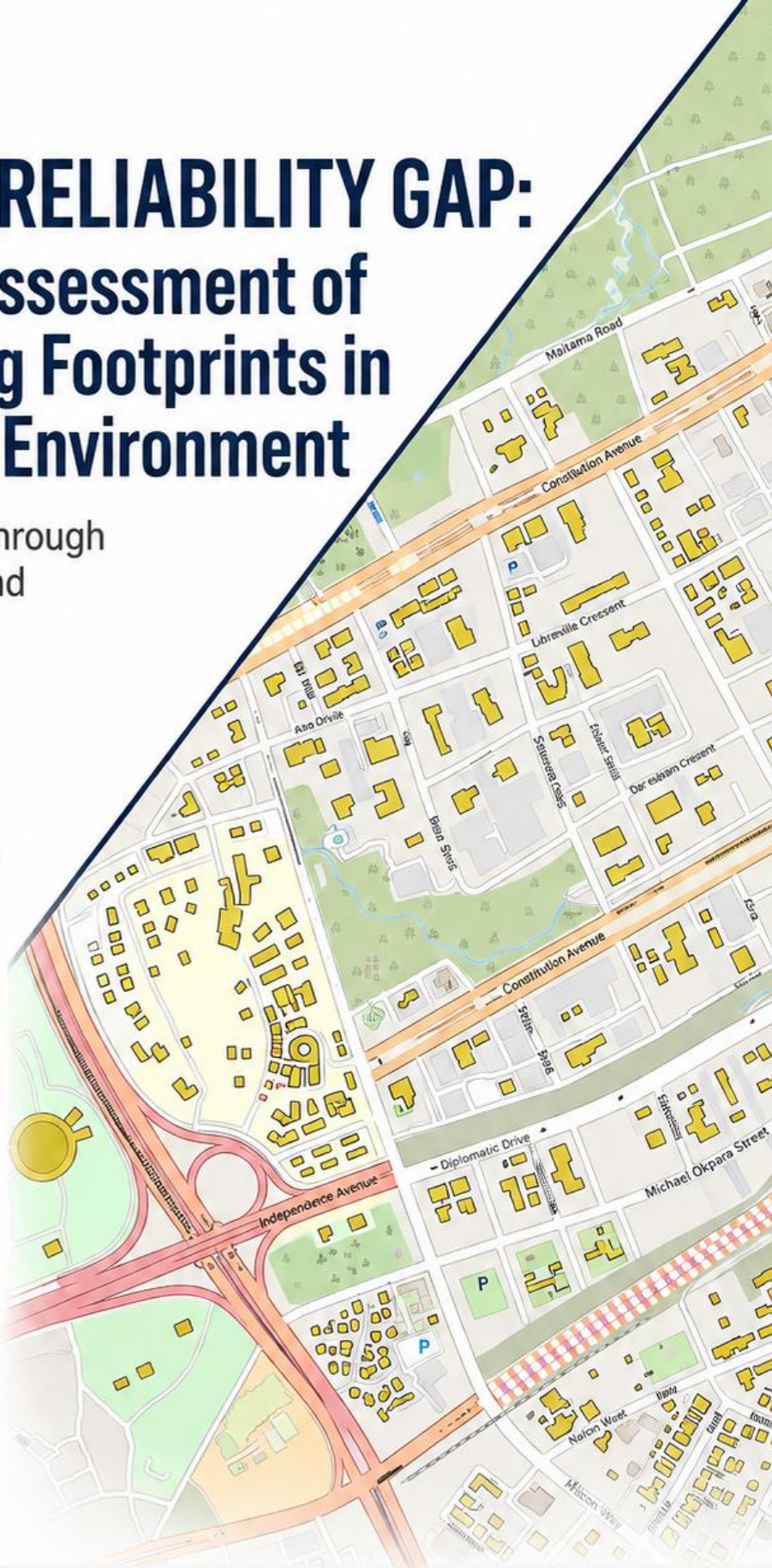
TECHNICAL REPORT

QUANTIFYING THE RELIABILITY GAP: A Spatial Accuracy Assessment of AI-Generated Building Footprints in a Data-Scarce Urban Environment

Evaluating Geospatial AI Systems Through Instance-Level Empirical Analysis and Human-Centered Oversight Design



KEY STUDY COMPONENTS

EMPIRICAL EVALUATION
Instance-level reliability assessment of AI-generated building footprints compared against OpenStreetMap reference data in Abuja, Nigeria

SPATIAL METRICS
Geometric agreement (IoU), positional accuracy (centroid offset), detection recall with IoU-threshold sensitivity analysis

GROUND-TRUTH VERIFICATION
Street-level imagery confirmed failure modes; independent validation beyond overhead imagery alone

HUMAN-IN-THE-LOOP DESIGN
Critical analysis of inconsistent AI behavior and implications for human oversight protocols in high-stakes geospatial decision systems

AI SAFETY FOCUS
Core finding: aggregate coverage (7.3× increase) does not equal per-instance reliability (56.4% recall); both metrics essential for trustworthy deployment

AUTHOR
Gamze Gül Mungan
Geospatial Data Specialist · AI Reliability Researcher

DATE
July 2026

STUDY AREA
Abuja, Nigeria
4.28 km² single defined study area

This work represents an independent technical study conducted for research and evaluation purposes.
It provides empirical evidence supporting governance frameworks (NIST AI RMF 1.0, EU AI Act Annex III) for high-risk AI systems.
The findings expressed are those of the author and do not necessarily reflect any organizational affiliation.

Quantifying the Reliability Gap:

A Spatial Accuracy Assessment of AI-Generated Building Footprints in a Data-Scarce Urban Environment (Abuja, Nigeria)

Technical Work Sample — AI Safety Fellowship Application

In one sentence: a deployed GeoAI system that generates 7.3× more building polygons than the OSM baseline simultaneously fails to detect 43.6% of independently known buildings aggregate coverage and per-instance reliability are distinct, safety-relevant properties.

Gamze Gül Mungan

Geospatial Data Specialist · AI Reliability Researcher

Türkiye
July 2026

Contents

Abstract

Artificial intelligence is increasingly used to generate building-footprint data in regions where OpenStreetMap (OSM) coverage is incomplete, with outputs feeding into applications such as digital twins, disaster risk mapping, and infrastructure planning. This study presents an empirical, instance-level reliability assessment of AI-generated building footprints (Mapflow.ai, satellite-imagery-based deep learning extraction) against OSM reference data, using a defined 4.28 km² study area in Abuja, Nigeria.

