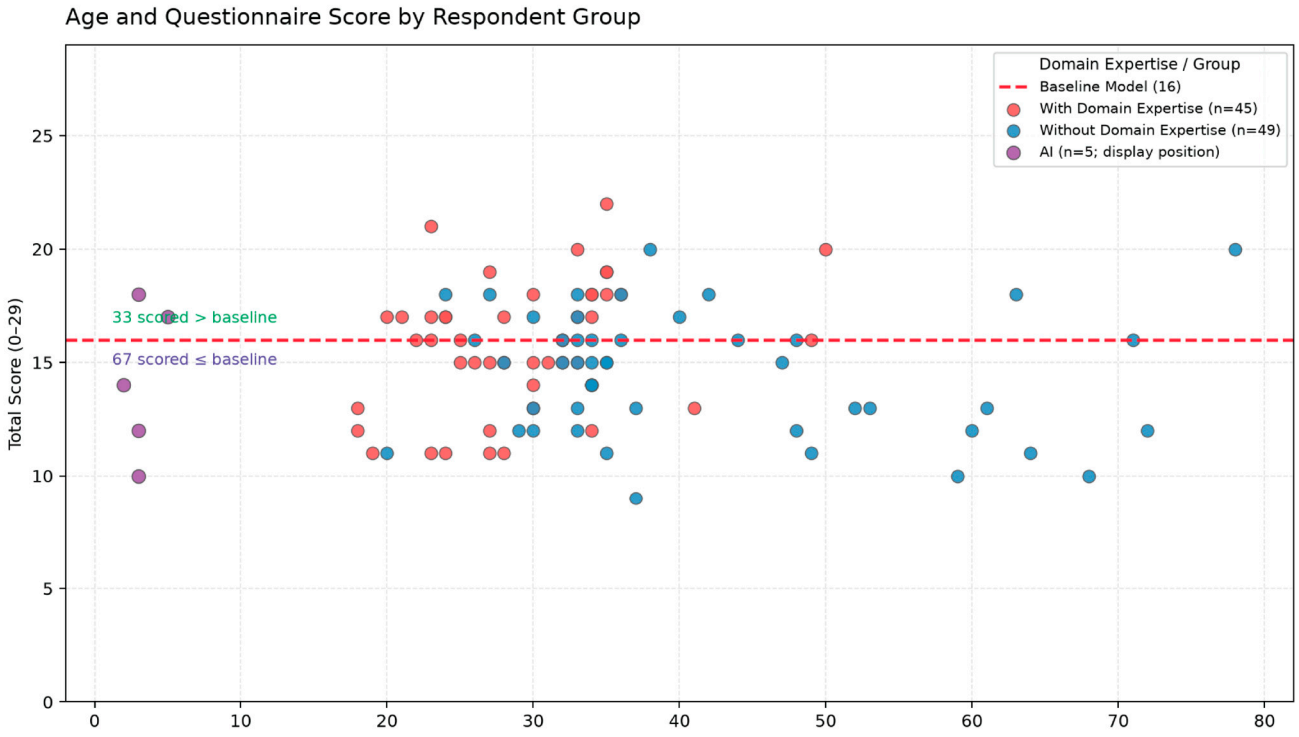

Figure 4. The score of all the evaluators.

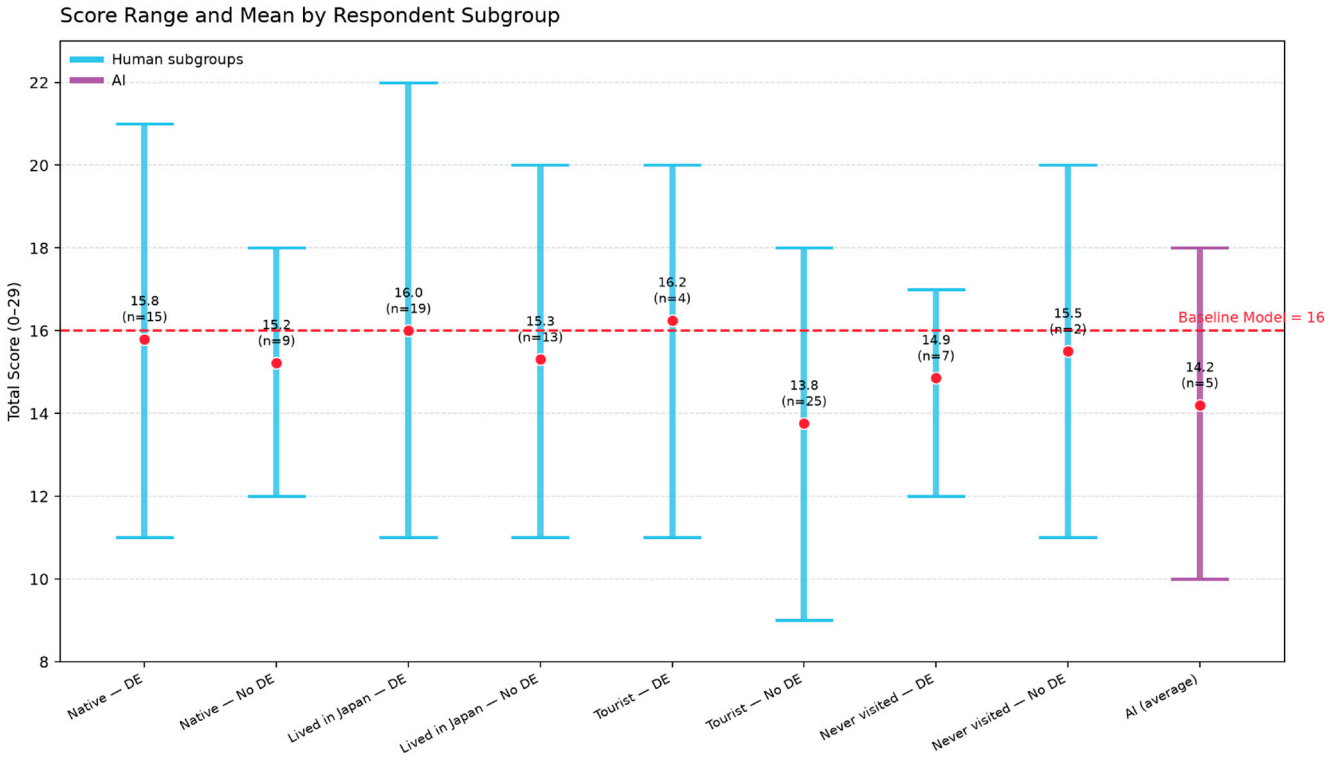

Figure 5. The score of all evaluators along with the minimum and maximum score for each subgroup.

3.2. Reasoning Score

We used feature importance scores derived from SHAP as the baseline representation of reasoning for each evaluator type. By analyzing these scores, we can measure the strength

of impact of each feature. Figure 6 shows the feature importance scores for each evaluator type in the overall context, while corresponding scores for Classes 0, 1, and 2 are shown in Figure A2. The features are arranged in descending order based on human-perceived impact. The green dotted lines highlight the score difference for features that the baseline model considers more influential than human evaluators do. The red dotted lines highlight features that human evaluators considered more influential than the baseline model did. By comparing differences in these scores across evaluator types, we can identify similarities and differences in their way of reasoning. Table 4 highlights the five features with the largest differences between humans and the baseline model and also shows the differences between humans and AI models for those features.

