Evaluating Cross-Model Consistency of LLM-Generated Sustainability Data for Points of Interest

No Author Given

No Institute Given

Abstract. The lack of reliable structured sustainability data for tourism points of interest (POIs) remains a key limitation for sustainability-aware recommendation systems. While Large Language Models (LLMs) have emerged as a potential solution for synthetic data generation, the reliability and consistency of their outputs remain largely unexplored. This paper evaluates the cross-model consistency of LLM-generated environmental sustainability data for POIs. We propose a framework combining correlation analysis, bucket-based comparison, and POI-level disagreement. The results show that LLMs exhibit moderate to high agreement in relative ranking patterns for variables such as noise level and water consumption. However, agreement is strongly variable-dependent and varies across categories, with lower consistency observed for more subjective criteria. Bucket-level analysis further reveals high near-category agreement but lower exact agreement, indicating differences in scale usage and classification boundaries. Furthermore, models differ in numerical magnitude, indicating that outputs are not directly interchangeable. This evaluation highlights that performance is variable-dependent. While high-performing models provide practical solutions for resource-constrained settings, robust decision-making requires variable-specific model selection. These insights establish a foundation for future integration of LLM-generated data into analytical and decision-making frameworks.

Keywords: Large Language Models (LLMs), Sustainability Data Generation, Points of Interest, Cross-Model Consistency.

1 Introduction

Global tourism generates significant economic benefits but also raises major environmental, cultural, and social challenges in host destinations [1, 2]. Traditional Point of Interest (POI) recommender systems [3], typically driven by popularity or user preferences, often fail to account for these sustainability issues, limiting their ability to support responsible tourism decisions [4].

In the works [5, 6], they introduced the foundations of a Multi-Criteria Recommender System (MC-RS) for sustainable tourism. In [5], they proposed a hierarchical organization of criteria, structuring sustainability into environmental,

cultural, social, and economic dimensions, following a multi-dimension perspective on sustainable tourism. Building on this, [6] introduced a multi-source data acquisition strategy, combining direct data, proxy variables, and Artificial Intelligence techniques such as Natural Language Processing and Large Language Models (LLMs) to populate this hierarchy. Despite these advances, a key limitation is the lack of reliable and fine-grained sustainability data at the POI level, where direct measurements are often unavailable or costly.

Recent research has explored integrating sustainability into tourism recommender systems, for example by promoting environmentally friendly POIs through re-ranking techniques [7] or by using aggregated sustainability indicators for destination recommendation [8]. In parallel, LLMs have shown strong potential in recommendation [9], data generation tasks [10] and tourism experimentation platforms that employ LLM-generated synthetic user data and reviews to support sustainable recommendation scenarios [11]. However, the consistency and reliability of such generated data remain largely unexplored.

In this paper, we focus on the environmental dimension and evaluate the cross-model consistency of LLM-generated sustainability data for POIs. We propose an evaluation framework combining correlation analysis (general and category-based), bucket comparison, and POI-level disagreement. The findings provide insights into the stability and variability of LLM-generated sustainability data and highlight the extent to which model outputs are interchangeable. In addition, a consistency-based model recommendation approach is introduced to support informed model selection. This study does not include direct integration into decision-making systems; instead, it provides a foundation for the future use of LLM-generated data in applications such as multi-criteria decision-making (MCDM) and recommendation systems.

The main contributions of this paper are threefold: (i) a comprehensive framework for evaluating cross-model consistency in LLM-generated sustainability data across multiple dimensions; (ii) a consistency-based model recommendation approach for selecting reliable models; and (iii) practical guidelines for using LLM-generated sustainability data under resource and variability constraints.

The remainder of the paper is organized as follows. Section 2 presents the methodology. Section 3 reports the results. Section 4 outlines practical guidelines for using LLM-generated data. Section 5 concludes the paper.

2 Methodology

2.1 Study Objective and Evaluation Scope

This study evaluates the consistency of environmental sustainability data generated by multiple LLMs for POIs in Barcelona. In data-scarce domains like tourism sustainability, LLMs offer a promising approach for generating structured variables; however, the reliability of such data depends on the degree of agreement across different models.