Using a greedy one-to-one spatial matching procedure and Intersection-over-Union (IoU), we find that AI-generated output substantially increases aggregate building coverage (7.3× the OSM building count in the study area) while simultaneously failing to detect 43.6% of buildings independently confirmed to exist in OSM (recall = 56.4% at a matching threshold of IoU ≥ 0.1). Among confirmed matches, geometric agreement is moderate (mean IoU = 0.458) and positional offset is consistent (mean = 6.03 m); recall declines substantially as the matching threshold is tightened, from 59.6% (any overlap) to 25.5% (IoU ≥ 0.5), indicating that detection completeness is highly sensitive to how strictly a "match" is defined.

A targeted, street-level (ground-level imagery) verification exercise further identified a confirmed case of false-positive misclassification, in which stacked concrete infrastructure was labeled as a building by the AI model. We argue that the central safety-relevant finding is not any single metric but the gap between aggregate coverage and per-instance reliability: a system that appears to substantially improve data completeness can simultaneously be unreliable for any specific, individually consequential decision. We discuss implications for human-oversight design in high-stakes geospatial decision systems and outline the study's limitations, including the single-city geographic scope of the study area and the unverified status of the large majority of AI-only detections.

1. Introduction

1.1 Problem Statement

Global building-footprint data in OpenStreetMap is unevenly distributed. Herfort et al. (2023), analyzing 13,189 urban agglomerations worldwide using a machine-learning completeness model, find that OSM building-footprint completeness exceeds 80% for cities home to only 16% of the global urban population, while remaining below 20% for cities home to 48% of that population — disproportionately concentrated in lower- and middle-income regions, including much of West Africa.

AI-assisted building extraction from satellite imagery ("GeoAI") is increasingly proposed to close this gap: deep-learning models can generate building footprints across areas where volunteer-contributed mapping is sparse or absent, in a fraction of the time required for manual digitization. However, AI-generated footprints are frequently intended for use in applications with direct real-world consequences — digital twin construction, disaster risk assessment, infrastructure and cadastral planning, and population estimation. In these contexts, the reliability of individual AI-generated geometries, not merely their aggregate coverage, is a safety-relevant property.

1.2 Research Question

This study asks: when AI-generated building footprints are compared against independently existing reference data, how reliable are they in terms of detection (recall), geometric accuracy, and positional accuracy — and is this reliability consistent across independently drawn samples? This is deliberately narrower than asking whether AI "can replace" OSM, or whether AI-generated data is usable at all. The goal is to quantify where and how much AI output diverges from known reference data, in order to inform what kind of human oversight is necessary before such data is used in high-stakes decisions.

Formally, we hypothesize (H1) that aggregate coverage gains from a deployed GeoAI system can co-exist with substantial per-instance unreliability — that is, that volume expansion and instance-level recall are empirically separable properties of the same output.

1.3 Relation to AI Safety

This work does not propose a new AI model, nor does it propose a new governance framework. It treats an existing, publicly available GeoAI tool (Mapflow.ai) as a black box and asks an empirical validation question: how much should a downstream decision-maker trust this specific, deployed system's output, and under what conditions? This framing — empirical reliability auditing of a deployed AI system in a concrete domain, using real, non-synthetic data — is intended to complement, rather than duplicate, the governance-level frameworks discussed in Section 2: those frameworks specify what should be evaluated (e.g., accuracy, robustness, human oversight); this study demonstrates, with a worked example, what such an evaluation actually finds when applied to a real system in a real, data-scarce environment.

2. Related Work

2.1 AI Governance Frameworks and High-Risk Classification

Several governance frameworks now specify evaluation requirements for AI systems used in consequential decisions. The NIST AI Risk Management Framework (AI RMF 1.0) organizes trustworthy-AI practice around four functions — Govern, Map, Measure, Manage — but is explicitly domain-agnostic and does not specify geospatial-specific evaluation procedures (NIST, 2023). The EU AI Act (Regulation (EU) 2024/1689), Annex III, classifies AI systems used in critical infrastructure management among its high-risk categories, imposing requirements including risk management, data governance, technical documentation, logging (auditability), transparency, human oversight, and accuracy/robustness assessment.

Matuszczyk et al. (2025) provide the most direct antecedent to the present study, mapping documented GeoAI bias mechanisms (representation, algorithmic, and aggregation bias) onto the EU AI Act's high-risk provisions and arguing that widely deployed GeoAI applications qualify as high-risk systems under the Act. Their contribution is a legal-technical mapping and literature synthesis; it does not include an original empirical accuracy audit of a specific deployed tool against reference data — a gap this study addresses directly with a single, concrete case.

2.2 Reliability and Explainability in GeoAI

Within GIScience, a growing literature addresses the reliability and interpretability of geospatial deep learning. Xing and Sieber (2023) identify three classes of challenge in applying explainable-AI (XAI) techniques to GeoAI: challenges specific to XAI computation itself, challenges arising from geographic data structures and scale, and "geosocial" challenges related to the incompleteness of geographic knowledge representation. Roussel and Böhm (2023) review geospatial XAI methods and note that map-based presentation is essential for making model explanations usable in a geographic context, rather than treating spatial data as an undifferentiated input modality. Li et al. (2024) survey the broader GeoAI research agenda and explicitly call for evaluation approaches that move beyond aggregate accuracy to consider robustness and reliability under real deployment conditions.

This literature establishes that aggregate-accuracy-only evaluation is a recognized, general limitation of current GeoAI practice. What remains comparatively scarce is empirical, instance-level evidence of what this limitation looks like in a specific, deployed, commercially available tool, applied to a real gap-filling use case in a data-scarce environment.

2.3 The OSM Completeness Gap as Context for AI-Assisted Mapping

The motivating context for AI-assisted building extraction is itself well documented: Herfort et al. (2023) is, to our knowledge, the most comprehensive global assessment of OSM building-footprint completeness, and directly motivates the use of AI tools such as Mapflow.ai in under-mapped regions. It is worth noting, and returned to in Section 7, that this same unevenness in OSM completeness is also a limitation of using OSM as a reference dataset for accuracy assessment: OSM itself is neither uniformly complete nor error-free.