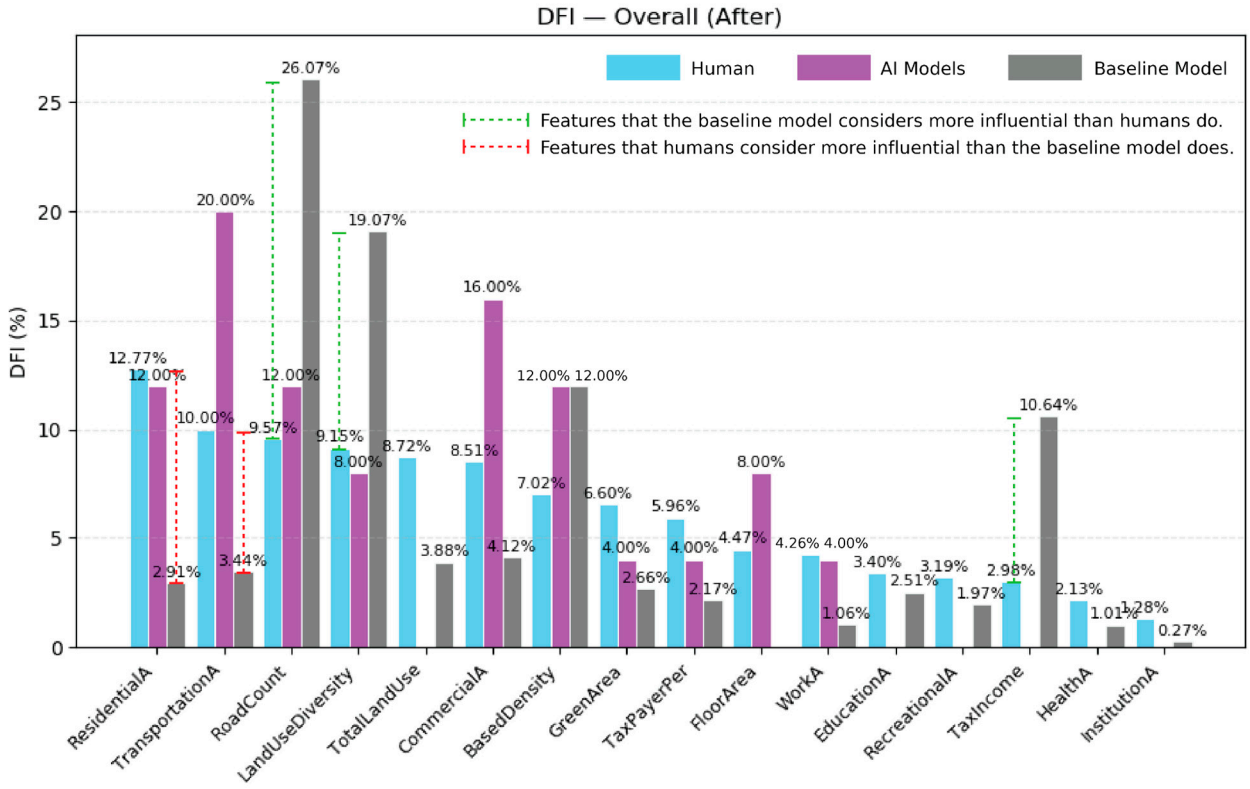

Figure 6. The feature importance score for humans (blue), AI models (purple), and the baseline model (grey).

From Table 4, the features that most frequently appear in the top five across evaluator types and contexts are RoadCount, LandUseDiversity, ResidentialA, CommercialA, and TransportationA. TaxIncome is the most divisive feature, appearing in the baseline model's top five but not in the others. Humans and the baseline model agree on the importance of LandUseDiversity and TotalLandUse, whereas AI models do not. Humans and AI models agree that ResidentialA, TransportationA, and CommercialA are most influential, but the baseline model does not. The most frequent top five appearances are ResidentialA for humans, TransportationA for AI models, and RoadCount for the baseline model. The detailed explanation of each input feature, how they were calculated, and their SHAP values are shown in Table A1.

Another behavioral pattern that can be identified is the allocation of importance. Human selections are spread more evenly across features, forming a gradual distribution. Selections by the AI models were more concentrated on a small number of features, although this pattern may partly reflect the small number of AI models tested. In contrast, the baseline model places more weight on only a few features across all contexts. In addition,

the weight ranking varies less across contexts for humans and AI models, but changes significantly for the baseline model.

Table 4. The top 5 features with the highest difference for each context between humans and the baseline ML model.

Context	#1	#2	#3	#4	#5
Overall	RoadCount Diff 16.49 H (9.57) ML (26.07) AIs (2.42)	LandUseDiversity Diff 9.91 H (9.14) ML (19.05) AIs (8.00)	ResidentialA Diff −9.85 H (12.76) ML (2.91) AIs (12.00)	TaxIncome Diff 7.66 H (2.97) ML (10.64) AIs (0)	TransportationA Diff −6.56 H (10.0) ML (3.44) AIs (20.00)
Class 0	RoadCount Diff 41.90 H (8.93) ML (50.83) AIs (12.00)	TaxIncome Diff 11.66 H (2.34) ML (14.01) AIs (0)	ResidentialA Diff −10.20 H (12.76) ML (2.55) AIs (16.00)	TransportationA Diff −6.29 H (9.57) ML (3.27) AIs (20.00)	LandUseDiversity Diff −6.12 H (10.00) ML (3.87) AIs (4.00)
Class 1	TaxIncome Diff 12.56 H (1.06) ML (13.62) AIs (0)	ResidentialA Diff −9.82 H (14.25) ML (4.43) AIs (16.00)	RoadCount Diff 9.75 H (10.21) ML (19.93) AIs (16.00)	BasedDensity Diff 6.08 H (4.68) ML (10.76) AIs (8.00)	TaxPayerPer Diff −5.79 H (8.08) ML (2.29) AIs (4.0)
Class 2	LandUseDiversity Diff 35.60 H (11.06) ML (46.66) AIs (12.00)	BasedDensity Diff 13.12 H (5.10) ML (18.22) AIs (12.00)	ResidentialA Diff −11.35 H (13.19) ML (1.84) AIs (12.00)	TaxPayerPer Diff −5.71 H (7.65) ML (1.94) AIs (4.00)	RoadCount Diff −5.41 H (7.23) ML (1.82) AIs (20.00)

Table 4 lists the five features with the largest weight differences between humans and the baseline model for each context. Overall, both RoadCount and LandUseDiversity appear in the top five for humans and the baseline model. However, their assigned importance differs significantly: +16.49 percentage points for RoadCount and +9.91 for LandUseDiversity, indicating that the baseline model considers these features more significant than humans do. In contrast, ResidentialA and TransportationA show differences of −9.85 and −6.56 percentage points, indicating that humans view these features as more influential than the baseline model does. These gaps reflect different ways of reasoning between humans and the baseline. For the same set of features, the differences between AI models and humans are smaller than those between humans and the baseline model, indicating closer alignment in reasoning between humans and AI models.

3.3. Confidence Score

Figure 7 shows the evaluator's confidence scores for every question in the questionnaire, along with the correctness of their responses. Answers enclosed in green boxes indicate correct responses. The blue horizontal dotted line separates the training questions from the main questionnaire. The question labels are color-coded according to difficulty: green represents Easy, yellow represents Medium, and red represents Hard questions. Figure 7 provides several insights. First, an average of 72% of human evaluators answered the training questions correctly without prior training. The question most frequently answered correctly by humans was S29Q1 (Kurobe-Unazukionsen, Toyama), with an 89% correct-response rate. The question with the fewest correct answers was S16Q1 (Hino, Shiga), with only five humans answering correctly, and both the baseline and AI models were also wrong.