Focusing on a diverse group of LLMs and a broad set of POIs, we analyze the values produced by every model for each environmental variable. Unlike

accuracy-based evaluations, this study does not rely on external ground truth. Instead, it adopts a consistency-oriented framework, using cross-model agreement as a proxy for stability and reliability. The evaluation considers three complementary dimensions: ranking, category-aware, buckets consistency, and magnitude variability, as summarized in Table 1.

The analysis includes multiple sustainability variables representing both quantitative and qualitative impacts. By conducting the evaluation under controlled conditions, we ensure that differences in outputs arise solely from model behaviour. Though this study does not include downstream applications (such as MCDM), it provides the foundational analysis necessary to support the future integration of LLM-generated data into decision-making frameworks.

Table 1. Summary of evaluation components and methodologies.

Component	**Objective**	**Method**
Cross-model correlation	Ranking agreement	Pairwise Spearman rank correlation (global and category wise)
Bucket-based comparison	Classification behaviour	Fine-grained discretization and match analysis
POI-level disagreement	Magnitude variability	Standard deviation and value range across models
Model recommendation	Model selection	Rank-based composite consistency score

2.2 Dataset and Model Outputs

The analysis is conducted on a dataset of 240 tourism POIs collected from TripAdvisor, spanning multiple categories (e.g., museums, nature and parks). For each POI, the values of sustainability variables are generated using multiple LLMs under a controlled setup. Seven variables are considered, covering both normalized scores and real-world estimates, as summarized in Table 2. A total of 9 LLMs [1] are used. All models are queried using the same prompt and identical input, ensuring that differences arise solely from model behaviour.

2.3 Cross-Model Consistency Analysis

Cross-model consistency refers to the similarity in how different models rank the impact of environmental variables across POIs. To evaluate this agreement

[1] Models abbreviations: claude-haiku-4.5 (claude), deepSeek-r1 (deepseek), gpt-4.1-mini (gpt-4.1), gpt-5.4-mini (gpt-5.4), gpt-5.4-nano (gpt-5.4N), gpt-OSS-20B (gpt-OSS), ministral-3-14B (ministral), phi-4 (phi-4), and qwen3-14B (qwen).

[2] Variables abbreviation: water Consumption score (water score), water Consumption actual (water actual), CO_2 Emissions score CO_2 (score), CO_2 Emissions actual (CO_2 actual), pollution control (pollution), biodiversity Protection (biodiversity).

Table 2. Environmental sustainability criteria description[2]

Criteria	Range	Description
Water (Score)	[0,1]	Relative water consumption for the POI type.
Water (Actual)	> 0	Estimated daily water consumption (liters).
CO_2 (Score)	[0,1]	Relative CO_2 emissions for the POI type.
CO_2 (Actual)	> 0	Estimated daily CO_2 emissions (kg).
Noise Level	[1-5]	Expected noise impact on nearby residents.
Pollution Control	[0,1]	Relative pollution control practices.
Biodiversity Protection	[0,1]	Relative biodiversity protection practices.

regardless of the specific numerical scales used by each model, we employ the Spearman rank correlation. This consistency analysis is conducted at two levels:

- **Global Level:** A global assessment where models are compared based on their rankings of all individual data points for every environmental variable.
- **Category Level:** A categorical assessment where the analysis is restricted to specific POI categories within the variable to identify where models exhibit stronger or weaker agreement based on POI category.

Procedure 1: Ranking Consistency Evaluation

1. Multi-Level Correlation: For each variable, compute Spearman rank similarity between all model pairs at the *Global* level (all POIs) and *Categorical* level (POI subsets). Aggregate results into comparison grids.
2. Consensus Calculation: Calculate each model's average agreement score against the group and determine mean correlations per category to identify context-dependent variability.
3. Output: A consistency profile mapping model alignment with the consensus across all data points and specific categories.