2.4 Positioning of This Study

This study does not propose a new trustworthy-AI framework, evaluation taxonomy, or governance instrument. Existing frameworks (NIST AI RMF, EU AI Act Annex III) already specify what dimensions a high-risk AI system should be evaluated on; existing GeoAI literature (Xing & Sieber, 2023; Roussel & Böhm, 2023; Li et al., 2024) already argues, at a conceptual level, that aggregate accuracy is insufficient evidence of deployment reliability. The contribution of this study is narrower and more concrete: a real, reproducible, instance-level empirical audit of one specific, commercially deployed GeoAI tool, in one specific data-scarce urban environment, including a case in which a failure mode (misclassification of non-building infrastructure) was confirmed through independent ground-level imagery rather than inferred from overhead imagery alone. We present this as a worked example of the kind of evidence that governance frameworks call for, rather than as a new framework in its own right.

3. Study Area and Data

3.1 Study Area

The primary analysis uses a defined study area of 4.28 km² (4,284,582.7 m²) in Abuja, Nigeria, spanning a mix of formally planned, road-grid neighborhoods and less regular urban fabric along Constitution and Independence Avenues. All results reported in this study refer to this single, defined boundary.

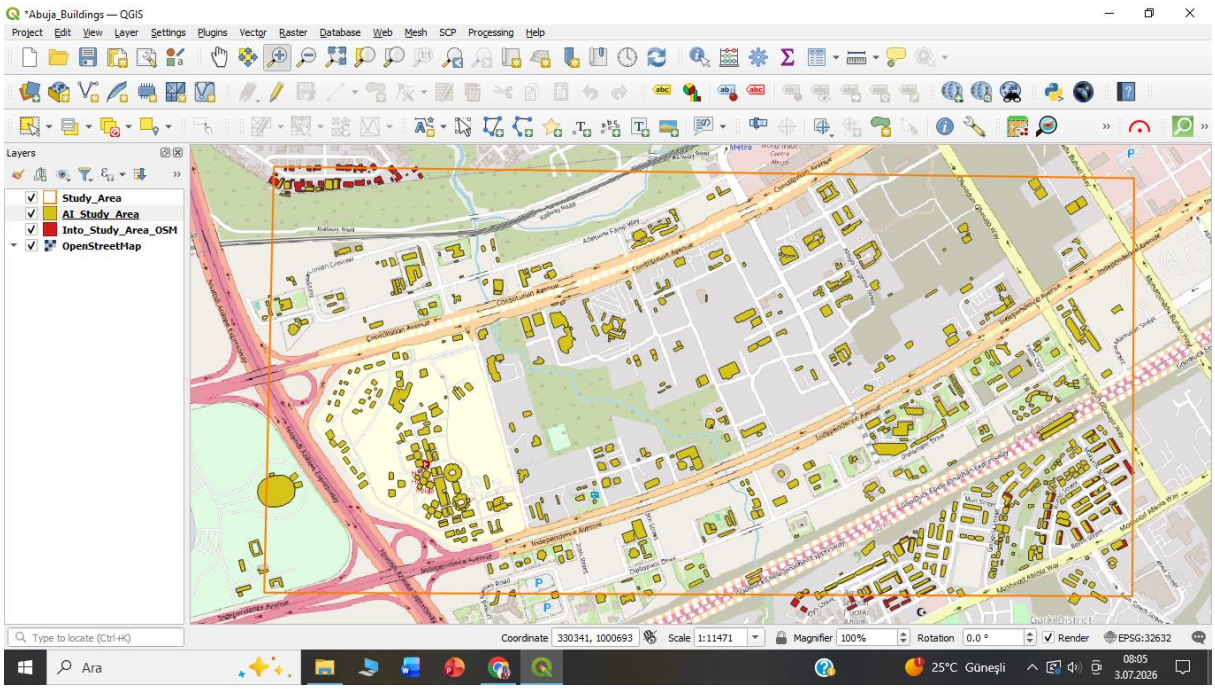


Figure 1. The 4.28 km² primary study area (orange boundary), Abuja. AI-generated building footprints (olive) are visibly denser than the OSM reference layer (red) across most of the area.

3.2 Data Sources

OpenStreetMap (OSM): Building-footprint polygons extracted via the QuickOSM plugin (Overpass API), tag building=*, clipped to the study area boundary. OSM building data in this study area is treated as the primary comparison baseline, not as an infallible ground truth a distinction discussed further in Section 7. AI-generated data: building footprints generated by Mapflow.ai (Geoalert), using the "Buildings" deep-learning model applied to Mapbox Satellite imagery, with a recorded processing/imagery date of 29 April 2026 for the primary study area layer. Output includes per-building attributes (class: housing, commercial, industrial, residential, other; estimated building height; processing metadata).

Supplementary ground-level verification: for a small number of specific, visually ambiguous cases identified during overhead-imagery review, Google Street View imagery was consulted to confirm or disconfirm the true nature of the object on the ground (Section 5.4.1). This is a targeted, qualitative verification method applied to individual cases, not a comprehensive manually digitized reference dataset covering the full study area.

	OSM	AI (Mapflow)
Buildings in primary study area	94	684
Data source	Volunteer digitization (Overpass API)	Deep-learning extraction, satellite imagery
Geometry style	Hand-digitized, regular edges	Pixel-based, irregular edges
Role in this study	Comparison baseline	System under evaluation

Table 1. OSM vs. AI-generated (Mapflow) building counts and characteristics in the primary study area.

3.3 Data Licensing

OSM data is distributed under the Open Database License (ODbL); it is used here for research/analysis purposes with attribution to OpenStreetMap contributors. Satellite basemap imagery displayed in figures is sourced from Mapbox Satellite (AI processing) and Google Earth Pro / Airbus (visual verification), and is reproduced here for non-commercial research illustration under the respective platforms' terms of use.

4. Methodology

4.1 Spatial Matching

Each OSM building polygon is matched against AI-generated polygons using a greedy one-to-one assignment procedure: all OSM–AI polygon pairs with non-zero spatial intersection are enumerated, Intersection-over-Union (IoU) is computed for each candidate pair, and pairs are assigned in descending order of IoU, with each OSM polygon and each AI polygon permitted to appear in at most one assigned pair. This prevents a single large AI polygon from being counted as a "match" for multiple distinct OSM buildings, which would otherwise artificially inflate recall.