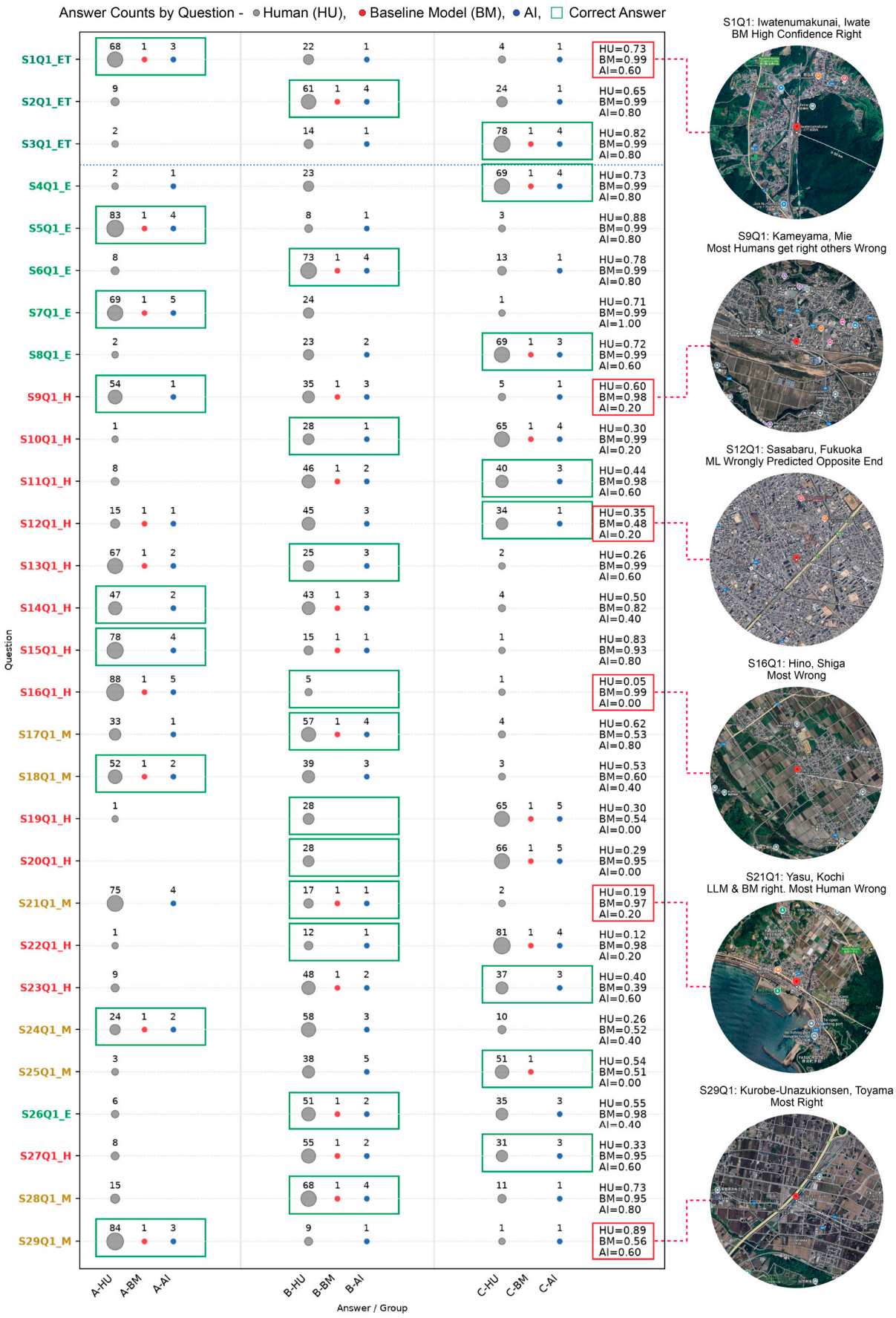

Figure 7. Correctness and confidence scores for each evaluator type by question, with example aerial images of neighborhoods selected for case studies.

Looking more closely at the confidence level and correctness, several behavior patterns emerge. There are cases where the model is both correct and highly confident, as in S1Q1 (Iwatenumakunai, Iwate), where it was 99% confident that the answer is Class 0, and it is correct. There are also cases where the model made right (S29Q1: Kurobe-Unazukionsen, Toyama) and wrong (S23Q1: Fukaya, Saitama) predictions with low confidence. However, discrepancies appear in S11Q1 (Sakurahommachi, Gifu), where the model was 98% confident in Class 1 but wrong, and in S12Q1 (Sasabaru, Fukuoka), where it predicted Class 0 while the correct answer was Class 2, the opposite end of the spectrum.

For human evaluators, no incorrect decision was made with high collective confidence. Their confidence scores generally are correlated with the correctness. For S21Q1 (Yasu, Kochi), 80% of human evaluators answered incorrectly and collective confidence was 19%. For 29Q1 (Kurobe-Unazukionsen), 89% answered correctly and collective confidence was also 89%. A similar pattern appears for AI models. Questions with average confidence above 60% were usually answered correctly. For example, in S7Q1 (Nahari, Kochi), all five AI models are 100% confident and all were correct; in S19Q1 (Kashiwa, Chiba) all five AI models were wrong with confidence near 0%.

These observations, together with the confidence scores for each evaluator group, provide insight into predictive behavior and indicate which questions each group found easier or harder. Figure 7 also shows the aerial images and associated feature values for several of the questions mentioned. These behavioral findings are discussed further in the discussion section.

4. Discussion

From the results, domain expertise appears to be one of the factors most associated with better performance. Those with domain expertise outperformed those without by an average of 1.3 points across all subgroups. Being a native Japanese evaluator shows an advantage among those without domain expertise but makes little difference among those with domain expertise. While unequal subgroup sizes may partly explain this pattern, a plausible reason is that non-experts find it harder to extract cues from the provided information, whereas native Japanese evaluators may recognize neighborhoods from their names, giving them a slight advantage. Conversely, among evaluators with domain expertise, prior contextual knowledge may introduce bias that leads to errors. On the other hand, those unfamiliar with Japanese neighborhoods may rely more objectively on the given aerial images and the values of each feature, sometimes resulting in higher scores.

For AI model performance, we initially expected strong results given their extensive pre-training with a large amount of dataset. In practice, only two out of five tested models outperformed the baseline model. Possible reasons include the uniqueness of the dataset used to train the baseline model, which AI models have not seen, rate limits during the questionnaire session, prompt fatigue from repeated similar inputs, and limited context windows. One observation, although not formally recorded, was that early questions often received more careful and accurate responses from the AI models. With repeated prompts, however, some models began to treat later inputs as identical and invested less effort in image analysis. Limited memory also prevented AI models from retaining their earlier answers, so they could not learn across items, unlike humans. While humans improved with experience, AI model performance sometimes declined over time.

Regarding reasoning, we see more similarity between humans and AI models than between humans and the baseline model. The differences appear in the weight assigned to each feature (Table 4). The baseline model assigned more weight to RoadCount and LandUseDiversity. Human evaluators emphasized ResidentialA and CommercialA, whereas AI models emphasized TransportationA. These differences may partly reflect the different

input formats. The baseline model depends solely on numerical values, whereas humans and AI models relied on images. Therefore, features that are easily recognized from aerial images, such as ResidentialA, CommercialA and GreenA, may seem more influential. It is not yet obvious whose reasoning is more appropriate, but contrasting feature priorities provide complementary perspectives for further investigation.

From a confidence perspective, humans and AI models showed similar behavior, whereas the baseline model differed. The confidence scores of humans and AI models generally corresponded with the correctness of their answers, while the baseline model showed several discrepancies. For example, in S11Q1, the baseline model was 98% confident but incorrect. The questions perceived as difficult also differed among evaluator types. In S29Q1, the baseline model had relatively low confidence, while most humans answered correctly with high confidence. In contrast, in S21Q1, the baseline model was highly confident and correct, whereas humans and AI models reported very low confidence. The similarity between humans and AI models may partly reflect the fact that AI models are trained on large amounts of human-generated data and may therefore exhibit some human-like response patterns. In contrast, the baseline model was trained specifically on the numerical dataset used in the previous study.

The confidence scores of the baseline model were derived from uncalibrated predicted class probabilities. Boosted decision-tree classifiers can produce poorly calibrated probability estimates, so the XGBoost probabilities used here may overstate certainty [49]. Probability calibration methods, such as sigmoid calibration or isotonic regression [50], could be applied to improve the reliability of these confidence estimates. Sigmoid calibration fits a logistic function to map the model's raw probability outputs to better calibrated probabilities, while isotonic regression uses a more flexible non-linear mapping based on the observed relationship between predicted probabilities and actual outcomes. However, calibration was not applied in this study because the objective was to examine the original behavior of the previously trained baseline model. Applying calibration may reduce some of the observed confidence discrepancies, particularly in cases where the model is highly confident but incorrect. However, probability calibration would primarily improve the reliability of the confidence estimates and would not necessarily improve classification accuracy.

Another notable characteristic is the difference in prediction patterns among humans, AI models, and the baseline model. Human and AI predictions generally showed more gradual changes, tending to move from Class 0 to Class 1 and then to Class 2. Therefore, even when the majority predicted incorrectly, their responses were usually limited to the directly adjacent class. In contrast, the baseline model does not always follow this pattern. Although rare, it occasionally made sudden changes from Class 0 to Class 2, or vice versa. One example is S12Q1 (Sasabaru, Fukuoka), where the baseline model predicted Class 0 even though the correct answer was Class 2. This behavior suggests that the baseline model sometimes made counterintuitive decisions that differed from human intuition.