2.4 Bucket-Based Consistency Analysis

This measures the level of agreement between models where generated values are discretized into specific classification levels. To assess whether LLMs are effective at fine-grained classification, variables are grouped into five ordered buckets, {*Very Low to Very High*}. For score-based variables, equal-width discretization is applied. Its [0,1] range is divided into intervals of size 0.20 corresponding to the five bucket levels. For actual variables, equal-width intervals are computed after capping values above the 95th percentile to reduce distortion caused by extreme estimates. This capping is applied only for bucket construction and does not remove observations from the analysis. For *Noise level*, the original ordinal scale is preserved. The analysis focuses on two complementary perspectives:

- **Bucket Distribution:** It identifies the proportion of points assigned to each bucket by every model to identify if LLMs favour specific buckets over others.
- **Bucket Agreement:** It evaluates the precision of the consensus within fine-grain scale. It distinguishes between *Exact Matches*, where models assign the same bucket, demonstrating high precision, and *Near Matches*, where model assignments fall within one adjacent bucket, showing general consistency.

Procedure 3: Bucket-Based Consistency Analysis

1. Discretization: Map scores into a five-level scale (*Very Low* to *Very High*). Apply equal-width discretization to continuous variables, 95th-percentile capped ranges for actual variables. Keep the ordinal scale for *Noise level*.
2. Distribution Assessment: Calculate the frequency of bucket assignments for every model to identify classification patterns.
3. Agreement Mapping: For each data point, identify **Exact matches** and **Near matches** across models.

2.5 POI-Level Disagreement Analysis

Magnitude variability measures the numerical differences in estimates provided by each model at specific POIs. This analysis provides a quantitative complement to ranking and bucket-based assessments by capturing the absolute spread of model outputs. The variability is quantified using two primary metrics: Standard Deviation, to assess the overall dispersion of estimates, and Range, to identify the total spread between the highest and lowest values.

These metrics are analyzed across two dimensions:

- **Variable-level:** Captures the average dispersion across all POIs for a specific environmental factor.
- **Category-level:** Captures the average variability within a specific POI category, i.e., which POI types (e.g., parks, museums) exhibit the highest or lowest consensus in model magnitude.

Procedure 4: Variability and Dispersion Measurement

1. Point-Level Calculation: For each POI and variable, compute the standard deviation (dispersion) and range (total spread) across all model estimates.
2. Variable-Level Aggregation: Calculate the mean variability across all points for each environmental variable to determine metric-specific consistency.
3. Contextual Analysis: Group results by POI category to identify location types exhibiting the highest model disagreement.

2.6 Consistency-Based Model Recommendation

A model recommendation framework is developed to identify the models most consistent with the collective ensemble behaviour. This framework prioritizes relative consistency, treating the ensemble consensus as the reference point for reliability. The recommendation is determined by evaluating three metrics:

- **Mean Correlation:** Measures the average ranking similarity between a model and all others (higher is better).
- **Distance from Consensus:** Calculates the average deviation of a model's output from the group mean (lower is better).
- **Statistical Divergence:** Uses the Wilcoxon signed-rank test to determine if a model's estimates differ significantly from the ensemble behaviour.

This rank-based composite approach allows for the selection of the most representative models at both the global and variable-specific levels.

It is important to note that the recommendation is based on relative consistency rather than external accuracy, as no ground-truth data is available. Therefore, the identified models should be interpreted as those most aligned with the collective ensemble behaviour rather than objectively more accurate. This provides a practical basis for model selection when using LLM-generated sustainability data.

3 Results

3.1 Cross-Model Consistency Analysis

Global Level: The consistency of sustainability estimates across the nine evaluated LLMs varies depending on the environmental variable Table 3. The highest agreement is observed for *Noise Level* ($\rho = 0.84$), indicating highly stable rankings across models. Strong consistency is also found for *Water Consumption*, both in score and actual values ($\rho = 0.66$–0.70). *CO_2 emissions* show moderate agreement ($\rho = 0.60$–0.64), suggesting relatively consistent ranking behaviour. In contrast, agreement decreases for more complex variables. *Biodiversity* ($\rho = 0.55$) exhibits moderate but less consistent patterns, while *Pollution* shows the lowest and most variable consistency ($\rho = 0.36$). This variability is further reflected in the wide range of pairwise correlations ($\rho = 0.08$–0.71), compared to more stable variables such as *Noise Level*. Although most models follow similar ranking trends, some pairwise comparisons exhibit substantially lower correlations (down to $\rho = 0.08$), particularly for variables with lower overall agreement. This highlights the influence of model-specific estimation behaviour on cross-model consistency.