4.2 Matching Threshold and Its Effect on Recall

A methodological choice that materially affects reported recall is the minimum IoU required to count a spatially intersecting pair as a "confirmed match" rather than incidental, negligible overlap. We report results at IoU ≥ 0.1 as our primary working definition of a confirmed match — a deliberately permissive threshold chosen because building footprints in this dataset are frequently offset or rotated (Section 5.4.1) without being wrong about a building's existence — and report a full sensitivity analysis across thresholds from 0.0 (any overlap) to 0.5 (Section 5.3, Figure 5) so that the reader can evaluate how conclusions depend on this choice.

4.3 Metrics

- Recall (coverage): proportion of OSM buildings (i.e., recall relative to the OSM baseline, not absolute recall; see Section 7) with at least one confirmed-match AI polygon (Section 4.2).

- Geometric agreement (IoU): for confirmed matches, intersection area divided by union area (1.0 = perfect overlap).
- Positional accuracy (centroid offset): Euclidean distance, in meters, between the centroids of matched OSM and AI polygons.
- AI-only detections: AI polygons with no OSM match under the one-to-one assignment. This quantity is reported descriptively; it is not converted into a "false positive rate" or "precision" figure, because — as Section 5.4.2 shows directly — a large fraction of AI-only detections are plausibly genuine buildings absent from OSM (legitimate gap-filling) rather than errors, and only a small, non-random subsample has been visually verified (Section 5.4). Reporting a precision figure computed by treating all unmatched AI polygons as false positives would misrepresent the evidence.

4.4 Coordinate Reference System

All geometries were projected to EPSG:32632 (UTM Zone 32N) prior to analysis to ensure metric-accurate area and distance calculations.

5. Results

5.1 Coverage and Recall

Within the primary study area, OSM contains 94 building polygons; the AI-generated layer contains 684 — 7.3 times as many. Under the one-to-one matching procedure (Section 4.1) with a confirmed-match threshold of IoU ≥ 0.1, 53 of the 94 OSM buildings (56.4%) have a confirmed AI counterpart, while 41 (43.6%) do not.

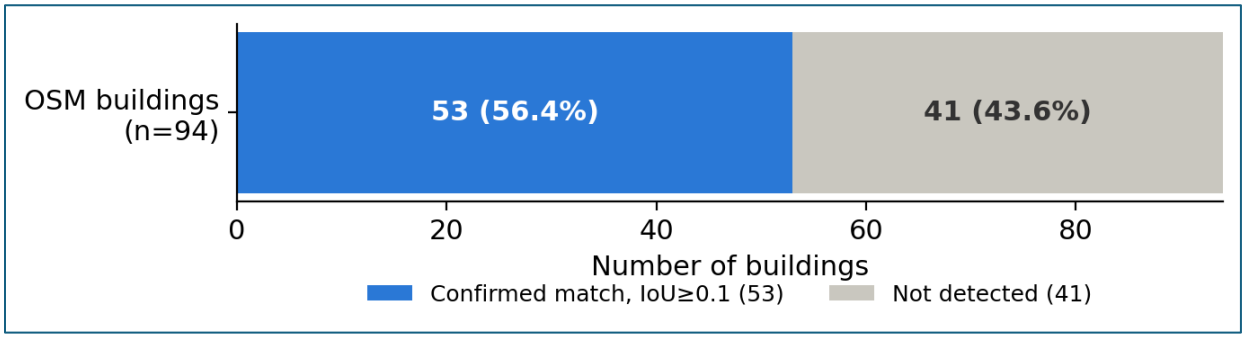


Figure 2. Recall of AI detection against the 94 OSM-mapped buildings in the primary study area, at a confirmed-match threshold of IoU ≥ 0.1.

This recall figure is distinct from, and should not be conflated with, the 7.3× aggregate volume increase: a system can simultaneously generate many more building polygons in total and fail to detect a large fraction of buildings that are independently known to exist. Figure 3 illustrates this qualitatively: a contiguous cluster of OSM-mapped buildings along a mapped street, the majority of which have no corresponding AI-generated polygon despite comparable size, shape, and imagery visibility. This pattern — missed detections clustering spatially rather than scattering uniformly — is relevant to any downstream use that aggregates AI-derived counts by neighborhood or block (e.g., population estimation).

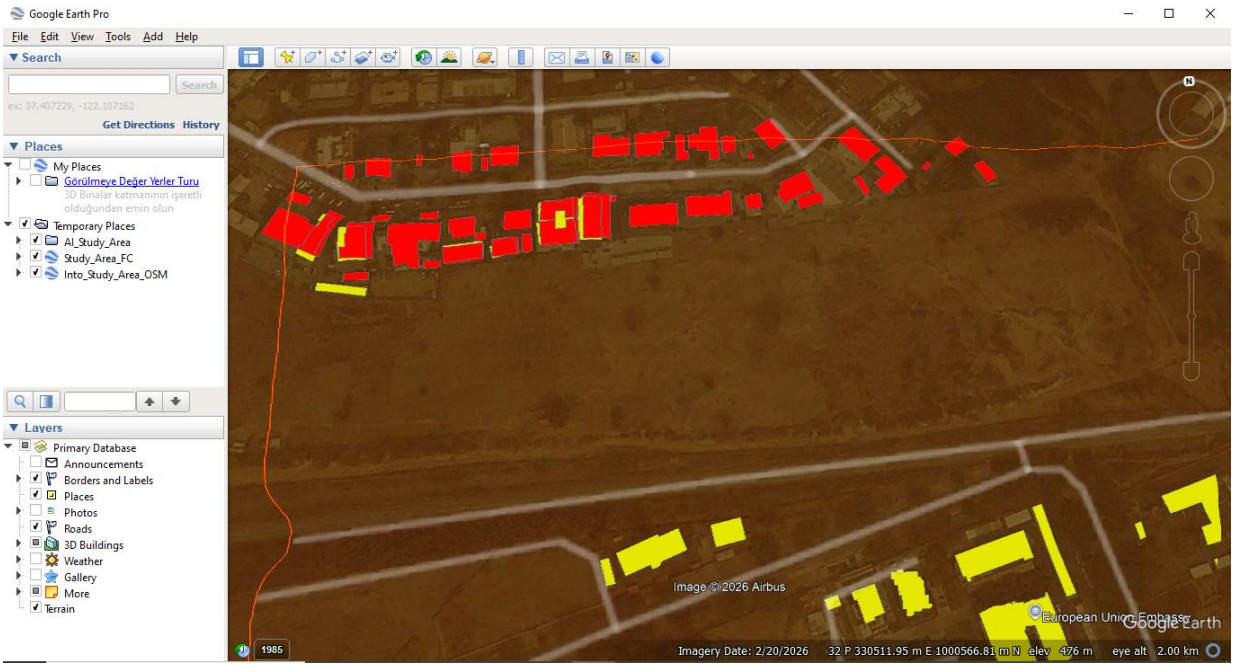


Figure 3. Qualitative illustration of the recall gap: OSM-mapped buildings (red) along a street, most without a corresponding AI-generated polygon (yellow).