4.1. Case Studies

By examining the unique cases identified in Figure 7, we can observe differences in evaluator behavior that may provide possible explanations for these patterns. However, these interpretations are exploratory and should be considered as hypotheses rather than definitive explanations.

The question most frequently answered correctly was S29Q1 (Kurobe-Unazukionsen, Toyama), a Class 0 neighborhood that the baseline model also predicted correctly, although with low confidence (0.56). The aerial image shows low building density and farmlands, which are clear indicators of Class 0. These features are easy to recognize, and the fact

that it was the last question likely helped, since evaluators could draw on experience from earlier questions. However, on closer inspection, this neighborhood is an onsen town that is usually busier on weekends and during seasonal tourism. This temporal pattern may not be fully recognizable by humans from a single image, but the baseline model may have detected hints of it from the feature values, which could explain its lower confidence.

The question most frequently answered incorrectly was S16Q1 (Hino, Shiga). The correct answer was Class 1, 93% of humans, all AI models, and the baseline model chose Class 0. Humans and AI models may have sensed the difficulty, which is why they reported low confidence, whereas the baseline model had 99% confidence. The aerial image shows mostly farmland and greenery with some building density, and the feature values appear relatively low except for LandUseDiversity, which suggests a low AHAPD neighborhood. A deeper review indicates that although the station is in a rural area with an agriculture-driven economy, there are nearby manufacturing facilities, pharmaceutical-related businesses, and supporting services such as business hotels. The main economic activities are not centered at the station but extend eastward. None of this context was provided to any of the evaluators, so widespread error is understandable. This case highlights a limitation: neighborhoods whose activity centers are not near the station or lie just outside the study area can be difficult to classify from imagery and their features value. Hence, the model was confidently wrong, whereas humans and AI models reported lower confidence.

In S9Q1 (Kameyama, Mie), most humans answered Class 0 correctly, while the AI models chose Class 1 and the baseline model chose Class 1 with high confidence (98%). The aerial image shows a mix of a developed town with larger buildings and farmland. The feature values also suggest a more developed neighborhood, with higher values for LandUseDiversity, RoadCount, and RecreationalA. At first glance it looks like Class 1, yet most humans chose Class 0. This may reflect human comparative reasoning: compared with the Class 1 example shown during training, this neighborhood appears emptier and less continuous, which leaned human decisions toward Class 0.

In S21Q1 (Yasu, Kochi), most humans (78%) and most AI models were wrong, while the baseline model was correct with 97% confidence. The aerial imagery shows sea, greenery, and farmland around a small port city, which resembles a low AHAPD neighborhood. However, the values of several features are relatively high, especially CommercialA, ResidentialA, and TotalLandUse, which pointed the trained model toward higher AHAPD; the baseline model predicted Class 1 with 97%confidence. Although not obvious from the aerial image, the area serves as a hub for fishing boats, with increasing population at specific times, especially in the early morning around 7:00 to 8:00. This temporal pattern aligns with the numerical indicators, which may lead to the model making a correct prediction.

In S12Q1 (Sasabaru, Fukuoka), the baseline model made an incorrect prediction at the opposite end of the spectrum, predicting Class 0 when the correct answer was Class 2, with 48% confidence. In contrast, most human evaluators and AI models answered correctly. One possible explanation is that recent growth in local activities may not be reflected in the data used to train the baseline model, either because the census data are outdated or because relevant information is missing from the OSM database. These changes, however, are more recognizable from the aerial imagery. Upon closer inspection, this neighborhood also serves as a transfer point between two train lines. Although the two stations are separated by a few hundred meters and are not directly connected, many commuters travel between them on foot. This pedestrian movement may contribute to higher AHAPD than would be expected based solely on the physical environment features.

From these examples, we see cases where visual cues are superior and cases where numerical features are more reliable. Each evaluator type has its own strengths and weaknesses. Performance in estimating AHAPD declines when the main activity centers lie

outside the study boundary or when economic activity is not centered around the station. Given these limitations, combining imagery with numerical features would likely improve the robustness of the baseline model.

4.2. Potential Application of the Outcome

The previously trained baseline model has the potential to be used for estimating ambient population density in locations where direct population measurements are limited or unavailable, because it relies on generally available tools and data sources. However, its applicability outside Japan has not yet been tested. Urban characteristics vary considerably across countries. Japan is generally characterized by a relatively compact urban form, high population density, strong public transport networks, and lower dependence on private cars. In contrast, many other countries may have more dispersed, less compact, and more auto-oriented urban patterns. Therefore, a model trained only on Japanese neighborhoods may not perform accurately when directly applied to different urban contexts. At the same time, some relationships among the input features may still remain useful. For example, in less compact neighborhoods, a larger commercial area may be accompanied by lower land-use diversity, which could influence the model output in a different but still meaningful way. The extent to which the current model can be generalized internationally requires further investigation.

Another potential application of the baseline model is its ability to quantify how different input features contribute to model decisions. Reasoning is generally more difficult to measure than accuracy, but model interpretation methods such as feature importance analysis can provide a quantitative representation of how individual urban features influence predictions. This makes the baseline model useful not only as a predictive tool, but also as an analytical tool for examining relationships between urban characteristics and outcomes such as ambient population density. A similar approach could potentially be applied to other urban outcomes, such as tax income or demographic indicators, provided that appropriate target data and input features are available.

In addition to the baseline model itself, this study proposes a questionnaire-based method for evaluating the real-world performance of ML models. The method also enables a three-way comparison among ML models, human evaluators, and multimodal AI models. Importantly, the comparison extends beyond accuracy to include reasoning and confidence, which are less commonly evaluated across different evaluator types. This provided a more diagnostic benchmark for identifying discrepancies, similarities, and behavioral patterns that may not be visible through accuracy alone. Because the framework uses a simple online questionnaire and widely available tools, it can be relatively easily adapted for further ML studies. By modifying the target classes, input information, and scoring methods, a similar approach could be applied to other spatial reasoning tasks for which no standardized benchmark currently exists.

4.3. Comparison with Previous Literature

Previous studies have applied machine learning and spatial data to estimate population and other urban conditions. For example, Zhao et al. [51] developed an XGBoost-based method for mapping population distribution using multisource spatial data. Akiyama et al. [52] developed machine-learning approaches, including XGBoost, to estimate the spatial distribution of vacant houses in Japan using municipal and building related data. These studies focus primarily on improving estimation accuracy and practical applicability rather than comparing the raw spatial reasoning capability of trained models with human or AI systems.

Studies that directly compare humans and machine-learning systems also tend to emphasize predictive accuracy. Measures such as reasoning, confidence, and characteristic prediction behavior are less commonly evaluated together. The present study extends this type of comparison by using the previously trained baseline model not only as an evaluator, but also as a diagnostic reference. Its predictions, confidence scores, and feature importance were used to identify particular cases where humans, multimodal AI models, and baseline models behave differently. This allows the comparison to move beyond the question of which evaluator achieved the highest score and toward understanding why their predictions may differ.

The reasoning results also show connections with previous urban literature. LandUseDiversity was one of the most influential features for the baseline model, consistent with the importance of mixed land use and diversity described in previous urban literature. Cervero and Kockelman identified density, diversity, and design as important dimensions relating the built environment to travel behavior [2]. Jacobs similarly emphasized mixed uses, short blocks, density, and concentration of activity as conditions contributing to urban vitality [3,4]. RoadCount, BasedDensity, and CommercialA, which were also influential in the baseline model, similarly reflect characteristics associated with accessibility, urban intensity, and activity concentration. These types of features also appear in previous population-estimation studies, including LandScan [1], which utilized transportation infrastructure, land cover, and indicators of economic activity when modeling population distribution. Humans and AI models also assigned importance to several of these features. However, some differences were notable. ResidentialA was considered relatively important by human evaluators but received considerably less importance from the baseline model, while TransportationA was particularly important to the AI models. These differences suggest that the evaluator types may emphasize different aspects of the same urban environment.

4.4. Limitations

This study has several limitations. The baseline model has no vision capability and was trained solely on 16 numerical features. It also does not account for the spatial distribution of each feature, which may be important for understanding neighborhood characteristics [5]. Aerial images, which contain additional spatial and visual cues, were not used during training. Developing a vision-capable model would require a different model architecture and training dataset and is therefore left for future work.