Category Level: Cross-model agreement also varies across POI categories (Table 4). The highest consistent agreement is observed for *nature and parks*, particularly for *Water Consumption*, *CO_2 emissions*, and *Noise Level* ($\rho \approx 0.84$–0.91). This indicates that models align more closely in well-defined environmental contexts. Moderate agreement appears in *sights and landmarks* and *spas and wellness*, while lower agreement is observed for *fun and games*, *museums*, and *nightlife*, reflecting greater divergence in model estimates. Across categories, *Biodiversity* and *Pollution* consistently exhibit weak agreement, whereas *Water Consumption* and *CO_2 emissions* remain comparatively stable.

Overall, these results indicate that model agreement is strongly dependent on both the environmental variable and the contextual characteristics of the POI.

Table 3. Summary of Mean Correlation by Sustainability Variable

Variable	Mean ρ	Min ρ	Max ρ
Water Consumption (Score)	0.66	0.42	0.81
Water Consumption (Actual)	0.70	0.56	0.81
CO_2 Emissions (Score)	0.64	0.49	0.82
CO_2 Emissions (Actual)	0.60	0.36	0.77
Noise Level	0.84	0.73	0.92
Pollution Control	0.36	0.08	0.71
Biodiversity Protection	0.55	0.35	0.71

Table 4. Cross-model mean agreement by category across sustainability variables.

Category	Water score	Water actual	CO_2 score	CO_2 actual	Noise level	Biodiv.	Pollu.
Fun and Games	0.53	0.57	0.60	0.57	0.49	0.33	0.09
Museums	0.47	0.61	0.47	0.38	0.49	0.44	0.27
Nature and Parks	0.91	0.90	0.89	0.84	0.88	0.47	0.16
Nightlife	0.42	0.55	0.55	0.56	0.50	0.26	0.07
Sights & Landmarks	0.66	0.69	0.73	0.66	0.80	0.46	0.19
Spas & Wellness	0.70	0.67	0.65	0.64	0.47	0.34	0.27

3.2 Distribution and Bucket-Level Consistency

Classification behaviour is evaluated through bucket distributions and agreement metrics (Fig. 1, Table 5). The distributions reveal clear differences in how models utilize the scoring scale. For *Water (score)* and *CO_2 (score)*, most models concentrate values in the low to medium range, although some models (e.g., *deepseek* and *ministral*) show more dispersed scoring behaviour. For *Water (actual)* and *CO_2 (actual)*, distributions vary across models, with values spanning multiple bucket levels. While several models concentrate estimates in lower ranges, others assign larger proportions to medium, high, and very high categories, indicating notable differences in quantitative scale utilization and estimation behaviour across models. *Noise* values remain comparatively balanced across buckets, reflecting variability in perceived noise intensity across models. For *Pollution Control*, most models concentrate scores in the low to medium range, although some models assign larger proportions of very low values, indicating systematic differences in scale usage. Similarly, *Biodiversity Protection* remains strongly skewed toward very low values across all models, indicating substantial score compression and limited differentiation.

Agreement results further highlight differences in classification behaviour. As shown in Table 5, near-match agreement remains high for score-based and ordinal variables (≈91–98%), indicating that models generally assign the same or adjacent categories. In contrast, *Water (actual)* and *CO_2 (actual)* exhibit lower agreement levels, with exact agreement of 50.8% and 38.1%, respectively, and near-match agreement of 69.5% and 64.2%. This indicates greater variability in how

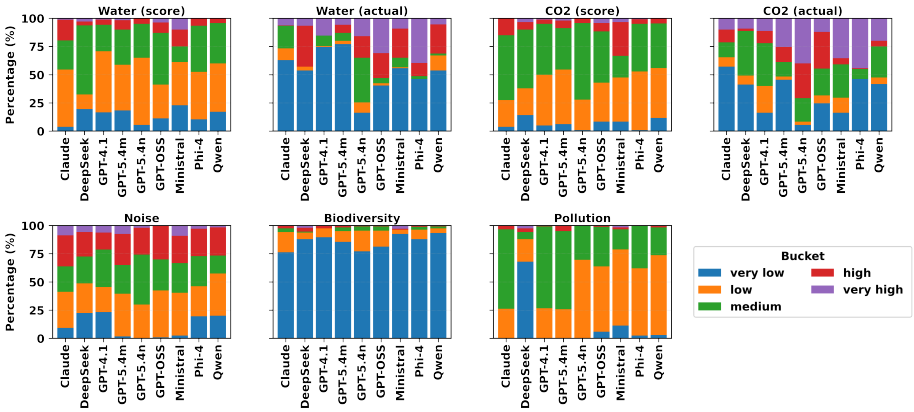


Fig. 1. Sustainability score distributions by model and variable across buckets.