5.2 Geometric Agreement

Among the 53 confirmed matches, mean IoU is 0.458 (median 0.425, SD 0.187). The distribution (Figure 4) is broad: a minority of matched pairs show strong geometric agreement (IoU > 0.7), while a substantial fraction fall in the 0.2–0.5 range — meaning that even where the AI model correctly identifies that a building exists at a location, the geometric footprint it produces frequently diverges substantially from the reference geometry. Whether the observed positional offset originates in the segmentation model itself or in georeferencing differences between the underlying imagery mosaics cannot be determined from the present data; we flag this attribution as unresolved.

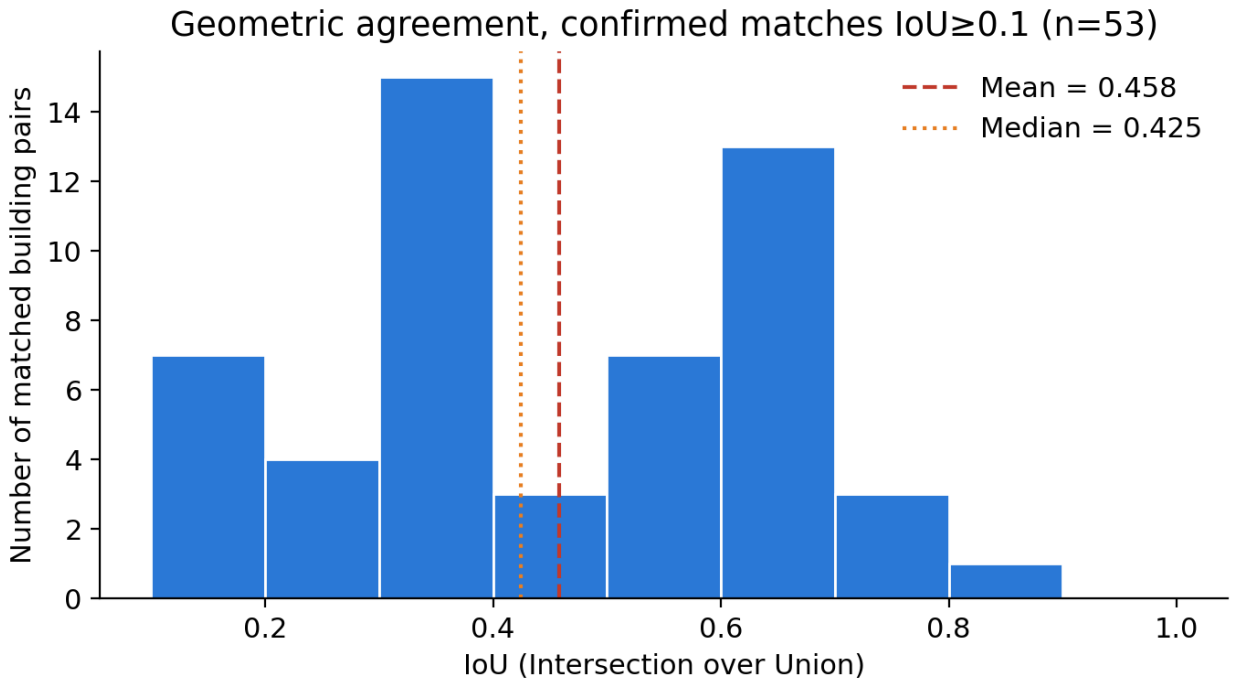


Figure 4. Distribution of IoU scores among confirmed matches (IoU ≥ 0.1, n = 53), primary study area.

5.3 Sensitivity of Recall to Matching Threshold

Because the definition of a "confirmed match" is a methodological choice (Section 4.2), Figure 5 and Table 2 report recall across a range of thresholds. Recall decreases substantially as the threshold is tightened — from 59.6% (any overlap) to 25.5% (IoU ≥ 0.5) in the primary study area — indicating that conclusions about detection completeness are sensitive to how strictly "detection" is defined, and that a stricter, COCO-style threshold (IoU ≥ 0.5, common in general computer vision benchmarks) would characterize this system's building-level recall as substantially lower than a permissive any-overlap definition would suggest.