Another limitation concerns the differences in information provided to each evaluator type. Although the study aimed to compare the three evaluator types as fairly as possible, the information provided to humans and AI models was not identical to that used by the baseline model. The baseline model was trained on numerical values and is therefore more suited to processing tabular information. Humans rely on visual interpretation, life experience, and for some evaluators, domain expertise, but may have more difficulty interpreting raw numerical values without prior training. Multimodal AI models can process both images and numerical information and benefit from large-scale pre-training, but they are also affected by factors such as prompting, rate limits, context limitations, and their limited ability to adapt during a single evaluation session.

For these reasons, humans and AI models were provided with aerial images in addition to the numerical features, while the baseline model relied only on numerical inputs. This difference may influence the comparative results. Neighborhoods with strong visual cues may favor humans and AI models, while neighborhoods where relevant information is less visually apparent but is captured in the numerical features may favor the baseline model. Because the questionnaire cases were selected based on baseline-model confidence and correctness rather than on the strength of visual cues, the results should be interpreted as a

diagnostic comparison under different but evaluator-appropriate input conditions rather than as a strictly controlled comparison of performance. Future studies could improve fairness by selecting approximately equal numbers of cases that favor visual information, tabular information, or both.

The sample size and subgroup distribution also present limitations. Although 94 human evaluators were included, the number of evaluators within some demographic subgroups was small and could not be controlled during recruitment (Table 3). As a result, individual high or low scores may have a stronger influence on the mean of smaller subgroups. The subgroup comparisons should therefore be interpreted cautiously. A larger and more balanced sample would provide more stable estimates of subgroup performance. Future studies could also compare the reasoning patterns of the highest-, average-, and lowest-performing evaluators within each subgroup. However, such detailed subgroup analysis was beyond the scope of the present study and could reduce the focus of the current comparison.

The tools used to evaluate reasoning also differed across evaluator types. For the baseline model, SHAP-based feature importance was used as a quantitative proxy for reasoning. No directly equivalent method exists for evaluating human or AI reasoning in the same form. Therefore, the questionnaire was designed to extract perceived feature importance from humans and AI models and convert these responses into a comparable metric. Although this approach allows reasoning patterns to be compared across evaluator types, the resulting measures are not identical and should be interpreted as approximations of feature-based reasoning rather than direct measurements of the underlying decision-making process. One possible direction for improving the reasoning score is to treat each evaluator's responses as analogous to a tree within an XGBoost system and recalculate feature contributions using a method more closely aligned with SHAP-based feature importance. This could provide a more consistent basis for comparison across evaluator types. However, this approach is still conceptual and has not yet been tested.

The questionnaire design also required a balance between thoroughness and accessibility. The test set for the baseline model included 180 neighborhoods, but using the entire set would have made the questionnaire too long and likely reduced human participation and subgroup diversity. Conversely, using too few questions would have limited the ability to distinguish performance patterns across evaluator types. We therefore selected 29 cases from the full set to balance the time required to complete the questionnaire with the need to capture a wider range of model behavior. An alternative approach could involve fewer evaluators completing a more detailed questionnaire or a larger group completing a shorter questionnaire. The present study adopted a middle approach to capture a broader range of perspectives.

AHAPD also does not fully capture temporal variations across the day, weekends, or seasons because it was represented by the average of selected time periods. These variations are important because residential neighborhoods may have higher populations at night, office districts during working hours, transportation hubs during commuting periods, and tourism areas during weekends or particular seasons. Using different temporal objectives could therefore reveal different relationships between urban features and ambient population density. The performance of each evaluator type may also change depending on the temporal objective being evaluated. Some temporal or functional characteristics may be easier to infer from aerial imagery, while others may be better represented by numerical features. As a result, the relative performance of humans, AI models, and the baseline model may vary when different forms of spatial reasoning are required.

The dataset itself was developed at the proof-of-concept stage. Data from OSM and MSS were collected manually and combined to create the training dataset. Because

this process was not conducted at scale, the training dataset remained limited and only one main AHAPD objective was explored. Future studies could expand the dataset and evaluate additional temporal objectives to examine how relationships among urban features, evaluator reasoning, and prediction performance change under different conditions.

The AI model evaluation is also limited by the stochastic nature of generative AI outputs. Each AI model was tested only once, meaning that the recorded result may not represent its average or best possible performance. However, the single-run design was consistent with the human evaluation procedure, in which each evaluator completed the questionnaire only once. Similarly, the comparison used the predictions from the single previously trained baseline model rather than repeated model-training runs. Future studies could conduct repeated AI sessions and multiple model-training runs to quantify variation within each evaluator type.

Finally, the study relies mainly on isolated aerial images centered on each neighborhood, providing limited spatial context beyond the defined study area. This may affect human and AI judgments in cases where important activities, urban functions, or connections are located outside the visible area. Future studies could provide additional regional context or multi-scale imagery while maintaining consistent information across questionnaire cases.

4.5. Future Studies

In addition to the improvement directions discussed in the Section 4.4, several areas should be explored in future studies to improve the baseline model and move it toward practical application. A priority is the development of a more adaptable data-collection pipeline. Because data quality and availability are critical for machine learning, an automated pipeline that can regularly retrieve and integrate updated OSM data would improve both the scale and reliability of the training dataset. Adding vision capability and incorporating the spatial distribution of land use could also make the baseline model more robust. With these improvements, future comparison could use more similar input conditions across evaluator types and provide a fairer evaluation.

Further analysis of additional cases from the dataset could also help identify situations in which the baseline model, humans, or AI models have particular advantages. Expanding the study area would be useful for neighborhoods where urban activity is not centered on the station or where important attractions and activity centers are located just outside the current study boundary. To further improve the benchmark, a complementary study with fewer evaluators but more detailed questions and deeper analysis could also be valuable. Exploring potential applications of the framework in professional planning and urban analysis contexts is another important direction.

In this study, uncalibrated predicted class probabilities were used to represent the baseline model's confidence. Future studies could apply probability calibration methods, such as sigmoid calibration or isotonic regression, to improve the reliability of these confidence scores. Calibration may reduce some of the discrepancies observed in cases where the baseline model produced highly confident but incorrect predictions. It would therefore be useful to compare the confidence patterns of a calibrated baseline model with those of human evaluators and AI models. However, calibration should not simply involve uniformly reducing confidence values, because the relationship between predicted probability and correctness may vary across cases. Probability calibration may improve the reliability and interpretation of confidence estimates, but it would not necessarily improve classification accuracy.

This study focuses entirely on Japanese neighborhoods, so the extent to which the findings can be generalized internationally remains uncertain. Japanese neighborhoods

have distinctive characteristics related to public transportation, urban compactness, land-use patterns, and cultural context that may differ from those in other countries. However, the same general framework could be applied elsewhere by training a country-specific machine learning model using locally available data and evaluating it with the same questionnaire-based methodology. One important challenge would be identifying a suitable proxy for AHAPD, since the MSS data used in this study are available in Japan but may not have a direct equivalent in other countries. Cross-country applications of the framework could help determine which findings are specific to the Japanese context and which are more broadly applicable.

5. Conclusions

This study evaluated the performance of a previously trained XGBoost baseline model by comparing it with human evaluators and multimodal AI models using a questionnaire-based benchmark. It also examined similarities and differences in their performance, reasoning, confidence, strengths, and weaknesses through the spatial reasoning task of estimating Average Hourly Ambient Population Density (AHAPD). The benchmark was designed to evaluate the three evaluator types across accuracy, reasoning, and confidence. It was first tested internally and refined before being distributed to external human evaluators with different demographic and professional backgrounds and varying levels of familiarity with Japan. The same benchmark was also distributed to 5 free-tier commercially available multimodal AI models. In total, 100 responses were analyzed: 94 human evaluators, 5 AI models, and 1 output from the baseline model.

Within this diagnostic benchmark, the baseline model scored 16 points. This was 6% higher than the average score of the 94 human evaluators, which was 15.1 points, and 12.7% higher than the average score of the 5 tested AI models, which was 14.2 points. The baseline model also performed close to human evaluators with domain expertise, who averaged 15.8 points. However, these results should be interpreted within the conditions of the benchmark. The questionnaire cases were intentionally selected based on baseline-model correctness and confidence, and human subgroup sizes varied because recruitment was not controlled. The AI result should also be interpreted cautiously, as only five free-tier multimodal AI models were tested, each in a single stochastic run under limitations such as rate limits, restricted context windows, and possible prompt fatigue. Therefore, the reported scores should not be treated as population-level estimates, overall test-set performance, or the upper-bound capability of current multimodel AI models. Nevertheless, the results suggest that a task-specific ML model trained on a relatively small dataset can remain competitive in a narrowly defined spatial reasoning task.