models assign quantitative ranges for actual sustainability estimates. Among the remaining variables, *Biodiversity Protection* shows the highest exact agreement (84.1%), largely reflecting consistently compressed distributions across models. *Pollution Control* exhibits classification variability, indicating substantial variability in how models classify more subjective sustainability dimensions.

Overall, the findings suggest that models exhibit comparable relative positioning of POIs while differing in scale utilization and classification precision.

Table 5. Exact and near-match bucket agreement (%) per sustainability variable.

Variable	Exact	Near	Agreement Pattern
Water (score)	48.62	94.00	Moderate consistency across models
Water (actual)	50.75	69.54	Higher variability in quantitative estimates
CO_2 (score)	52.80	95.74	High near-match; moderate exact agreement
CO_2 (actual)	38.06	64.22	Lowest agreement for quantitative estimates
Noise	58.28	98.03	Highest near-match consistency
Biodiversity	84.13	97.67	High consensus with concentrated distributions
Pollution	45.88	91.03	High variability in subjective classification

3.3 POI-Level Disagreement Analysis

Disagreement is evaluated using the standard deviation and range of values assigned by different models for each sustainability variable. Table 6 summarizes the mean disagreement across variables. The analysis indicates that disagreement is highly dependent on the variable. The most significant variability occurs in actual-value variables, notably *Water Consumption* (mean std $\approx$ 6893, range

$\approx$ 19,306) and *CO_2 emissions* (mean std $\approx$ 184, range $\approx$ 518). These high values reflect differences in magnitude and scale adopted by different models. Among score variables, *Pollution Control* (mean std $\approx$ 0.112) and *Water Consumption* ($\approx$ 0.102) show relatively higher variability, indicating greater divergence in model assessments. *CO_2 emissions* ($\approx$ 0.096) and *Noise Level* ($\approx$ 0.459 in discrete scale) exhibit moderate disagreement, suggesting partial consistency across POIs. The lowest disagreement is observed for *Biodiversity Protection* (mean std $\approx$ 0.059), indicating relatively consistent outputs across models, likely due to the score compression observed in the distribution analysis.

Category-level analysis further shows that disagreement is POI-dependent (Table 7). Higher disagreement is generally observed for *Nature and Parks* and *Sights and Landmarks*, particularly for actual-value variables, while lower disagreement appears in categories such as *Museums* and *Nightlife*. This suggests that both the sustainability criterion and tourism context influence the degree of model divergence.

Table 6. Aggregate POI-level disagreement metrics across sustainability variables.

Variable	**Mean Std.**	**Mean Range**	**Disagreement Level**
Water (Score)	0.102	0.305	Moderate divergence
Water (Actual)	6893.137	19306.042	High (Scale-dependent)
CO_2 (Score)	0.096	0.296	Moderate divergence
CO_2 (Actual)	183.653	517.780	Moderate (Scale-dependent)
Noise Level	0.459	1.188	Significant spread
Pollution	0.112	0.352	Higher divergence
Biodiversity	0.059	0.182	Lowest disagreement

Table 7. POI categories with the highest and lowest disagreement mean Std.