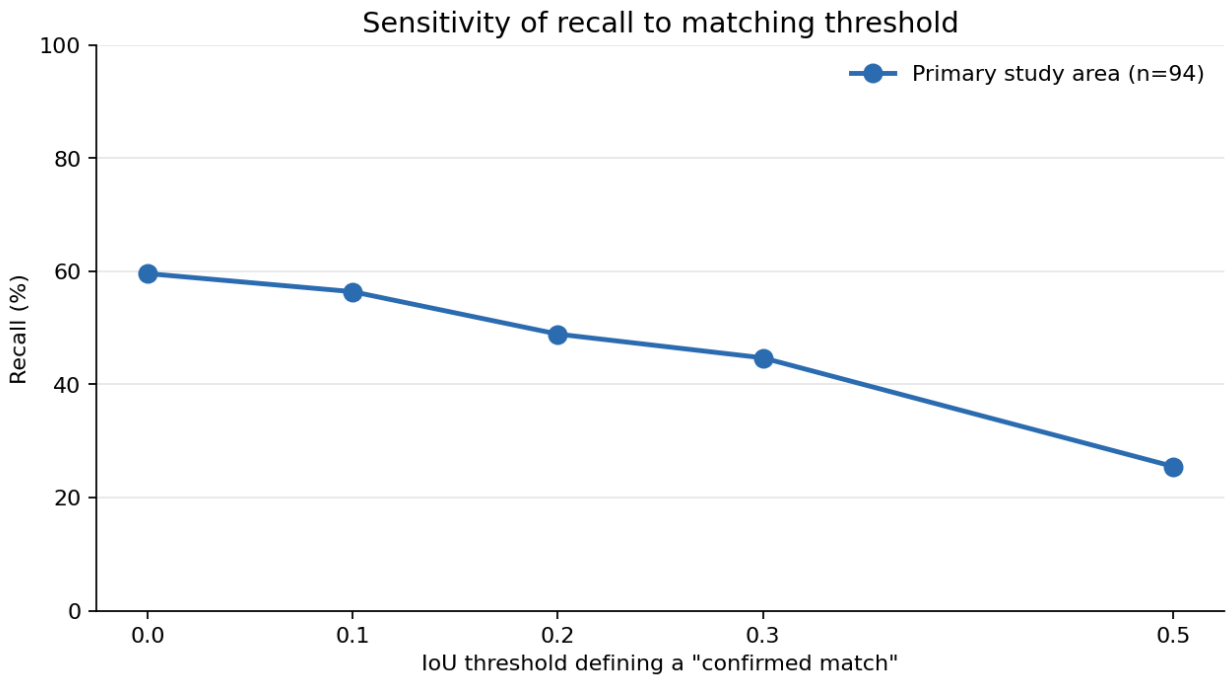


Figure 5. Recall as a function of the IoU threshold defining a confirmed match, primary study area.

IoU threshold	**Primary area recall (n=94)**
≥ 0.0 (any overlap)	59.6%
≥ 0.1 (working definition)	56.4%
≥ 0.2	48.9%
≥ 0.3	44.7%
≥ 0.5 (COCO-style)	25.5%

Table 2. Sensitivity of recall to the IoU threshold defining a confirmed match.

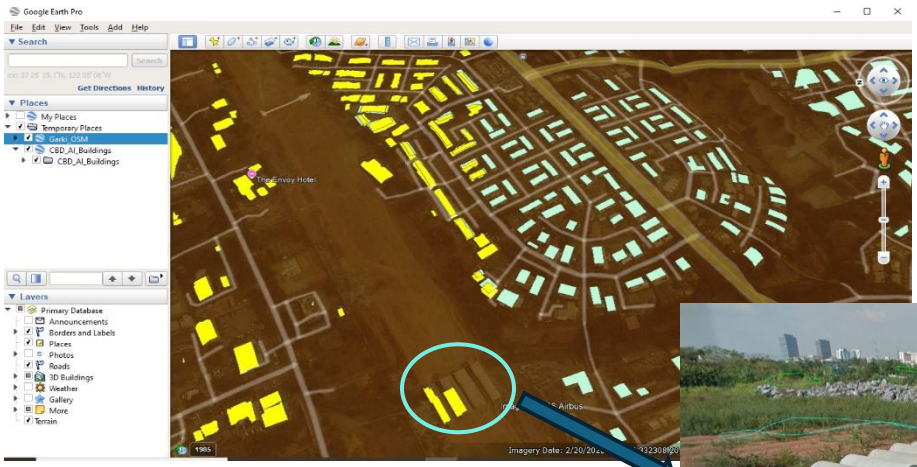


Figure 6. Three adjacent, elongated objects, later confirmed via Street View (Figure 7) to be stacked concrete pipes. The AI model generated misaligned "building" polygons for two of the three; the third was not detected.

Figure 7. Street View ground-truth confirmation: the objects circled in overhead imagery are stacked concrete culvert/pipe segments, not buildings

5.4 Qualitative Failure Modes

Beyond the aggregate statistics, visual inspection of individual cases — including one case cross-checked against ground-level Street View imagery — revealed distinct, non-overlapping failure patterns.

5.4.1 Misclassification of Non-Building Infrastructure (Ground-Truthed)

A cluster of three adjacent, visually similar elongated rectangular objects was initially treated as a case of ambiguous or ground-truth-uncertain structures. Street View imagery of the same location subsequently confirmed that these objects are stacked concrete culvert/pipe segments — construction material, not buildings. The AI model generated "building" polygons, with additional positional/rotational misalignment, for two of the three pipe stacks, and did not generate a polygon for the third.

This is the only instance in this study independently confirmed against ground-level, rather than only overhead, imagery, and it demonstrates a false-positive misclassification rather than a simple detection gap. It also carries a methodological implication: overhead-imagery-only visual verification (as used for the remaining cases in this section) cannot always reliably distinguish building-like objects from non-building infrastructure. Ground-level verification is a meaningfully stronger validation method but is impractical to apply at scale, reinforcing the case for structured human-in-the-loop review rather than either full automation or exhaustive manual re-verification.

5.4.2 Joint Blind Spots — Buildings Missed by Both Sources

Independently of Section 5.1's recall gap (buildings in OSM that AI missed), visual inspection identified structures missed by both OSM and the AI model simultaneously. These are visually regular, standard-form buildings with no obvious distinguishing feature, which weakens a simple explanation based on unusual building morphology.

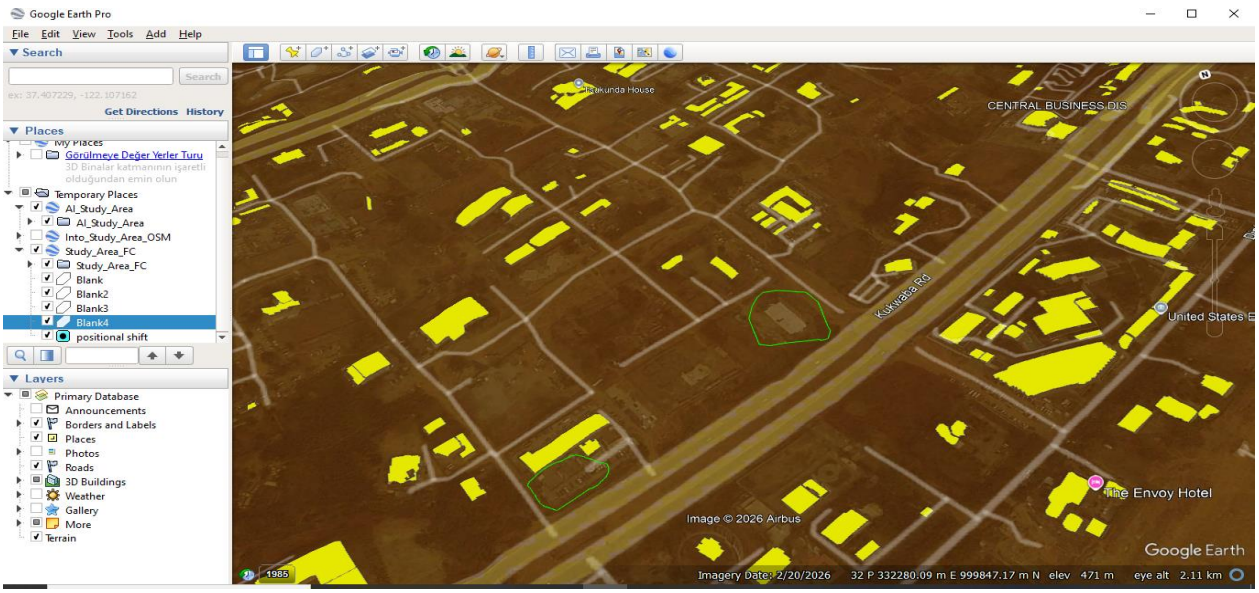


Figure 8. Two standard rectangular buildings (green outlines, manually identified against satellite imagery) present in neither the OSM reference layer nor the AI-generated layer.