The reasoning analysis revealed clear differences among evaluator types. Humans tended to emphasize features such as ResidentialA, CommercialA, and GreenArea, while AI models prioritized TransportationA and the baseline model assigned more weight to RoadCount and LandUseDiversity. Human evaluators and AI models showed greater similarity to each other, possibly because both use visual and numerical information, while the baseline model relied only on tabular features. This allows the baseline model to provide a complementary perspective based on patterns that may not be immediately apparent through visual interpretation.

Confidence analysis also revealed different behavioral patterns. Human and AI confidence generally corresponded with response correctness, while the baseline model sometimes produced highly confident but incorrect predictions. Because its confidence was represented by uncalibrated predicted class probabilities, some of these discrepancies may reflect probability miscalibration. Comparing confidence across evaluator types also helped

identify cases of agreement, disagreement, and differing levels of certainty, providing insight beyond accuracy alone.

The main contribution of this study is therefore not simply the comparison of final accuracy scores. It is the development of a questionnaire-based diagnostic benchmark that enables a task-specific ML model, human intuition, and multimodal AI models to be compared across accuracy, reasoning, and confidence in the same spatial reasoning task. This framework could also be adapted to future task-specific comparisons among different evaluator types. By examining feature priorities, confidence patterns, and discrepancy cases, the benchmark provides a more detailed understanding of how different evaluator types make decisions.

Overall, the baseline model demonstrated performance comparable to the average human evaluator with domain expertise under the conditions of this benchmark. At the same time, the differences observed in reasoning and confidence suggest that no single evaluator type consistently has an advantage across all cases. Humans and multimodal AI models benefit from visual interpretation, while the baseline model can identify numerical relationships that may be less intuitive. Combining these complementary forms of spatial reasoning may therefore provide a useful direction for improving future ambient population density estimation models. Future work should expand the dataset, improve the data-collection pipeline, incorporate spatial and visual information, evaluate probability calibration, conduct repeated AI evaluations, and test the framework in other geographic and temporal contexts.

Author Contributions: Conceptualization, P.R. and J.T.; methodology, P.R.; software, P.R.; validation, P.R. and J.T.; formal analysis, P.R.; investigation, P.R.; resources, P.R.; data curation, P.R.; writing—original draft preparation, P.R.; writing—review and editing, P.R.; visualization, P.R.; supervision, J.T.; project administration, P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: All research involving human participants was conducted in accordance with the principles of the Declaration of Helsinki. Formal ethical approval from a local Institutional Review Board (IRB) or equivalent ethics committee was not sought because the study involved a non-medical, minimal-risk online questionnaire. All responses included in the analysis were from adult evaluators from multiple countries. Electronic informed consent was obtained before participation. No directly identifying information was included in the analytical dataset. The study fell outside the scope of Japan's Ethical Guidelines for Life Science and Medical Research Involving Human Subjects because it did not concern health, disease, medical treatment, biological samples, or genetic information. In addition, the study was not conducted or supported by the U.S. Department of Health and Human Services and was therefore not subjected to 45 CFR Part 46 under the HHS framework. If the Common Rule criteria were applied, the study would likely fall within the exemption for survey research under 45 CFR 46.104(d)(2)(i). Accordingly, formal prior approval from an IRB or Ethics Committee was not sought.

Informed Consent Statement: Electronic informed consent was obtained from all human participants before they began the questionnaire. The introductory section explained the purpose and content of the study, the voluntary nature of participation, and the intended academic use of the collected data. Participants were free to discontinue the questionnaire at any time. Only responses from participants who provided informed consent were included in the analysis.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: The authors would like to thank all evaluators who spent their valuable time completing the questionnaire. Their participation provided valuable data for this research. We are grateful for all the feedback, questions, and comments during the post-questionnaire section,

some of which became meaningful discussion topics. We would like to thank our peers, colleagues, and advisors for all their support during the whole process, from design, testing, conducting the questionnaire, to the very end of the research. In this paper, generative AI (ChatGPT-5) was used as a brainstorming assistant to improve the clarity of ideas after the authors had outlined the research scope and drafted the manuscript. After the initial results were received, AI was also used as a coding assistant for data analysis. Five generative AIs (GPT-4o, GPT-5, Gemini 2.5 Flash, Claude Sonnet 4, and Grok-3) were included as an evaluator type for performance comparison between humans, AI and ML, which is the main focus of the study. Finally, ChatGPT-5 was used as a language improvement assistant at the end of the study, under the authors' close supervision at every step.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AHAPD	Average Hourly Ambient Population Density
ACQ	Average Confidence for Question Q
DFI	Derived Feature Importance
ML	Machine Learning
MSS	Mobile Spatial Statistics

Appendix A

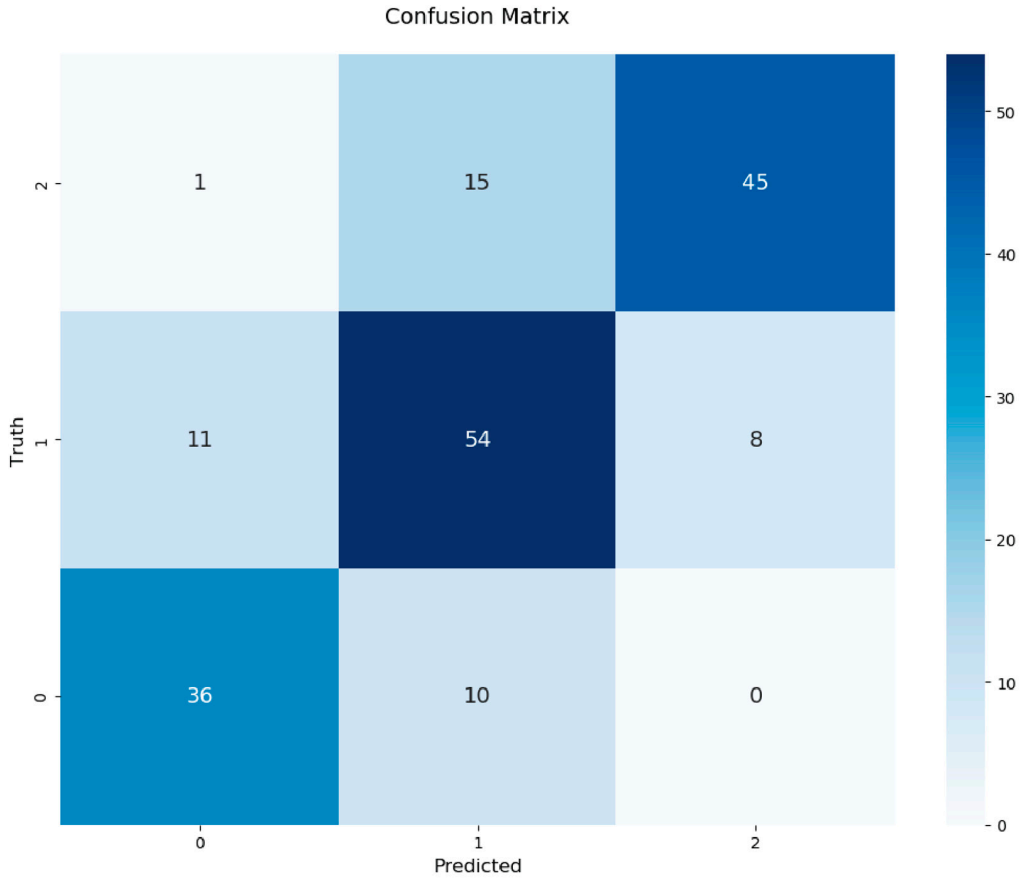

Figure A1. Confusion Matrix showing all predictions made by the baseline model on the testing dataset. The neighborhoods used in the questionnaire were selected from this confusion matrix.