Variable	**Highest Disagreement (Std)**	**Lowest Disagreement (Std)**
Water (score)	Sights & Landmarks (0.12)	Fun & Games (0.08)
Water (actual)	Nature & Parks (22852.36)	Fun & Games (3157.92)
CO_2 (score)	Nightlife (0.11)	Museums (0.07)
CO_2 (actual)	Nature & Parks (336.56)	Museums (133.70)
Noise Level	Sights & Landmarks (0.56)	Nightlife (0.35)
Pollution	Nature & Parks (0.18)	Nightlife (0.09)
Biodiversity	Nature & Parks (0.17)	Fun & Games

3.4 Model Consistency-Based Recommendation

The recommendation framework combines three consistency metrics: average correlation with other models (*Corr*), average distance from the ensemble consensus (*Dist*), and the number of statistically significant pairwise differences identified by the Wilcoxon test (*SigP*). Models are ranked for each metric individually, producing rank scores $\mathbf{R}_C$, $\mathbf{R}_D$, and $\mathbf{R}_S$. The final composite score (*Comp.*) is computed as the average of these rank positions, where lower values indicate stronger overall consistency with the ensemble behaviour. Table 8 presents the overall model consistency rankings. *Phi-4* achieves the highest overall consistency, followed by *claude*, *gpt-4.1*, and *gpt-5.4*, which exhibit comparable performance. To account for variable-specific differences, Table 9 reports the top two consistent models for each environmental criterion using the same composite ranking procedure. The results show that consistency is variable-dependent. For example, *Claude* performs best for *Noise Level* and *Water (actual)*, while *gpt-4.1* is consistent for *Water (score)* and CO_2 *(actual)*. Similarly, *gpt-5.4* excels for CO_2 *(score)* and *Pollution*, whereas *phi-4* is the strongest for *Biodiversity*.

These findings indicate that model selection depends on the target variable. While models such as *phi-4* provide strong overall consistency, variable-specific selection offers greater robustness in the absence of ground-truth data.

Table 8. Overall model consistency rankings based on composite consistency scores.

Model	Corr	Dist	SigP	$\mathbf{R}_C$	$\mathbf{R}_D$	$\mathbf{R}_S$	Comp.	Rank
phi-4	0.618	0.086	35	6	3	1	3.333	1
claude	0.686	0.095	38	1	5	5	3.667	2
gpt-4.1	0.662	0.082	41	2	2	7	3.667	2
gpt-5.4	0.653	0.074	41	3	1	7	3.667	2
qwen	0.619	0.093	37	5	4	3	4.000	5
ministral	0.631	0.099	39	4	6	6	5.333	6
deepseek	0.523	0.110	35	9	8	1	6.000	7
gpt-5.4-n	0.594	0.198	37	8	9	3	6.667	8
gpt-oss	0.615	0.103	41	7	7	7	7.000	9

4 Guidelines for Using LLM-Generated Data

Based on our findings, we propose the following guidelines for utilizing LLM-glenerated sustainability data in decision-making:

– **Prefer ranking-based methods:** LLM-generated data shows stronger consistency in relative ordering than in absolute values. Ranking-oriented methods are more robust to cross-model variability.

Table 9. Top-2 models per variable based on composite ranking. Pairs separated by |.

Variable	Top Models	Corr	Dist	SigP	Rank
Water score	gpt-4.1, gpt-5.4	0.71 | 0.70	0.08 | 0.07	5 | 5	1 | 2
Water actual	claude, phi4	0.76 | 0.73	0.05 | 0.06	6 | 6	1 | 2
CO_2 score	gpt-5.4, gpt-4.1	0.69 | 0.69	0.07 | 0.07	7 | 6	1 | 1
CO_2 actual	gpt-4.1, phi4	0.65 | 0.61	0.07 | 0.07	6 | 5	1 | 2
Noise Level	claude, gpt-4.1	0.87 | 0.87	0.08 | 0.08	4 | 5	1 | 2
Pollution	gpt-5.4, claude	0.44 | 0.48	0.09 | 0.10	6 | 7	1 | 2
Biodiversity	phi4, claude	0.59 | 0.62	0.05 | 0.09	5 | 5	1 | 2

- **Use a single consistent model in resource-constrained settings:** When computational cost or response time is critical (e.g., large-scale tourism platforms or real-time recommendation systems), selecting a model with relatively high consistency (e.g., Phi-4 or GPT-4.1) provides a practical solution.
- **Adopt variable-specific model selection when higher robustness is required:** Model performance varies across criteria. For example, Claude shows stronger agreement for Noise and Water-related variables, while GPT-4.1 performs well for CO2-related criteria. Tailoring model selection by variable can improve stability.
- **Apply aggregation selectively:** Combining outputs from multiple models can reduce variability but increases computational cost. Aggregation should be reserved for variables with lower agreement (e.g., Pollution or Biodiversity), rather than applied uniformly.
- **Be cautious with absolute values and thresholds:** Differences in scale usage and magnitude may affect value-sensitive decision methods. Additional normalization or calibration may be required before applying threshold-based MCDM techniques (which is not in the case of ELECTRE-H).