5.4.3 Unclassified / Ambiguous Structures

A separate category of joint blind spot involves structures that are visually irregular and not confidently classifiable — possibly agricultural, temporary, or informal in nature — and which were also absent from both OSM and AI outputs. We do not classify the specific nature of these structures with confidence; we note only that this is a second, distinct route by which structures fall outside both data sources' coverage, alongside the standard-form case above.

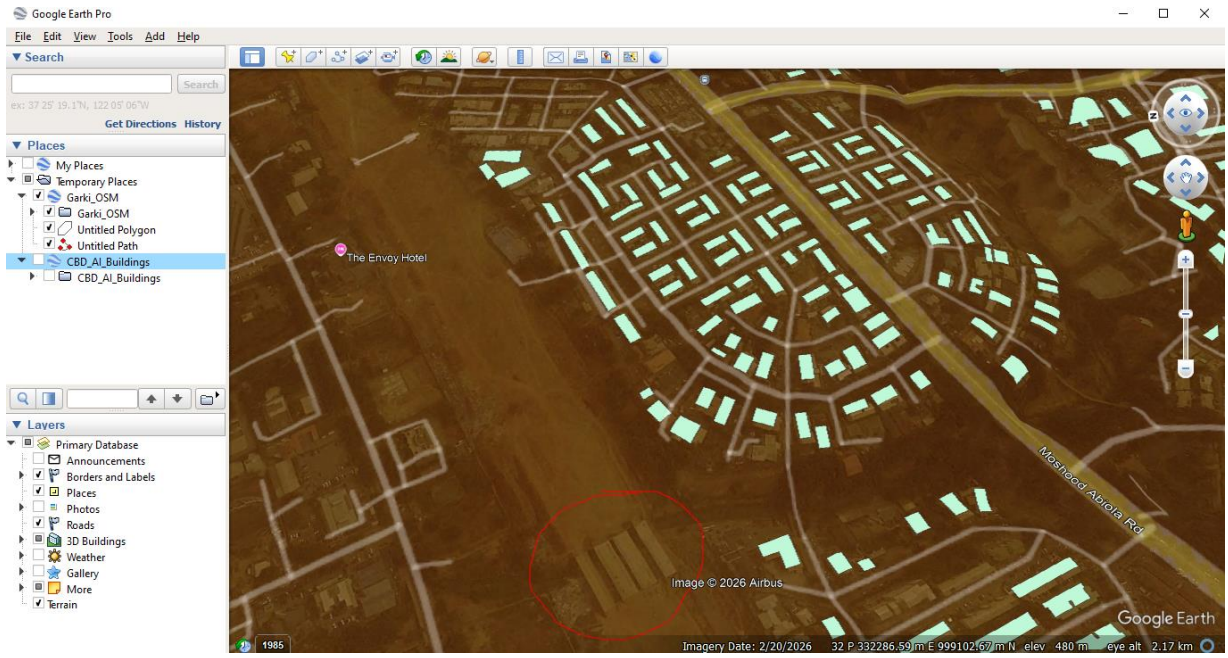


Figure 9. A cluster of irregular, low-profile structures (red circle), unclassified with confidence, absent from both OSM and AI outputs.

5.4.4 Additional Suspected False Positives (Unverified)

In a separate gap-filling exercise (Garki district), AI-generated polygons appeared in locations without a clearly identifiable structure in overhead satellite imagery. Unlike Section 5.4.1, these instances have not been confirmed via ground-level imagery and remain an open item, discussed further in Section 7 and Section 8.

6. Discussion

6.1 The Aggregate-vs-Per-Instance Reliability Gap

The central finding of this study is a distinction that is easy to overlook when evaluating AI-generated geospatial data at a glance: aggregate coverage and per-instance reliability are not the same property, and one does not imply the other. An AI system that generates 7.3× more building polygons than an existing reference dataset appears, on its face, to represent a substantial improvement in data completeness. But the same system, in the same area, fails to confirm-detect over 40% of buildings independently known to exist. A decision system relying on AI-generated footprints to answer a question like "does a building exist at this specific location relevant to this specific decision" cannot assume that high aggregate volume implies high answer reliability for any particular location.

This gap is not specific to geography. The same structure — strong aggregate metrics coexisting with unpredictable instance-level failure — is the central evaluation problem for any deployed perception system: medical image screening, autonomous-vehicle object detection, or biosecurity content filtering. What the geospatial setting adds is unusually cheap ground truth (overhead and street-level imagery), which makes the gap directly measurable here in a way that can inform oversight design in domains where verification is far more expensive.

6.2 Implications for High-Stakes Geoscience Decision Systems

The building-level failure modes documented in Sections 5.1–5.4 have distinct implications depending on the downstream application:

- Digital twin construction: A digital twin intended to represent existing urban infrastructure would, on this evidence, systematically under-represent buildings in a pattern that clusters spatially (Figure 3) rather than distributing randomly — meaning specific neighborhoods, not a random subset of buildings citywide, would be disproportionately mis-represented.
- Disaster risk and emergency response mapping: A recall of 56.4% for known buildings — falling to 25.5% under a stricter IoU ≥ 0.5 matching threshold (Section 5.3) — means that population-at-risk or exposure estimates built directly on AI-only building counts could be substantially incomplete, particularly given that the finding in Section 5.4.2 (standard buildings missed by both sources) shows that even combining OSM and AI does not guarantee full coverage.
- Cadastral and taxation use: The mean 6.03 m positional offset observed among confirmed matches (Section 5.2) is large relative to typical urban parcel dimensions and could plausibly place a structure on the wrong side of a parcel or zoning boundary if AI-generated geometry were used without correction.
- Population estimation from building counts: Because a majority of AI-only detections have not been verified as genuine buildings (Section 4.3), models that convert raw AI building counts into population estimates risk compounding an unquantified error term from a base layer whose false-positive rate is, at present, only spot-checked rather than systematically measured.