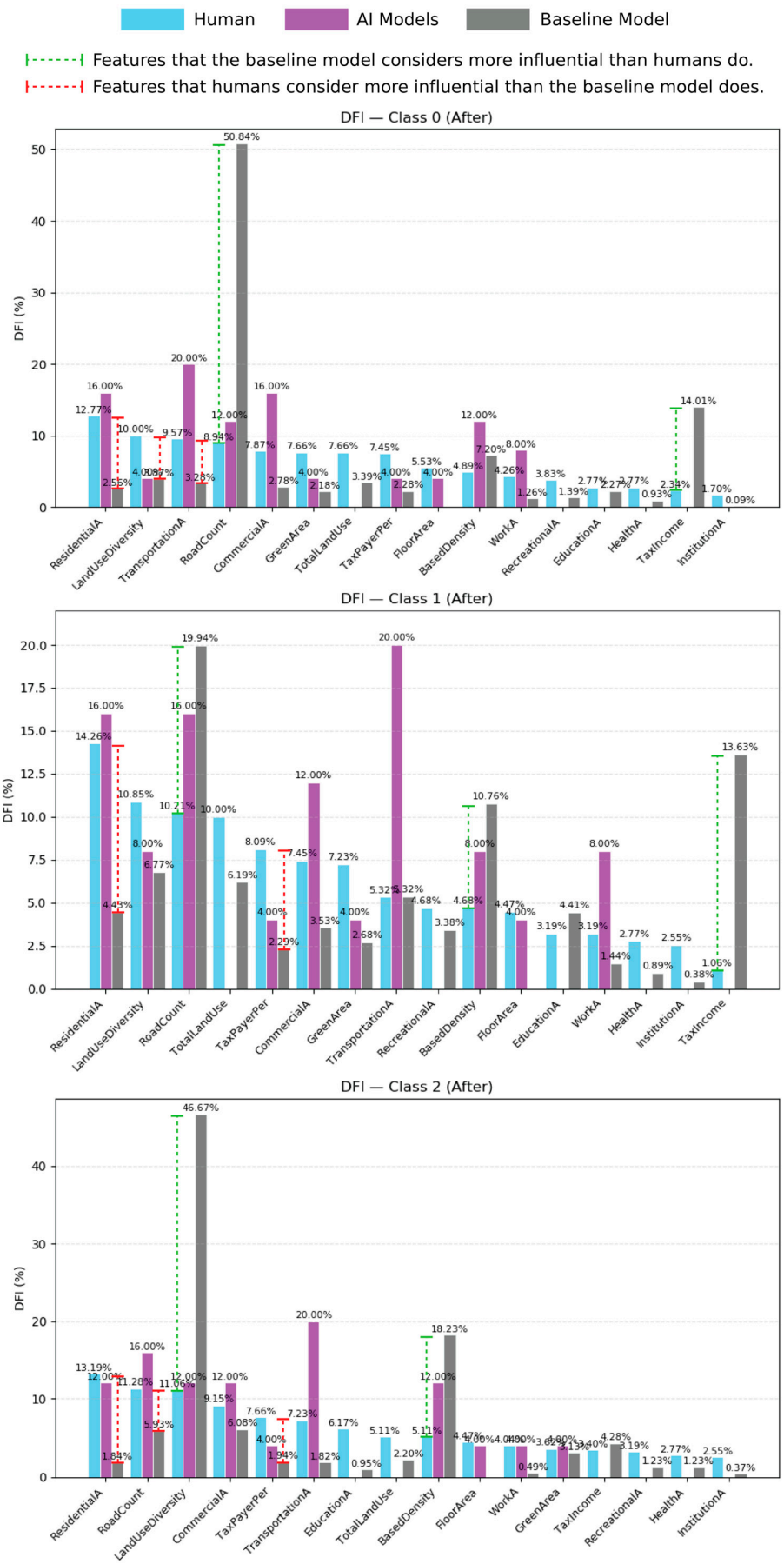

Figure A2. Comparison of the Derived Feature Importance scores for each evaluator type in the context of Classes 0, 1, and 2.

Table A1. Definitions, SHAP values, correlation directions, and data sources for the 16 input features used by the baseline model.

Feature Name.	Description	Mean Absolute SHAP (Class '2')	Correlation	Source
CommercialA	Total floor area of commercial land use (bank, bar, cafe, commercial, restaurant, retail, etc.)	0.276235	+	GIS & OSM
EducationA	Total floor area of educational land use including (college, language_school, kindergarten, university, etc.)	0.047702	–	GIS & OSM
HealthA	Total floor area of health related land use including (clinic, dentist, doctors, hospital, nursing_home, etc.)	0.021966	–	GIS & OSM
InstitutionA	Total floor area of institutional land use including (civic, fire_station, government, police, townhall, etc.)	0.016998	–	GIS & OSM
RecreationalA	Total floor area of recreational land use (arts_centre, barn, bbq, church, community_centre, etc.)	0.032891	–	GIS & OSM
ResidentialA	Total floor area of residential type of land use (apartments, detached, dormitory, house, residential, etc.)	0.102843	+	GIS & OSM
TransportationA	Total floor area of transportation land use (bus_station, ferry_terminal, train_station, etc.)	0.091451	+	GIS & OSM
WorkA	Total floor area of work related land use (coworking_space, industrial, office, warehouse, etc.)	0.016481	+	GIS & OSM
FloorArea	Total footprint × Number of Floors of all buildings within the study area.	0.190040	+	GIS & OSM
RoadCount	Total number of road segment within the road networks within the study area	0.476555	+	GIS & OSM
GreenArea	Total floor area of leisure type of land use (park, garden, pitch, playground, etc.)	0.148018	+	GIS & OSM
LandUseDiversity	Total types of land use within the study area	0.952107	+	GIS & OSM
TotalLandUse	Total count of all land use types within the study area	0.109048	+	GIS & OSM
BasedDensity	Municipality Population/Total Land Area (ha)	0.818544	+	e-Stats
TaxPayerPer	Number of Taxpayer/Municipality Population	0.141565	+	e-Stats
TaxIncome	Tax revenue of the Municipality	0.214959	+	e-Stats

References

1. Dobson, J.E.; Bright, E.A.; Coleman, P.R.; Durfee, R.C.; Worley, B.A. LandScan: A Global Population Database for Estimating Populations at Risk. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 849–858. [CrossRef]
2. Cervero, R.; Kockelman, K. Travel Demand and the 3Ds: Density, Diversity, and Design. *Transp. Res. Part Transp. Environ.* **1997**, *2*, 199–219. [CrossRef]
3. Row, A.T. Review of The Death and Life of Great American Cities. *Yale Law J.* **1962**, *71*, 1597–1602. [CrossRef] [PubMed]
4. Šćepanović, S.; Joglekar, S.; Law, S.; Quercia, D. Jane Jacobs in the Sky: Predicting Urban Vitality with Open Satellite Data 2021. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–25.
5. Krugman, P. *A Dynamic Spatial Model*; National Bureau of Economic Research: Cambridge, MA, USA, 1992; p. w4219.
6. Fujita, M.; Thisse, J.-F. New Economic Geography: An Appraisal on the Occasion of Paul Krugman's 2008 Nobel Prize in Economic Sciences. *Reg. Sci. Urban Econ.* **2009**, *39*, 109–119. [CrossRef]
7. Rojradtanasiri, P.; Tamura, J.; Kobayashi, M. Estimating Ambient Population Density Using Physical Features from GIS and Machine Learning: A Study Based on Japanese Neighborhood. *J. Asian Archit. Build. Eng.* **2024**, *24*, 5787–5802. [CrossRef]
8. Spatial Without Compromise QGIS. Available online: https://www.qgis.org/ (accessed on 3 July 2026).
9. Trimaille, E. QuickOSM—Download OSM Data Thanks to the Overpass API. Available online: https://plugins.qgis.org/plugins/QuickOSM/ (accessed on 17 September 2025).
10. NTT DOCOMO, Inc. Mobile Spatial Statistics Population Map 「モバイル空間統計 人口マップ」 "Mobile Spatial Statistics Population Map". 2026. Available online: https://mobakumap.jp/ (accessed on 9 December 2025).
11. Terada, M.; Nagata, T.; Kobayashi, M. Population Estimation Technology for Mobile Spatial Statistics. *NTT DOCOMO Tech. J.* **2013**, *14*, 10–15.
12. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2006.
13. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco, CA, USA, 2016; pp. 785–794.
14. View Data | Municipality Data | System of Social and Demographic Statistics(SSDS) | Search by Areas. Available online: https://www.e-stat.go.jp/en/regional-statistics/ssdsview/municipality (accessed on 27 June 2024).