5 Conclusion

This study evaluated the consistency of LLM-generated sustainability data for tourism POIs in the absence of ground-truth information. The results show that models achieve relatively high agreement for score-based and ordinal variables such as *Noise Level*, while greater variability is observed for actual quantitative estimates and more subjective criteria such as *Pollution Control*. Differences in numerical magnitudes and classification boundaries further indicate that model outputs are not directly interchangeable. Our findings highlight that consistency is strongly variable-dependent. Although some models provide strong general alignment with ensemble behaviour, more robust performance can be achieved through a variable-specific model selection strategy. These insights provide practical guidance for integrating sustainability data generated by LLMs into decision-making systems such as MCDM and recommendation systems. Future work will investigate the impact of model variability on downstream decision processes.

References

1. Baena, V., Cerviño, J.: Tourism in the era of social responsibility and sustainability: Understanding international tourists' destination choices. Sustainability **16**(19), 8509 (2024). https://doi.org/10.3390/su16198509
2. Said, A.: Recommender systems for social good: The role of accountability and sustainability. In: Proceedings of the International Workshop on Recommender Systems for Sustainability and Social Good, pp. 1–4. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-87654-7_1
3. Zhang, Q., Yang, P., Yu, J., Wang, H., He, X., Yiu, S.M., Yin, H.: A survey on point-of-interest recommendation: Models, architectures, and security. IEEE Transactions on Knowledge and Data Engineering (2025). https://doi.org/10.1109/TKDE.2025.3551292
4. Merinov, P., Ricci, F.: Simulating the impact of recommendation salience on tourists experienced utility. In: Proceedings of the International Workshop on Recommender Systems for Sustainability and Social Good, pp. 35–51. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-87654-7_4
5. Salmony, M.Y., Moreno, A., Valls, A.: Towards sustainable tourism: A novel comprehensive hierarchical multi-criteria evaluation framework for points-of-interest recommendations. In: Artificial Intelligence Research and Development, pp. 12–21. IOS Press (2025)
6. Salmony, M.Y., Moreno, A., Valls, A.: Advancing sustainable tourism recommendations through innovative hierarchical evaluation and data acquisition methods. In: Martínez, L., et al. (eds.) Intelligent Data Engineering and Automated Learning – IDEAL 2025, LNCS, vol. 16239, pp. 59–70. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-10489-2_6
7. Sánchez Pérez, P., Bellogín, A.: Towards sustainability optimization in touristic route recommendation. In: Intelligent Management of Data and Information in Decision Making: Proceedings of the 16th FLINS Conference on Computational Intelligence in Decision and Control & the 19th ISKE Conference on Intelligence Systems and Knowledge Engineering (FLINS-ISKE 2024), pp. 9–16 (2024). https://doi.org/10.1142/9789811294631_0002
8. Banerjee, A., Satish, A., Wörndl, W.: Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation. In: Proceedings of the International Workshop on Recommender Systems for Sustainability and Social Good, pp. 19–34. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-87654-7_3
9. Wu, L., Zheng, Z., Qiu, Z. et al. A survey on large language models for recommendation. World Wide Web 27, 60 (2024). https://doi.org/10.1007/s11280-024-01291-2
10. Nadăș, M., Dioșan, L., Tomescu, A.: Synthetic data generation using large language models: Advances in text and code. IEEE Access (2025). https://doi.org/10.1109/ACCESS.2025.3589503
11. Abluton, A., Barrera-Leon, L., Benetti, S., Canonico, M., Cerutti, S., Desimoni, F., Portinale, L., et al.: Design and implementation of a tourism experimentation platform for context-aware and sustainable recommendations. Applied Sciences **16**(8), 3937 (2026). https://doi.org/10.3390/app16083937