6.3 Why Inconsistency, Not Just Error Rate, Matters

The ground-truthed case in Section 5.4.1 — where the AI model misclassified stacked concrete pipes as buildings in two of three near-identical instances, and only avoided the error in the third by what appears to be chance rather than a discernible rule — is arguably more consequential for human-oversight design than the aggregate recall or geometric-agreement statistics on their own. A predictable failure pattern (e.g., "this model reliably fails on elongated non-building infrastructure") would allow a validation protocol to target review effort efficiently. An inconsistent failure pattern, by contrast, implies that no subset of AI output can be treated as reliably self-verifying: every output requires potential verification, with direct implications for the cost and design of human-in-the-loop review processes in operational deployments.

6.4 Relationship to Existing Governance Frameworks

This study's findings are consistent with, and provide concrete empirical support for, the general requirements already articulated by the NIST AI RMF and the EU AI Act Annex III (Section 2.1): both frameworks require accuracy/robustness evidence and human oversight for high-risk AI systems, but neither specifies a domain-specific methodology for producing that evidence in a geospatial context. Matuszczyk et al. (2025) argue on legal-technical grounds that widely deployed GeoAI applications meet the EU AI Act's high-risk criteria; this study supplies one worked example of what an accuracy/robustness audit responsive to that classification could look like in practice, applied to a single real system rather than derived from a literature synthesis.

7. Limitations

- Geographic scope: the study area is located within a single city (Abuja, Nigeria). Findings should not be generalized to other GeoAI models, other cities, or other urban morphologies without further replication.
- OSM as a reference baseline, not ground truth: OSM building data is itself incomplete and subject to the same volunteer-mapping unevenness documented by Herfort et al. (2023). Recall figures in this study should be read as "AI performance relative to what OSM has mapped," not as an absolute measure against a surveyed ground truth. Section 5.4.2's finding of buildings missed by both sources is direct evidence that OSM itself under-represents the true building population in this area.
- Unverified false-positive rate: 91.8% of AI-generated buildings have no OSM counterpart. Only a small number of these have been visually cross-checked (Section 5.4.1, 5.4.4), and only one against ground-level imagery. The true proportion that represents genuine gap-filling versus misclassification is not established by this study and is the single largest open question for future work (Section 8).
- Matching methodology: the one-to-one greedy matching procedure (Section 4.1) resolves the many-to-one ambiguity present in an earlier, uncorrected version of this analysis, but a small number of geometrically complex cases (e.g., a single large AI polygon spanning what OSM maps as multiple adjacent buildings) may still be imperfectly represented by any one-to-one scheme.
- Temporal mismatch between sources: the AI layer derives from imagery dated 29 April 2026, while OSM features were digitized at unknown, generally earlier dates from potentially different imagery.

Some apparent detection failures may therefore reflect construction or demolition between capture dates rather than model error; this component cannot be separated with the present data.

- No field survey: "reference data" in this study refers to OSM and visual interpretation of overhead and, in one case, ground-level imagery — not a surveyed cadastral or field-verified ground truth.

8. Conclusion and Future Work

This study presented an instance-level empirical reliability audit of AI-generated building footprints from a commercially available GeoAI tool (Mapflow.ai), compared against OSM reference data in a data-scarce urban environment (Abuja, Nigeria). The central finding — that a system generating 7.3× more building polygons than an existing baseline simultaneously fails to confirm-detect over 40% of independently known buildings — illustrates a general AI safety concern in concrete, quantified form: aggregate output volume and per-instance reliability are distinct properties, and evidence of one is not evidence of the other. A ground-truthed misclassification case (stacked concrete pipes labeled as buildings) further shows that this system's errors are not confined to omission, and that overhead-imagery-only verification is itself an imperfect tool for characterizing those errors.

We do not propose a new evaluation framework; we present this case as a worked example of the kind of instance-level, ground-truth-anchored evidence that existing governance frameworks (NIST AI RMF, EU AI Act Annex III) call for but do not, by themselves, produce. Three directions for future work follow directly from the limitations identified in Section 7: (1) a stratified, systematic visual and, where feasible, ground-level verification of a random sample of AI-only detections, to establish a defensible false-positive rate rather than the current spot-checked evidence; (2) replication in a second, independent study area (ideally in a different city), to test whether the geometric-agreement and positional-offset patterns observed here generalize beyond this single site; and (3) extension of the matching methodology to handle complex one-to-many building configurations more precisely than the current one-to-one scheme allows.

References

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Annex III — High-risk AI systems.

Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nature Communications, 14, 3985. https://doi.org/10.1038/s41467-023-39698-6

Li, W., Arundel, S., Gao, S., Goodchild, M. F., Hu, Y., Wang, S., et al. (2024). GeoAI for science and the science of GeoAI. Journal of Spatial Information Science, (29), 1–17.

Matuszczyk, N., Barnes, C. R., Roy, S., Gupta, R., Ozel, B., & Mitra, A. (2025). From bias to accountability: How the EU AI Act confronts challenges in European GeoAI auditing. arXiv:2505.18236.

National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

Roussel, C., & Böhm, S. (2023). Geospatial XAI: A review. ISPRS International Journal of Geo-Information, 12(9), 355.

Xing, J., & Sieber, R. (2023). The challenges of integrating explainable artificial intelligence into GeoAI. Transactions in GIS, 27(3), 626–645. https://doi.org/10.1111/tgis.13045

Appendix A: Data Provenance and Reproducibility

- OSM extraction method: QuickOSM plugin, Overpass API, tag building=*.
- AI extraction method: Mapflow.ai, "Buildings" model, Mapbox Satellite imagery; recorded imagery/processing date 29 April 2026 (primary study area layer).
- Analysis software: Python (GeoPandas, Shapely, pandas), QGIS, ArcMap.
- Coordinate reference system: EPSG:32632 (UTM Zone 32N).
- Matching procedure: greedy one-to-one assignment by descending IoU (Section 4.1); confirmed-match threshold IoU ≥ 0.1 (Section 4.2), with full sensitivity analysis reported (Section 5.3).
- Analysis code and processed geometries (GeoJSON/Shapefile layers referenced in this report) are available upon request.