15. F1_Score. Available online: https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html (accessed on 27 June 2024).
16. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
17. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [CrossRef]
18. IBM. What Is an AI Model? | IBM. Available online: https://www.ibm.com/think/topics/ai-model (accessed on 3 July 2026).
19. Meta AI. Introducing Llama 3.1: Our Most Capable Models to Date. Available online: https://ai.meta.com/blog/meta-llama-3-1/ (accessed on 3 July 2026).
20. Large AI Model Empowered Multimodal Semantic Communications. Available online: https://ieeexplore.ieee.org/abstract/document/10670195?casa_token=_y_QWEC6YnwAAAAA:ulpttGfp5cfEl_QGZkR8HleoRp-ZN-sWyaeTG_JLCr9J4t8CkRIvAeko7wIk37YPkNzBhZM-SLo (accessed on 15 August 2026).
21. Artificial Analysis. AIME 2025 Benchmark Leaderboard. Available online: https://artificialanalysis.ai/evaluations/aime-2025 (accessed on 2 July 2026).
22. Center for AI Safety; Phan, L.; Gatti, A.; Li, N.; Khoja, A.; Kim, R.; Ren, R.; Hausenloy, J.; Zhang, O.; Mazeika, M.; et al. A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities. *Nature* **2026**, *649*, 1139–1146. [CrossRef] [PubMed]
23. Tschisgale, P.; Maus, H.; Kieser, F.; Kroehs, B.; Petersen, S.; Wulff, P. Evaluating GPT- and Reasoning-Based Large Language Models on Physics Olympiad Problems: Surpassing Human Performance and Implications for Educational Assessment 2025. *Phys. Rev. Phys. Educ. Res.* **2025**, *21*, 020115. [CrossRef]
24. Koubaa, A.; Qureshi, B.; Ammar, A.; Khan, Z.; Boulila, W.; Ghouti, L. Humans Are Still Better than ChatGPT: Case of the IEEEXtreme Competition. *Heliyon* **2023**, *9*, e21624. [CrossRef] [PubMed]
25. Katz, D.M.; Bommarito, M.J.; Gao, S.; Arredondo, P. GPT-4 Passes the Bar Exam. *Philos. Trans. A Math. Phys. Eng. Sci.* **2024**, *382*, 20230254. [CrossRef] [PubMed]
26. Salinas, M.P.; Sepúlveda, J.; Hidalgo, L.; Peirano, D.; Morel, M.; Uribe, P.; Rotemberg, V.; Briones, J.; Mery, D.; Navarrete-Dechent, C. A Systematic Review and Meta-Analysis of Artificial Intelligence versus Clinicians for Skin Cancer Diagnosis. *npj Digit. Med.* **2024**, *7*, 125. [CrossRef] [PubMed]
27. Shehu, H.A.; Browne, W.; Eisenbarth, H. A Comparison of Humans and Machine Learning Classifiers Detecting Emotion from Faces of People with Different Coverings 2021. *Appl. Soft Comput.* **2022**, *130*, 109701.
28. Kandul, S.; Micheli, V.; Beck, J.; Burri, T.; Fleuret, F.; Kneer, M.; Christen, M. Human Control Redressed: Comparing AI and Human Predictability in a Real-Effort Task. *Comput. Hum. Behav. Rep.* **2023**, *10*, 100290. [CrossRef]
29. Yan, J.; Yan, P.; Chen, Y.; Li, J.; Zhu, X.; Zhang, Y. Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels 2024. *arXiv* **2024**, arXiv:2411.13775.
30. Cowley, H.P.; Natter, M.; Gray-Roncal, K.; Rhodes, R.E.; Johnson, E.C.; Drenkow, N.; Shead, T.M.; Chance, F.S.; Wester, B.; Gray-Roncal, W. A Framework for Rigorous Evaluation of Human Performance in Human and Machine Learning Comparison Studies. *Sci. Rep.* **2022**, *12*, 5444. [CrossRef] [PubMed]
31. OpenAI. Rate Limits | OpenAI API. Available online: https://developers.openai.com/api/docs/guides/rate-limits (accessed on 3 July 2026).
32. Bergmann, D. What Is a Context Window? | IBM. Available online: https://www.ibm.com/think/topics/context-window (accessed on 3 July 2026).
33. Dong, Z.; Li, J.; Men, X.; Zhao, W.X.; Wang, B.; Tian, Z.; Chen, W.; Wen, J.-R. Exploring Context Window of Large Language Models via Decomposed Positional Vectors. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 10320–10347. [CrossRef]
34. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* **2024**, *630*, 625–630. [CrossRef] [PubMed]
35. Sarmadi, H.; Wahab, I.; Hall, O.; Rögnvaldsson, T.; Ohlsson, M. Human Bias and CNNs' Superior Insights in Satellite Based Poverty Mapping. *Sci. Rep.* **2024**, *14*, 22878. [CrossRef] [PubMed]
36. IBM. What Are Convolutional Neural Networks? | IBM. Available online: https://www.ibm.com/think/topics/convolutional-neural-networks (accessed on 3 July 2026).
37. Ahn, D.; Yang, J.; Cha, M.; Yang, H.; Kim, J.; Park, S.; Han, S.; Lee, E.; Lee, S.; Park, S. A Human-Machine Collaborative Approach Measures Economic Development Using Satellite Imagery. *Nat. Commun.* **2023**, *14*, 6811. [CrossRef] [PubMed]
38. Panczak, R.; Charles-Edwards, E.; Corcoran, J. Estimating Temporary Populations: A Systematic Review of the Empirical Literature. *Humanit. Soc. Sci. Commun.* **2020**, *6*, 87. [CrossRef]
39. OpenAI. What Is ChatGPT: FAQ. Available online: https://help.openai.com/en/articles/12677804-what-is-chatgpt-faq (accessed on 16 August 2026).

40. Rawat, S.; Suresh, V.K. GPT Takes the SAT: Tracing Changes in Test Difficulty and Students' Math Performance. Cincinnati, United States. Publisher SSRN. 2025. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4915452 (accessed on 3 July 2026).
41. The SAT—SAT Suite | College Board. Available online: https://satsuite.collegeboard.org/sat (accessed on 2 July 2026).
42. Workspace, G. Google Forms: Online Form Builder. Available online: https://workspace.google.com/products/forms/ (accessed on 3 July 2026).
43. Confusion_Matrix. Available online: https://scikit-learn/stable/modules/generated/sklearn.metrics.confusion_matrix.html (accessed on 27 June 2024).
44. Hello GPT-4o. Available online: https://openai.com/index/hello-gpt-4o/ (accessed on 3 July 2026).
45. GPT-5 Is Here. Available online: https://openai.com/gpt-5/ (accessed on 3 July 2026).
46. Gemini 2.5 Flash | Gemini API. Available online: https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash (accessed on 3 July 2026).
47. Claude Sonnet. Available online: https://www.anthropic.com/claude/sonnet (accessed on 3 July 2026).
48. Grok 3 Beta—The Age of Reasoning Agents. Available online: https://x.ai/news/grok-3 (accessed on 3 July 2026).
49. Niculescu-Mizil, A.; Caruana, R. Obtaining Calibrated Probabilities from Boosting. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*; AUAI Press: Arlington, VA, USA, 2005; pp. 413–420.
50. 1.16. Probability Calibration. Available online: https://scikit-learn.org/stable/modules/calibration.html (accessed on 15 August 2026).
51. Zhao, X.; Xia, N.; Xu, Y.; Huang, X.; Li, M. Mapping Population Distribution Based on XGBoost Using Multisource Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11567–11580. [CrossRef]
52. Akiyama, Y.; Baba, H.; Ono, Y.; Takaoka, H. Sophistication of Monitoring Method for Spatial Distribution of Vacant Houses Using Machine Learning: A study on the estimation method of spatial distribution of vacant houses using municipal public data (Part 3). *J. Archit. Plan. Trans. AIJ* **2021**, *86*, 2136–2146. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.